# CERTIFYING SELECTIVE GENERATION FOR CULTURAL TEXT-TO-IMAGE: FEASIBILITY, ORDERING, AND RELEASE GATES BEYOND CALIBRATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Text-to-image systems deployed in cultural heritage settings can produce images that are visually plausible yet culturally unfaithful, substituting generic or neighboring traditions for the requested one. A common mitigation is selective generation: sample several candidates, score them with a verifier, and emit the highest-scoring one. But what evidence justifies releasing such a system? Risk-control methods can calibrate a frozen score, yet release has two logically prior failure modes: the candidate pool may contain no culturally acceptable output, or an acceptable output may exist but the verifier ranks another candidate above it. We formalize this distinction with request-level estimands—infeasibility $I$, mixed-pool mass $M$, and within-mixed-pool selection success $\gamma$—which obey the exact decomposition $R_{\mathrm{sel}} = I + M(1-\gamma)$. The identity is elementary; its value is attributional: the two actionable sources of risk become separately measurable. On top of this attribution layer we define a frozen-policy certification protocol: the operating point is chosen on development data only, the policy is frozen, and release is accepted if and only if an exact one-shot upper bound on the total selected-failure count $x = f + o$ meets the target; Bonferroni component lower bounds only attribute rejections. Common release metrics do not identify the certified quantity: with four candidates, 80.5% pairwise accuracy is compatible with a top-one hit rate from 0% to 100%, and flattening into per-request risk control cannot recover the decomposition. On CulturalFrames, at the prespecified endpoint, the frozen SigLIP verifier's ordering-error component has a one-sided 95% lower bound of 15.5%—above the full 10% risk budget—and a bounded Qwen2.5-VL repair fails its one-shot held-out gate. LiveCodeBench fails feasibility ($\hat{I} = 0.292 > 0.10$); three benchmark rows pass, including a RewardBench certificate with exact upper risk bound 2.83%. Synthetic validation confirms the gates open on non-degenerate $K > 2$ pools, and a risk–coverage frontier view certifies coverage $\approx 0.90$ at target $0.10$. Our cultural audit is a negative, endpoint-conditional diagnosis; we do not claim a benchmark-native positive certificate for non-degenerate cultural pools.

## 1 INTRODUCTION

Generative systems are increasingly used to depict cultural heritage: traditional attire, rituals, architecture, crafts, and everyday practices. In this setting, errors are not random. A text-to-image system can return an image that is visually polished and yet culturally unfaithful—substituting similar-looking but wrong symbols, rules, and content drawn from a neighboring or generic tradition for the tradition that was requested. Benchmarks with request-level human judgments document exactly this pattern across many cultures, and show that automatic verifiers often fail to detect it (Nayak et al., 2025; Rege et al., 2025; Liu et al., 2024; Malakouti et al., 2026). For institutions and creators working with living traditions, releasing such a system without evidence is not a neutral act: misrepresentation at scale is itself a cultural harm. A risk certificate is only as meaningful as the endpoint it controls: trustworthy cultural generation must be grounded in inheritor- or community-authorized cultural rules—the people whose tradition is depicted are entitled to specify what counts as faithful. We treat an authorized endpoint as a *precondition* of certification (Gate 1, Section 5), not an



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.

output of it; we do not solve the endpoint-authority problem here, but make it explicit as a gate that statistical machinery cannot bypass.

A common mitigation is selective generation: produce several candidates, score them with a verifier, and emit the highest-scoring one—the Best-of-$N$ pattern used in reward-model selection, verifier-guided generation, and self-consistency. But what evidence suffices to release such a system? A common answer is calibration: choose a threshold so that selected outputs satisfy a target error rate. Calibration, however, presupposes that the selection problem is meaningful, and two failures occur before it matters: the pool may contain no culturally acceptable candidate, or the verifier may rank a failing candidate above an acceptable one. The first calls for better candidate generation; the second for a better verifier. A threshold cannot create a missing candidate, and a strictly increasing recalibration cannot change the argmax of a frozen score. Common evaluation practices obscure this distinction: global rank correlation averages over requests that may never emit, pairwise accuracy summarizes comparisons irrelevant to the top-one decision, and labels collected only for selected outputs say nothing about discarded alternatives. High values of these metrics can coexist with unacceptable selected risk.

We study selective generation as a certification problem, with cultural text-to-image generation as the primary deployment scenario. For each emitted request-level candidate pool, we distinguish infeasible pools, mixed pools containing both acceptable and failing candidates, and safe pools. Let $I$ be the probability of an infeasible emitted pool, $M$ the probability of a mixed emitted pool, and $\gamma$ the probability that the frozen selector chooses an acceptable candidate conditional on a mixed emitted pool. Then

$$R_{\mathrm{sel}} = I + M(1 - \gamma). \tag{1}$$

The identity follows from total probability; its contribution is the operational separation it enables as an *attribution layer*: $I$ measures a pool-level limitation that no selector can repair, while $\gamma$ measures a selector-level limitation that the candidate budget cannot repair for already-emitted pools. Acceptance never plugs component estimates into Eq. (1) (Section 2.4). The formulation is domain-general, and we audit four further domains to demonstrate generality.

**Contributions.** (1) **A frozen-policy one-shot certification protocol for selective generation**: all design choices are made on a development split only; the frozen policy is scored once on disjoint evaluation requests; release is accepted iff the exact one-sided upper bound on the total selected-failure count $x = f + o$ satisfies $U_{1-\delta}(x, n) \leq \alpha$; Bonferroni component lower bounds only attribute rejections (Section 3.1). (2) **An exact decomposition of selected risk as an attribution layer** (Eq. (1)): a sharp feasibility condition for frozen score-argmax policies (Corollary 1), and a precise account of what flattening into per-request risk control cannot recover (Section 2.5). (3) **Identification limits of common metrics**: sharp bounds linking pairwise accuracy to top-one success, and bounds for selected-only labels. (4) **A cross-domain audit plus synthetic validation**: on CulturalFrames, mixed-pool ordering errors alone exceed the 10% budget—a negative, endpoint-conditional audit; LiveCodeBench fails feasibility; three rows pass their one-shot gates; synthetic validation shows the gates open on non-degenerate $K > 2$ pools, and a risk–coverage frontier replaces the arbitrary fixed-coverage point (Sections 3.4–3.5).

## 2 Selective Generation as a Certification Problem

### 2.1 Formal setup

Requests $r \sim P$ arrive independently. For request $r$, a generator produces $K$ candidates $y_1, \ldots, y_K$. A frozen score $s$ ranks candidates, and the selector emits $k^* = \arg\max_k s(y_k)$ when the maximum score exceeds a threshold $\tau$. Each candidate has a binary endpoint label, with $Y_k = 1$ denoting failure against a prespecified endpoint. A development split chooses the score arm and threshold; the resulting policy is evaluated on disjoint requests. All population quantities below are conditional on the development-frozen policy.

Let $E$ denote the emission event. An emitted pool is *infeasible* if all candidates fail, *mixed* if it contains both passing and failing candidates, and *safe* if all candidates pass. Define $I = \Pr(\text{infeasible} \mid E)$, $M = \Pr(\text{mixed} \mid E)$, $\gamma = \Pr(Y_{k^*} = 0 \mid \text{mixed}, E)$, and $R_{\mathrm{sel}} = \Pr(Y_{k^*} = 1 \mid E)$. These

are request-level quantities; $\gamma$ is estimated only on emitted mixed pools, so its denominator can be much smaller than the total number of emitted requests.

**Proposition 1** (Exact decomposition)**.** *For every generator, selector, score distribution, and threshold policy, $R_{\mathrm{sel}} = I + M(1-\gamma)$ (Eq. (1)), with no independence across candidates, fixed-$K$ assumption, or parametric score model required.*

*Proof.* Partition the emitted requests into infeasible, mixed, and safe pools. Infeasible pools fail with probability one, safe pools with probability zero, and mixed pools with probability $1-\gamma$; total probability gives the result. □

### 2.2 What the decomposition implies

For a target risk $\alpha$, Eq. (1) immediately gives a feasibility condition.

**Corollary 1** (Frozen-policy feasibility condition)**.** *For a frozen score-argmax policy, if $M > 0$, a target $R_{\mathrm{sel}} \le \alpha$ is attainable for the fixed emitted set if and only if*

$$I \le \alpha, \qquad \gamma \ge \gamma_{\min} := 1 - \frac{\alpha - I}{M}. \tag{2}$$

*If $M = 0$, the condition reduces to $I \le \alpha$.*

This is a deterministic property of the frozen policy, not a test statistic: we do not accept systems by plugging estimates into Eq. (2) (Section 2.4). Recalibration by a strictly increasing transformation preserves within-pool ordering (Rabanser and Papernot, 2025), and threshold changes cannot repair a fixed ordering error within a pool. The floor is sharp over unrestricted selectors on the same emitted pools, and if $I > \alpha$ even an oracle selector cannot meet the target.

### 2.3 Why common metrics do not identify release risk

**Global rank correlation.** Global rank correlation can be made high by concordance on requests that never emit, while the emitted tail is systematically misordered (and vice versa), so no distribution-free implication from global rank correlation alone bounds $R_{\mathrm{sel}}$.

**Pairwise accuracy does not identify top-one success.** Let $A$ be request-balanced all-pairs accuracy, $H$ the human-best hit rate, and $C = \binom{K}{2}$. For a unique human order and fixed score ties,

$$\max\{0,\ 1 - C(1-A)\} \ \le\ H \ \le\ \min\left\{1,\ \frac{CA}{K-1}\right\}, \tag{3}$$

because a top-one miss requires at least one wrong pair and a top-one hit requires at least $K-1$ correct comparisons; both bounds are attainable (Appendix A.1). Thus, for $K = 4$ and $A = 80.5\%$, the identified interval is $[0, 100\%]$, and guaranteeing $H \ge 90\%$ from pairwise accuracy alone requires $A \ge 98.3\%$.

**Selected-only labels do not identify improvement.** Suppose a same-generator pool has $K$ candidates with failure means $\bar{y}_k \in [0,1]$, and the frozen selector chooses $k^*$. Let $\bar{y}_s$ be the selected-output failure mean and $\Delta = \bar{y}_{k^*} - \bar{y}_{\mathrm{pool}}$ the selected-minus-pool contrast. Observing only $\bar{y}_s$ yields the sharp bounds

$$-\frac{K-1}{K}(1-\bar{y}_s) \ \le\ \Delta \ \le\ \frac{K-1}{K}\bar{y}_s. \tag{4}$$

For $K > 1$ and $0 < \bar{y}_s < 1$, this interval straddles zero, so selected outputs alone cannot establish either improvement or harm over random selection from the same pool. These results motivate measuring emitted pool types and mixed-pool outcomes directly rather than inferring them from proxy metrics.

### 2.4 Acceptance by total risk; attribution by component bounds

Safe pools never fail, infeasible pools always fail, and a mixed pool fails exactly when the selector orders a failing candidate on top. The total number of selected failures on $n$ emitted evaluation

requests is therefore directly observable as $x = f + o$ (with $\hat{R_{\text{sel}}} = x/n$), where $f$ is the number of emitted infeasible pools and $o$ the number of emitted mixed pools whose selected candidate fails. This count is more direct than any plug-in combination of component estimates, and it is exactly the quantity bounded by Proposition 1. The release decision uses an exact one-sided upper bound on the total selected risk: accept iff $U_{1-\delta}(x, n) \leq \alpha$. We do *not* accept by plugging in $\hat{I} + \hat{M}(1 - \hat{\gamma})$: because $I$, $M$, and $\gamma$ are estimated from the same evaluation sample, comparing a marginal interval for $\hat{\gamma}$ with a plug-in $\hat{\gamma}_{\min}$ is not a valid joint test.

Attribution is a separate, diagnostic step: with one-sided lower bounds at levels $\delta_f$ and $\delta_o$, $\delta_f + \delta_o \leq \delta$, declare feasibility failure if $L_{1-\delta_f}(f, n) > \alpha$ and ordering failure if $L_{1-\delta_o}(o, n) > \alpha$, since ordering failures alone lower-bound selected risk. Diagnostics do not require estimating $\gamma_{\min}$ and do not certify acceptability (Appendix B).

We distinguish an accept certificate from diagnostic attribution. Acceptance is based on an exact upper bound for total selected risk. Component-wise estimates explain why a rejected system failed, but plug-in combinations of $\hat{I}, \hat{M}, \hat{\gamma}$ are not used to accept.

### 2.5 Why not flatten the problem into standard risk control?

Could the pipeline be collapsed into a single per-request loss $L = Y_{k^*}$ on emitted requests, after which standard distribution-free risk control—Learn-Then-Test, risk-controlling prediction sets, or conformal risk control (Angelopoulos et al., 2021; Bates et al., 2021; Angelopoulos et al., 2022)—could certify $R_{\text{sel}} \leq \alpha$ directly, without ever defining $I$, $M$, or $\gamma$? Such flattening is statistically valid for the final release statement, and our protocol is fully compatible with it: the acceptance rule of Section 2.4 *is* a flattened exact bound on the total selected risk. What flattening cannot recover is the decomposition itself: conditional on the emission event $E$, the loss distribution depends on the component triple only through $R_{\text{sel}} = I + M(1 - \gamma)$, so any two triples with the same $R_{\text{sel}}$ induce identical conditional loss distributions, certificates, power, and accept/reject decisions for any procedure that observes only $(E, L)$ (Remark 1, Appendix A.4). For instance, $(I, M, \gamma) = (0.03, 0.60, 0.80)$ and $(0.09, 0.60, 0.90)$ both induce $R_{\text{sel}} = 0.15$; a flattened procedure rejects both identically, yet the first system needs a better selector and the second a better candidate generator. No amount of evaluation data resolves this ambiguity from $(E, L)$ alone, because the identifying information—the request-level pool type—has been discarded. The decomposition is thus an attribution layer underneath the risk-control machinery: flattened control decides whether the target is met; the decomposition decides what to fix when it is not (Section 3.4).

## 3 Audit Protocol and Cross-Domain Results

### 3.1 Protocol

Each audit row uses request-grouped candidate scores and binary endpoint labels. A hash-fixed 20% development split selects the score arm and threshold at a nominal 50% coverage target; the disjoint 80% split is scored exactly once by the frozen policy. We report two-sided 95% Clopper–Pearson intervals for $\gamma$ and one-sided exact 95% upper bounds for risk, where "exact" means exact under the conditional independence premise of the evaluation requests given the frozen policy; generalization to future deployment requests additionally requires the usual i.i.d.-type assumption, which curated benchmarks only approximate. We report denominators throughout and, for the cultural rows, a dependence sensitivity analysis.

**Protocol 1** (Frozen-policy certification)**. *Inputs.*** *Request-grouped candidate scores, endpoint labels, target risk $\alpha$, familywise level $\delta$.*

1. Development only. *On the hash-fixed development split, choose the score arm, selector class, threshold grid, and final operating point $\tau^\star$ or coverage-risk pair. No evaluation labels are used for this choice.*

2. Evaluation only. *Freeze the policy $(s, \tau^\star)$ and every exclusion rule. Score the disjoint evaluation requests once.*

Table 1: Development-frozen operating points on disjoint evaluation requests. $x = f + o$ is the total selected-failure count and $U_{95}$ its one-sided exact 95% upper bound; accept/reject uses only $U_{95}(x, n)$ versus $\alpha = 0.10$. Component estimates and $\gamma_{\min}$ are diagnostics ($\gamma_{\min} > 1$ reported as infeasible). The Qwen arm is diagnostic: its sealed test remains unscored.

| Audit | $n_{\text{emit}}$ | $n_\gamma$ | $\hat{I}$ | $\hat{M}$ | $\hat{\gamma}$ [95% CI] | $\gamma_{\min}$ | $x\ (= f + o)$ | Risk ($U_{95}$) | Decision |
|---|---|---|---|---|---|---|---|---|---|
| CulturalSigLIP | 94 | 78 | .074 | .830 | .731 [.618,.825] | .969 | 28 (7+21) | 29.8% (38.5%) | Reject |
| RewardBench | 1,186 | 1,186 | .000 | 1.000 | .980 [.970,.987] | .900 | 24 (0+24) | 2.0% (2.8%) | Accept |
| HumanEvalPack | 369 | 369 | .000 | 1.000 | .978 [.958,.991] | .900 | 8 (0+8) | 2.2% (3.9%) | Accept |
| LiveCodeBench | 89 | 29 | .292 | .326 | .793 [.603,.920] | infeas. | 32 (26+6) | 36.0% (45.1%) | Reject |
| ScienceQA | 210 | 210 | .000 | 1.000 | .995 [.974,1.000] | .900 | 1 (0+1) | 0.5% (2.2%) | Accept |
| CulturalQwen† | 147 | 109 | .020 | .741 | .807 [.721,.877] | .893 | 24 (3+21) | 16.3% (22.2%) | Reject |

3. One-shot certificate. *Let $n$ be emitted evaluation requests, $f$ emitted infeasible pools, and $o$ emitted mixed pools whose selected candidate fails. Report $x = f + o$, $\hat{R}_{\text{sel}} = x/n$, and exact intervals with stated denominators.*

4. Accept/reject. *Accept only if the one-sided exact upper bound satisfies $U_{1-\delta}(x, n) \leq \alpha$. Do not accept by plugging in $\hat{I} + \hat{M}(1 - \hat{\gamma})$.*

5. Attribution. *Report component estimates and simultaneous lower bounds as diagnostics. If rejected, attribute only when a component lower bound crosses its stated gate, e.g. infeasibility if $L_{1-\delta_f}(f, n) > \alpha$, or ordering if $L_{1-\delta_o}(o, n) > \alpha$ with $\delta_f + \delta_o \leq \delta$.*

**Threshold sweeps that reuse held-out outcomes are diagnostic frontier traces, not certificates. A certified operating point must be selected on development data and certified once on untouched evaluation data. Reporting multiple thresholds requires either a pre-registered grid with simultaneous exact bounds, or an additional sample-splitting step that separates threshold selection from final certification.**

### 3.2 Results

Table 1 is the central empirical summary. Decisions use only the one-shot accept rule of Protocol 1—accept iff $U_{95}(x, n) \leq \alpha$ at $\alpha = 0.10$; component estimates $\hat{I}, \hat{M}, \hat{\gamma}$ and the feasibility condition $\gamma_{\min}$ of Corollary 1 are diagnostics, not a plug-in accept rule. Three rows are accepted; LiveCodeBench fails feasibility before selector quality can matter; CulturalSigLIP has a large ordering-error component; the Qwen row is a one-shot development-selected repair whose sealed test was not scored.

**Feasibility failure.** LiveCodeBench is upstream-limited: $\hat{I} = 0.292$ ($f = 26$ of $n = 89$) exceeds the full $\alpha = 0.10$ budget, and the one-sided 97.5% lower bound on $I$ is 20.1%, so the gate $L_{1-\delta_f}(f, n) > \alpha$ closes this row with attribution *feasibility*. No selector can repair those pools; $\hat{\gamma}$ uses only 29 mixed pools and is diagnostic only.

**Ordering failure.** CulturalSigLIP emits 94 evaluation requests: 7 infeasible, 78 mixed, 9 safe. The selector passes in 57 of 78 mixed pools ($\hat{\gamma} = .731$). The 21 mixed-pool ordering errors are $21/94 = 22.3\%$ of emitted requests, with one-sided 95% lower bound 15.5% (Bonferroni 97.5% bound 14.4%)—already above the entire 10% budget. The ordering component alone closes the gate ($L_{1-\delta_o}(o, n) > \alpha$), independently of the unresolved feasibility bound ($U_{95}(I) = .135$) and of the total-risk decision, which also rejects ($U_{95}(28, 94) = 38.5\%$).

**Positive sanity check.** RewardBench emits 1,186 of 2,423 evaluation requests and records 24 failures: empirical risk 2.02%, exact upper bound 2.83%—accepted at $\alpha = 0.10$. All emitted pools are mixed and $\hat{\gamma} = 1162/1186 = .980$ [.970, .987], above the .900 feasibility condition: the protocol is not vacuously fail-closed. This $K = 2$ all-mixed sanity check does not exercise the full pool structure; Section 3.4 supplies the $K > 2$ evidence.

Table 2: CulturalFrames cohort flow. Counts must reconcile to the frozen 203-pool audit cohort; "tbc" marks counts to be completed from audit logs.

| Stage | Requests retained | Requests excluded |
|---|---|---|
| Released judgments | 10,412 judgments over 981 requests | |
| Complete $K = 4$ candidate sets | tbc | tbc |
| Complete endpoint labels | tbc | tbc |
| Complete frozen verifier scores | tbc | tbc |
| De-duplication / incomplete records | 203 | tbc |
| Final audit cohort | 203 pools: 17 infeasible, 154 mixed, 32 safe | |

**Other benchmark rows.** HumanEvalPack and ScienceQA are also accepted by the one-shot gate; ScienceQA has 209/210 successful selections, giving the exact interval $[.974, 1.000]$. The common .900 floor in these three rows follows from $\hat{I} = 0$ and $\hat{M} = 1$, where the decomposition degenerates to a direct selection-risk statement.

**Qwen repair.** The Qwen2.5-VL repair arm improves the development-selected ordering estimate, but one-shot evaluation gives $\hat{\gamma} = .807$ $[.721, .877]$, below the .893 condition. Feasibility is certified ($f = 3/147$, 97.5% upper bound .058), but total risk is above target ($x = 24/147$, $U_{95} = 22.2\%$), so the gate rejects. We treat this arm as diagnostic evidence that better supervision can improve ordering without producing a release certificate; its sealed test was never scored.

### 3.3 Cohort accounting and endpoint semantics

The CulturalFrames release contains 10,412 rater judgments over 981 requests; the audit cohort contains 203 complete $K = 4$ pools after deterministic cohort construction (incomplete candidate sets or endpoint judgments, unavailable verifier scores, and duplicate records are excluded before the split), hash-split 20%/80% for development and evaluation, fixed before evaluation access and label-independent with respect to the held-out outcome (Table 2; counts not recoverable from the released artifacts are marked for completion from audit logs).

We compare retained and excluded requests on all pre-exclusion covariates available without using held-out endpoint labels: country/region, request source, candidate count, score availability, and rater coverage. Exact binomial intervals are conditional on the frozen evaluation cohort and assume independent requests. Because CulturalFrames judgments share raters and requests cluster within countries, we additionally report country- and rater-clustered bootstrap intervals (to be completed from audit logs; the present submission reports the effective-sample-size sensitivity of Section 7 instead); nominal sample size is not treated as effective sample size.

The binomial certificate is valid under the stated independence premise. Under rater/country clustering the intervals are optimistic; the ordering-failure count remains interpretable as an observed audit statistic, but broad certification claims require clustered confirmatory inference.

Endpoints are benchmark-specific: CulturalFrames passes a request if mean rater prompt alignment is at least .75; RewardBench and HumanEvalPack use the released preference endpoint; LiveCodeBench requires all unit tests to pass; ScienceQA requires the answer key to match. The decomposition is portable across endpoints; their numerical scores are not directly comparable.

### 3.4 Synthetic validation on $K > 2$ pools with known ground truth

The positive rows in Table 1 are $K = 2$ all-mixed cases, where the decomposition degenerates. To verify the gates on non-degenerate structure, we run the unmodified protocol on synthetic pools with ground-truth components at $\alpha = 0.10$, $n = 2{,}000$ evaluation requests per regime (Appendix C.3, Table 4). The gates open when the target is met (regime P, true $(.03, .60, .95)$: all three pool types present, interior floor $\gamma_{\min} = .887$, $U_{95} = 7.2\%$), and close with the correct attribution otherwise—regime O on ordering alone, regime F on feasibility despite an excellent selector ($\hat{\gamma} = .974$). The plug-in decomposition matches the directly estimated risk in every regime, as Proposition 1 requires, and regime O instantiates Remark 1: flattened control reaches the identical verdict ($U_{95} = 16.4\%$) with no way to distinguish the observationally equivalent feasibility failure $(.09, .60, .90)$. These

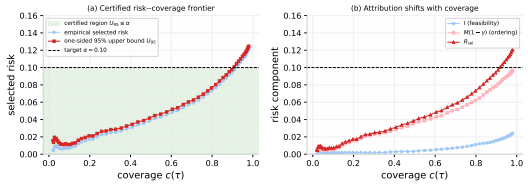


Figure 1: Empirical/simultaneous frontier under a fully specified synthetic generating process ($K = 4$, $I = 0.03$, $M = 0.60$, scores $s_k = \mu(Y_k) + \varepsilon_k$, $\varepsilon_k \sim \mathcal{N}(0, 0.75^2)$, induced $\gamma \approx 0.83$, $n = 20{,}000$); on real benchmarks, outcome-reusing threshold sweeps are exploratory and are not release certificates. (a) Full emission has risk 13.2%, above target $\alpha = 0.10$; raising the threshold lowers coverage but also both risk components, and the exact upper bound first meets the target at coverage $\approx 0.90$—the largest certifiable coverage. (b) Attribution shifts along the frontier: at high coverage the ordering component dominates, while abstention suppresses infeasible pools fastest. A single fixed-coverage operating point (e.g. 50%) discards most of the certifiable region.

regimes are synthetic; they complement, not replace, a benchmark-native $K > 2$ positive certificate (Section 7).

### 3.5 The certificate is a risk–coverage frontier, not a point

Every estimand in Section 2 is conditional on the emission event $E_\tau$ induced by threshold $\tau$; varying $\tau$ traces a frontier of policies $\pi_\tau$ with coverage $c(\tau) = \Pr(E_\tau)$ and selected risk $R_{\text{sel}}(\tau)$. On a fully specified generating process we trace the simultaneous-gate frontier $\mathcal{F} = \{(c(\tau), U_{95}(\tau)) : \tau \text{ passes the one-shot gate at level } \delta\}$, so that a release claim is a pair—"at coverage $c$, certified risk at most $\bar{R}$"—rather than a scalar at an arbitrarily fixed coverage; the largest certifiable coverage $c^* = \sup\{c : (c, \bar{R}) \in \mathcal{F},\ \bar{R} \le \alpha\}$ summarizes the deployed system. On real benchmarks the same sweep is only a diagnostic frontier trace: thresholds reuse held-out outcomes, so no traced point is a certificate unless pre-selected on development data and certified once on disjoint evaluation data (Protocol 1).

Figure 1 traces the frontier for a frozen score-based policy on synthetic pools with ground-truth components $I = 0.03$, $M = 0.60$, and an induced $\gamma \approx 0.83$. Full emission fails the target (risk 13.2%). Because the score is informative, raising $\tau$ preferentially abstains on infeasible and misordered pools: $\hat{I}$ falls from .024 to .004 and $\hat{\gamma}$ rises from .84 to .94 as coverage drops from 0.98 to 0.46. The exact upper bound first meets $\alpha = 0.10$ at coverage $\approx 0.90$, so the frontier certifies a deployable policy that the conventional fixed 50%-coverage operating point neither identifies nor needs. Two practical messages follow: select the operating point on the development frontier and certify it once on disjoint evaluation data, rather than fixing it by convention; and report the coverage side of the trade per subgroup, since abstention is not automatically benign (Section 4). The CulturalFrames sweep of Section 4 is exactly such a frontier trace on real data—diagnostic, not certified.

## 4 The Cultural Heritage Audit: An Ordering Bottleneck

Cultural text-to-image generation is a stringent test of selective generation: a visually plausible image can still substitute a generic or neighboring tradition. CulturalFrames provides request-level human alignment judgments (Nayak et al., 2025); related benchmarks expose long-tail gaps (Rege et al., 2025; Liu et al., 2024). We audit the ordering component of a frozen verifier pipeline, conditional on the prespecified endpoint—we do not claim the endpoint constitutes universal cultural validity. The frozen SigLIP policy targets 50% coverage; across two cross-fit halves, held-out coverage is 47.7%/53.0% and primary-endpoint risk is 28.1%/27.1%. Table 1 gives the stronger component-wise diagnosis: mixed-pool ordering errors alone exceed the full 10% budget.

**Endpoint sensitivity.** Under a post-hoc severe endpoint below .50, the same frozen policies produce 4.1%/6.6% risk and can open. We report this only as sensitivity: neither the primary .75 threshold nor the severe one is community-authorized. The cultural ordering verdict is endpoint-conditional; we release no cultural verdict as deployment guidance.

**Threshold and recalibration stress test.** Strictly increasing recalibration preserves every within-pool argmax (Appendix A.3); changing a threshold moves the emission set but does not repair the ordering of a fixed pool. Enumerating the 322 distinct scalar threshold policies on held-out requests at at least 50% empirical coverage—the diagnostic frontier of Section 3.5—the best observed risks remain 27.2%/26.4%: no point on the audited trace meets the target. An outcome-aware oracle transfers to 28.3%/29.0% test risk, with lower bounds 23.6%/24.8%, still above target. These searches reuse held-out outcomes and are diagnostic frontier traces, not certificates (Section 3.1).

**Repair and candidate controls.** The Qwen2.5-VL repair crosses pointwise, pairwise, and listwise objectives with retrieval on/off; development-only selection chooses the pointwise retrieval arm. Development $\gamma$ is .909 $[.843, .954]$, above its .864 floor, but one-shot evaluation falls to .807 $[.721, .877]$, below .893: development success cannot substitute for held-out certification.

**Subgroup coverage.** Rescueable failures occur across all ten represented countries, and country-level coverage differs under a shared threshold. Human-best candidates demonstrate candidate availability but use outcome information unavailable at runtime. A release certificate should report subgroup coverage—per-subgroup frontiers rather than a single aggregate curve—and specify an authorized refusal policy.

## 5 From Statistical Certificate to Release Contract

The statistical core embeds in a release workflow with five gates: *(1) endpoint validity*—for cultural deployments, the endpoint is fixed by, or grounded in, the rules of the tradition-bearers whose heritage is depicted (Section 1); *(2) feasibility*; *(3) ordering*; *(4) one-shot statistical power*; *(5) runtime enforcement of the certified frozen policy*. Gates 2–4 are the statistical core of this paper; Gate 1 is a governance question that internal calibration cannot settle and Gate 5 an engineering/provenance question (Appendix E). A fail-closed contract stores the endpoint definition, policy revisions, certified operating point, split specification, source hashes, denominators, bounds, and gate decision; missing or stale evidence causes abstention. The limitation is also the safety property: enforcement cannot manufacture a valid endpoint, a missing candidate, correct ordering, or statistical power.

## 6 Related Work

Our problem intersects four lines: distribution-free risk control (Angelopoulos et al., 2021; Bates et al., 2021; Angelopoulos et al., 2022); selective prediction and the reject option (Chow, 1970; El-Yaniv and Wiener, 2010); Best-of-$N$ and reward-model selection (Huang et al., 2025; Yu et al., 2026; Khairi et al., 2026); and cultural text-to-image benchmarks (Nayak et al., 2025; Rege et al., 2025; Malakouti et al., 2026). We ask the complementary post-generation question: given a frozen, already-deployed pool and selector, what can be certified before release, and can failure be attributed to the pool or to the selector? Appendix F gives the extended discussion.

## 7 Limitations

**Dependence.** Exact binomial inference assumes independent evaluation requests conditional on the frozen policy. CulturalFrames judgments share a rater pool and requests cluster within ten countries, so nominal denominators may overstate effective sample size. Halving the effective denominators widens the $\gamma$ intervals to approximately $[.55, .85]$ for SigLIP and $[.67, .90]$ for Qwen; the direct ordering-failure lower bound remains above the 10% budget (Section 3.3).

**$K = 2$ degeneracy, endpoint validity, and scope.** The three positive rows in Table 1 are $K = 2$ all-mixed cases, where the decomposition degenerates; the synthetic validation shows the gates open on non-degenerate $K > 2$ pools, but it is no substitute for a benchmark-native $K > 2$ positive certificate, which a future release should include before claiming broad success on complex real pools. The cultural result is conditional on the prespecified prompt-alignment endpoint, whose validity is not equivalent to community authorization; the endpoint-sensitivity analysis shows the verdict can

flip under a different convention. The RewardBench certificate is benchmark-scoped, not a universal safety guarantee. Candidate pools are observational, and human-best rows are diagnostics, not deployable causal effects; outcome-reusing threshold searches are diagnostic frontier traces, never certificates (Section 3.5).

## 8 CONCLUSION

Culturally faithful generation cannot be certified by calibration alone: a release decision must distinguish whether the candidate pool contained an acceptable candidate and whether the selector chose it. The exact decomposition $R_{\mathrm{sel}} = I + M(1-\gamma)$ makes these failure modes measurable as an attribution layer that flattened per-request risk control provably cannot recover, and sharp identification bounds show why pairwise accuracy and selected-only labels cannot substitute for direct evidence about the deployed top-one decision. Our cultural audit is a negative, endpoint-conditional diagnosis—mixed-pool ordering failures alone exceed the 10% risk budget, and we claim no benchmark-native positive certificate for non-degenerate cultural pools—while three benchmark rows pass their one-shot gates and synthetic validation confirms the gates open on non-degenerate $K > 2$ pools. The practical rule is modest: fix an authorized cultural endpoint, inspect feasibility, measure ordering with its true denominator, select the operating point on the development frontier, certify once on disjoint evaluation data, and enforce only the policy for which evidence exists.

### AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing of the manuscript (grammar, clarity, and LaTeX formatting) and for debugging and documentation of the analysis code. We have not used generative AI tools for research ideation, experimental design, data collection, or the generation of results. All AI-assisted outputs were reviewed and verified by the authors; in particular, all mathematical statements, proofs, tables, and reported numbers were checked by the authors against the source data and derivations. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This paper audits benchmark data and publicly released models and involves no new human subjects data collection. Cultural judgments are treated as released benchmark labels rather than community authorization. Abstention policies can redistribute error and coverage burdens; therefore subgroup coverage and endpoint governance are reported as explicit release considerations. No personally identifiable information is used.

### REPRODUCIBILITY STATEMENT

The manuscript reports every denominator used for the main empirical estimates and uses a fixed 20% development/80% evaluation split at the request level. Exact Clopper–Pearson intervals and one-sided exact risk bounds are used throughout the main certificate; acceptance uses the total selected-failure count $x = f + o$ and the one-shot gate $U_{1-\delta}(x, n) \leq \alpha$ of Protocol 1. The synthetic validation and frontier experiments use fully specified generating processes with fixed seeds (Appendix C.3) and are reproducible end-to-end from the stated parameters. The audit tool, split specification, release-contract schema, and training configurations for the Qwen repair are intended to accompany the anonymized supplementary material. The theoretical results are proved in Appendix A.

## REFERENCES

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

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. arXiv:2110.01052, 2021.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. arXiv:2208.02814, 2022.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021.

Casacuberta, S. and Kanade, V. Selective omniprediction and fair abstention. In *NeurIPS*, 2025.

Chow, C. K. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, (4):247–254, 1957.

Chow, C. K. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Cox, D. R. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444, 1975.

El-Yaniv, R. and Wiener, Y. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Fithian, W., Sun, D., and Taylor, J. Optimal inference after model selection. arXiv:1410.2597, 2014.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. In *ICML*, 2023.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *NeurIPS*, 2017.

Geifman, Y. and El-Yaniv, R. SelectiveNet: A deep neural network with an integrated reject option. In *ICML*, 2019.

Heng, A. and Soh, H. K. Know when to abstain: Optimal selective classification with likelihood ratios. In *ICLR*, 2026.

Huang, A., Block, A., Liu, Q., Jiang, N., Krishnamurthy, A., and Foster, D. J. Is Best-of-N the best of them? Coverage, scaling, and optimality in inference-time alignment. In *ICML*, 2025.

Imbens, G. W. and Manski, C. F. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Jeong, S., Choi, I., Yun, Y., and Kim, J. Culture-TRIP: Culturally-aware text-to-image generation with iterative prompt refinement. In *NAACL*, 2025.

Jiang, L., Chen, R., Gao, C., and Niu, D. RAISE: Requirement-adaptive evolutionary refinement for training-free text-to-image alignment. In *CVPR*, 2026.

Joren, H., Zhang, J., Ferng, C.-S., Juan, D.-C., Taly, A., and Rashtchian, C. Sufficient context: A new lens for retrieval augmented generation systems. In *ICLR*, 2025.

Jung, J., Brahman, F., and Choi, Y. Trust or escalate: LLM judges with provable guarantees for human agreement. In *ICLR*, 2025.

Khairi, A., Dsouza, D., Fadaee, M., and Kreutzer, J. Making, not taking, the Best of N. In *ICLR*, 2026.

Kirichenko, P., Ibrahim, M., Chaudhuri, K., and Bell, S. J. AbstentionBench: Reasoning LLMs fail on unanswerable questions. In *NeurIPS*, 2025.

Lee, M., Kim, K., Kim, T., and Park, S. Selective generation for controllable language models. In *NeurIPS*, 2024.

Liu, B., Wang, L., Lyu, C., Zhang, Y., Su, J., Shi, S., and Tu, Z. On the cultural gap in text-to-image generation. In *ECAI*, 2024.

Malakouti, S., Gong, B., and Kovashka, A. Culture in action: Evaluating text-to-image models through social activities. In *ICLR*, 2026.

Manski, C. F. *Partial Identification of Probability Distributions*. Springer, 2003.

Manvi, R., Hong, J., Seyde, T., Labonne, M., Lechner, M., and Levine, S. Zero-overhead introspection for adaptive test-time compute. In *ICLR*, 2026.

Mohri, C. and Hashimoto, T. Language models with conformal factuality guarantees. In *ICML*, 2024.

Nayak, S., Bhatia, M., Zhang, X., Rieser, V., Hendricks, L. A., van Steenkiste, S., Goyal, Y., Stanczak, K., and Agrawal, A. CulturalFrames: Assessing cultural expectation alignment in text-to-image models and evaluation metrics. In *Findings of EMNLP*, 2025.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. Conformal language modeling. In *ICLR*, 2024.

Rabanser, S. and Papernot, N. What does it take to build a performant selective classifier? In *NeurIPS*, 2025.

Rege, A., Nie, Z., Ramesh, M., Raskar, U., Yu, Z., Kusupati, A., Lee, Y. J., and Vinayak, R. K. CuRe: Cultural gaps in the long tail of text-to-image systems. In *ICCV*, 2025.

Snell, C., Lee, J., Xu, K., and Kumar, A. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *ICLR*, 2025.

Venkatesh, K., Dalva, Y., Lourentzou, I., and Yanardag, P. RAVEL: Rare concept generation and editing via graph-driven relational guidance. arXiv:2412.09614, 2025.

Wang, Q., Fan, Y., and Wang, X. E. SAFER: Risk-constrained sample-then-filter in large language models. In *ICLR*, 2026.

Yin, T., Ton, J.-F., Guo, R., Yao, Y., Liu, M., and Liu, Y. Fair classifiers that abstain without harm. In *ICLR*, 2024.

Yu, Z., Wu, S., and Block, A. From curiosity to caution: Mitigating reward hacking for Best-of-N with pessimism. In *ICLR*, 2026.

# A Proofs and Selection Diagnostics

## A.1 Proof of the pairwise-to-top-one bounds

For each request, let the human order be unique. If the selector misses the human-best candidate, at least one pairwise comparison involving that candidate must be wrong. Therefore the per-request number of wrong comparisons is at least one on every miss, giving $1 - H \leq C(1 - A)$ and hence $H \geq 1 - C(1 - A)$. Conversely, on every top-one hit the selected candidate must beat each of the other $K - 1$ candidates, so the total number of correct pairwise comparisons is at least $(K - 1)H$. Since the expected number of correct comparisons is $CA$, we obtain $(K - 1)H \leq CA$, yielding the upper bound. Sharpness follows by constructing mixtures of requests with exactly one wrong comparison on misses and exactly $K - 1$ correct comparisons on hits, with the remaining pairwise outcomes assigned to attain the desired average.

## A.2 Corrected proof of the selected-only bounds

**A.2 Corrected proof.** Fix $K$ and the selected failure mean $\bar{y}_s$. Let $U = \sum_{k \neq k^*} Y_k$ be the total failure mass of the unselected candidates. Since the pool mean is $(\bar{y}_s + U)/K$,

$$\Delta = \bar{y}_s - \frac{\bar{y}_s + U}{K} = \frac{K - 1}{K}\bar{y}_s - \frac{U}{K}.$$

To minimize $\Delta$, maximize $U$ by setting $Y_k = 1$ for every unselected candidate; feasibility requires only that the selected mean remains $\bar{y}_s$. This gives

$$\Delta_{\min} = \frac{K - 1}{K}\bar{y}_s - \frac{K - 1}{K} = -\frac{K - 1}{K}(1 - \bar{y}_s).$$

To maximize $\Delta$, minimize $U$ by setting $Y_k = 0$ for every unselected candidate, giving

$$\Delta_{\max} = \frac{K-1}{K}\bar{y}_s.$$

Both constructions are feasible, so the bounds are sharp.

### A.3 Monotone recalibration invariance

Let $g$ be strictly increasing. Then $\arg\max_k g(s(y_k)) = \arg\max_k s(y_k)$. Thus every within-pool selection is unchanged by monotone recalibration. A threshold on $g \circ s$ corresponds to an upper set of the original score and can change the emission event, but it cannot change which candidate wins an emitted pool.

### A.4 Corrected flattening statement (proof of Remark 1)

*Remark* 1 (Observational equivalence under flattened control). Let $L = Y_{k^*}$ be the flattened per-request loss on emitted requests. Conditional on the emission event $E$, the distribution of $L$ depends on the component triple only through $R_{\text{sel}} = I + M(1-\gamma)$: for any two triples $(I, M, \gamma) \neq (I', M', \gamma')$ with $I + M(1-\gamma) = I' + M(1-\gamma')$, the induced conditional loss distributions given $E$—and hence the certificates, power, and accept/reject decisions of any procedure that observes only $(E, L)$ and certifies $\Pr(L = 1 \mid E) \leq \alpha$—are identical conditional on $E$. They need not induce the same joint distribution of $(E, L)$ unless $\Pr(E)$ is also held fixed.

**A.4 Corrected flattening statement.** On the emission event $E$, the flattened loss is $L = Y_{k^*} \in \{0, 1\}$ and

$$\Pr(L = 1 \mid E) = R_{\text{sel}} = I + M(1-\gamma)$$

by Proposition 1. A Bernoulli law is determined by its conditional mean, so any two triples with the same $R_{\text{sel}}$ induce the same conditional distribution of $L \mid E$, and therefore the same conditional risk-control decision for any procedure that certifies $\Pr(L = 1 \mid E) \leq \alpha$. They need not induce the same joint distribution of $(E, L)$ unless $\Pr(E)$ is also held fixed. Threshold changes alter $\Pr(E)$; hence policies with different thresholds should be compared as (coverage, selected risk) pairs, not as identically distributed $(E, L)$ tuples unless emission probabilities are matched.

## B Exact Binomial Intervals and Gates

For a binomial count $x$ out of $n$, the two-sided Clopper–Pearson interval is obtained by inversion of exact binomial tails. For an upper one-sided bound at level $1 - \delta$ we use the corresponding beta quantile, $U_{1-\delta}(x, n) = \text{Beta}^{-1}(1 - \delta;\, x + 1,\, n - x)$, and analogously for lower bounds $L_{1-\delta}(x, n) = \text{Beta}^{-1}(\delta;\, x,\, n - x + 1)$. The accept/reject decision uses a single exact upper bound on the total selected-failure count $x = f + o$ and therefore involves no multiplicity adjustment: accept iff $U_{1-\delta}(x, n) \leq \alpha$. Attribution is separate and diagnostic: component claims use Bonferroni-adjusted one-sided levels when multiple claims are combined, with levels $\delta_f, \delta_o$ satisfying $\delta_f + \delta_o \leq \delta$ (with two component claims at familywise level $0.05$, each uses a one-sided 97.5% interval). Declare feasibility failure iff $L_{1-\delta_f}(f, n) > \alpha$ and ordering failure iff $L_{1-\delta_o}(o, n) > \alpha$; ordering failures alone are a lower bound on selected risk, since safe pools never fail and infeasible pools always fail. In particular, an ordering failure can be established without estimating $\gamma_{\min}$ by counting mixed-pool ordering errors directly. These diagnostics do not certify acceptability, and no plug-in combination of $\hat{I}, \hat{M}, \hat{\gamma}$ is used to accept.

## C Audit Cohort and Implementation Details

### C.1 CulturalFrames cohort flow

Table 3 summarizes the full pool construction for every audit row before the development/evaluation split. The released cultural set contains 981 requests and 10,412 rater judgments. The audit cohort construction is deterministic and label-independent: retain complete $K = 4$ candidate pools; require

Table 3: Full pool construction before the development/evaluation split.

| Dataset | Score | Endpoint | $K$ | Pools | I/M/S |
|---|---|---|---|---|---|
| Cultural | T2I SigLIP cosine | alignment $\geq .75$ | 4 | 203 | 17/154/32 |
| RewardBench | SkyworkRM | preference | 2 | 2,985 | 0/2,985/0 |
| HumanEvalPack | SkyworkRM | preference | 2 | 984 | 0/984/0 |
| LiveCodeBench | Qwen3 log-prob. | all tests pass | 10 | 264 | 110/70/84 |
| ScienceQA | Qwen log-prob. | answer key | 2 | 500 | 0/500/0 |

complete request-level endpoint aggregation; require all frozen verifier scores needed for the policy; remove duplicate or incomplete records. This produces 203 audit pools, with 17 infeasible, 154 mixed, and 32 safe pools before the split (Table 2). A request-level hash assigns 20% to development and 80% to evaluation. The development portion is used once for arm and threshold selection; the evaluation portion is untouched until scoring. Per-stage retained/excluded counts that cannot be recovered from the released artifacts are marked "tbc" (to be completed from audit logs) in Table 2 and will be filled from the frozen pipeline logs before any camera-ready claim that depends on them.

### C.2 Threshold and repair rules

On each development split, select the largest score threshold attaining nominal 50% coverage. The Qwen repair crosses pointwise, pairwise, and listwise objectives with retrieval on/off; the arm is selected on development only. The sealed evaluation set is not scored for the second-shot protocol. A staged second-shot re-opening rule requires a development-only verifier to reach a prespecified predicted opening probability, complete retrieval, and a timestamped policy before any sealed evaluation access.

### C.3 Synthetic validation and frontier: generating processes and seeds

Table 4: Synthetic validation with known ground truth, $K = 4$, $n = 2{,}000$ emitted evaluation requests per regime, $\alpha = 0.10$. Acceptance uses only the one-shot gate $U_{95}(x, n) \leq \alpha$. Attribution uses Bonferroni one-sided 97.5% lower bounds ($\delta_f = \delta_o = 0.025$): a component is blamed only when its lower bound crosses its stated gate. "Budget" is the remaining risk budget $\alpha - \hat{I}$ available to the ordering component, shown for reference. In regime O, flattened per-request risk control on the same data yields the identical $U_{95}$ (16.4%) and the same rejection, but no attribution.

| Regime | true $(I, M, \gamma)$ | true $R_{\text{sel}}$ | $\hat{I}$ | $\hat{M}$ | $\hat{\gamma}$ [95% CI] | $\gamma_{\min}$ | Risk ($U_{95}$) | Gate decision |
|---|---|---|---|---|---|---|---|---|
| P: pass | (.03,.60,.95) | .060 | .031 | .612 | .949 [.935,.960] | .887 | 6.3% (7.2%) | open |
| O: ordering | (.03,.60,.80) | .150 | .036 | .589 | .806 [.783,.829] | .891 | 15.0% (16.4%) | closed: ordering* |
| F: feasibility | (.15,.60,.97) | .168 | .152 | .605 | .974 [.964,.983] | infeas. | 16.8% (18.2%) | closed: feasibility† |

*ordering-error lower 97.5% bound $= 10.0\%$, at the $\alpha$ gate and above the remaining budget $6.4\%$. †lower 97.5% bound on $I = 13.7\% > \alpha$.

**Validation regimes (Table 4).** For each request, draw the pool type i.i.d. from (infeasible, mixed, safe) with probabilities $(I, M, 1 - I - M)$. Infeasible pools have all $K = 4$ candidates failing; safe pools have all passing; mixed pools assign a uniformly random count $J \sim \text{Uniform}\{1, \ldots, K - 1\}$ of passing candidates to uniformly random positions. On a mixed pool, the selector selects an acceptable candidate with probability $\gamma$ (uniform among acceptable ones), else a failing one (uniform among failing ones); on infeasible and safe pools the choice does not affect the label. All requests emit. Regimes: P $(I, M, \gamma) = (.03, .60, .95)$, O $(.03, .60, .80)$, F $(.15, .60, .97)$; $n = 2{,}000$ evaluation requests per regime; RNG seeds 11, 22, 33 respectively. Realized pool counts: P $= 62/1224/714$, O $= 72/1177/751$, F $= 304/1209/487$ (infeasible/mixed/safe).

**Frontier (Figure 1).** $n = 20{,}000$ requests, $K = 4$, pool types as above with $I = 0.03$, $M = 0.60$. Candidate scores are $s_k = \mu(Y_k) + \varepsilon_k$ with $\mu(0) = 1$, $\mu(1) = 0$, $\varepsilon_k \overset{\text{iid}}{\sim} \mathcal{N}(0, 0.75^2)$; the policy emits iff $\max_k s_k \geq \tau$ and selects the argmax. The induced within-mixed-pool success is $\gamma \approx 0.83$; the

full-emission selected risk is 13.2%. We sweep 60 thresholds spanning the 2nd–98th percentiles of the pool-max distribution and report only operating points with at least 200 emitted requests. RNG seed 7.

**Consistency check.** In all three validation regimes, the plug-in decomposition $\hat{I}+\hat{M}(1-\hat{\gamma})$ equals the directly estimated selected risk up to rounding, as required by Proposition 1.

## D Additional Cultural Diagnostics

A separate CuRe audit covers 28,800 images. A small observed FLUX selected-minus-pool gain of 0.0006 does not transfer to SD 1.5 or OpenCLIP-H; candidate-budget, automatic-QA, and frozen FLUX.2[klein] selector intervals include zero. These tests are bounded transfer diagnostics rather than universal claims about verifier quality.

## E Extended Release-Contract Discussion

Gate 1 is a measurement and governance question that cannot be settled by internal calibration; our cultural endpoint deliberately leaves it open, and we do not claim that any benchmark label constitutes community authorization. Gate 5 is an engineering/provenance question: in our RewardBench sanity check, the frozen policy is replayed under a hash-bound schema, but we do not claim a complete runtime enforcement evaluation with failure injection, revocation, or stale artifacts.

## F Extended Related Work

**Cultural T2I evaluation.** CulturalFrames, CuRe, C3, CULTIVATE, and related work expose cultural and long-tail gaps in text-to-image generation (Nayak et al., 2025; Rege et al., 2025; Liu et al., 2024; Malakouti et al., 2026). Retrieval and iterative refinement can improve candidate generation (Jeong et al., 2025; Venkatesh et al., 2025; Jiang et al., 2026); our analysis asks the distinct question of whether a frozen verifier selects the acceptable member of the already-emitted pool, and whether the resulting system can be certified for release.

**Selective prediction and the reject option.** Risk–coverage tradeoffs date to Chow's reject-option formulation (Chow, 1957; 1970) and were formalized for noise-free selective classification by El-Yaniv and Wiener (2010). Deep selective classifiers integrate rejection into prediction (Geifman and El-Yaniv, 2017; 2019). Recent work revisits optimal abstention through likelihood-ratio tests (Heng and Soh, 2026), while broader work studies abstention under fairness or limited-selectivity constraints (Yin et al., 2024; Casacuberta and Kanade, 2025). Our decomposition is consistent with this lineage: conditional on a frozen score and threshold, selective risk is fixed. The pool-structured generative setting adds a request-level candidate set and therefore an infeasible-pool event that has no analogue when a single hypothesis always produces an output.

**Conformal and distribution-free risk control.** Learn-Then-Test, distribution-free risk-controlling prediction sets, and conformal risk control provide finite-sample guarantees for threshold or policy selection (Angelopoulos et al., 2021; Bates et al., 2021; Angelopoulos et al., 2022; Angelopoulos and Bates, 2023). Conformal language modeling and conformal factuality extend distribution-free guarantees to generation (Quach et al., 2024; Mohri and Hashimoto, 2024). These methods are complementary: they control risk under specified scores, policies, and endpoints. Our claim is narrower and concerns the informal use of calibration or ranking metrics as evidence that a selective-generation system is releasable before feasibility and ordering have been established; Section 2.5 makes precise what such flattened control cannot recover.

**Selective generation and selective evaluation.** Recent work studies selective generation for controllable language models, guided abstention, and risk-constrained sample-then-filter pipelines. Lee et al. (2024) use conformal ideas to control an entailment-based generation error rate; Sufficient Context studies abstention for RAG systems when the available context is inadequate (Joren et al., 2025); and SAFER explicitly separates sampling and filtering risk (Wang et al., 2026). Cascaded selective

evaluation and related judge-calibration work provide guarantees of human agreement through selective evaluation and escalation (Jung et al., 2025). Our focus is the complementary post-generation question: given a frozen candidate pool and selector, can the resulting selection risk be diagnosed and certified, and can one tell whether failure comes from the pool or the selector?

**Best-of-N, test-time scaling, and reward-model selection.** Test-time compute scaling makes multi-candidate selection a central deployment pattern (Snell et al., 2025). Recent theoretical and empirical work shows that Best-of-$N$ can suffer reward hacking as the candidate budget grows, and studies pessimistic alternatives (Huang et al., 2025; Yu et al., 2026). FusioN instead argues that candidate pools need not be collapsed by a single winner and proposes combining information across candidates (Khairi et al., 2026); adaptive test-time compute similarly asks when to continue sampling rather than committing to a fixed budget (Manvi et al., 2026). Reward-model overoptimization provides an earlier warning that proxy scores can drift from target outcomes (Gao et al., 2023). Our identification results formalize one evaluation gap in this setting: even accurate pairwise scores do not, by themselves, identify the final top-one hit rate, and a high-scoring pool is not evidence that an acceptable candidate was present or selected.

**Evaluation reliability and uncertainty.** LLM-judge work increasingly treats confidence and selective escalation as first-class components of evaluation (Jung et al., 2025). Recent benchmarks also show that abstention remains difficult for reasoning models, including on unanswerable or underspecified questions (Kirichenko et al., 2025). We use these results as motivation for keeping endpoint validity and verifier reliability separate: our certificate is conditional on an authorized endpoint and does not treat judge confidence as ground-truth validity.

**Partial identification and selective inference.** The bounds in Eqs. (3)–(4) are partial-identification statements in the spirit of Manski (2003); interval construction for partially identified parameters follows Imbens and Manski (2004). The development/evaluation separation follows the data-splitting tradition (Cox, 1975; Fithian et al., 2014).

## G Reproducibility Checklist

- Development/evaluation split is request-level and fixed before evaluation access.
- Main risk intervals are exact one-sided binomial bounds; acceptance uses the total selected-failure count $x = f + o$ with gate $U_{1-\delta}(x, n) \leq \alpha$.
- Every $\gamma$ estimate reports its mixed-pool denominator.
- Cultural cohort construction is fixed before held-out evaluation and excludes incomplete/duplicate records; the cohort flow reconciles to the frozen 203-pool audit cohort (Table 2).
- Qwen repair selection is development-only; its sealed test remains unscored.
- Outcome-reusing threshold sweeps and human-best selections are diagnostic frontier traces and are never used as certificates.
- Synthetic validation and frontier experiments use fully specified generating processes with fixed seeds (Appendix C.3).
- Certified operating points are selected on development data; the frontier is reported as a coverage–risk claim, not a single point.