RIGHT PROOF, WRONG THEOREM: A RELATIONAL BOTTLENECK IN VERIFIER-GUIDED SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Best-of-N selection samples N candidate solutions and returns the one preferred by a verifier. It improves only if the selected candidate solves the requested problem. A coherent proof or program may instead solve a nearby problem, a failure we call a *relational mismatch*. We certify such errors in code and mathematics outputs using claim neighborhoods frozen before artifact inspection. Across two frozen generators, the upper score tail of our primary verifier contains fewer local defects but more relational mismatches. For the primary generator, raw best-of-N validity first rises and then falls: from 88.41% at $N = 32$ to 80.34% at $N = 256$ in code, and from 82.03% at $N = 16$ to 76.56% at $N = 64$ in mathematics. Exact finite-pool selection weights forecast these two curves with weighted mean absolute errors of 0.71 and 1.09 percentage points. A development-frozen interior-rank selector then improves fixed-budget validity by 4.49 to 8.98 percentage points across four domain–generator settings, without additional generation or verifier calls. Finally, in the primary code setting, relational training improves held-out claim–artifact binding and natural deployment; replacing ordinary relational ranking with a crossed relational objective adds 1.65 percentage points. These results show that recognizing a good artifact is insufficient: reliable verifier-guided scaling also requires binding the artifact to the claim it must satisfy.

1 INTRODUCTION

Test-time computation has become a scaling axis for reasoning systems (Guo et al., 2025; Kimi Team et al., 2025; Muennighoff et al., 2025; Snell et al., 2025). A strategy samples N candidate solutions, scores them with a learned verifier, and returns the highest-scoring candidate (Cobbe et al., 2021; Chen et al., 2021; Lightman et al., 2024). More samples increase the chance that a valid solution exists, but they improve the returned answer only if the verifier ranks that solution correctly.

We study a specific failure of this ranking. A proof may establish the converse of the requested implication or rely on an assumption absent from the theorem. A program may correctly implement a nearby boundary convention rather than the specified one. Such artifacts can be coherent and valid for another task while failing the task that was asked. We call this failure a *relational mismatch*: the artifact itself may be plausible, but its relation to the requested claim is wrong. This differs from a local defect such as an invalid inference, arithmetic mistake, runtime failure, or malformed program.

Relational mismatches are consequential under best-of-N selection. As N grows, selection concentrates on the verifier's score tail, where rare ranking errors can dominate deployment. This connects reward overoptimization (Manheim & Garrabrant, 2018; Gao et al., 2023) with partial-input shortcuts and contrastive evaluation (Gururangan et al., 2018; Poliak et al., 2018; Gardner et al., 2020; Ribeiro et al., 2020): a verifier may recognize a convincing artifact without binding it to the claim it must satisfy. Existing verifier and pass@k work primarily evaluates candidate-level discrimination or the existence of at least one successful candidate (Cobbe et al., 2021; Chen et al., 2021; Lightman et al., 2024). Our estimand instead concerns the validity of the artifact returned by a specified selector at a specified search budget. Throughout the paper, G1 denotes `Qwen/Qwen2.5-7B-Instruct`, G2 denotes `meta-llama/Llama-3.1-8B-Instruct`, and V0 denotes our primary trainable outcome verifier; their configurations are given in Sections 3 and G. A notation such as Code/G1/V0

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.

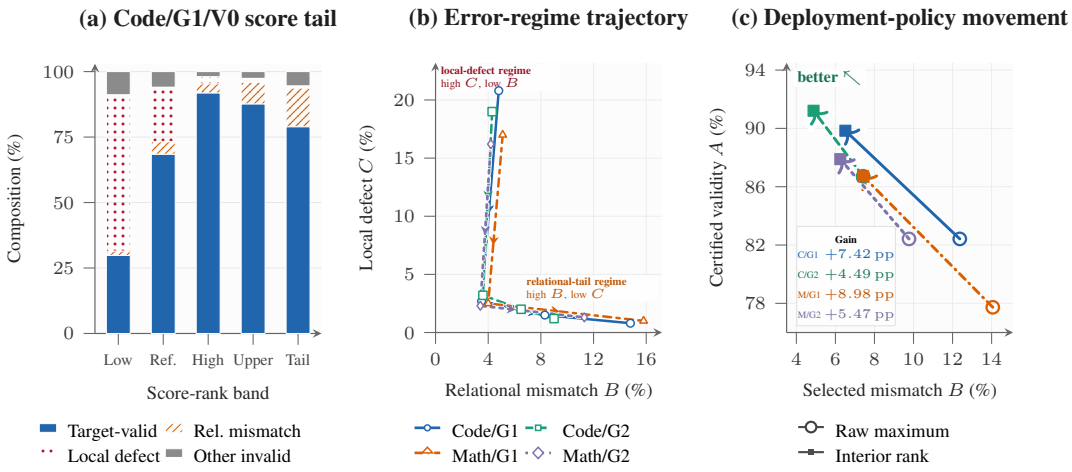


Figure 1: **Verifier-guided scaling enriches relational mismatches; interior-rank selection improves deployment validity.** Panel (a) shows score-rank composition for Code/G1/V0: local defects decrease while relational mismatches increase in the deepest tail. Panel (b) gives the error-regime trajectories for all four domain–generator settings. Panel (c) shows that the development-frozen interior-rank selector moves each deployment point toward higher validity and fewer mismatches. Open circles denote raw maximum selection and filled squares denote interior-rank selection.

identifies the domain, generator, and verifier, respectively. For the overview figure, A denotes certified target-valid artifacts, B denotes certified relational mismatches, and C denotes certified local defects; the mutually exclusive taxonomy is defined in Section 2. We use $N^\star$ for the primary deployment budget (128 in code and 32 in mathematics), and pp denotes percentage points.

We test this hypothesis on natural code and mathematics outputs from two frozen generators. For each target, neighboring claims are frozen before the corresponding artifacts or verifier scores are inspected. Across all four domain–generator settings, the verifier removes local defects more effectively than relational mismatches. Consequently, target validity peaks before the deepest score tail and declines thereafter (Figure 1). For the primary generator, raw best-of-N validity rises to 88.41% at $N = 32$ and falls to 80.34% at $N = 256$ in code. In mathematics, it rises to 82.03% at $N = 16$ and falls to 76.56% at $N = 64$. A bulk-matched comparison verifier achieves approximately eight percentage points higher validity at the primary deployment budgets, showing that similar average accuracy and AUROC can hide materially different search behavior. Our contributions are:

- **Relational-mismatch certification.** We distinguish target-valid artifacts, artifacts valid for frozen neighboring claims, local defects, and unresolved or residual errors.
- **Prospective deployment forecasting.** Exact finite-pool selection weights convert independently certified score-tail profiles into forecasts of selected validity under best-of-N.
- **Selection and training repairs.** A frozen interior-rank selector improves fixed-budget validity by 4.49 to 8.98 pp without additional generation or verifier calls. In the primary Code/G1/V0 training experiment, relational supervision improves both controlled binding and natural deployment.

The central deployment question is therefore not only whether a verifier recognizes a good artifact, but whether its ranking returns one that satisfies the requested claim.

2 Relational Validity and Selection

Let T be a requested mathematical claim or executable specification, let $P \sim \pi(\cdot \mid T)$ be a generated artifact, and let $S = s(T, P)$ be its verifier score. Unless stated otherwise, a best-of-N candidate block consists of $P_1, \ldots, P_N$ drawn conditionally independently and identically distributed from $\pi(\cdot \mid T)$, together with independent saved tie randomizers. Semantic validity is the relation

$$V(T, P) \in \{0, 1\},$$

rather than a property of P alone. The certification protocol records

$$\widehat{V}(T, P) \in \{1, 0, ?\},$$

where 1 and 0 denote resolved valid and invalid judgments under the protocol, and ? denotes unresolved target validity. Each artifact is assigned one of five mutually exclusive recorded outcomes:

$$K \in \mathcal{K} = \{A, B, C, U_0, U_?\}.$$

Class A has $\widehat{V}(T, P) = 1$. Class B has $\widehat{V}(T, P) = 0$ but, without modifying the artifact, is certified valid for at least one frozen neighboring target. Class C has a local defect that invalidates it under the target and every registered applicable neighbor. Class U_0 contains other resolved target-invalid artifacts, and class $U_?$ has unresolved target validity. For each target, a typed grammar enumerates a finite neighborhood $\mathcal{N}(T)$ before the artifacts or verifier scores are inspected. Its transformations include implication direction, quantifiers, hypotheses, conclusions, domains, boundaries, and neighboring executable contracts. A relational-mismatch certificate requires

$$\widehat{V}(T, P) = 0 \quad \text{and} \quad \exists T' \in \mathcal{N}(T) : \widehat{V}(T', P) = 1, \tag{1}$$

with the artifact held fixed.

Example. Consider the requested claim

$$T: \qquad 8 \mid (n^2 - 1) \Longrightarrow n \text{ is odd.}$$

One natural candidate begins by writing $n = 2k + 1$ and derives

$$n^2 - 1 = 4k(k + 1),$$

which is divisible by 8 because one of k and $k + 1$ is even. The argument is mathematically sound, but it assumes that n is odd and proves the reverse implication. It therefore does not establish T. Before this candidate was generated, the implication-direction neighbor

$$T': \qquad n \text{ is odd} \Longrightarrow 8 \mid (n^2 - 1)$$

had already been registered. The unchanged artifact is invalid for T and valid for T', so it receives class B. V0 assigns the invalid target pairing score 0.9737 and the valid neighbor pairing score 0.9698. Both implications happen to be true, but a proof of one direction is not a proof of the other. This example illustrates a relational error that artifact quality alone cannot identify.

The complete grammar, decision order, and evidence requirements appear in Section D.

For each target T, let

$$F_T(s) = \mathbb{P}(S \le s \mid T), \qquad F_T(s^-) = \lim_{u \uparrow s} F_T(u).$$

To compare scores across targets and handle score atoms, define the randomized within-target score rank

$$Q = F_T(S^-) + Z\{F_T(S) - F_T(S^-)\}, \qquad Z \sim \mathrm{Unif}[0, 1],$$

where Z is independent of the candidate record conditional on T. Then $Q \mid T \sim \mathrm{Unif}[0, 1]$. For almost every $q \in [0, 1]$, let

$$\eta_c(q, T) \coloneqq \mathbb{P}(K = c \mid Q = q, T)$$

denote a version of the regular conditional class probability. Define the target-averaged class profile

$$\bar{\eta}_c(q) \coloneqq \mathbb{E}_T[\eta_c(q, T)]\,.$$

All statements involving $\eta_c(q, T)$ are understood up to almost-everywhere equivalence in q. For a score-rank band I, define $p_c(I) \coloneqq \mathbb{P}(K = c \mid Q \in I)$. In tables, A, B, and C also abbreviate their respective band prevalences. Thus $B/(1 - A)$ means $p_B(I)/\{1 - p_A(I)\}$, with unresolved $U_?$ cases retained in the denominator. The reference band is $[.5, .875)$ in both domains. The deepest tail is $[.9921875, 1]$ in code and $[.96875, 1]$ in mathematics; the peak band is $[.875, .96875)$ in both. Unless specified otherwise, Δ denotes tail minus reference, whereas peak-to-tail validity uses the peak band as its baseline. Reported differences are in percentage points. Reported deployment validity counts certified class-A selections; unresolved artifacts are not thereby declared semantically invalid.

Raw best-of-N selection returns the maximum rank, so

$$R_c^{\mathrm{raw}}(N) = N \int_0^1 \bar{\eta}_c(q) q^{N-1}\, dq. \tag{2}$$

As N grows, the weight Nq^{N-1} concentrates near $q = 1$. Search can therefore amplify an upper-tail error that is nearly invisible to candidate-level averages.

If target validity is nondecreasing in score rank, raw best-of-N validity is nondecreasing; a decline therefore identifies a failure of the ranking rather than candidate availability. Conversely, arbitrarily similar bulk score–label laws can have opposite large-N deployment behavior. Formal statements and proofs are given in Propositions 2 and 3. Our primary deployment estimand is

$$R_A^{\rho}(N) = \mathbb{P}\{\text{policy } \rho \text{ returns a class-}A \text{ artifact from } N \text{ candidates}\}.$$

Pass@k records whether the candidate set contains a valid output. The quantity $R_A^{\rho}(N)$ records whether the deployed policy selects one.

3 Experimental Design

Systems and source provenance. We study unmodified outputs from two frozen general-purpose instruction models: G1 is `Qwen/Qwen2.5-7B-Instruct` (Qwen et al., 2024), and G2 is `meta-llama/Llama-3.1-8B-Instruct` (Grattafiori et al., 2024). Both generate code and natural-language mathematics solutions under the same fixed sampling configuration. V0 is a trainable Qwen2.5-7B outcome verifier and is the primary system for tail analysis, forecasting, selection, and training. V1 is an independently trained Llama-3.1-8B verifier selected using only development-set bulk metrics and serves as a bulk-matched comparator. The auxiliary API-based scorers introduced in Section G use GPT-4o and GPT-4o mini; these endpoints are documented in the GPT-4o system card and the official GPT-4o mini model announcement (OpenAI, 2024a;b). The code sources used for task construction or verifier training include MBPP (Austin et al., 2021), HumanEval (Chen et al., 2021), and APPS (Hendrycks et al., 2021a). The mathematics sources include MATH (Hendrycks et al., 2021b), PRM800K (Lightman et al., 2024), miniF2F (Zheng et al., 2022), and the Lean mathematical library (The mathlib Community, 2020). These source references describe provenance; the reported code results are not presented as official scores on the original public benchmarks. Architectures, decoding parameters, prompts, training sources, and checkpoint selection are detailed in Section G.

Data separation. Development data determine policies, hyperparameters, and checkpoint selection. Independent forecast pools provide score-stratified certification labels. Frozen forecasts and policies are then evaluated on newly generated, source-group-disjoint prospective targets. A source group contains all rewrites, neighboring claims, templates, and relational blocks derived from one source object. Prospective outcomes are never used to fit a forecast or select a checkpoint.

The code forecast pools contain 128 G1 targets and 64 G2 targets, each with 2,048 fully certified candidates. The mathematics pools contain 64 and 48 targets with 512 candidates each; 4,096 and 3,072 candidates are labeled under known inclusion probabilities. The deepest score strata contain 2,048 and 1,024 certified artifacts in code and 1,024 and 768 in mathematics. At $N^\star$, the corresponding inclusion-probability-adjusted selection-weight effective-support diagnostics for raw maximum selection are 4,078, 2,034, 1,176, and 870; these describe weight concentration rather than independent target counts. A *selection unit* is one target–repetition candidate block on which a policy returns exactly one artifact. Prospective evaluation contains 768, 512, 256, and 256 such selection units for Code/G1, Code/G2, Mathematics/G1, and Mathematics/G2, respectively. Code uses

$$N \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}, \qquad N^\star = 128,$$

and mathematics uses

$$N \in \{1, 2, 4, 8, 16, 32, 64\}, \qquad N^\star = 32.$$

Smaller budgets are prefixes of the same candidate blocks, enabling paired comparisons.

Certification. Code artifacts are evaluated under frozen parsing, compilation, termination, interface, and bounded-domain execution procedures. The unchanged program is tested against the target and applicable neighboring contracts. Mathematics uses independent target-first assessment and

Table 1: **Relational mismatches replace local defects in the score tail, and interior-rank selection improves deployment.** The first three rows compare the deepest tail with $Q \in [.5, .875)$; the fourth compares tail with peak-band validity. Brackets are simultaneous two-sided clustered 95% intervals. The final row gives prospective gains at $N^\star$, with one-sided clustered 95% lower bounds. All effects are percentage points.

	Code/G1	Code/G2	Math/G1	Math/G2
Local defects: ΔC	-20.0 $[-23.84, -16.42]$	-17.8 $[-21.82, -13.91]$	-16.0 $[-21.14, -11.08]$	-14.9 $[-20.61, -9.46]$
Mismatches: ΔB	$+10.0$ $[6.79, 13.31]$	$+4.7$ $[1.13, 8.24]$	$+10.7$ $[5.61, 15.69]$	$+7.1$ $[1.42, 12.73]$
Mismatch share: $\Delta\{B/(1-A)\}$	$+55.0$ $[43.18, 65.27]$	$+44.1$ $[31.12, 56.03]$	$+47.4$ $[33.46, 59.89]$	$+42.1$ $[26.73, 56.41]$
Peak-to-tail validity: ΔA	-12.9 $[-17.61, -8.36]$	-6.1 $[-10.68, -1.57]$	-14.7 $[-21.31, -8.24]$	-11.0 $[-18.39, -3.91]$
Interior-rank gain [LCB]	$+7.42$ $[4.61]$	$+4.49$ $[0.82]$	$+8.98$ $[3.42]$	$+5.47$ $[0.73]$

adjudication. Reviewers do not see verifier identity, scores, strata, or neighboring claims until target invalidity has been established. Unresolved target validity remains in $U_?$. Code forecast pools are fully labeled. Mathematics candidates are sampled for labeling within target-specific score strata, and their recorded inclusion probabilities enter the deployment estimator. Complete strata, label support, execution controls, annotation rules, and experiment chronology appear in Sections D and F.

Inference. Tail contrasts use simultaneous two-sided target- or source-group-clustered 95% confidence intervals. Policy and training comparisons use the stated one-sided clustered 95% lower confidence bounds (LCBs). Nested budgets and policies acting on the same candidate sets are resampled jointly. All confirmation endpoints, budget grids, and evaluated policies were frozen before the corresponding prospective outcomes were generated.

4 Relational Mismatches in the Score Tail

4.1 Tail substitution

V0 removes local defects before it resolves claim–artifact mismatches. In Code/G1, class C falls from 20.80% in the reference band $Q \in [.5, .875)$ to 0.80% in the deepest tail, while class B rises from 4.80% to 14.80%. Validity peaks at 91.80% in an interior band and falls to 78.90% in the tail. Mathematics/G1 exhibits the same substitution: C falls from 17.00% to 1.00%, B rises from 5.10% to 15.80%, and validity falls from a peak of 89.50% to 74.80%.

The joint geometry is visible in Figure 1(b). The initial movement is predominantly vertical: large reductions in class C occur before class B becomes strongly enriched. The trajectories then turn toward the mismatch-dominated lower-right region, showing that the extreme tail is not merely less defective; it contains a different kind of error. Across the four settings, the trajectories show a similar shift from local defects toward relational mismatches.

Table 1 reports the pre-specified contrasts in all four settings. The simultaneous intervals exclude zero in the expected direction in all four settings. Code/G1 and both mathematics settings meet all four registered effect-size thresholds; Code/G2 meets the directional criteria but not the magnitude thresholds for mismatch enrichment and peak-to-tail validity decline. The mismatch share $B/(1-A)$ retains unresolved mathematics cases in its denominator. Its increase shows that relational mismatches explain a growing share of all outputs not certified as valid.

The enrichment spans several transformation families rather than a single edit type. In Code/G1, relation or boundary substitutions rise from 1.71% to 5.86%, while neighboring-object or contract mismatches rise from 0.96% to 4.47%. In Mathematics/G1, implication-direction, strengthened-hypothesis, and domain substitutions increase. The aggregate increase is driven by several transformation families, with small declines in a few others. Decompositions appear in Section H.2.

Interpreting the tail shift. The classes describe outcomes certified by our protocol, not an exhaustive taxonomy of reasoning errors. The increase in class B requires high-scoring artifacts to be certified as valid for neighbors registered before artifact inspection; it does not follow from the class decision order. The binding interpretation is further supported by artifact-fixed and crossed-block comparisons, which test the verifier's response to the target while holding the artifact fixed. Relational training improves these controlled comparisons and reduces class-B selections in natural deployment. These findings support incomplete target–artifact binding, although tail composition alone cannot distinguish that mechanism from greater coherence among high-scoring near-misses.

4.2 Best-of-N validity turns over

The tail substitution changes deployment. For Code/G1, raw validity rises from 52.47% at $N = 1$ to 88.41% at $N = 32$, then falls to 82.42% at $N = 128$ and 80.34% at $N = 256$. For Mathematics/G1, it rises from 53.52% at $N = 1$ to 82.03% at $N = 16$, then falls to 77.73% at $N = 32$ and 76.56% at $N = 64$. Because the candidate sets are nested, every larger budget retains the earlier candidates; the decline is caused by ranking rather than availability. Relational mismatches are the largest resolved error class among selected outputs at $N^\star$. They account for 70.37%, 55.88%, 69.23%, and 60.98% of resolved invalid selections in Code/G1, Code/G2, Mathematics/G1, and Mathematics/G2.

4.3 Direct evidence for incomplete binding

Artifact-fixed comparisons ask whether V0 scores a natural class-B artifact more highly under the frozen claim it actually satisfies than under the requested target it fails. For a class-B artifact P, requested target T, and certified neighbor T', define rescue credit as

$$r(T, T', P) = \mathbf{1}\{s(T', P) > s(T, P)\} + \frac{1}{2}\mathbf{1}\{s(T', P) = s(T, P)\}.$$

Rescue accuracy is the average of this credit over the registered class-B frame. For a symmetric relational block with valid diagonal pairs (T_1, P_1), (T_2, P_2) and invalid off-diagonal pairs (T_1, P_2), (T_2, P_1), claim-fixed (CF) accuracy compares artifacts while holding the claim fixed, and artifact-fixed (AF) accuracy compares claims while holding the artifact fixed. Pairwise CF and AF comparisons award half credit to score ties. Strict paired-block accuracy (PBA) is

$$\mathbf{1}\{\min(\ell_{11}, \ell_{22}) > \max(\ell_{12}, \ell_{21})\},$$

so a block is correct only when every valid diagonal cell scores above every invalid off-diagonal cell. In code, V0 obtains 63.33% rescue accuracy and 28.33% strict paired-block accuracy, compared with 89.17% and 76.67% for a reference judge. In mathematics, the results are 62.50% and 27.50%, versus 87.50% and 75.00%. The artifact-only scorer has zero strict block accuracy because its score cannot change with the claim. V0 performs better than this control but substantially worse than the reference judge, indicating partial and inconsistent claim sensitivity on these blocks.

Bulk metrics do not reveal the deployment gap. In code, V0 and V1 have bulk accuracies of 75.58% and 74.86% and areas under the receiver operating characteristic curve (AUROCs) of 0.776 and 0.783, yet achieve 82.42% and 90.10% validity at $N = 128$. In mathematics, their accuracies are 72.18% and 72.91% and their AUROCs are 0.739 and 0.733, while deployment validity is 77.73% and 85.94% at $N = 32$. The registered equivalence margins are $1.50\,\mathrm{pp}$ for accuracy and 0.015 for AUROC. For a two-sided confidence interval $[L, U]$ for a verifier difference, define its confidence-envelope magnitude as

$$\mathcal{E}([L, U]) \coloneqq \max\{|L|, |U|\}.$$

The registered bulk-matching criterion requires this magnitude to lie within the equivalence margin. For V0 minus V1, the code differences are $0.72\,\mathrm{pp}$ in accuracy and -0.007 in AUROC, with confidence-envelope magnitudes of $1.32\,\mathrm{pp}$ and 0.012, respectively. The mathematics differences are $-0.73\,\mathrm{pp}$ and 0.006, with confidence-envelope magnitudes of $1.41\,\mathrm{pp}$ and 0.013. Both comparisons therefore satisfy the bulk-matching criteria. Brier scores were not included in the equivalence criterion. A strictly increasing recalibration preserves score ordering and therefore leaves raw best-of-N decisions unchanged.

Independent audits confirm $286/300$ natural class-B labels in code, or 95.33% with a 95% Wilson interval of $[92.33, 97.20]\%$, and $282/300$ in mathematics, or 94.00% with an interval of

$[90.72, 96.17]\%$. A mathematics target-validity audit confirms $390/400$ decisions. Score-stratified audits address whether class-B certification becomes easier in the high-score tail. All four tail-minus-reference recall intervals lie within the registered $[-10, 10]$ pp equivalence region. In the precision audit, tail-minus-reference differences are -0.28 pp in code, with a Newcombe interval of $[-4.11, 3.45]$ pp, and 0.45 pp in mathematics, with an interval of $[-4.81, 8.37]$ pp. Precision and recall are similar across the audited score strata. These diagnostics are not deployment-weighted error-rate estimates. Audit frames, class profiles, and family-level counts appear in Section H.

5 Forecasting and Repairing Selection

5.1 Finite-pool deployment forecasts

A forecast pool can predict best-of-N deployment without repeatedly sampling candidate subsets. For a pool of M candidates ordered from lowest to highest under a frozen ranking, the probability that pool rank j becomes the k-th ascending member of a uniformly sampled size-N subset is

$$w_j^{(k)}(N; M) = \frac{\binom{j-1}{k-1}\binom{M-j}{N-k}}{\binom{M}{N}}. \tag{3}$$

Raw maximum selection uses $k = N$; an interior-rank selector uses its declared k_N. Let $a_t \geq 0$ denote the registered population weight of target t, normalized so that

$$\sum_t a_t = 1.$$

Unless otherwise stated, targets are weighted uniformly within the registered evaluation population. Let I_{ti} indicate whether candidate i for target t was labeled, let $\pi_{ti} > 0$ be its known inclusion probability, and let $w_{ti}^{\rho}(N)$ be its policy-selection weight. The selected rate of class c is estimated with the Horvitz–Thompson (HT) construction (Horvitz & Thompson, 1952):

$$\widehat{R}_c^{\rho}(N) = \sum_t a_t \sum_i \frac{I_{ti}}{\pi_{ti}} \mathbf{1}\{K_{ti} = c\} w_{ti}^{\rho}(N). \tag{4}$$

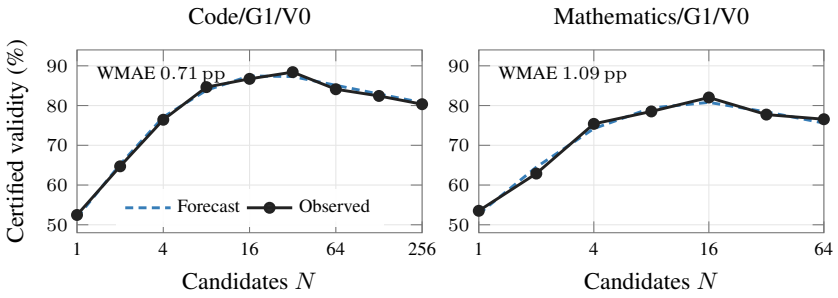
For certified code pools, $\pi_{ti} = 1$. Under conditionally i.i.d. generation and a frozen ordering, this estimator is design-unbiased for the full-pool subset average and, over pool generation, for fresh-N deployment (Proposition 4). Figure 2 compares forecasts frozen before generation with outcomes on new targets. The weighted mean absolute errors (WMAEs) are 0.71 pp for Code/G1/V0 and 1.09 pp for Mathematics/G1/V0, and both forecasts recover the interior optimum and subsequent decline. At $N^{\star}$, predicted versus observed mismatch rates are 12.15% versus 12.37% in code and 13.50% versus 14.06% in mathematics. The same pools also forecast the gains from interior-rank selection, as shown in Figure 2. Across all twelve domain–generator–verifier settings, design-based WMAE ranges from 0.71 to 2.91 pp (Table 21). Flexible binary prediction has lower realized WMAE in ten of twelve settings, while exact weighting performs best in the two primary V0 settings. The HT estimator provides a design-unbiased baseline under the registered sampling design. Monotone prediction misses the validity turnover in the primary V0 curves.

5.2 Interior-rank selection

The measured validity profile peaks before the extreme score tail. We therefore select the development-frozen ascending rank

$$k_N = \mathrm{clip}_{[1,N]}\left(\left\lfloor (N+1)\alpha + \frac{1}{2} \right\rfloor\right), \qquad \alpha = 0.9375. \tag{5}$$

Here, $\mathrm{clip}_{[1,N]}(x) = \min\{N, \max\{1, x\}\}$ truncates x to the interval $[1, N]$. At the primary budgets, this returns rank 121 of 128 in code and rank 31 of 32 in mathematics. Let τ be the deepest band's within-target rank mass. Under conditionally i.i.d. sampling, the number of candidates in that band is $\mathrm{Binomial}(N, \tau)$. Raw maximum enters the band if at least one candidate does; ascending rank k_N requires at least $N - k_N + 1$. At $N^{\star}$, raw versus interior-rank exposure is 63.4% versus $< 0.01\%$ in code ($\tau = 1/128,\ k_{128} = 121$), and 63.8% versus 26.4% in mathematics ($\tau = 1/32,\ k_{32} = 31$). The



Interior-rank gain at $N^\star$ (pp)	Code/G1	Code/G2	Math/G1	Math/G2
Forecast	7.90	3.85	7.50	4.80
Observed	7.42	4.49	8.98	5.47

Figure 2: **Independent tail measurements predict both prospective best-of-N validity and the benefit of changing selection.** Forecasts and policies were frozen before candidate generation. The lower panel compares predicted and observed interior-rank gains at $N^\star = 128$ for code and $N^\star = 32$ for mathematics. Complete counts and secondary settings appear in Tables 20 and 21.

policy uses the same candidates, scores, and tie randomizers as raw maximum selection. Prospective validity gains are 7.42 pp, 4.49 pp, 8.98 pp, and 5.47 pp in Code/G1, Code/G2, Mathematics/G1, and Mathematics/G2, respectively (Table 1). Figure 1(c) compares the selected class composition of the two policies. In every setting, the interior-rank selector has higher certified validity and selects fewer class-B artifacts. The largest change in error composition comes from the reduction in relational mismatches. Local defects and residual errors do not increase enough to offset this gain.

Class-A counts increase from 633 to 690 out of 768, from 444 to 467 out of 512, and from 199 to 222 and 211 to 225 out of 256. Mismatch counts fall from 95 to 50, 38 to 25, 36 to 19, and 25 to 16. Paired sensitivity analyses for unresolved mathematics cases retain positive lower bounds. The fixed-budget comparison is most relevant when the larger candidate sets are already generated or their coverage is valuable. When the generation budget is adjustable, smaller raw searches are competitive: their best observed results are 88.41% at $N = 32$ in Code/G1 and 82.03% at $N = 16$ in Mathematics/G1. At the primary budgets, interior-rank selection reaches 89.84% and 86.72%, respectively. A reference-calibrated non-monotone selector also improves fixed-budget validity but requires an auxiliary candidate pool. Its definition, full cost accounting, unresolved-case analyses, and budget-matched comparisons appear in Section I.

6 Relational Training

We next ask whether relational supervision can improve V0's ranking in the Code/G1 setting. M0 denotes the original verifier. M1 is trained with additional defect-oriented examples and hard negatives. M2 uses the cells of relational blocks as classification examples. M3 adds bidirectional ranking losses to the M2 training data. M4 uses the same data and replaces the M3 ranking term with the crossed objective defined below. M2–M4 use identical relational examples, standard examples, sequence exposure, optimization steps, class ratios, checkpoint-selection rules, and hyperparameter-search budgets. For a symmetric relational block, diagonal pairs (T_1, P_1) and (T_2, P_2) are valid and off-diagonal pairs (T_1, P_2) and (T_2, P_1) are invalid. If ℓ_{ij} is the verifier logit for (T_i, P_j), define the four positive-versus-negative margins

$$m_1 = \ell_{11} - \ell_{12}, \qquad m_2 = \ell_{22} - \ell_{21},$$
$$m_3 = \ell_{11} - \ell_{21}, \qquad m_4 = \ell_{22} - \ell_{12}.$$

M3 uses the ordinary bidirectional ranking loss

$$\mathcal{L}_{\text{rank}} = \frac{1}{4} \sum_{r=1}^{4} \text{softplus}(-m_r).$$

Here

$$\text{softplus}(x) = \log(1 + e^x).$$

Table 2: **Relational supervision improves binding and deployment.** M1–M4 entries are means and sample standard deviations over three matched training seeds. M0 is one fixed checkpoint. Natural evaluation uses the same prospective Code/G1 candidate frame for every checkpoint.

Model	Natural validity	Selected mismatch	Unseen-family PBA
M0	82.42	12.37	25.00
M1	81.77 ± 1.17	14.93 ± 1.11	25.22 ± 2.17
M2	87.46 ± 1.08	8.29 ± 0.76	52.67 ± 2.52
M3	88.67 ± 1.58	7.42 ± 1.24	61.22 ± 3.10
M4	90.32 ± 1.01	6.16 ± 0.87	65.33 ± 2.00

M4 replaces this ranking term with the crossed relational objective

$$\mathcal{L}_{\text{cross}} = \text{softplus}\left(-\frac{\ell_{11} + \ell_{22} - \ell_{12} - \ell_{21}}{2}\right). \tag{6}$$

Under the conceptual decomposition

$$\ell(T, P) = a(T) + b(P) + g(T, P),$$

the crossed contrast cancels the claim-only term $a(T)$ and the artifact-only term $b(P)$, leaving only interaction information. The matched M4–M3 comparison tests the incremental value of this aggregation while holding relational data and training resources fixed.

Table 2 shows that defect-oriented training alone does not improve natural deployment: M1 reaches 81.77% validity and increases selected mismatches to 14.93%. Relational supervision produces the larger gains, with M2, M3, and M4 reaching respectively 87.46%, 88.67%, and 90.32% natural validity, while unseen PBA rises from 25.00% for M0 to 52.67%, 61.22%, and 65.33%. Relative to M3, M4 improves natural validity by $1.65\,\text{pp}$, with a one-sided 95% lower bound of $0.47\,\text{pp}$, and improves unseen claim-fixed and artifact-fixed accuracy by $1.94\,\text{pp}$ and $2.00\,\text{pp}$. Relative to M0, M4 reduces selected mismatches by $6.21\,\text{pp}$, while defect control remains within the registered $1\,\text{pp}$ noninferiority margin. M0–M4 measures the combined effect of relational supervision and the crossed objective; the matched M3–M4 comparison isolates the change in objective aggregation. M4 and interior-rank selection offer complementary repairs, with no established superiority of M4 over the selector applied to V0. Complete seed-level counts and family analyses appear in Section J.

7 Discussion and Conclusion

We studied a failure mode of verifier-guided search in which a high-scoring proof or program addresses a nearby task. Under V0, these relational mismatches become more common near the top of the score distribution in both code and mathematics. Local defects become rarer in the same region. As a result, raw best-of-N validity eventually decreases when the candidate budget grows. Candidate-level accuracy, AUROC, calibration, and pass@k do not directly measure the validity of the selected output. Deployment evaluation should also report the validity and error composition of the output returned at the intended candidate budget. The comparison between V0 and V1 illustrates the issue: similar bulk metrics can coexist with different best-of-N outcomes. The intervention results suggest two practical responses. A selector can avoid a known unreliable part of the score tail, as illustrated by the interior-rank policy. Relational supervision can also improve how the verifier responds to changes in the target. The first intervention changes how an existing score is used. The second changes the learned ranking itself. Our evidence comes from two approximately 8B generators, two domains, one primary verifier, and a Code/G1 training intervention. Code certification covers bounded contracts, while mathematics relies on expert review. The forecasts assume conditionally independent generation, fixed rankings, and label support wherever a policy places weight. Adaptive search, iterative refinement, and generator training remain important settings for future evaluation. Within the evaluated setting, increasing the candidate budget exposes properties of the verifier's upper-score ranking that are largely absent from aggregate metrics. Measuring that ranking at the intended deployment budget is therefore important for reliable test-time scaling.

AI USE STATEMENT

Generative AI tools assisted with language editing, experimental code, and the preparation of LaTeX figures. The authors reviewed the resulting text, code, and figures and remain responsible for all claims and reported results.

ETHICS STATEMENT

Mathematics annotation uses independent assessment, adjudication, abstention, and pseudonymous reviewer identifiers. Generated programs are executed in a restricted environment with fixed resource, filesystem, and network controls. The principal deployment risk is misplaced confidence in a high-scoring artifact that fails the requested task; we mitigate this by reporting certification scope, selected-error composition, budgets, and cost assumptions. Further details appear in Section L.

REPRODUCIBILITY STATEMENT

The supplementary artifact contains generation requests, raw outputs, parsed artifacts, execution logs, verifier scores, tie randomizers, annotation records, inclusion probabilities, frozen forecasts, policy versions, checkpoint manifests, and scripts for regenerating every table and figure. Detailed experimental, certification, theoretical, and statistical procedures appear in Sections D to G and K.

REFERENCES

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, August 2021. URL https://arxiv.org/abs/2108.07732.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. Evaluating large language models trained on code, July 2021. URL https://arxiv.org/abs/2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, October 2021. URL https://arxiv.org/abs/2110.14168.

R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh and London, 1935. URL https://books.google.com/books?id=-EsNAQAAIAAJ.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models, July 2024. URL https://arxiv.org/abs/2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017/.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with APPS. In Joaquin Vanschoren and Serena Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021a. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c24cd76e1ce41366a4bbe8a49b02a028-Abstract-round2.html.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Serena Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, September 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, December 1952. doi: 10.1080/01621459.1952.10483446. URL https://doi.org/10.1080/01621459.1952.10483446.

Kimi Team, Angang Du, Bofei Gao, et al. Kimi k1.5: Scaling reinforcement learning with LLMs, January 2025. URL https://arxiv.org/abs/2501.12599.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's Law, March 2018. URL https://arxiv.org/abs/1803.04585.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

OpenAI. GPT-4o system card, October 2024a. URL https://arxiv.org/abs/2410.21276.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence, July 2024b. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL https://aclanthology.org/S18-2023/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen2.5 technical report, December 2024. URL https://arxiv.org/abs/2412.15115.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

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*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1b623663fd9b874366f3ce019fdfdd44-Abstract-Conference.html.

The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, pp. 367–381. Association for Computing Machinery, 2020. doi: 10.1145/3372885.3373824. URL https://doi.org/10.1145/3372885.3373824.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10495–10516, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.547. URL https://aclanthology.org/2025.findings-acl.547/.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. miniF2F: A cross-system benchmark for formal olympiad-level mathematics. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9ZPegFuFTFv.

A NOTATION

Table 3: **Notation used in the main paper and appendix.** Probabilities are taken over the registered target distribution, candidate generation, and any policy randomization unless a conditioning event is shown.

Symbol	Meaning
T, P	Requested target and generated artifact, respectively.
$\pi(\cdot \mid T)$	Frozen conditional generator distribution.
$V(T, P)$	Semantic validity of artifact P for target T.
$\widehat{V}(T, P)$	Protocol judgment in $\{1, 0, ?\}$: valid, invalid, or unresolved.
$K \in \mathcal{K}$	Recorded class in $\{A, B, C, U_0, U_?\}$.
A	Certified target-valid artifact.
B	Target-invalid artifact certified valid for at least one frozen neighboring target.
C	Artifact with a certified local defect that invalidates it for the target and all registered applicable neighbors.
U_0, $U_?$	Other resolved target-invalid and unresolved artifacts.
$\mathcal{N}(T)$	Proper neighborhood of target T, frozen before artifact inspection.
$S = s(T, P)$	Verifier score for a target–artifact pair.
Q	Randomized within-target score rank in $[0, 1]$.
$\eta_c(q, T)$	Conditional probability of class c at rank q for target T.
$\bar{\eta}_c(q)$	Target-averaged class profile $\mathbb{E}_T[\eta_c(q, T)]$.
N, $N^\star$	Candidate budget and primary deployment budget.
M, L, R	Forecast-pool size, auxiliary reference-pool size, and number of prospective repetitions per target.
ρ	A frozen candidate-selection policy.
$R_c^\rho(N)$	Probability that policy ρ selects a class-c artifact from N candidates.
$R_V^\rho(N)$	Semantic validity probability of the selected artifact.
$G^\rho(N)$	Certified-validity gain over raw maximum selection.
$w_j^{(k)}(N; M)$	Probability that finite-pool rank j is the k-th ascending member of a uniformly sampled size-N subset.
I_{ti}, π_{ti}	Label-inclusion indicator and known inclusion probability for candidate i of target t.
a_t	Registered population weight of target t, normalized to sum to one.
ESS_{HT}	Inclusion-probability-adjusted selection-weight concentration diagnostic.
G1, G2	`Qwen/Qwen2.5-7B-Instruct` and `meta-llama/Llama-3.1-8B-Instruct`.
V0, V1, V2	Primary trainable verifier, bulk-matched comparator, and discretized GPT-4o mini scorer.
M0–M4	Original verifier and four matched training conditions.
S1, S2	Interior-rank and reference-calibrated selection policies.
CF, AF, PBA	Claim-fixed accuracy, artifact-fixed accuracy, and strict paired-block accuracy.
pp	Percentage points.

B RELATED WORK AND POSITIONING

Verifier-guided generation. Learned outcome and process verifiers are widely used to select among generated mathematical solutions and programs (Cobbe et al., 2021; Chen et al., 2021; Lightman et al., 2024; Zhang et al., 2025). Recent systems also treat inference computation as a scaling axis (Guo et al., 2025; Kimi Team et al., 2025; Muennighoff et al., 2025; Snell et al., 2025). Our setting holds the generator fixed and evaluates the artifact returned by a declared selector as its candidate budget changes. Unlike pass@k, which asks whether a set contains a valid candidate, $R_A^\rho(N)$ asks whether the deployed policy returns one.

Reward overoptimization. Optimizing an imperfect evaluator can amplify disagreement with the intended objective (Manheim & Garrabrant, 2018; Gao et al., 2023). We isolate one concrete form of this phenomenon: an artifact that is valid for a frozen neighboring claim but invalid for the requested claim. The primary experiments vary selection pressure without changing the generator distribution. Adaptive search or policy optimization can also change that distribution and therefore requires separate analysis.

Partial-input shortcuts and relational evaluation. Models can exploit one input field while appearing to solve a relational task (Gururangan et al., 2018; Poliak et al., 2018). Contrast sets and behavioral tests evaluate sensitivity to controlled input changes (Gardner et al., 2020; Ribeiro et al., 2020). Our natural outputs estimate deployment prevalence, while artifact-fixed and crossed-block comparisons test claim–artifact sensitivity. Keeping these roles separate avoids estimating natural prevalence from constructed diagnostic blocks.

Calibration and finite-budget decisions. Calibration concerns the relationship between confidence and empirical correctness (Guo et al., 2017), but a strictly increasing recalibration preserves candidate ordering and hence raw best-of-N selection. Our deployment estimators instead combine classical order-statistic weights, without-replacement subset averaging, and Horvitz–Thompson estimation (Hoeffding, 1948; Horvitz & Thompson, 1952; Chen et al., 2021). We use these classical identities to estimate deployment-level selection rates and validate the resulting forecasts prospectively.

Crossed relational training. Crossed comparisons remove additive row and column effects, following a standard experimental-design principle (Fisher, 1935). For claim–artifact logits, the contrast

$$\ell_{11} + \ell_{22} - \ell_{12} - \ell_{21}$$

is invariant to additive claim-only and artifact-only terms. Our matched M4–M3 experiment tests the practical effect of aggregating relational margins through this contrast while holding data and training resources fixed. It does not uniquely identify invariance, rather than optimization or regularization differences, as the sole cause of the observed gain.

C Limitations and Scope

System coverage. The primary natural-tail study evaluates one trainable verifier, V0, under two approximately 8B generators and two domains. V1 shows that a bulk-matched verifier can behave differently in deployment, and V2 provides a separate score-saturation case. Its prevalence across verifier architectures remains to be measured. The complete training intervention is limited to Code/G1/V0.

Task interpretation. Code targets are normalized finite-domain contracts derived from MBPP/HumanEval-style sources rather than unmodified benchmark instances. Their results are not official benchmark scores. Mathematics targets use formal sources to fix semantics, but the evaluated outputs are natural-language proofs rather than formally checked proof terms.

Certification scope. The reported outcomes are protocol-certified classes. Code validity is relative to the registered parser, bounded input domain, evaluator, and execution environment. Mathematics relies on expert assessment and can remain unresolved. Audits estimate reliability in sampled populations but do not prove error-free certification in every score region.

Frozen neighborhoods. Class B depends on a finite claim-edit grammar fixed before artifact inspection. Freezing prevents post hoc construction of a rescuing claim, but the grammar may omit meaningful alternatives. Class-B prevalence is conditional on the declared grammar: omitted neighbors can leave mismatches in U_0, while broader neighborhoods change which artifacts qualify as mismatches.

Mechanistic interpretation. High-scoring invalid artifacts may be both more coherent and more likely to satisfy a neighboring claim. Artifact-fixed comparisons, crossed blocks, and relational training strengthen the binding interpretation. The contribution of mismatch reduction to each deployment gain remains uncertain.

Selection assumptions. The forecasting identities require conditionally independent candidates, saved independent tie randomizers, a frozen ordering, and positive label-inclusion probability wherever policy weight is nonzero. Beam search, iterative refinement, adaptive generation, or utilities rebuilt inside each selected subset require a different estimator or direct policy evaluation.

Table 4: **Registered claim transformations.** Every edit must satisfy domain-specific type and applicability constraints.

Transformation	Admissible change
Implication direction	Reverse an admissible implication or use a registered converse.
Quantifier change	Apply a typed quantifier reorder or registered quantifier edit.
Parameter specialization	Replace a general statement with a registered special case.
Hypothesis strengthening	Add or strengthen a predeclared assumption.
Conclusion weakening	Replace the conclusion with a registered weaker conclusion.
Relation or boundary change	Modify a comparison, boundary convention, or contract condition.
Domain substitution	Replace the domain with an admissible alternative.
Neighboring object or contract	Use a predeclared alternative object or executable specification.

Policy scope and cost. The fitted fraction $\alpha = 0.9375$ is not universally optimal and should be re-estimated after changing the target distribution, generator, verifier, or budget. Fixed-budget gains also differ from compute-optimal gains: a smaller raw budget may be competitive when generation cost is freely adjustable. Auxiliary pools, tokens, latency, memory, and hardware should be included in deployment comparisons.

Training inference. Three matched seeds characterize optimization variation only within the evaluated setup. M4–M3 isolates the change in loss aggregation under matched data and resources; M4–M0 additionally includes relational supervision. Broader claims require more architectures, domains, and training seeds.

D Certification Protocol

D.1 Targets, artifacts, and recorded outcomes

For target T and artifact P, semantic validity is

$$V(T, P) \in \{0, 1\},$$

whereas the protocol records

$$\widehat{V}(T, P) \in \{1, 0, ?\}.$$

The value ? denotes unresolved target validity. Generation draws are the sampling units; repeated strings retain distinct draw identifiers. Parsing failures, compilation failures, non-termination, interface violations, and scoring failures remain in the candidate frame.

Each artifact receives one mutually exclusive class:

$$K \in \mathcal{K} = \{A, B, C, U_0, U_?\}.$$

Class A is certified target-valid. Class B is target-invalid but certified valid, without artifact modification, for at least one frozen neighbor. Class C has a specific local defect that invalidates it for the target and all registered applicable neighbors. Class U_0 contains other resolved target-invalid artifacts, and $U_?$ contains unresolved cases.

D.2 Frozen claim neighborhoods

A typed grammar $\mathcal{G}$ specifies admissible transformations and their applicability conditions. For edit depth d, define

$$L^d_{\mathcal{G}}(T) = \{T\} \cup \{\text{claims reachable from } T \text{ by at most } d \text{ valid edits}\},$$

and

$$\mathcal{N}^d_{\mathcal{G}}(T) = L^d_{\mathcal{G}}(T) \setminus \{T\}.$$

The set is enumerated before inspecting the corresponding artifacts or verifier scores. The main text abbreviates it as $\mathcal{N}(T)$.

Each neighbor records its provenance, such as library derivation, task-author construction, independent-pilot observation, or grammar generation. Provenance determines entry into the frozen set; the unchanged artifact still requires a separate validity certificate. Logical implication is respected: if an artifact validly establishes both the target and a weaker neighbor, it remains class A.

D.3 Decision order

Certification follows this fixed order.

1. **Assess the requested target.** If $\widehat{V}(T, P) = 1$, assign A. If the prescribed review cannot resolve target validity, assign $U_?$.
2. **Test applicable frozen neighbors.** For a target-invalid artifact, assign B when
$$\widehat{V}(T, P) = 0, \qquad \exists T' \in \mathcal{N}(T) : \widehat{V}(T', P) = 1.$$
The artifact cannot be edited, repaired, supplemented, or reinterpreted.
3. **Seek a local-defect certificate.** If no neighbor is certified valid, assign C only when a specific defect invalidates the artifact for the target and every registered applicable neighbor.
4. **Retain other resolved errors.** Assign U_0 when target invalidity is established but neither the B nor C criterion is met.

Secondary tags record information not represented by the primary class, including circularity, unlocalized semantic jumps, implicit false lemmas, formatting exploitation, or derivations outside the frozen neighborhood.

D.4 Domain-specific evidence

Mathematics. Reviewers first assess the requested target without seeing neighboring claims, verifier identity, score, or score stratum. Neighbors are revealed only after target invalidity is established. Independent initial reviews are followed by adjudication, and unresolved cases remain $U_?$.

Code. The protocol records parsing, compilation, termination, interface, and bounded-domain execution outcomes. The unchanged program is evaluated against the requested contract and all applicable neighbors under the registered sandbox and resource limits.

A certificate refers to evidence under this protocol, not unrestricted semantic truth. Independent audits assess sampled decisions.

D.5 Certified and semantic validity

Let $P_{\rho,N}$ be the artifact returned by policy ρ. Semantic deployment validity is
$$R_V^\rho(N) = \mathbb{P}\{V(T, P_{\rho,N}) = 1\}.$$
If resolved judgments are sound, then
$$R_V^\rho(N) \in \left[R_A^\rho(N), R_A^\rho(N) + R_{U_?}^\rho(N)\right].$$
Define the certified-validity gain
$$G^\rho(N) = R_A^\rho(N) - R_A^{\mathrm{raw}}(N).$$
It follows that
$$R_V^\rho(N) - R_V^{\mathrm{raw}}(N) \in \left[G^\rho(N) - R_{U_?}^{\mathrm{raw}}(N), G^\rho(N) + R_{U_?}^\rho(N)\right].$$
Paired selected-artifact identities can tighten these marginal bounds. Errors in resolved judgments require a separate allowance and are not covered by varying only $U_?$.

E Theory and Proofs

E.1 Randomized score ranks

For a fixed target, let
$$F_T(u) = \mathbb{P}(S \le u \mid T).$$
With an independent saved randomizer $Z \sim \mathrm{Unif}[0, 1]$, define
$$Q = F_T(S^-) + Z\{F_T(S) - F_T(S^-)\}.$$

The randomized probability-integral transform gives

$$Q \mid T \sim \mathrm{Unif}[0,1],$$

including when the score distribution has atoms. Ordering by Q agrees almost surely with ordering by score followed by the saved tie randomizer. The population transform is analytical; finite-pool estimators require only the observed frozen ordering.

E.2 Order-statistic selection

Let $Q_{(k:N)}$ be the k-th ascending order statistic among N conditionally independent candidate ranks.

Proposition 1 (Selected class rates)**.** *A policy returning the k-th ascending rank satisfies*

$$R_c^{(k)}(N \mid T) = \int_0^1 \eta_c(q,T) \frac{N!}{(k-1)!(N-k)!} q^{k-1}(1-q)^{N-k}\, dq.$$

Averaging over targets replaces $\eta_c(q,T)$ with $\bar{\eta}_c(q)$. For maximum-score selection,

$$R_c^{\mathrm{raw}}(N) = N \int_0^1 \bar{\eta}_c(q) q^{N-1}\, dq.$$

Proof. Conditional on T, ranks are uniform and $Q_{(k:N)}$ has the displayed beta density. Given selected rank q, the class probability is $\eta_c(q,T)$. Integration and target averaging give the result. □

If $k_N/(N+1) \to \alpha \in (0,1)$, then

$$Q_{(k_N:N)} \to \alpha \quad \text{in probability.}$$

When $\eta_A(\cdot,T)$ is continuous at α for almost every target,

$$R_A^{(k_N)}(N) \longrightarrow \mathbb{E}_T[\eta_A(\alpha,T)].$$

This identifies the asymptotic rank region of a fixed-fraction policy; it does not imply finite-budget optimality.

E.3 When raw scaling is monotone

Proposition 2 (When more search helps)**.** *If $\eta_A(q,T)$ is nondecreasing in q for almost every target T, then $R_A^{\mathrm{raw}}(N)$ is nondecreasing in N.*

Proof. Couple candidate sets as prefixes of one infinite conditionally independent sequence. Their maximum rank is nondecreasing in N. Applying the nondecreasing function $\eta_A(\cdot,T)$, taking conditional expectations, and then averaging over targets proves the claim. □

Thus, under nested candidate sets, declining raw selected validity identifies a ranking failure rather than declining candidate availability.

E.4 Bulk–tail non-identification

Proposition 3 (Bulk–tail non-identification)**.** *For every $\varepsilon > 0$ and integer $m \geq 1$, there exist two score–label laws with identical class priors and score marginals and total-variation distance at most ε, such that every bounded k-sample statistic with $k \leq m$ differs in expectation by at most*

$$2k\varepsilon \|h\|_\infty.$$

Nevertheless, their raw best-of-N validities converge to 1 and 0, respectively.

Proof. Set

$$\delta = \min\{\varepsilon/2, 1/4\}.$$

Both laws assign mass $1-2\delta$ to a shared component with $S \sim \mathrm{Unif}[0,1]$ and an independent balanced label in $\{A,B\}$.

Under $\mathcal{L}_{\text{good}}$, assign mass δ to

$$(A, S \sim \text{Unif}[2, 3])$$

and mass δ to

$$(B, S \sim \text{Unif}[-2, -1]).$$

Under $\mathcal{L}_{\text{bad}}$, exchange these two labels while preserving the score distributions. The laws have identical score marginals and class priors, and

$$\text{TV}(\mathcal{L}_{\text{good}}, \mathcal{L}_{\text{bad}}) = 2\delta \leq \varepsilon.$$

Tensorization gives

$$\text{TV}\left(\mathcal{L}_{\text{good}}^{\otimes k}, \mathcal{L}_{\text{bad}}^{\otimes k}\right) \leq k\varepsilon,$$

which implies the stated bound for a bounded kernel h.

The selected validities are

$$R^{\text{raw}}_{A,\text{good}}(N) = 1 - \frac{1}{2}(1-\delta)^N - \frac{1}{2}\delta^N$$

and

$$R^{\text{raw}}_{A,\text{bad}}(N) = \frac{1}{2}(1-\delta)^N + \frac{1}{2}\delta^N.$$

Their limits are 1 and 0. □

More generally, if two candidate-record laws have total-variation distance τ, applying the same selection functional to N independent records changes the expectation of any selected binary outcome by at most

$$1 - (1-\tau)^N \leq N\tau.$$

Hence a small candidate-level discrepancy can produce an order-one deployment discrepancy at a sufficiently large budget.

E.5 Finite-pool selection weights

Consider an independent pool of size M, ordered from lowest to highest under a frozen ranking. If a size-N subset is sampled uniformly without replacement, pool rank j is its k-th ascending member with probability

$$w_j^{(k)}(N; M) = \frac{\binom{j-1}{k-1}\binom{M-j}{N-k}}{\binom{M}{N}}.$$

For raw maximum selection,

$$w_j^{\max}(N; M) = \frac{\binom{j-1}{N-1}}{\binom{M}{N}}.$$

Proof. A subset selecting rank j as its k-th member must contain exactly $k-1$ lower ranks and $N-k$ higher ranks. The numerator counts such subsets and the denominator counts all size-N subsets. □

The weights sum to one. A uniformly sampled subset of N distinct indices from an independent M-sample has the same joint distribution as a fresh independent N-sample. Therefore, for a frozen policy ordering and independent auxiliary information H,

$$\mathbb{E}\left[\sum_{i=1}^{M} \mathbf{1}\{K_i = c\} w_i^{\rho}(N) \,\middle|\, T, H\right] = R_c^{\rho}(N \mid T, H).$$

This identity does not apply if a candidate utility is refitted inside each evaluated subset.

E.6 UNEQUAL-PROBABILITY LABEL SAMPLING

Let $\mathcal{F}$ be the full candidate frame. For candidate i of target t, let

$$\pi_{ti} = \mathbb{P}(I_{ti} = 1 \mid \mathcal{F})$$

be its known first-order label-inclusion probability. Positive inclusion probability is required whenever policy weight is nonzero.

Theorem 4 (Finite-budget deployment estimation)**.** *For each target t, let M_t be the size of its independent forecast pool and suppose $N \le M_t$. Assume conditionally independent generation, a frozen candidate ordering, a fixed selected rank for each N, and frozen tie breaking. Any auxiliary reference information must be independent of the selectable pool. Then*

$$\widehat{R}_c^{\rho}(N) = \sum_t a_t \sum_i \frac{I_{ti}}{\pi_{ti}} \mathbf{1}\{K_{ti} = c\} w_{ti}^{\rho}(N)$$

is design-unbiased for the fully labeled pool average over uniformly sampled size-N subsets. Averaging over pool generation gives the corresponding fresh-N deployment rate under the registered target design.

Proof. Conditional on the full frame,

$$\mathbb{E}[I_{ti}/\pi_{ti} \mid \mathcal{F}] = 1.$$

Thus the conditional expectation of the estimator is

$$\sum_t a_t \sum_i \mathbf{1}\{K_{ti} = c\} w_{ti}^{\rho}(N),$$

the fully labeled subset average. The finite-pool identity above equates its expectation over candidate generation with fresh-N deployment. Averaging over targets and independent auxiliary information completes the proof. □

For $Y_{ti} = \mathbf{1}\{K_{ti} = A\}$, the paired gain estimator is

$$\widehat{G}^{\rho}(N) = \sum_t a_t \sum_i \frac{I_{ti}}{\pi_{ti}} Y_{ti} \left[w_{ti}^{\rho}(N) - w_{ti}^{\mathrm{raw}}(N) \right].$$

Linearity gives the same unbiasedness result. Because only $\mathbf{1}\{K_{ti} = A\}$ enters, splitting or merging the non-A classes does not change the certified-validity estimate.

E.7 VARIANCE AND EFFECTIVE SUPPORT

Absorb the target index into i, and write

$$b_i = a_{t(i)} w_i^{\rho}(N) \mathbf{1}\{K_i = c\}.$$

With joint label-inclusion probabilities

$$\pi_{ij} = \mathbb{P}(I_i = I_j = 1 \mid \mathcal{F}), \qquad \pi_{ii} = \pi_i,$$

the conditional label-design variance is

$$\mathrm{Var}_L(\widehat{R} \mid \mathcal{F}) = \sum_{i,j} b_i b_j \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}.$$

For independent Bernoulli inclusion, this reduces to

$$\sum_i b_i^2 \frac{1 - \pi_i}{\pi_i}.$$

We report the weight-concentration diagnostics

$$\mathrm{ESS}_w^{\rho}(N) = \left[\sum_{t,i} \{a_t w_{ti}^{\rho}(N)\}^2 \right]^{-1}$$

and

$$\mathrm{ESS}_{HT}^{\rho}(N) = \left[\sum_{t,i} \frac{\{a_t w_{ti}^{\rho}(N)\}^2}{\pi_{ti}}\right]^{-1}.$$

For fully labeled frames, $\mathrm{ESS}_{HT} = \mathrm{ESS}_w$. These quantities summarize weight concentration. Variance is estimated separately at the design or target level.

E.8 Maximum-score atoms

Suppose the score distribution for target T has mass $m_T \in (0,1)$ at its maximum score, with validity a_T inside that atom. Under independent random tie breaking,

$$R_A^{\mathrm{raw}}(N \mid T) = \{1-(1-m_T)^N\}a_T + (1-m_T)^N R_{A,<}^{\mathrm{raw}}(N \mid T),$$

where $R_{A,<}^{\mathrm{raw}}$ conditions on scores below the maximum atom. Consequently,

$$|R_A^{\mathrm{raw}}(N \mid T) - a_T| \le (1-m_T)^N.$$

The budget required to encounter the atom with probability at least 0.95 is

$$N_{95}(T) = \left\lceil \frac{\log(0.05)}{\log(1-m_T)} \right\rceil.$$

Thus finite-budget deployment mixes atom and below-atom selection, while large-budget validity is determined by the atom.

E.9 Crossed-contrast invariance

Under the conceptual decomposition

$$\ell(T,P) = a(T) + b(P) + g(T,P),$$

a symmetric relational block satisfies

$$\Delta_\ell = \ell_{11} + \ell_{22} - \ell_{12} - \ell_{21} = g_{11} + g_{22} - g_{12} - g_{21}.$$

The crossed contrast therefore cancels additive claim-only and artifact-only terms.

For margins

$$m_1 = \ell_{11} - \ell_{12}, \qquad m_2 = \ell_{22} - \ell_{21},$$
$$m_3 = \ell_{11} - \ell_{21}, \qquad m_4 = \ell_{22} - \ell_{12},$$

their mean is

$$\frac{1}{4}\sum_{r=1}^{4} m_r = \frac{\Delta_\ell}{2}.$$

Let $\phi(m) = \mathrm{softplus}(-m)$. Convexity gives

$$\frac{1}{4}\sum_{r=1}^{4} \phi(m_r) \ge \phi\left(\frac{\Delta_\ell}{2}\right).$$

Hence M3 penalizes individual margins before averaging, whereas M4 aggregates margins before applying the logistic penalty. The matched experiment isolates this change in objective aggregation under fixed data and training resources. It does not imply that every individual margin is positive, nor that invariance is the unique explanation for M4's empirical gain.

F Experimental Details

F.1 Study chronology and data separation

Discovery data informed the confirmation protocol but are not pooled with confirmation results. Development data determine policies, hyperparameters, and checkpoint selection. Independent

Table 5: **Confirmation design.** n_F and n_E are forecast and prospective target counts; M is the forecast-pool size; L is the independent reference-pool size used by S2; and R is the number of prospective repetitions per target.

	Code/G1	Code/G2	Math/G1	Math/G2
Forecast targets n_F	128	64	64	48
Prospective targets n_E	192	128	128	128
Forecast candidates per target M	2,048	2,048	512	512
S2 reference candidates L	512	512	128	128
Prospective repetitions R	4	4	2	2
Primary budget $N^\star$	128	128	32	32
Largest evaluated budget $N_{\max}$	256	256	64	64
Forecast seed	12101	12102	12101	12102
Evaluation seed	13101	13102	13101	13102

Table 6: **Confirmation labeling support.** ESS_{HT} measures selection-weight concentration and is not an independent-target count.

	Code/G1	Code/G2	Math/G1	Math/G2
Label-sampling seed	14201	14202	14203	14204
Forecast targets	128	64	64	48
Pool artifacts	262,144	131,072	32,768	24,576
Labeled artifacts	262,144	131,072	4,096	3,072
Deepest-tail labels	2,048	1,024	1,024	768
$\mathrm{ESS}_{HT}(N^\star)$	4,078	2,034	1,176	870
$\mathrm{ESS}_{HT}(N_{\max})$	2,051	1,022	612	451

forecast pools provide score-stratified certification labels. Forecasts and policies are then frozen before prospective candidate generation.

Splits are defined by source group rather than displayed task identifier. A source group contains all rewrites, neighboring claims, templates, and relational blocks derived from one source object. Development, forecast, and prospective evaluation use disjoint source groups for the primary comparisons.

A prospective *selection unit* is one target–repetition block from which a policy returns exactly one artifact. The four settings contain 768, 512, 256, and 256 selection units, respectively. Smaller budgets are prefixes of the same largest-budget block, so nested budgets and policies are compared jointly rather than treated as independent samples.

F.2 Candidate and score records

Each candidate record contains the domain, generator, target, stage, repetition, draw identifier, generation seed, raw output, parsed artifact, and parsing or execution status. Repeated strings remain distinct draws.

The verifier receives only its fixed instruction, the requested target, and the complete artifact. It does not receive neighbors, certification labels, score strata, or competing candidates. Independent saved randomizers resolve score ties and, where applicable, utility ties. Failed parsing, execution, or scoring calls remain in the candidate frame under the frozen fallback rules.

F.3 Label sampling and support

Code forecast pools are fully certified. Mathematics uses known-probability sampling within target-specific V0 score strata. The same item-level inclusion probability is retained when another verifier orders the artifact.

The code score-rank strata are

$$[0, .5), \quad [.5, .875), \quad [.875, .96875), \quad [.96875, .9921875), \quad [.9921875, 1].$$

The mathematics strata are

$$[0, .5), \quad [.5, .875), \quad [.875, .96875), \quad [.96875, 1].$$

Every candidate with nonzero policy weight has positive label-inclusion probability. Point estimation uses first-order inclusion probabilities; variance calculations retain joint-inclusion information when required by the sampling design.

F.4 Selection policies

Raw selection. Return the highest-scoring candidate, breaking ties with the saved score randomizer.

Interior-rank selection, S1. Return the ascending rank in Equation (5), with $\alpha = 0.9375$ for the primary V0 confirmation study. The parameter is selected on development data and remains fixed during forecast and prospective evaluation.

Reference-calibrated selection, S2. For each target, generate an independent reference pool H of size L. For score s, let $L_<(s)$ and $L_=(s)$ count reference scores below and equal to s. With a saved independent randomizer Z_H,

$$\widehat{Q}_H(s) = \frac{L_<(s) + Z_H\{L_=(s) + 1\}}{L + 1}.$$

A development-fitted non-monotone bin model maps $\widehat{Q}_H(s)$ to estimated validity, and S2 returns the candidate with maximum mapped utility. The reference pool and mapping remain fixed during subset evaluation. S2 is charged $N + L$ generated and scored candidates unless reference cost is explicitly amortized.

F.5 Forecast error and inference

For registered budget grid $\mathcal{B}$, forecast error is

$$\mathrm{WMAE} = \sum_{\substack{N \in \mathcal{B} \\ N \geq 2}} \omega_N \left| \widehat{R}_A(N) - R_A^{\mathrm{obs}}(N) \right|,$$

with equal weights over the log-spaced budgets:

$$\omega_N = \begin{cases} 1/8, & \text{code}, \\ 1/6, & \text{mathematics}. \end{cases}$$

The $N = 1$ result is displayed but excluded from WMAE.

Tail contrasts use simultaneous two-sided target- or source-group-clustered 95% intervals. Policy and training comparisons use the stated one-sided clustered 95% lower confidence bounds. Nested budgets, verifiers, and policies acting on the same candidate frame are resampled jointly.

F.6 Matched training design

The training conditions are:

Condition	Training setup
M0	Original V0 checkpoint.
M1	Defect-oriented examples or hard negatives.
M2	Relational block cells with cell-wise classification.
M3	M2 plus ordinary bidirectional margin ranking.
M4	M2 with the crossed objective replacing the M3 ranking term.

M2–M4 share relational examples, standard examples, optimization steps, sequence exposure, class ratios, checkpoint-selection rules, and hyperparameter-search budgets. Their matched seeds are

Table 7: **Frozen generators and decoding configuration.** Each request returns one candidate and uses the model's native chat template. Maximum lengths count newly generated tokens.

Label	Public checkpoint	Scale	Temp.	Top-p	Code max.	Math max.	Beam search
G1	`Qwen/Qwen2.5-7B-Instruct`	7.6B	0.8	0.95	2,048	4,096	No
G2	`meta-llama/Llama-3.1-8B-Instruct`	8.0B	0.8	0.95	2,048	4,096	No

Table 8: **Ordinary V0 training configuration.**

Configuration	Value
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
Effective batch size	64
Warmup fraction	0.03
Epochs	2
Maximum input tokens	8,192
Checkpoint selection	Lowest development BCE

11101, 11102, and 11103. Every checkpoint is evaluated on the same prospective Code/G1 candidate frame.

Per checkpoint, controlled evaluation contains 300 seen-family PBA blocks, 300 unseen-family PBA blocks, 600 unseen claim-fixed comparisons, 600 unseen artifact-fixed comparisons, and 1,200 defect-control pairs.

G Systems, Tasks, and Configuration

G.1 Generators and decoding

No additional top-k truncation or repetition penalty is used. Generation stops at EOS or the length limit. Requests do not share history and use no self-refinement, multi-answer prompt, or verifier-guided resampling. Truncated and malformed outputs are not replaced by new draws.

G.2 V0 and training sources

V0 uses a Qwen2.5-7B-Instruct backbone with a learned scalar head on the last valid input-token state:

$$\ell(T, P) = w^\top h_{\text{last}}(T, P) + b, \qquad s(T, P) = \sigma\{\ell(T, P)\}.$$

Candidate ordering uses the raw logit; threshold and Brier calculations use the sigmoid score. V0 is an outcome verifier, not an aggregation of step-level rewards.

Ordinary code supervision uses APPS introductory tasks and source-group-disjoint MBPP-style tasks labeled through execution. Mathematics supervision uses filtered PRM800K/MATH-style complete-solution records. Incompletely assessed trajectories are not automatically positive. Ordinary training excludes the frozen neighbors, class-B labels, and crossed relational blocks used in this study.

V0/M0 denotes the resulting trained checkpoint, not the public Qwen initialization. The artifact records trained-weight, tokenizer, data-manifest, and configuration hashes.

G.3 Comparator and diagnostic scorers

V1. V1 uses a Llama-3.1-8B-Instruct backbone with an independently trained linear outcome head. Its task types and ordinary-supervision sources are aligned with V0, but it does not consume V0 scores or prospective labels.

V1 was selected before confirmation generation from three candidate backbones: Qwen2.5-3B-Instruct, Llama-3.1-8B-Instruct, and Gemma-2-9B-it. Selection used only development accuracy and

Table 9: **Task and certification frames.** Public datasets indicate provenance; reported results are not official benchmark scores.

Domain	Source and output form	Evaluation scope
Code	MBPP/HumanEval-style finite-domain contracts; Python 3.10 pure functions with supplied signatures.	Exhaustive registered domains, typically at most about 4,096 inputs per task. Evaluation checks parsing, interface, termination, exact return types, and values in a network-disabled sandbox.
Mathematics	Natural-language rewrites of miniF2F/mathlib-style propositions and manually written elementary lemmas.	Complete natural-language proofs in number theory, algebra, inequalities, sets, quantifiers, finite combinatorics, and basic analysis. Primary neighborhoods use one typed edit ($d = 1$).

AUROC. For candidate verifier v,

$$D(v) = \sum_{d \in \{\text{code},\text{math}\}} \left[\left\{ \frac{\text{Acc}_d(v) - \text{Acc}_d(V0)}{0.015} \right\}^2 + \left\{ \frac{\text{AUC}_d(v) - \text{AUC}_d(V0)}{0.015} \right\}^2 \right].$$

The Llama verifier minimized $D(v)$. Tail composition, deployment curves, and prospective labels were not used for screening.

V2 and reference scorers. V2 is `gpt-4o-mini-2024-07-18`, prompted at temperature zero to return an integer score from 0 to 10, with at most 128 generated tokens. Its finite score set creates the maximum-score atom analyzed in Section E.

The stronger diagnostic reference judge is `gpt-4o-2024-08-06`. It supports binding diagnostics but does not replace executable or human certification. The artifact-only control replaces the target field with a fixed placeholder, making its score independent of the paired claim.

G.4 Task frames

Code neighbors include strict versus non-strict boundaries, first versus last positions, ordered versus unordered pairs, empty-input conventions, universal versus existential conditions, and strengthened preconditions. Mathematics neighbors include implication direction, quantifier order, hypothesis strength, conclusion strength, domain, and parameter changes. All targets and neighbors derived from one source share a `source_group_id`.

G.5 Worked mathematics certificate

Consider the requested target

$$T: \qquad 8 \mid (n^2 - 1) \Longrightarrow n \text{ is odd}, \qquad n \in \mathbb{Z}.$$

A natural candidate states:

> Write $n = 2k + 1$. Then
> $$n^2 - 1 = (2k + 1)^2 - 1 = 4k(k + 1).$$
> One of k and $k + 1$ is even, so $4k(k + 1)$ is divisible by 8. Hence $n^2 \equiv 1 \pmod 8$.

The first sentence assumes the requested conclusion and proves only the reverse implication. The artifact therefore does not prove T. Before generation, the following implication-direction neighbor had been registered:

$$T': \qquad n \text{ is odd} \Longrightarrow 8 \mid (n^2 - 1).$$

The unchanged artifact is valid for T', so

$$\widehat{V}(T, P) = 0, \qquad \widehat{V}(T', P) = 1, \qquad K(T, P) = B.$$

The invalid requested-target pairing receives the slightly higher V0 score. Aggregate artifact-fixed results are reported below.

Table 10: **V0 scores for the worked artifact-fixed pair.** Both claims are true, but a proof of one direction is not a proof of the other.

Pair	Logit	Sigmoid	Certification
(T, P)	3.61	0.9737	Invalid for requested target
(T', P)	3.47	0.9698	Valid for frozen neighbor

Table 11: **V0 class composition by within-target score-rank stratum.** Code pools are fully labeled; mathematics entries are Horvitz–Thompson estimates. All values are percentages.

Domain	Generator	Rank stratum	A	B	C	U_0	$U_?$
Code	G1	$[0, .5)$	29.8	1.6	59.7	8.9	0.0
		$[.5, .875)$	68.4	4.8	20.8	6.0	0.0
		$[.875, .96875)$	91.8	3.5	2.7	2.0	0.0
		$[.96875, .9921875)$	87.6	8.3	1.5	2.6	0.0
		$[.9921875, 1]$	78.9	14.8	0.8	5.5	0.0
	G2	$[0, .5)$	30.5	1.8	58.0	9.7	0.0
		$[.5, .875)$	69.9	4.3	19.0	6.8	0.0
		$[.875, .96875)$	90.7	3.6	3.2	2.5	0.0
		$[.96875, .9921875)$	88.2	6.5	2.0	3.3	0.0
		$[.9921875, 1]$	84.6	9.0	1.2	5.2	0.0
Math	G1	$[0, .5)$	34.0	1.7	52.0	9.5	2.8
		$[.5, .875)$	66.7	5.1	17.0	8.9	2.3
		$[.875, .96875)$	89.5	4.0	2.5	2.5	1.5
		$[.96875, 1]$	74.8	15.8	1.0	6.0	2.4
	G2	$[0, .5)$	35.2	1.5	50.4	10.1	2.8
		$[.5, .875)$	69.6	4.2	16.2	7.8	2.2
		$[.875, .96875)$	90.8	3.4	2.3	2.1	1.4
		$[.96875, 1]$	79.8	11.3	1.3	5.8	1.8

H Full Natural-Tail Results

H.1 Score-stratified class profiles

The simultaneous contrasts are reported in Table 1. The registered magnitude thresholds were

$$\Delta C \leq -8\,\text{pp}, \qquad \Delta B \geq 5\,\text{pp}, \qquad \Delta\{B/(1-A)\} \geq 20\,\text{pp},$$

and

$$\Delta A_{\text{peak}\to\text{tail}} \leq -8\,\text{pp}.$$

Code/G1 and both mathematics settings satisfy all four thresholds. Code/G2 has the registered directions and intervals excluding zero but falls below the magnitude thresholds for mismatch enrichment and peak-to-tail validity decline.

H.2 Neighbor-family decomposition

An artifact can satisfy multiple frozen neighbors. For descriptive accounting, its primary family is the minimum-edit-depth neighbor, breaking ties by frozen neighbor identifier. Other valid neighbors remain secondary tags; neither verifier score nor observed effect size determines the assignment.

Several families contribute to the aggregate increase, but not every family increases. Parameter specialization declines slightly in code, and conclusion weakening declines slightly in mathematics. The table is descriptive; no family-level multiplicity procedure was registered.

H.3 Independent certification audits

A separate score-stratified audit estimates class-B detection recall among target-invalid artifacts.

Table 12: **Class-B prevalence by primary transformation family.** Each cell gives reference-band $\rightarrow$ tail prevalence as a percentage of all candidates in the corresponding stratum. A dash means that the domain-specific catalog did not instantiate the family.

Transformation family	Code/G1	Code/G2	Math/G1	Math/G2
Implication direction	—	—	$0.87 \rightarrow 3.26$	$0.74 \rightarrow 2.21$
Quantifier change	$0.62 \rightarrow 1.44$	$0.55 \rightarrow 0.87$	$0.64 \rightarrow 2.05$	$0.53 \rightarrow 1.52$
Parameter specialization	$0.39 \rightarrow 0.34$	$0.36 \rightarrow 0.28$	$0.72 \rightarrow 1.97$	$0.58 \rightarrow 1.47$
Hypothesis strengthening	$0.58 \rightarrow 1.27$	$0.47 \rightarrow 0.94$	$0.98 \rightarrow 3.17$	$0.83 \rightarrow 2.12$
Conclusion weakening	—	—	$0.53 \rightarrow 0.47$	$0.45 \rightarrow 0.39$
Relation or boundary change	$1.71 \rightarrow 5.86$	$1.49 \rightarrow 3.57$	$0.43 \rightarrow 1.19$	$0.36 \rightarrow 0.89$
Domain substitution	$0.54 \rightarrow 1.42$	$0.49 \rightarrow 0.93$	$0.58 \rightarrow 1.85$	$0.49 \rightarrow 1.41$
Neighboring object or contract	$0.96 \rightarrow 4.47$	$0.94 \rightarrow 2.41$	$0.35 \rightarrow 1.84$	$0.22 \rightarrow 1.29$
All class B	$4.8 \rightarrow 14.8$	$4.3 \rightarrow 9.0$	$5.1 \rightarrow 15.8$	$4.2 \rightarrow 11.3$

Table 13: **Independent pooled certification audits.** Intervals are two-sided 95% Wilson intervals.

Audit population	Domain	Confirmed/audited	Rate [95% CI]
Natural class-B labels	Code	$286/300$	95.33 $[92.33, 97.20]$
Natural class-B labels	Mathematics	$282/300$	94.00 $[90.72, 96.17]$
Target-validity decisions	Mathematics	$390/400$	97.50 $[95.46, 98.64]$

All recall-difference intervals lie within the registered $[-10, 10]$ pp equivalence region.

A distinct audit starts from candidates already assigned class B and measures certificate precision. A certificate is confirmed only if the artifact fails the requested target, the unchanged artifact satisfies a frozen neighbor, and that neighbor predates artifact inspection.

Audited certificate precision and recall are similar across the reference and tail strata. The audit intervals describe the sampled records; deployment inference uses source-group clustering.

H.4 Natural binding diagnostics

For a natural class-B artifact P, requested target T, and certified neighbor T', rescue accuracy awards one point when $s(T', P) > s(T, P)$, half for a tie, and zero otherwise. CF and AF comparisons similarly award half credit to ties. Strict PBA requires every valid diagonal score to exceed every invalid off-diagonal score.

The artifact-only scorer necessarily has zero strict PBA because it cannot distinguish claims paired with the same artifact. V0 is claim-sensitive but remains far below the reference judge, supporting partial rather than absent binding.

H.5 Bulk-matched verifier comparison

For V0 minus V1, the code differences are 0.72 pp in accuracy and -0.007 in AUROC, with 95% confidence-envelope magnitudes 1.32 pp and 0.012. The mathematics differences are -0.73 pp and 0.006, with envelope magnitudes 1.41 pp and 0.013. These lie within the registered margins of 1.50 pp for accuracy and 0.015 for AUROC. No Brier-score equivalence claim is made.

H.6 Auxiliary score-saturation case

Large-budget outcomes are close to the validity of the maximum-score atom, as predicted by the decomposition in Section E. This score-resolution limitation is distinct from V0's non-monotone high-resolution ranking.

Table 14: **Score-stratified class-B recall audits.** Count triples are audited target-invalid artifacts, adjudicated true mismatches, and detected mismatches. Differences are tail minus reference.

Setting	Reference	Tail	Difference [95% CI]
Code/G1	$400/72/68$ 94.44%	$240/164/154$ 93.90%	-0.54 $[-7.16, 5.43]$
Code/G2	$400/66/61$ 92.42%	$240/151/138$ 91.39%	-1.03 $[-8.44, 6.05]$
Math/G1	$640/109/98$ 89.91%	$240/149/135$ 90.60%	0.69 $[-7.23, 8.34]$
Math/G2	$560/95/84$ 88.42%	$240/134/121$ 90.30%	1.88 $[-6.55, 9.62]$

Table 15: **Score-stratified class-B certificate precision.** Differences are tail minus reference. Difference intervals use Newcombe's score method.

Domain	Reference count	Reference precision	Tail count	Tail precision	Difference [95% CI]
Code	$267/280$	95.36%	$251/264$	95.08%	$-0.28\ [-4.11, 3.45]$ pp
Mathematics	$75/80$	93.75%	$211/224$	94.20%	$0.45\ [-4.81, 8.37]$ pp

H.7 Discovery-to-confirmation comparison

Discovery and confirmation use separate source groups and seeds. Confirmation shows the same directions with somewhat smaller deep-tail mismatch rates and validity declines.

The confirmation claim concerns replicated directions and pre-specified effect sizes. A cross-stage interval containing zero does not establish equivalence.

I Full Forecasting and Selection Results

I.1 Primary deployment curves

The prospective curves recover both the interior optimum and the subsequent decline. Code/G1 WMAE is 0.71 pp, and Mathematics/G1 WMAE is 1.09 pp.

I.2 Prediction errors across verifiers

The flexible predictor has lower realized WMAE in ten of twelve settings, while the exact-weight estimator performs best in the two primary V0 settings. This illustrates the distinction between design unbiasedness and finite-sample point prediction. The monotone predictor performs poorly in all four V0 settings because it cannot represent validity turnover.

Merging all non-A classes changes the implemented $\widehat{R}_A(N)$ by at most 3.3×10^{-13}, below the registered numerical tolerance of 10^{-12}.

I.3 Selected-class forecasts

The forecast pools also predict S1 gains of 7.90, 3.85, 7.50, and 4.80 pp, compared with observed gains of 7.42, 4.49, 8.98, and 5.47 pp.

I.4 Paired policy gains

All policies move mass from class B toward class A. The composition identity

$$\Delta R_A = -\Delta R_B - \Delta R_C - \Delta R_{U_0} - \Delta R_{U_?}$$

is descriptive class accounting; it does not by itself establish causal mediation.

Table 16: **Natural claim–artifact binding diagnostics.** CF and AF denote claim-fixed and artifact-fixed accuracy; PBA is strict. Entries are percentages.

Domain	Scorer	B pairs	Rescue	CF	AF	PBA
Code	Artifact-only	120	50.0	60.8	50.0	0.0
	V0	120	63.3	65.8	60.8	28.3
	Reference judge	120	89.2	92.5	87.5	76.7
Mathematics	Artifact-only	80	50.0	58.8	50.0	0.0
	V0	80	62.5	65.0	60.0	27.5
	Reference judge	80	87.5	90.0	86.3	75.0

Table 17: **Bulk and deployment performance for G1.** Deployment uses $N^\star = 128$ in code and $N^\star = 32$ in mathematics.

Domain	Verifier	Bulk acc.	AUROC	Brier	Valid count	Validity
Code	V0	75.58	0.776	0.189	$633/768$	82.42
	V1	74.86	0.783	0.185	$692/768$	90.10
Mathematics	V0	72.18	0.739	0.203	$199/256$	77.73
	V1	72.91	0.733	0.206	$220/256$	85.94

I.5 Unresolved-case sensitivity

A common all-valid or all-invalid assignment need not minimize a paired policy difference. The robust calculation therefore preserves artifact identities.

I.6 Compute-aware comparisons

Fixed-$N^\star$ gains assume that deployment already generates the declared budget. On the observed G1 grids, the best raw points are 88.41% at $N = 32$ in code and 82.03% at $N = 16$ in mathematics. S1 at the larger primary budgets reaches 89.84% and 86.72%, corresponding to gains of $1.43\,\mathrm{pp}$ and $4.69\,\mathrm{pp}$ over the best observed smaller-budget raw policies.

S2 additionally requires a reference pool. The following comparison charges its full cost without amortization.

At matched total cost, S2 is $0.39\,\mathrm{pp}$ above S1 in code and $2.73\,\mathrm{pp}$ below S1 in mathematics. These S2–S1 differences are descriptive and do not have paired confidence intervals.

For V1, development selects $\alpha = 1$, so S1 equals raw selection: validity is 90.10% in code and 85.94% in mathematics, consistent with Table 17. S2 changes these values by $-0.65\,\mathrm{pp}$ and $0.78\,\mathrm{pp}$, respectively. A development-label permutation control reduces V0 validity by $4.56\,\mathrm{pp}$ in code and $3.12\,\mathrm{pp}$ in mathematics; this checks implementation dependence on development labels but is not a permutation-test null distribution.

J Full Relational-Training Results

J.1 Seed-level natural and controlled results

All checkpoints act on the same prospective Code/G1 candidate frame at $N^\star = 128$. Each checkpoint has 768 natural selection units. Controlled denominators are 300 for seen and unseen PBA, 600 for CF and AF, and 1,200 for defect recognition.

Defect-oriented M1 improves its matched defect control but does not improve natural deployment. Relational supervision produces the larger changes in natural validity, mismatch rate, and unseen-family binding. Means and sample standard deviations appear in Table 2.

Table 18: **Discovery-stage G1/V2 score saturation.** Atom mass, N_{95}, and distinct-score counts are target medians.

Statistic	Code	Mathematics
Maximum-score atom mass	15.00%	22.00%
Validity within maximum atom	82.10%	77.30%
Median N_{95}	19	13
Median distinct tail scores	1	1
$R_A(N^\star)$	82.16%	76.95%
$R_A(N_{\max})$	82.03%	77.34%

Table 19: **G1/V0 discovery and confirmation results.** Differences are confirmation minus discovery. Intervals compare independent study stages and are not equivalence tests.

Domain	Estimand	Discovery	Confirmation	Difference [95% CI]
Code	Deep-tail A	74.3	78.9	4.6 $[-0.3, 9.5]$
	Deep-tail B	18.9	14.8	-4.1 $[-8.3, 0.1]$
	Peak-to-tail ΔA	-16.7	-12.9	3.8 $[-2.1, 9.7]$
Mathematics	Deep-tail A	68.9	74.8	5.9 $[-2.0, 13.8]$
	Deep-tail B	20.1	15.8	-4.3 $[-10.8, 2.2]$
	Peak-to-tail ΔA	-18.1	-14.7	3.4 $[-5.4, 12.2]$

J.2 Natural relation-family generalization

The natural evaluation is divided into 384 seen-family and 384 held-out family selection units.

The corresponding held-out-family means are 80.47% for M0, 85.16% for M2, 86.55% for M3, and 88.28% for M4. Thus, the relational gains are not confined to transformation families seen during training.

J.3 Forecasting repaired checkpoints

The same forecast and prospective splits are used for every checkpoint. Policy weights are recomputed after each checkpoint changes the score ordering.

J.4 Paired training effects

The defect-control bounds exceed the registered -1 pp noninferiority margin. M4 has a positive point difference over S1 applied to the original V0, but the lower bound does not establish superiority. Selection and training therefore remain complementary repairs.

The M4–M3 comparison isolates the change from separately penalized margins to crossed margin aggregation under matched data and resources. Because the objectives also differ in the placement of the nonlinear penalty, the experiment does not uniquely attribute the gain to additive-effect cancellation.

K Statistical and Implementation Validation

K.1 Estimator checks

The checks verify the estimator implementation. The study design and audits provide separate evidence for candidate independence, transport to the evaluation population, and certification accuracy.

Table 20: **Complete G1/V0 confirmation curves.** Forecast and observed values are percentages. Counts give certified-valid selections over all prospective selection units. $N = 1$ is excluded from WMAE.

	Code			Mathematics		
N	Forecast	Valid count	Observed	Forecast	Valid count	Observed
1	51.83	$403/768$	52.47	52.7	$137/256$	53.52
2	65.2	$497/768$	64.71	64.5	$161/256$	62.89
4	77.1	$587/768$	76.43	74.2	$193/256$	75.39
8	83.9	$650/768$	84.64	79.4	$201/256$	78.52
16	87.5	$666/768$	86.72	80.8	$210/256$	82.03
32	87.3	$679/768$	88.41	78.4	$199/256$	77.73
64	85.1	$646/768$	84.11	75.6	$196/256$	76.56
128	82.9	$633/768$	82.42		Not evaluated	
256	80.8	$617/768$	80.34		Not evaluated	

Table 21: **Confirmation WMAE across twelve settings.** Entries are percentage points. Binary and class-resolved Horvitz–Thompson estimators target the same class-A event.

Setting	Verifier	HT binary	HT class-resolved	Flexible binary	Monotone
Code/G1	V0	0.71	0.71	0.83	4.35
	V1	1.32	1.32	1.18	1.25
	V2	1.45	1.45	1.39	1.51
Code/G2	V0	1.58	1.58	1.31	2.74
	V1	2.21	2.21	1.63	1.71
	V2	1.73	1.73	1.59	1.82
Math/G1	V0	1.09	1.09	1.26	3.62
	V1	2.04	2.04	1.88	1.95
	V2	2.37	2.37	2.16	2.28
Math/G2	V0	2.18	2.18	1.91	3.29
	V1	2.76	2.76	2.31	2.45
	V2	2.91	2.91	2.67	2.84

K.2 Known-truth interval simulations

These simulations assess the implemented procedure under four specified regimes; coverage under other data-generating processes was not evaluated.

K.3 Annotation and provenance checks

The discovery mathematics frame contains 4,500 unique artifacts and 9,000 primary assessments. Primary-class agreement is 89.29%, with Cohen's $\kappa = 0.82$; target-validity agreement is 93.91%. After adjudication, 93 artifacts (2.07%) remain unresolved.

On a 200-artifact discovery gold set, target-validity accuracy is $190/200$. Class-B precision is $45/48 = 93.75\%$, and recall is $45/50 = 90.00\%$. These discovery diagnostics are separate from the confirmation audits in Table 13.

The integrity audit found:

- no primary class-B certificate requiring artifact modification;
- no class-B certificate based on a non-frozen neighbor;
- no forecast or prospective labels used to fit evaluated selectors;
- no family assignments made after inspecting evaluated outcomes; and
- no unresolved case silently converted to a resolved label.

Table 22: **V0 forecasts of selected validity and mismatch at $N^\star$.** Forecasts were frozen before prospective generation.

Setting	Forecast A	Observed A	Forecast B	Observed B
Code/G1	82.9	82.4	12.2	12.4
Code/G2	86.3	86.7	8.4	7.4
Math/G1	78.4	77.7	13.5	14.1
Math/G2	81.7	82.4	10.5	9.8

Table 23: **Confirmation policy gains at $N^\star$.** S1 is interior-rank selection and S2 is reference-calibrated selection. Wins and losses count paired changes in class-A selection relative to raw maximum selection. LCBs are one-sided clustered 95% lower confidence bounds.

Setting	Policy	Units	Wins/losses	Forecast gain	Observed gain	LCB
Code/G1	S1	768	$82/25$	7.9	7.4	4.6
	S2	768	$91/25$	9.0	8.6	5.4
Code/G2	S1	512	$38/15$	3.9	4.5	0.8
	S2	512	$42/15$	5.6	5.3	1.3
Math/G1	S1	256	$31/8$	7.5	9.0	3.4
	S2	256	$29/9$	8.2	7.8	2.2
Math/G2	S1	256	$24/10$	4.8	5.5	0.7
	S2	256	$26/10$	5.6	6.3	1.1

All mathematics artifacts selected differently by compared policies were independently re-audited. The supplementary manifest links model revisions, candidate records, parser and oracle versions, certification records, forecasts, policies, and paired comparisons. Hashes and timestamps establish that primary forecasts and policies preceded prospective candidate generation.

K.4 Clustering and multiplicity

The resampling unit is the target when targets are independent and the source group otherwise. Repeated prospective blocks, nested budgets, policies, and verifiers are resampled jointly.

Natural-tail intervals are simultaneous over the registered primary contrast family. Policy and training lower bounds apply only to their stated endpoint comparisons and do not imply simultaneous coverage over all reported metrics, policies, budgets, and checkpoints. Discovery, family-level decompositions, and other secondary analyses are descriptive unless an inferential procedure is explicitly stated.

L Reproducibility, Ethics, and Artifact Release

L.1 Artifact contents

The anonymous supplementary artifact contains the linked components summarized below.

Distinct draw identifiers are retained even when raw outputs or parsed artifacts coincide. Publicly releasable model, data, parser, and environment versions are recorded in machine-readable manifests. For proprietary API components, the artifact provides exact endpoint versions, prompts, decoding parameters, request metadata, permitted cached outputs, and the resulting reproduction limitations.

L.2 Human annotation

Mathematics annotation uses independent assessment, adjudication, and explicit abstention. Reviewers first assess the requested target while blinded to verifier scores, score strata, and neighboring claims. Annotation interfaces limit task duration and preserve unresolved outcomes when evidence is insufficient. Annotators were compensated according to the time and expertise required by the task.

Table 24: **Selected-class counts at $N^\star$.** Code/G1 has 768 units, Code/G2 has 512, and each mathematics setting has 256.

Setting	Policy	A	B	C	U_0	$U_?$
Code/G1	Raw	633	95	11	29	0
	S1	690	50	8	20	0
	S2	699	43	8	18	0
Code/G2	Raw	444	38	8	22	0
	S1	467	25	6	14	0
	S2	471	22	6	13	0
Math/G1	Raw	199	36	4	12	5
	S1	222	19	3	8	4
	S2	219	21	3	9	4
Math/G2	Raw	211	25	4	12	4
	S1	225	16	3	9	3
	S2	227	14	3	9	3

Table 25: **Mathematics policy gains under alternative unresolved outcomes.** Entries are percentage points. Robust bounds retain paired selected-artifact identities.

	G1/S1	G1/S2	G2/S1	G2/S2
All unresolved invalid	8.98	7.81	5.47	6.25
All unresolved valid	8.59	7.42	5.08	5.86
Outer lower bound	7.03	5.86	3.91	4.69
Outer upper bound	10.16	9.38	6.64	7.42
Robust 95% LCB	2.80	1.92	0.31	0.54

Released records use pseudonymous assessor identifiers while preserving the links needed to audit agreement and adjudication.

L.3 Executable-code safety

Generated programs are executed in a network-disabled environment with fixed filesystem, process, time, and memory controls. Non-termination and resource violations are retained as outcomes. The artifact records the sandbox image and evaluator configuration.

Passing the registered exhaustive finite-domain evaluation establishes correctness only for the declared contract and environment; it is not presented as unrestricted software correctness.

L.4 Use and transport risks

Verifier-guided search can amplify confidence in outputs favored by an incomplete evaluator. The principal risk is transporting a forecast or selection policy to a different target distribution, generator, verifier, or budget without remeasurement.

The artifact therefore reports deployment budgets, score resolution, selected-error composition, certification scope, unresolved outcomes, and candidate-cost assumptions alongside performance results. Interior-rank selection should be re-estimated whenever these deployment conditions change.

L.5 Release metadata

The artifact records model and checkpoint identifiers, applicable licenses, prompts, decoding parameters, training configurations, compute accounting, data manifests, repository structure, and file hashes. Symbolic labels such as G1, V0, and M4 are linked to concrete component identifiers in the experiment manifest. All authors remain responsible for the scientific claims, annotation protocol, released records, and final manuscript.

Table 26: **G1/V0 validity under matched candidate accounting.** For budget-matched raw and S1, all $N + L$ candidates are selectable.

Domain	Policy	Selectable	Reference	Total	Validity
Code	Raw	128	0	128	82.42
	S1	128	0	128	89.84
	S2	128	512	640	91.02
	Raw, matched total	640	0	640	79.43
	S1, matched total	640	0	640	90.63
Mathematics	Raw	32	0	32	77.73
	S1	32	0	32	86.72
	S2	32	128	160	85.55
	Raw, matched total	160	0	160	74.22
	S1, matched total	160	0	160	88.28

Table 27: **Complete confirmation results by checkpoint.** Natural columns are selected class counts. Controlled columns are numbers correct at the denominators stated above. M1–M4 use matched training seeds.

Model	Seed	A	B	C	U_0	Seen PBA	Unseen PBA	CF	AF	Defect
M0	0	633	95	11	29	90	75	384	366	1128
M1	11101	628	115	5	20	94	76	392	377	1150
	11102	637	106	4	21	91	82	401	385	1154
	11103	619	123	5	21	88	69	383	369	1146
M2	11101	665	68	11	24	205	151	472	458	1136
	11102	681	57	9	21	218	166	485	470	1129
	11103	669	66	10	23	198	157	476	449	1138
M3	11101	683	56	9	20	225	181	502	490	1139
	11102	692	48	10	18	237	194	514	501	1144
	11103	668	67	11	22	213	176	495	483	1132
M4	11101	696	44	9	19	238	196	516	503	1136
	11102	700	43	9	16	245	202	520	509	1139
	11103	685	55	10	18	226	190	510	498	1133

Table 28: **Natural validity by relation-family split.** Entries are class-A counts.

Model	Seed	Seen $A/384$	Held-out $A/384$	Total $A/768$
M0	0	324	309	633
M2	11101	341	324	665
	11102	350	331	681
	11103	343	326	669
M3	11101	350	333	683
	11102	354	338	692
	11103	342	326	668
M4	11101	356	340	696
	11102	358	342	700
	11103	350	335	685

Table 29: **Forecasting repaired Code/G1 verifiers.** Forecast and observed values at $N^\star$ are percentages; WMAE is measured over the eight registered budgets with $N \geq 2$.

Model	Seed	Forecast at $N^\star$	Observed at $N^\star$	WMAE
M3	11101	89.50	88.93	1.18
	11102	89.40	90.10	1.36
	11103	87.80	86.98	1.55
M4	11101	90.20	90.63	1.12
	11102	90.65	91.15	1.21
	11103	89.75	89.19	1.33

Table 30: **Confirmation training effects.** Differences and one-sided 95% lower confidence bounds are percentage points. Comparisons preserve matched training seeds and paired target- or block-level records.

Comparison	Difference	95% LCB
M4 minus M3: natural validity	1.65	0.47
M4 minus M2: natural validity	2.86	1.18
M4 minus M3: held-out natural validity	1.74	0.38
M4 minus M3: unseen CF	1.94	0.56
M4 minus M3: unseen AF	2.00	0.61
M0 minus M4: selected-B rate	6.21	3.28
M4 minus M0: defect control	0.67	-0.18
M4 minus M3: defect control	-0.19	-0.72
M4 minus original-V0 S1: natural validity	0.48	-0.51

Table 31: **Finite-pool estimator checks.** Exhaustive small-pool enumeration verifies the implemented order-statistic weights and bookkeeping.

Diagnostic	Tolerance	Observed
Maximum-selection weight error	10^{-10}	3.7×10^{-14}
Fixed-rank weight error	10^{-10}	1.1×10^{-14}
Frozen-utility weight error	10^{-10}	4.5×10^{-14}
Absolute weight-sum error	10^{-10}	2.2×10^{-16}
Change after merging non-A classes	10^{-12}	3.3×10^{-13}
Missing required inclusion probabilities	0	0
Nonzero-weight items with zero inclusion probability	0	0
Subset evaluations that refit a policy	0	0

Table 32: **Known-truth whole-curve interval validation.** Each regime uses 1,000 simulations at nominal simultaneous 95% coverage. Signed mean error is in percentage points.

Regime	Seed	Curves covered	Coverage	Signed mean error
Monotone validity	17201	944	94.4%	0.07 pp
Validity turnover	17202	951	95.1%	−0.11 pp
Rare upper-tail errors	17203	938	93.8%	0.09 pp
Maximum-score atom	17204	957	95.7%	−0.06 pp

Table 33: **Supplementary artifact structure.**

Component	Contents
Study manifests	Task sources, source groups, frozen neighborhoods, model and tokenizer versions, parsers, and execution oracles.
Generation records	Requests, seeds, raw outputs, parsed artifacts, failure states, execution logs, and sandbox identifiers.
Scoring records	Verifier inputs, raw logits and scores, extraction versions, and saved score- and utility-tie randomizers.
Certification records	Target judgments, neighbor certificates, local defects, adjudication, confidence, provenance, and unresolved outcomes.
Sampling records	Full-frame label masks, inclusion probabilities, target weights, and policy-selection weights.
Forecast records	Immutable forecast files, hashes, policy versions, and prospective selected-candidate records.
Training records	M0–M4 configurations, matched seeds, checkpoint manifests, and controlled relational blocks.
Reproduction scripts	Scripts that regenerate manuscript tables and figures from the candidate-, annotation-, block-, and estimator-level records.