

Verifier Selection With Limited Candidate-Pool Headroom

A local engineering pilot

Cisco Caceres · 6 October 2026 UTC · Draft technical report v1.0

Abstract

We evaluated budgeted selection on four generated candidates for each of 120 frozen HumanEval tasks. A pinned local code generator and scalar judge produced 960 accounted requests; isolated execution retained all 480 hidden-test outcomes and 196 public-test outcomes on a fixed 49-task subset. First-candidate success was 109/120 (90.83%). Soft best-of-n reached 91.39% at three larger token caps, a three-seed task-mean difference of +0.56 percentage points, with a paired descriptive interval of [0,+1.67 points]. Other unfiltered policies matched first candidate; all policies on the public-qualified subset achieved 47/49. A post-hoc pool diagnostic found only two mixed-outcome tasks, limiting oracle improvement to one task. These findings diagnose the feasibility of selection experiments on this collected pool. They do not test the separate joint-verifier hypothesis, establish superiority, novelty or frontier transfer, or constitute outside peer review.

1. Question and scope

The engineering question is: **can a bounded local collection, public-only selection and isolated grading pipeline produce complete, auditable policy comparisons, and does this candidate pool leave enough headroom to make those comparisons informative?**

This report concerns completed exploratory execution. It does not report the proposed study of correlated cross-verifier errors under natural task-family shift. That study requires additional verifier families, new task distributions, matched ablations, adequate error diversity, a priced power plan and prospective registration. None is supplied by relabeling this pilot.

The implementation integrates an existing dataset, existing models, scalar judging, classic selection policies, token accounting and disposable VM execution. Original work here is the bounded collection/contract handling, provenance and failure ledger, qualified extraction/grading integration, policy replay, reconciliation and reporting. It is not a new model architecture, novel learning algorithm, frontier-scale training result or an exact reproduction of another paper's system.

2. Data, split and reference qualification

We used [HumanEval's upstream source](#) at commit `6d43fb980f9fee3c892a914eda09951f772ad10d`, retrieved 6 October 2026 and licensed MIT. The 164-record compressed dataset SHA256 is `b796127e635a67f93fb35c04f4cb03cf06f38c8072ee7cee8833d7bee06979ef`.

Tasks were sorted by SHA256 of `cisco-verifier-engineering-v1:` plus task ID. The first 120 form this pilot, followed by 20 development, 20 calibration and four diagnostic records. Exact IDs and individual prompt/reference/test commitments are recorded in the frozen manifest. The reserved calibration split does not imply that this pilot's scalar judge was calibrated; no learned calibrator or joint-error estimator was fitted here.

Before collection, canonical hidden references passed on all 120 pilot tasks. Public examples were eligible only where reference extraction/execution qualified: 49 of 57 prompt-example tasks. All 120 stayed in the unfiltered analysis. The public subset is a reference-qualification subset, not a random sample or a filter based on generated candidate outcomes.

The original inline public extractor and its case payload were not retained. A prospective reconstruction directly decoded the trusted upstream entry-function docstring with AST and `doctest.DocTestParser`; it did not obtain tests from candidates or normalize their source. Direct example counts matched 55/57 historical records. HumanEval/10 and /32 parsed three

and two examples, while old metadata recorded extraction failures. New qualification passed 120/120 hidden and 51/57 public references. The original 49-task eligibility remained frozen: newly passing tasks were not added. Exact historical case-byte parity is not claimed.

HumanEval is a familiar public benchmark introduced in 2021. Task assignment does not establish absence of model-training exposure or provide fresh-data generalization. Deterministic hidden-test success means passing the declared suite, not proving general program correctness.

3. Models and collection contract

Collection began at 2026-10-06 04:11:16.534820 UTC from source commit

ab5cd6b56c2d38059e09687ffc12b8cec6778d4e . The execution owner identified the local Warden installation as Ollama 0.35.1. Model digests and request settings were pinned; a full serving-binary/environment lock is not supplied here. Exact serving tags are pinned by digest:

Role	Tag	Ollama digest
Generator	qwen3-coder:30b	06c1097efce0431c2045fe7b2e5108366e43bee1b4603a7aded8f21689e90bca
Judge	qwen3.5:9b	6488c96fa5faab64bb65cbd30d4289e20e6130ef535a93ef9a49f42eda893ea7

Digest pins identify the serving artifacts, not an independent model-family comparison. Both roles belong to the Qwen model ecosystem; independence of their errors is not assumed or established.

For each task, four candidates were generated in fixed sample order 0–3. Each generation was followed by its judge call. Both roles used temperature 0.2, context cap 4,096, `think=false`, nonstreaming responses and seed 20261006 + sample ; keep-alive was five minutes. Generator output was capped at 512 tokens and judge output at 128. Calls had a 120-second hard deadline, with ten-second metadata checks, a four-hour run ceiling, no retries, 960-call maximum, 3,932,160 prompt-token ceiling and 307,200 returned-token ceiling. Pins were checked before each call and returned model/completion metadata checked afterward.

The generator was asked for structured JSON containing a `code` string representing a full Python module. JSON parsing, AST parsing and exactly one matching top-level entry-function name were checked separately. Exact signature equivalence was not enforced at this contract stage; execution supplied the functional test. No Markdown stripping, source repair or regeneration was allowed. Invalid generation responses remained unchanged failures. The judge received the specification and either parsed source or the original failed response with its contract status; it returned structured JSON with a finite score in [0,1]. Invalid judge JSON would produce an explicit abstention, not an invented numeric score. Four candidates failed source contracts; zero judges abstained. Infrastructure, digest and unknown-usage errors were fail-closed conditions.

Earlier development contracts are historical engineering evidence, not silently discarded warmups: fourteen body-contract responses failed because they used Markdown; a separately frozen structured-module diagnostic later produced twelve passing candidates on six development tasks. Neither diagnostic is pooled with this 120-task analysis.

4. Isolated execution and outcome denominators

Usable generated source executed only through an audited host/guest grader in owned disposable VMs from a verified base image. Generated Python was not executed on the host. Grading reports record no NIC, no host mounts and no observed root-escape sentinel. Case and runner/image hashes, ordered identities, completion records and outcome coverage were reconciled across five candidate batches, each with at most 180 cases. Reports retained reference and negative controls, including premature zero exits.

The declared guest limits were 256 MiB candidate memory, 16 tasks/processes, three seconds CPU and five seconds inner runtime wall time. The host measured elapsed wall time around the outer wrapper, bounded by 15 seconds. CPU usage and local GPU execution time were not measured. These engineering safeguards do not establish malicious grading

integrity: candidate and tests share an interpreter, so a sufficiently adversarial candidate could attempt to forge a completion marker.

All 480 hidden slots were retained. Public execution covered 49 tasks $\times$ four candidates = 196 slots. Four invalid-source candidates, all outside the public subset, received failure labels without execution; hence four automatic failure slots, not eight duplicated hidden/public failures. Passing candidates numbered 437/480 hidden and 188/196 public. Qualification reference outcomes and negative controls are separate from these candidate denominators.

5. Frozen selection and token-budget replay

After outcome reconciliation, selection consumed only sanitized generation/judge token usage, scalar scores and qualified public-test metadata. No hidden labels entered the selection inputs. All 9,288 decisions were frozen to SHA256 `da9a0cc203fbc44e7511a384cf2257f11becc18c06ab72d8735e44bdf4d4d25` before hidden-label analysis, as recorded by the execution owner. This is an internal commitment, not independent timestamp custody or formal external preregistration.

Four input-plus-output token caps were used: 4,608, 9,216, 18,432 and 36,864. Each replay action reserved its conservative worst-case cost before acquisition: 4,608 for generation and 4,224 for judging. Actual consumed tokens were charged after each permitted action. These ceilings can prevent an action despite low observed average usage. Public execution used a separate 15-second outer-wall reservation; it was not converted to tokens, dollars or measured CPU seconds. Policies shared each task's same ordered four-candidate pool, observed usage and judge scores.

Policy	Prespecified behavior
Single	Select the first candidate without a judge or public-test requirement.
Best-of-n	Acquire eligible candidates within the cap and choose highest observed judge score; earliest tie.
Soft best-of-n	Same acquisition, then sample scored candidates with weights proportional to $\exp(\text{score})$, inverse temperature 1.
Bounded best-of-Poisson	Request $1 + \text{Poisson}(3)$ candidates, truncate at four and the token cap, then choose highest score.
First-public	On the 49-task subset, stop at the first public-test pass; no judge required.
Uniform-public	On that subset, select uniformly among acquired public-test passes; no judge required.

The first four policies ran on all 120 tasks and also on the 49-task subset. Public filtering applied to the other policies on that subset, but **single remained an unfiltered first-candidate comparator**. Subset membership and filtering must not be conflated. If no eligible candidate existed, selection abstained and counted as failure. An eligible but unscored fallback could be selected after budget exhaustion. Invalid-judge abstention was prespecified to stop acquisition and select highest scored eligible candidate, or earliest eligible fallback; none occurred in this run. No threshold policy was added.

Replay seeds 0,1,2 affect tie/random policy behavior within the fixed pools; they do not create independent generation pools or independent benchmark samples. Bounded Poisson and soft selection are engineering baselines with disclosed defaults, not claims of exact reproduction of published robust-selection methods.

6. Descriptive analysis and results

Each task's success is the mean of its three replay-seed binary outcomes. Tasks receive equal weight. Paired effects compare each policy against single on the identical subset and cap. We used 10,000 bootstrap resamples of paired task means, seed 20261006, with percentile endpoints at indices $\text{floor}(0.025 \times 9,999)$ and $\text{floor}(0.975 \times 9,999)$. The resampling cluster is the task; shared source/family dependence and generator-pool variation are not modeled.

These conditional descriptive intervals do not receive a confirmatory superiority interpretation. Multiple policy/cap comparisons are exploratory and unadjusted. A zero interval for paired differences can simply reflect identical observed task outcomes; it does not establish equivalence or a population-level zero effect.

First candidate passed 109/120 tasks. Soft best-of-n improved only at the three larger caps, averaging 329 successful seed decisions out of 360 (91.3889%); its task-mean effect was +0.5556 percentage points with [0,+1.6667] descriptive interval. The lowest cap and other unfiltered policies matched first candidate. All public-subset policies passed 47/49 tasks. The full 40-group table follows; costs are counterfactual replay consumption, not actual collection cost.

Subset	Cap (tokens)	Policy	Success	Effect vs single (pp)	Descriptive 95% interval (pp)	Mean consumed tokens	Cap-exhausted seeds
120 unfiltered	4,608	best_of_n	90.83%	+0.00	[+0.00, +0.00]	566.7	360/360
120 unfiltered	4,608	bounded_best_of_poisson	90.83%	+0.00	[+0.00, +0.00]	566.7	360/360
120 unfiltered	4,608	single	90.83%	+0.00	[+0.00, +0.00]	395.0	0/360
120 unfiltered	4,608	soft_best_of_n	90.83%	+0.00	[+0.00, +0.00]	566.7	360/360
120 unfiltered	9,216	best_of_n	90.83%	+0.00	[+0.00, +0.00]	3,109.4	9/360
120 unfiltered	9,216	bounded_best_of_poisson	90.83%	+0.00	[+0.00, +0.00]	3,109.4	9/360
120 unfiltered	9,216	single	90.83%	+0.00	[+0.00, +0.00]	395.0	0/360
120 unfiltered	9,216	soft_best_of_n	91.39%	+0.56	[+0.00, +1.67]	3,109.4	9/360
120 unfiltered	18,432	best_of_n	90.83%	+0.00	[+0.00, +0.00]	3,142.5	0/360
120 unfiltered	18,432	bounded_best_of_poisson	90.83%	+0.00	[+0.00, +0.00]	3,142.5	0/360
120 unfiltered	18,432	single	90.83%	+0.00	[+0.00, +0.00]	395.0	0/360
120 unfiltered	18,432	soft_best_of_n	91.39%	+0.56	[+0.00, +1.67]	3,142.5	0/360
120 unfiltered	36,864	best_of_n	90.83%	+0.00	[+0.00, +0.00]	3,142.5	0/360
120 unfiltered	36,864	bounded_best_of_poisson	90.83%	+0.00	[+0.00, +0.00]	3,142.5	0/360
120 unfiltered	36,864	single	90.83%	+0.00	[+0.00, +0.00]	395.0	0/360
120 unfiltered	36,864	soft_best_of_n	91.39%	+0.56	[+0.00, +1.67]	3,142.5	0/360
49 public-qualified	4,608	best_of_n	95.92%	+0.00	[+0.00, +0.00]	541.9	147/147
49 public-qualified	4,608	bounded_best_of_poisson	95.92%	+0.00	[+0.00, +0.00]	541.9	147/147
49 public-qualified	4,608	first_public	95.92%	+0.00	[+0.00, +0.00]	328.1	6/147
49 public-qualified	4,608	single	95.92%	+0.00	[+0.00, +0.00]	328.1	0/147
49 public-qualified	4,608	soft_best_of_n	95.92%	+0.00	[+0.00, +0.00]	541.9	147/147
49 public-qualified	4,608	uniform_public	95.92%	+0.00	[+0.00, +0.00]	328.1	147/147
49 public-qualified	9,216	best_of_n	95.92%	+0.00	[+0.00, +0.00]	2,540.5	0/147

Subset	Cap (tokens)	Policy	Success	Effect vs single (pp)	Descriptive 95% interval (pp)	Mean consumed tokens	Cap-exhausted seeds
49 public-qualified	9,216	bounded_best_of_poisson	95.92%	+0.00	[+0.00, +0.00]	2,540.5	0/147
49 public-qualified	9,216	first_public	95.92%	+0.00	[+0.00, +0.00]	372.8	0/147
49 public-qualified	9,216	single	95.92%	+0.00	[+0.00, +0.00]	328.1	0/147
49 public-qualified	9,216	soft_best_of_n	95.92%	+0.00	[+0.00, +0.00]	2,540.5	0/147
49 public-qualified	9,216	uniform_public	95.92%	+0.00	[+0.00, +0.00]	1,304.6	0/147
49 public-qualified	18,432	best_of_n	95.92%	+0.00	[+0.00, +0.00]	2,540.5	0/147
49 public-qualified	18,432	bounded_best_of_poisson	95.92%	+0.00	[+0.00, +0.00]	2,540.5	0/147
49 public-qualified	18,432	first_public	95.92%	+0.00	[+0.00, +0.00]	372.8	0/147
49 public-qualified	18,432	single	95.92%	+0.00	[+0.00, +0.00]	328.1	0/147
49 public-qualified	18,432	soft_best_of_n	95.92%	+0.00	[+0.00, +0.00]	2,540.5	0/147
49 public-qualified	18,432	uniform_public	95.92%	+0.00	[+0.00, +0.00]	1,304.6	0/147
49 public-qualified	36,864	best_of_n	95.92%	+0.00	[+0.00, +0.00]	2,540.5	0/147
49 public-qualified	36,864	bounded_best_of_poisson	95.92%	+0.00	[+0.00, +0.00]	2,540.5	0/147
49 public-qualified	36,864	first_public	95.92%	+0.00	[+0.00, +0.00]	372.8	0/147
49 public-qualified	36,864	single	95.92%	+0.00	[+0.00, +0.00]	328.1	0/147
49 public-qualified	36,864	soft_best_of_n	95.92%	+0.00	[+0.00, +0.00]	2,540.5	0/147
49 public-qualified	36,864	uniform_public	95.92%	+0.00	[+0.00, +0.00]	1,304.6	0/147

The companion machine-readable aggregate is `examples/verifier-reliability/results/2026-10-06-local-pilot-summary.json` , SHA256 `3ea3b1a6fe2daf407fbb7fd1bf98e329cee69eb03c983534bb3c295cf1404e0e` . It preserves full precision, paired task counts, usage, controls and the disclosed post-hoc diagnostic. A repository-relative identity does not itself imply public source access.

6.1 Post-hoc pool headroom

After frozen policy analysis, a boolean-outcome diagnostic found 108 tasks with four passing candidates, ten with none and two with mixed outcomes. Oracle selection of any passing candidate reaches 110/120. Relative to 109/120 first-candidate success, the maximum observed improvement is one task, or 0.8333 percentage points. This is a ceiling for these exact pools, not a universal bound on selection and not a prespecified primary endpoint.

Most tasks therefore provide no within-pool selection distinction. This is a descriptive explanation of why policies can consume more tokens without changing observed success. SoftBoN's three-seed improvement is confined to variation available in the mixed pools; no causal claim about its general advantage follows. The pilot cannot tell whether these findings persist with different samples, larger pools, model families, genuinely new tasks or feedback-conditioned generation.

7. Actual costs and failures

Actual collection accounted for all 960 requests: 276,882 prompt tokens and 100,216 returned tokens, 377,098 total input-plus-output tokens. Four source-contract failures were retained without repair or regeneration. Zero judges abstained. Qualified grading reconciled every candidate slot, including automatic failures. Reported external paid API, rented-compute and training spend was \$0; local electricity, amortized hardware and labor were not measured. Local inference is not claimed to have zero total economic cost.

The table counts only actions consumed by each counterfactual policy on cached observations. It does not refund the actual expense of collecting unused candidates or judges. The minimum-cap scored policies can acquire an eligible candidate but lack sufficient worst-case reservation for a further action, explaining nonzero cap-exhaustion counts despite small measured consumption. A cap-exhausted policy still uses its prescribed fallback; exhaustion is not automatically an incorrect answer.

The four contract failures are distinct from VM execution failures and model reasoning errors. This report does not infer a specific cause, such as truncation, from a failed contract alone. Earlier formatting failures remain separately versioned. No outcome-dependent repair, hidden-label selection, policy retuning or favorable-seed selection was performed after the frozen decision commitment.

8. Verification, access and limitations

A separate same-owner agent independently checked only completed metadata, not raw responses or candidate/test code. It matched source closure, manifest/coverage commitments, all 9,288 decisions, all 40 groups, actual usage, paired counts and every seeded task bootstrap. That recomputation disclosed the larger-cap SoftBoN improvement and the single-policy public-filter distinction. It is an internal assisted check, not external peer review, independent replication or attestation of trusted timestamp custody.

The source/data closure includes the frozen collector and plan, task manifest, licensed dataset, extraction/grading methods, replay/statistical helpers and prospective evaluator. Historical and prospective identities are not conflated: the new extractor is explicitly newly qualified, and missing historical case bytes remain missing. Independently frozen renderer-input hashes additionally bind publication metadata to grading/decision/analysis artifacts.

The main limits are:

- Familiar public 2021 tasks and possible training contamination; no fresh-data or natural-family-shift inference.
- A fixed generator, fixed scalar judge and four-candidate pools with minimal observed headroom; no learned calibration or correlated cross-verifier experiment.
- Three cached policy seeds, not independent generation replications; conditional task bootstrap with multiple unadjusted exploratory comparisons.
- Limited public-example coverage and an originally selected 49-task subset; reference qualification is not a proof of hidden-suite adequacy.
- VM isolation with same-interpreter grading-integrity limitations; no mathematical proof of malicious-code safety.
- Token caps and outer wall-time diagnostics, not measured FLOPs, GPU time, CPU usage or total economic costs.

Full integration source and scientific artifacts are currently private or gated. Public counts, hashes and this report do not reconstruct raw generations, exact task-level decisions or independent VM execution. A permitted reproducibility bundle must supply exact revisions and manifests, pinned serving artifacts/dependencies, owner-approved sanitized decisions, separately supplied post-freeze boolean outcomes and applicable code/data rights. A scoped review license is necessary; no unrestricted integration-code license is inferred from HumanEval's MIT license or a neighboring repository's license.

Raw hidden tests, generated source, credentials, agent state/transcripts and private outputs are excluded from this publication draft.

For an authorized holder of completed private metadata, the frozen evaluator stages permit coverage reconciliation, public-only selection, independent decision-file hashing and then hidden-label analysis. Reproducing published arithmetic from aggregates is weaker than independently regenerating candidates or qualifying grading. A later outside reviewer should record identity/relationship, source and input hashes, environment, exact commands, failures and any amendments. No reviewer has been commissioned for this report.

9. Relation to existing work and next experiment

HumanEval supplies the benchmark and reference tests [1]. The limits of resampling with imperfect verifiers are already studied [2]. Robust inference-time reward selection, including Best-of-Poisson/HedgeTune, is also existing work [3]; our truncated Poisson baseline is not its complete method or guarantee. Other reviewed allocation/judging methods such as ADAP, CAPS, DaJ and GRACE overlap the broader proposed agenda. This report makes no novelty claim from combining an existing generator, judge and bounded replay.

The actionable finding is a **feasibility constraint**: this particular pilot pool offers only one task of oracle improvement over first candidate. A future selection study needs sufficient correct/incorrect diversity, new task families, applicable matched baselines and an outcome suite qualified before unblinding. It should account separately for real pool collection and counterfactual consumption, retain every failure and power the actual primary contrast. Increasing benchmark difficulty or changing generation settings after seeing this result would create a new prospectively versioned experiment, not a repaired confirmation of this one.

The proposed joint-error/natural-shift study remains unrun. Its next decision is whether a defensible, affordable and adequately powered contrast survives full related-work review. Neither the observed SoftBoN difference nor the minimal headroom justifies a claim of general superiority, equivalence or frontier-model research success.

References

1. Chen et al. (2021), [Evaluating Large Language Models Trained on Code](#); official HumanEval source.
2. Stroebel, Kapoor and Narayanan, [The Limits of Inference Scaling Through Resampling](#), v3.
3. Khalaf et al., [Inference-Time Reward Hacking in Large Language Models](#).
4. Reviewed overlap: [ADAP](#), [CAPS](#), [DaJ](#), [GRACE](#). Full baseline feasibility and any novelty distinction remain future gates.

Artifact commitments and authorship

Artifact	SHA256
Public aggregate	3ea3b1a6fe2daf407fbb7fd1bf98e329cee69eb03c983534bb3c295cf1404e0e
Frozen decisions	da9a0cc203fbc44e7511a384cf2257f11becc18c06ab72d8735e44bdfd4d4d25
Completed analysis	4fcd3b98255aa15d7a6f58a303a159fb143f9a411c4bc4426c4d424bc67df3cb
Grading coverage	a1633eba0b239bd81c4b984a56de3267f60d0f330c0e325d957895fec9e330c2
Candidate case manifest	d5b3420c07d57099c630bdd457e55c614ee13ad03d97eeec02f4bb388ab98c9b
Same-owner separate-agent metadata audit	3516212a250e466ae0dd5819a7a44f5d4d545559c1a79fe2fb81cab7c24a445

Cisco Caceres is responsible for implementation decisions and claims. AI assistance supported implementation, literature screening, arithmetic/provenance review and drafting. Models, benchmark, serving software, language/runtime and established selection/statistical methods are upstream contributions credited separately. This draft has no outside reviewer endorsement, venue acceptance or peer-review status.