# SIEVE: When Small Language Models Look in the Mirror
# Limits of Inference-Only Self-Verification across Model Families and Tasks

**Doguk Kim** [1]

## Abstract

Agentic pipelines increasingly delegate correctness checking to the generator model, replacing human oversight with *intrinsic self-verification*. We ask when this delegation is safe for small language models (SLMs). We present **SIEVE**, a controlled study of six instruction-tuned SLMs (Qwen3.5 and Gemma 4, 2B–31B) on GSM8K and HumanEval under four conditions: no verification (A), naive self-check (B), role-separated verification (C), and a high-capability external verifier (D, Gemini 3 Flash Preview). Three findings emerge. *(i)* On GSM8K, naive self-check produces a non-positive $\Delta$ for every tested model—including the AWQ-quantized 27B/31B tier—(down to $-15.5$pp for Qwen3.5-4B), with magnitude governed by family alignment rather than scale. *(ii)* Nominal gains for Gemma on HumanEval under self-verification have bootstrap CIs that overlap the baseline and remain unconfirmed. *(iii)* The external verifier is not a universal upper bound: with the prompt fixed, iteration-1 false-rejection rates on Condition-A-correct GSM8K answers span $0.0\%$ to $10.9\%$ across targets, while false-acceptance of incorrect answers is high even on the panel's thickest-support cells (e.g., $79.3\%$, $165/208$ incorrect Qwen3.5-9B GSM8K answers). We frame these as reproducible failure triggers and argue that SLM self-verification reliability is not a simple function of scale.

## 1. Introduction

A growing number of deployed agent loops resolve uncertainty by asking the generator to check itself: produce an answer, prompt the same model to look for errors, and act on whichever revision survives. The appeal is operational—one model, one inference budget, no oracle—and the cost is a strong assumption: that the same parameters that emitted a flawed answer can also detect the flaw. Recent work has questioned this assumption at frontier scale (Huang et al., 2024), yet most of the systems that perform such loops in practice run on small models (SLMs, $\leq$31B parameters) where the assumption is far less examined.

We treat this gap as a concrete failure-mode question. Rather than asking whether self-verification "works" on average, we ask: under what model, family, task, and verifier choices does inference-only self-verification *degrade* accuracy, and is the degradation a property of scale or of something more local?

We introduce **SIEVE** (**S**elf-verification **I**nvestigation across **E**mpirical **V**ariants and **E**valuations), a controlled study designed around a single change of variable per axis. We hold the prompt skeleton fixed and vary (i) the model family (Qwen3.5 vs. Gemma 4), (ii) the size tier (2B–31B effective parameters), (iii) the verification protocol (no check, naive self-check, role-separated self-check, high-capability external verifier), and (iv) the task structure (math reasoning vs. code generation). All experiments are run with greedy decoding on a single RTX 3090 Ti, with quantization (AWQ-4bit) only for the 27B/31B tier. The setup is small enough to be reproduced by other groups and large enough that family- and tier-level patterns become visible.

Our central observations are:

1. **Naive self-check on GSM8K is uniformly negative in direction and family-specific in magnitude.** Every model in our panel loses accuracy on GSM8K under naive self-check (Condition B), but the loss spans $-0.7$pp to $-15.5$pp at comparable parameter counts. Family alignment, not scale, is the dominant explanatory variable in this range.

2. **Role separation does not consistently rescue self-verification.** Role-separated prompts (Condition C) help some models on HumanEval and *hurt* others on GSM8K, with no monotonic relationship to size. We

[1]Department of Artificial Intelligence, Inha University, Republic of Korea. Correspondence to: Doguk Kim <dgkim@inha.ac.kr>.

Presented at the ICML 2026 Workshop on Failure Modes in Agentic AI (FAGEN), Seoul, South Korea. Copyright 2026 by the author(s).

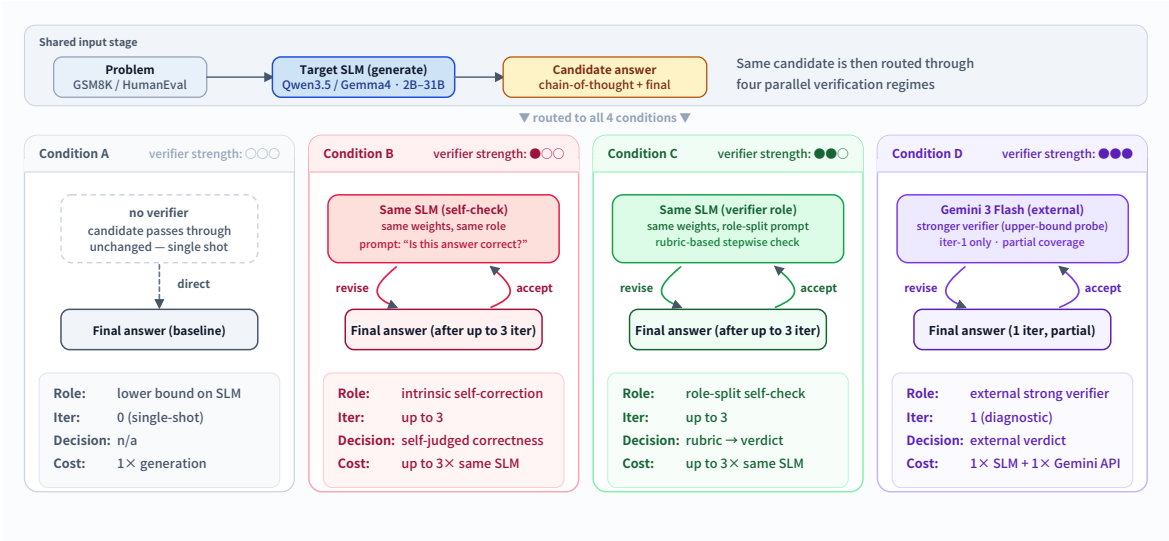



*Figure 1.* SIEVE study design. Each problem is routed through a fixed generation skeleton (target SLM, greedy decoding) and one of four verification conditions: (A) none, (B) same-model self-check, (C) same-model verification in a separate role, and (D) Gemini 3 Flash Preview as external verifier. B/C run up to three verify–revise rounds; D is an iteration-1 verifier-stage diagnostic. Six SLMs (Qwen3.5 {4B, 9B, 27B-AWQ}, Gemma 4 {E2B, E4B, 31B-AWQ}) are evaluated on GSM8K and HumanEval.

document a particularly clear inversion in Qwen3.5-9B, where Condition C is worse than Condition B on GSM8K.

3. **A high-capability external verifier exhibits target-dependent and direction-asymmetric failures.** With Gemini 3 Flash Preview as a fixed external verifier and a fixed verification prompt, iteration-1 false-rejection rates on Condition-A-correct GSM8K answers span 0.0% to 10.9% depending on the target model, though the upper end (Qwen3.5-4B) is partly driven by missing answer-format suffixes rather than by verifier strictness (§4.5). On the same panel the verifier accepts a large fraction of incorrect answers even where the incorrect-item support is thick—79.3% (165/208) for Qwen3.5-9B GSM8K, 70.1% (82/117) for Gemma4-E2B GSM8K—reaching 89.5% on the high-accuracy Gemma4-31B GSM8K cell, where only 38 incorrect items support the estimate (34/38; $\kappa$=0.14). The "strong verifier" upper bound depends on the target, and a single external verifier is reported here (§5.4).

We frame SIEVE around **reproducible triggers**—specific model×task×condition cells where the failure replicates with non-overlapping bootstrap CIs and a degradation-heavy flip count (e.g., Qwen3.5-4B GSM8K Condition B at 259 vs. 55 flips, a 4.7:1 harm-to-help ratio)—together with **trace-level diagnostics** (flip-direction decomposition, difficulty stratification, per-target verifier confusion matrices), and an honest accounting of coverage: Qwen3.5-27B-AWQ is reported across all four conditions on both benchmarks, while the heavier Gemma4-31B-AWQ is limited to Conditions A/B (GSM8K) and D for compute-budget reasons (§5.4). We mark all excluded cells as "—" rather than impute.

## 2. Related Work

**Intrinsic self-correction at scale.** Madaan et al. (2023) introduced iterative self-feedback as a generic refinement mechanism. Huang et al. (2024) argued that, absent oracle labels or external feedback, self-correction in large models is at best neutral and frequently harmful; Stechly et al. (2023) reached structurally similar conclusions for reasoning, planning, and constraint satisfaction. Our contribution is not to revisit the frontier-scale claim but to ask whether the phenomenon holds for SLMs at 2B–31B—the regime closer to deployed agent settings—and to separate family from scale as explanatory variables.

**Self-correction in small models.** The closest prior work is Zhang et al. (2024), who show that SLMs ($\leq$13B) require strong external verifiers for meaningful self-correction gains; Kamoi et al. (2024) catalogue when self-correction reliably helps (external tools, knowledge, fine-tuning) and characterize intrinsic self-correction as limited. SIEVE complements both: rather than averaging over models, we vary one axis at a time and report cell-level outcomes, exposing the family×task interactions that aggregate numbers hide.

**Tool-integrated and trained verification.** Outside the inference-only regime, tool-integrated verification (Kang et al., 2025) and verifier-trained models (Kumar et al., 2024)

narrow the gap. DeepSeek-AI (2025) report strong results from generator–verifier dual training on mathematical reasoning. These directions are orthogonal to ours: SIEVE measures the floor of inference-only self-verification, against which trained or tool-augmented systems can be benchmarked.

**Calibration, self-preference, and confirmation.** Classical calibration work (Guo et al., 2017) established that confidence does not in general track accuracy, and Kadavath et al. (2022) showed that LLMs predict their own correctness largely at scale. Self-preference effects (Panickssery et al., 2024) and CoT confirmation bias (Wan et al., 2025) provide candidate mechanisms for the asymmetric flip rates we observe; we do not claim a mechanism, only the trace-level signatures consistent with them. We do not compute a parametric calibration error (our pipeline did not log confidence; §5.4); the asymmetric flip rates serve as its behavioral proxy.

**Generation–verification asymmetry.** Chen et al. (2026) find that generation skill does not transfer to self-verification skill. Our family-specific magnitudes are consistent with this asymmetry being itself family- and alignment-dependent rather than purely a property of scale.

## 3. Method

### 3.1. Models

We evaluate six instruction-tuned models drawn from two families and three size tiers, summarized in Table 1. The two families were chosen because they share an instruction-tuning regime broadly typical of modern open SLMs but differ in alignment recipe; the size tiers were chosen so that family pairs are matched within $\pm 10\%$ of effective parameter count. AWQ 4-bit quantization (Lin et al., 2023; Frantar et al., 2022) is applied at the 27B/31B tier to fit a single RTX 3090 Ti (24GB VRAM); smaller tiers run in BF16. We use vLLM (Kwon et al., 2023) with greedy decoding (temperature 0.0) and a 4096-token context window throughout. The Phase 3 BF16 runs predate vLLM 0.19.x strict context-length validation; we re-audited those JSONLs for silent-truncation markers (`finish_reason: length`, embedded `[truncated]` tags) and found none, so the BF16 results are unaffected by the truncation behavior documented for the AWQ-tier Conditions B/C in §5.4. Model weights for the Qwen3.5 family (Qwen Team, 2026) and the Gemma 4 family (Gemma Team, 2026) are obtained from the families' public releases; the Condition-D verifier is Gemini 3 Flash Preview (Gemini Team, 2026).

*Table 1.* Models evaluated in SIEVE. Effective parameter count for Gemma 4 follows the family's published per-activation count.

| Family | Variant | Eff. params | Precision |
|---|---|---|---|
| Qwen3.5 | 4B | 4B | BF16 |
| Qwen3.5 | 9B | 9B | BF16 |
| Qwen3.5 | 27B | 27B | AWQ-4bit |
| Gemma 4 | E2B | 2B | BF16 |
| Gemma 4 | E4B | 4B | BF16 |
| Gemma 4 | 31B | 31B | AWQ-4bit |

### 3.2. Benchmarks

**GSM8K** (Cobbe et al., 2021): 1,319 grade-school math problems with verifiable final-answer extraction. We stratify by reference solution length into Easy ($\leq$3 steps), Medium (4–6), and Hard (>6).

**HumanEval** (Chen et al., 2021): 164 Python function-completion problems with a hidden test suite. We use pass@1 with the canonical test runner.

### 3.3. Verification Conditions

Each condition is implemented as a fixed prompt skeleton; only the verifier identity and the role assignment change across conditions. Iterations refer to verify–revise rounds; iteration 0 is the bare generation.

- **Condition A (No Verification).** Single-shot generation. Iteration 0 only. This is the operational baseline.
- **Condition B (Naive Self-Check).** The same model is prompted to inspect its own previous answer and either confirm or revise. Up to 3 iterations.
- **Condition C (Role-Separated Self-Verification).** The same model is queried with a verifier-style prompt that asks for an explicit `VERDICT: CORRECT/INCORRECT` judgement on the candidate solution, then a separate revision call is made if the verdict is `INCORRECT`. Up to 3 iterations.
- **Condition D (External High-Capability Verifier, partial).** The verification step is delegated to Gemini 3 Flash Preview with the same role-separated prompt as Condition C; the target SLM still performs revision. Gemini 3 Flash Preview is the low-latency variant of the Gemini 3 family, used as the strongest external verifier within our API budget; we treat D as an iteration-1 verifier-stage diagnostic rather than a definitive upper bound (§4.5), leaving multi-iteration refinement and additional verifiers to an extended version.

The full prompt texts are in Appendix A.

### 3.4. Metrics

We report (i) accuracy at iteration 0 and at the final iteration, (ii) $\Delta$ relative to Condition A, (iii) item-level bootstrap 95% confidence intervals over the question set with 10,000 resamples (each resample draws items with replacement and recomputes accuracy), (iv) flip counts decomposed into correct→incorrect (degradation) and incorrect→correct (correction), and (v) for Condition D, the verifier false-rejection rate on Condition-A-correct items and the false-acceptance rate on Condition-A-incorrect items.

### 3.5. Scope and Excluded Cells

Among the AWQ-quantized tier, Qwen3.5-27B-AWQ is evaluated under all four conditions (GSM8K and HumanEval), whereas Gemma4-31B-AWQ is restricted to Conditions A/B on GSM8K plus D: its GSM8K Condition C run repeatedly stalled during vLLM engine startup, and step times of $\sim$30 s/it on the available KV-cache budget would have required $>$33 wall-clock hours per run without parallelism, which itself triggered VRAM spills during pilots (§5.4). Gemma4-31B-AWQ on HumanEval is omitted because the KV-cache budget could not host both the full prompt set and the AWQ-4bit model. All such cells are listed as "—" rather than imputed.

## 4. Experiments

### 4.1. Main Accuracy

Table 2 reports accuracy at iteration 0 (Condition A) and at the final iteration for B, C, and the iteration-1 verifier-stage accuracy for D. Bootstrap 95% confidence intervals appear in Appendix B.

Three observations follow from Table 2.

**(R1) The direction of Condition-B effect on GSM8K is universal in our panel.** For every model with a completed Condition-B cell on GSM8K, accuracy at the final iteration is at or below the Condition-A baseline. The 95% bootstrap CIs (Appendix B) confirm that the drops for Qwen3.5-4B ($-15.5$pp), Gemma4-E2B ($-5.6$pp), and the AWQ-quantized Qwen3.5-27B ($-9.5$pp) and Gemma4-31B ($-4.7$pp; CI $[91.0, 93.8]$ vs. A $[96.1, 98.0]$) tiers are non-overlapping with their A baselines; the smaller drops for Gemma4-E4B and Qwen3.5-9B overlap. The pattern therefore extends to the largest (31B) tier in our panel rather than dissolving with scale.

**(R2) The magnitude of the effect is family-specific, not scale-monotonic.** At $\sim$4B parameters, Qwen3.5-4B loses 15.5pp on GSM8K under Condition B while Gemma4-E4B loses 0.7pp—a 21$\times$ ratio at matched scale. Within the Qwen family, the 9B drop ($-1.4$pp) is smaller than the 4B drop, and this is not confounded by a stronger 9B baseline: Qwen3.5-9B's Condition-A accuracy ($84.2\%$) is in fact slightly *below* the 4B baseline ($86.0\%$), so the smaller drop reflects fewer harmful revisions rather than reduced headroom to fall. Within the Gemma family, the BF16 tiers (E2B $-5.6$pp; E4B $-0.7$pp) are likewise non-monotonic in scale. Scaling does not reliably "solve" inference-only self-check in either family.

**(R3) Condition C does not consistently improve over Condition B.** On GSM8K, Condition C is better than B for Qwen3.5-4B ($-5.8$ vs. $-15.5$) and Gemma4-E2B ($-0.2$ vs. $-5.6$), comparable for Gemma4-E4B ($-1.7$ vs. $-0.7$), and *worse* for Qwen3.5-9B ($-11.0$ vs. $-1.4$). On HumanEval, Condition C produces both the largest nominal gain in the panel (Gemma4-E4B $+4.9$pp) and the largest nominal drop (Qwen3.5-4B $-15.2$pp). Role separation is therefore not a reliable mitigation; on GSM8K it is in fact *correction-starved* (§4.3), the within-model echo of the external verifier's near-constant-classifier behavior (§4.5). The 9B inversion is statistically robust—its Condition C GSM8K bootstrap CI $[70.8, 75.7]$ does not overlap its Condition B CI $[80.8, 84.8]$—and it is itself non-monotonic in scale: the 4B and 27B-AWQ tiers both show Condition C *less* harmful than B, whereas only the intermediate 9B tier shows the reverse. Why this amplification peaks mid-scale is left for an extended evaluation with seed variation.

### 4.2. Difficulty-Stratified Effects on GSM8K

Aggregate accuracy obscures where the harm concentrates. Table 3 stratifies the GSM8K results by reference-solution length. The clearest pattern is that Condition B's harm is not concentrated in any single difficulty bin in a family-independent way.

One cell stands out. Qwen3.5-9B Hard shows a $+3.8$pp gain under Condition B—the only positive entry in the GSM8K Condition-B column—but the bootstrap CIs ([54.3, 73.3] for A vs. [59.0, 76.2] for B) overlap, so we treat this as suggestive rather than confirmed. Across the BF16 tiers we did not observe a difficulty bin in which Condition B reliably improves on Condition A; the CI-confirmed harm concentrates in the Easy/Medium bins, where Condition A leaves more room to fall (e.g., Gemma4-E2B: Easy $-4.1$pp and Medium $-8.6$pp confirmed, Hard $-4.8$pp overlapping), and only Gemma4-E4B is genuinely near-flat across bins. The AWQ-quantized tier is excluded from Table 3 because Gemma4-31B-AWQ lacks Condition C and Qwen3.5-27B-AWQ's per-bin counts are sensitive to the context-truncation caveat of §5.4; we note, however, that the Gemma4-31B-AWQ aggregate Condition-B drop ($-4.7$pp) averages a strongly difficulty-dependent effect—$-3.1$pp on Easy and $-3.7$pp on Medium but $-21.0$pp on Hard ($94.3\% \rightarrow 73.3\%$; CIs $[89.5, 98.1]$ vs. $[64.8, 81.9]$, non-overlapping). The largest-tier degradation of finding (R1) is therefore concentrated

*Table 2.* Final target-model accuracy (%) and Δ from Condition A under Conditions A, B, and C. The Condition D column is *not* a refined target-model accuracy: it is the iteration-1 verifier accuracy $(\text{TP}+\text{TN})/n$ (Gemini 3 Flash Preview as judge), not directly comparable to A/B/C and reported without a Δ; per-target decomposition in Table 4. "—" marks cells excluded for compute budget (§5.4).

| Model | GSM8K | | | | HumanEval | | | |
|---|---|---|---|---|---|---|---|---|
| | A | B (Δ) | C (Δ) | D‡ | A | B (Δ) | C (Δ) | D‡ |
| Qwen3.5-4B | 86.0 | 70.5 (−15.5) | 80.2 (−5.8) | 84.6 | 68.3 | 47.0 (−21.3) | 53.0 (−15.2) | 74.4 |
| Qwen3.5-9B | 84.2 | 82.9 (−1.4) | 73.2 (−11.0) | 87.2 | 62.2 | 58.5 (−3.7) | 51.8 (−10.4) | 64.0 |
| Qwen3.5-27B-AWQ | 90.8 | 81.3 (−9.5) | 84.8 (−6.0) | 93.0 | 72.6 | 65.2 (−7.4) | 67.1 (−5.5) | 71.3 |
| Gemma4-E2B | 91.1 | 85.5 (−5.6) | 91.0 (−0.2) | 93.7 | 81.7 | 84.8 (+3.0) | 88.4 (+6.7) | 92.1 |
| Gemma4-E4B | 93.3 | 92.6 (−0.7) | 91.6 (−1.7) | 96.7 | 89.0 | 91.5 (+2.4) | 93.9 (+4.9) | 95.1 |
| Gemma4-31B-AWQ | 97.1 | 92.4 (−4.7) | — | 96.7 | — | — | — | — |

‡ Condition D entries are iteration-1 verifier accuracy, not refined model accuracy (see caption).

*Table 3.* Condition-B Δ on GSM8K by difficulty (pp). Difficulty bins: Easy (≤3 steps), Medium (4–6), Hard (>6).

| Model | Easy | Medium | Hard |
|---|---|---|---|
| Qwen3.5-4B | −16.4 | −14.5 | −12.4 |
| Qwen3.5-9B | −2.8 | ±0.0 | +3.8 |
| Gemma4-E2B | −4.1 | −8.6 | −4.8 |
| Gemma4-E4B | −1.4 | +0.5 | ±0.0 |

where multi-step reasoning is hardest, which sharpens rather than dilutes the scale claim.

### 4.3. Flip-Direction Asymmetry

Aggregate Δ values can be produced by very different mixtures of beneficial and harmful flips. We decompose each cell into correct→incorrect (degradation) and incorrect→correct (correction) flips between iteration 0 (Condition A) and the final iteration of the relevant condition.

Two cells anchor the asymmetry. On GSM8K Condition B, Qwen3.5-4B accumulates 259 degradation flips against 55 corrections, a 4.7:1 absolute-count ratio (§5 decomposes this into per-item rates and base-rate effects). On HumanEval Condition B, Qwen3.5-4B loses 56 of its 112 Condition-A-correct items, a 34.1% in-cell degradation rate.

Iterations compound rather than cancel the harm. For Qwen3.5-9B GSM8K Condition C, the cumulative correct→incorrect count rises from 81 at iteration 1 to 155 at iteration 3 against only 10 cumulative corrections. We saw no panel cell on GSM8K in which additional iterations produced a net positive trajectory after iteration 1.

Role separation is also *correction-starved* on GSM8K, which sharpens the mechanism behind R3. Across the four BF16 models, Condition C rescues only 1.6–10.1% of the answers that began incorrect (24 corrections against 269 degradations, an 11.2:1 ratio), whereas the same models under Condition B rescue 11.2–40.9%. The verdict-style prompt thus issues its INCORRECT verdicts predominantly on answers that were already correct while almost never catching genuinely wrong ones—the within-model counterpart of the near-constant-classifier behavior of the external verifier (§4.5)—which explains why role separation hurts on GSM8K despite the intuition that an explicit verifier role should improve error detection.

Figure 2 visualizes the rate-level signature across the panel. The dominant pattern across GSM8K is degradation-heavy: every model has more correct→incorrect than incorrect→correct flips under Condition B, including Qwen3.5-9B (whose flip imbalance is the narrowest in the panel at 103 vs. 85, so its accuracy aggregate reads as essentially flat despite a net-negative flip direction). HumanEval is mixed, with Gemma cells turning slightly green under Condition C while every Qwen cell—including the Qwen3.5-27B-AWQ tier (B +0.10, C +0.09)—remains red, so the degradation-heavy signature persists into the largest quantized model.

### 4.4. Iteration Dynamics

Figure 3 traces accuracy across iterations 0–3 under Conditions B and C. Qwen3.5-4B GSM8K under Condition B drops monotonically with no recovery; Gemma4-E4B HumanEval under Condition C rises at iteration 1 and plateaus. Qwen3.5-27B-AWQ follows the same degradation-then-plateau pattern as the BF16 Qwen tiers on GSM8K, though its trajectory should be read alongside the context-truncation caveat of §5.4. No panel cell showed reliable additional gains beyond iteration 1.

### 4.5. Condition D: Verifier-Target Sensitivity

Condition D probes an empirical upper bound for the verification step by replacing the SLM verifier with Gemini 3 Flash Preview, holding the prompt fixed. We compute the iteration-1 verifier confusion matrix against Condition-A ground truth on each target SLM; two failure modes emerge that an aggregate "strong-verifier accuracy" number would hide.

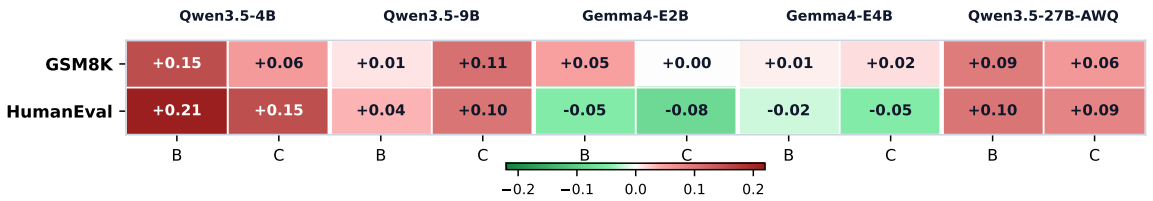
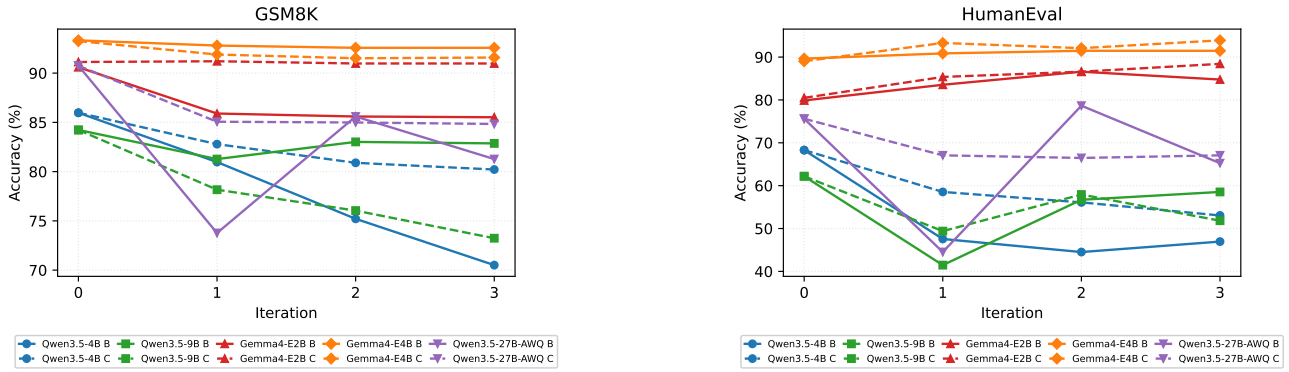



*Figure 2.* Per-item flip asymmetry. Each cell reports (correct→incorrect rate) − (incorrect→correct rate) between iteration 0 (Condition A) and the final iteration under Condition B or C. Positive values (red) indicate net degradation; negative (green) net correction. Cell values closely track the sign and, on GSM8K, the magnitude of the corresponding $\Delta$ in Table 2; on HumanEval ($n$=164) a handful of cells diverge somewhat in magnitude because the two statistics are computed differently (population-level accuracy change vs. per-item flip rate) and are not expected to match exactly at this sample size. Gemma4-31B-AWQ is omitted (no Condition C on either benchmark). Panel covers two benchmarks and five target SLMs.




*(a)* GSM8K. *(b)* HumanEval.

*Figure 3.* Accuracy across iterations under Conditions B and C for the four BF16 tiers plus Qwen3.5-27B-AWQ. Iteration 0 corresponds to Condition A. Gemma4-31B-AWQ is omitted (no Condition C on either benchmark); the Qwen3.5-27B-AWQ curves are subject to the context-truncation caveat of §5.4 (per-item truncation rate 37.9% for Condition B on GSM8K).

**Strictness vs. leniency are direction-asymmetric across families.** Table 4 reports per-target precision metrics. On Qwen3.5-4B GSM8K, the verifier rejects 10.94% (124/1,134) of Condition-A-correct answers as INCORRECT; manual inspection of these cases shows the verifier flagging solutions that lack the canonical `#### <answer>` suffix even when the chain-of-thought arrives at the correct number, so the elevated rejection rate reflects target-side output-format compliance as much as verifier strictness (§5.4). The verifier is simultaneously *lenient* on Gemma4-E2B GSM8K, accepting 70.1% of Condition-A-incorrect answers as correct (false positives 82/117). The two biases coexist on the same verifier, with the same prompt, on the same task; only the target SLM changes. The most extreme false-acceptance rate, 89.5% on Gemma4-31B GSM8K, must be read against its denominator: because that target is 96.7% accurate under Condition A, only $n_{\text{inc}}$=38 incorrect items remain (34/38), so the estimate is high-variance and the cell's $\kappa$=0.14 indicates a verifier behaving almost as a constant "CORRECT" classifier. The same lenient direction is visible on cells with much thicker support—79.3% (165/208) for Qwen3.5-9B GSM8K, 86.7% (39/45) for Qwen3.5-27B HumanEval, and 42.7% (79/185) for Qwen3.5-4B GSM8K itself—so the finding does not hinge on the thinnest cell; the high-accuracy targets simply yield the noisiest, and rhetorically largest, individual estimates.

**The two biases compose into nearly degenerate cases.** Qwen3.5-4B HumanEval is illustrative: the verifier accepts 78.8% (41/52) of Condition-A-incorrect solutions while rejecting only 0.89% (1/112) of correct ones. Aggregate verifier accuracy reads 74.4% ($\kappa$=0.26), but the underlying pass rate masks how close the verifier comes to a constant "CORRECT" classifier on this cell.

**Implication for "strong verifier" baselines.** On the Gemma cells, the iteration-1 verdicts would refuse to flag most incorrect candidates and induce no useful refinement; on Qwen3.5-4B GSM8K, they would trigger spurious refinements on nearly one in five correct candidates. Either failure misleads if reported as a single number: verifier-target sensitivity is a first-order axis for self-verification claims, not a higher-order correction.

*Table 4.* Condition-D iteration-1 verifier behavior. "FN" = false-rejection rate (Condition-A-correct items marked INCORRECT); "FP" = false-acceptance rate (Condition-A-incorrect items marked CORRECT); $n_{\text{inc}}$ = FP denominator—FP rates on small $n_{\text{inc}}$ are high-variance and should be read with this column. Cohen's $\kappa$ measures verifier-truth agreement; values near zero indicate a near-constant classifier on that cell.

| Target | Bench | FN | FP | $n_{\text{inc}}$ | Acc | $\kappa$ |
|---|---|---|---|---|---|---|
| Gemma4-E2B | GSM8K | 0.08% | 70.1% | 117 | 93.7% | 0.43 |
| Gemma4-E2B | HumanEval | 2.24% | 33.3% | 30 | 92.1% | 0.71 |
| Gemma4-E4B | GSM8K | 0.49% | 41.6% | 89 | 96.7% | 0.69 |
| Gemma4-E4B | HumanEval | 2.74% | 22.2% | 18 | 95.1% | 0.75 |
| Gemma4-31B | GSM8K | 0.70% | 89.5% | 38 | 96.7% | 0.14 |
| Qwen3.5-4B | GSM8K | 10.94% | 42.7% | 185 | 84.6% | 0.42 |
| Qwen3.5-4B | HumanEval | 0.89% | 78.8% | 52 | 74.4% | 0.26 |
| Qwen3.5-9B | GSM8K | 0.36% | 79.3% | 208 | 87.2% | 0.30 |
| Qwen3.5-9B | HumanEval | 6.86% | 83.9% | 62 | 64.0% | 0.11 |
| Qwen3.5-27B | GSM8K | 0.0% | 76.9% | 121 | 92.9% | 0.35 |
| Qwen3.5-27B | HumanEval | 6.72% | 86.7% | 45 | 71.3% | 0.08 |

$n_{\text{inc}}$ is small wherever the target is highly accurate under Condition A (few incorrect items remain); the Gemma4-31B GSM8K FP rate of $89.5\%$ rests on $n_{\text{inc}}$=38 ($34/38$) and is correspondingly high-variance, consistent with its near-zero $\kappa$.

# 5. Discussion

## 5.1. The Asymmetry of Self-Doubt

The clearest panel-level signal is that the Qwen3.5 family revises its answers far more often than the Gemma 4 family at matched scale. On GSM8K Condition B, Qwen3.5-4B issues degradation flips on 259 of its 1,134 Condition-A-correct items ($22.8\%$) against 55 corrections out of 185 Condition-A-incorrect items ($29.7\%$); per-item, repair is slightly more likely than harm, but Condition A is correct $86\%$ of the time, so the population accuracy still drops sharply. The $4.7$:1 absolute-count ratio is a base-rate consequence of this asymmetry, not evidence of an intrinsic bias toward harmful revisions. Gemma4-E4B at matched scale flips only 20 against 11—one-thirteenth the volume—revealing that the same prompt produces qualitatively different revision propensities in the two families. The lone nominal positive entry (Qwen3.5-9B GSM8K Hard, $+3.8$pp) has overlapping bootstrap CIs and is treated as unconfirmed (§4.2).

This asymmetry persists across Condition D, but with the verifier role flipped: judging Qwen3.5-4B GSM8K, the external verifier takes the strict posture ($10.94\%$ false rejection), which vanishes for the larger Qwen models ($0.36\%$ at 9B, $0.0\%$ at 27B). As discussed in §4.5, the elevated 4B rate is driven in part by output-format non-compliance—a target-side, not verifier-side, signal that nonetheless degrades any pipeline taking the verdict at face value.

## 5.2. Code vs. Math: Structure, Not Salvation

The HumanEval Gemma cells (B $+2.4$/$+3.0$pp; C $+4.9$/$+6.7$pp) are the only positive-delta entries in the table, but their 95% bootstrap CIs overlap the baseline (Appendix B; e.g., Gemma4-E4B Condition B [86.6, 95.7] vs. A [84.1, 93.3]), so we cannot reject the null of zero improvement for code self-verification in this panel. The structural property of code—partial verifiability through explicit syntactic constraints—may be doing real work, but our sample size cannot establish it.

## 5.3. Synthesis

Taken together, the three patterns above sketch a consistent picture of inference-only self-verification in the 2B–31B range and clarify what can and cannot be claimed from the present panel.

**What can be concluded from the present panel.** For SLMs in the 2B–31B range, intrinsic inference-only self-verification on math reasoning is reliably non-improving and frequently harmful, with magnitude determined more by family alignment than by parameter count. Role separation is not a robust mitigation, and a high-capability external verifier is not a free upper bound: its error profile depends on the target SLM in ways large enough to alter downstream conclusions.

**What cannot be concluded.** We cannot conclude that the negative direction extends to open-ended generation or multi-turn agent tasks, rule out that family-specific magnitudes compress at >31B or under different decoding temperatures, or claim on the present sample size that HumanEval self-verification works for Gemma.

**Implications for agent design.** For correctness-critical agent loops at SLM scale, replacing HITL oversight with same-model self-check is unsafe in the cells we examined. The implication is not "do not self-verify"; it is that the verifier choice and per-target false-rejection profile must be reported alongside aggregate accuracy. "Strong verifier" should be a measurement rather than a label.

**Candidate fixes (future work).** The trace-level diagnostics isolate two lightweight mitigations worth testing in the extended version: (i) *format-aware verification*—a verifier prompt that ignores presentation and judges only the underlying answer should narrow the Condition-D family gap driven by format non-compliance (§4.5); and (ii) *conditional self-verification*—a gate that disables refinement where the Condition-A baseline is high and verifier $\kappa$ is low would avoid the population-level harm we observe. We register these as the smallest interventions the present diagnostics motivate.

### 5.4. Limitations

Our study is bounded by a single 24GB GPU, which forced AWQ-4bit quantization at the 27B/31B tier. Within that tier, Qwen3.5-27B-AWQ is evaluated under all four conditions, whereas the heavier Gemma4-31B-AWQ is limited to Conditions A/B on GSM8K plus Condition-D diagnostics (§3). Scale-related claims therefore draw on the full Qwen3.5 ladder (4B/9B/27B-AWQ) and the BF16 Gemma tiers. AWQ quantization may itself bias self-verification behavior at the largest tier; we do not have BF16 controls at 27B/31B and cannot disentangle this from intrinsic family alignment.

The context window is held fixed at 4096 tokens to fit Conditions B/C iteration histories on the available KV-cache budget. GSM8K traces fit comfortably for the BF16 tiers ($\leq$9B; maximum observed chain-of-thought 2,861 tokens, no truncation). Qwen3.5-27B-AWQ, however, emits substantially longer rollouts, so when its previous-iteration answer is wrapped into the self-check prompt our pipeline truncates the head of that answer to fit `max_model_len`. For the Condition B GSM8K cell we report a *per-item* truncation rate (items truncated in $\geq$1 of 3 iterations) of $37.9\%$ ($500/1{,}319$) and a *per-call* rate of $17.4\%$ ($690/3{,}957$); these differ by $\approx 2.2\times$ because long items truncate across multiple iterations. Truncated answers are marked explicitly and retained in all metrics; no BF16 tier truncates, so this caveat applies only to the AWQ rows of Table 2. Our protocol initially scoped a confidence-calibration analysis (ECE) as a headline metric, but the pipeline did not log confidence at inference time, so we report accuracy- and flip-based diagnostics only; the direction-asymmetric flip rates (§4, §4.5) are a behavioral footprint of the same miscalibration ECE would quantify parametrically. We use greedy decoding without seed variation (vLLM batch ordering can introduce small non-determinism we do not characterize), and we do not include a self-consistency baseline (Wang et al., 2023); whether self-verification adds value *over* self-consistency is the single most informative direction for an extended evaluation. Condition D uses a single external verifier at iteration 1 only; replication with additional verifiers and multi-iteration refinement remain for the extended version. HumanEval at $n$=164 is too small to confirm the Gemma positive deltas. Finally, "family alignment" refers throughout to the families' instruction-tuning and post-training recipe rather than to the underlying architecture.

## 6. Conclusion

SIEVE characterizes inference-only self-verification in SLMs as a failure mode whose direction is consistent across our panel but whose magnitude is family-dependent rather than scale-determined. We document three reproducible triggers—naive self-check on GSM8K, role-separated verification on Qwen3.5-9B, and target-dependent verifier false-negatives in Condition D—and report them as cells rather than aggregates. The practical takeaway: "does self-verification help?" is the wrong question for SLMs at this scale; the right question is for which family, on which task, with which verifier, at which false-negative cost.

## References


Chen, M. et al. (2021). Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*. https://arxiv.org/abs/2107.03374.

Chen, Y., Wang, Y., Zhang, Y., Ye, Z., Cai, Z., Shi, Y., Gu, Q., Su, H., Cai, X., Wang, X., Zhang, A., and Chua, T. (2026). Learning to Self-Verify Makes Language Models Better Reasoners. *arXiv preprint arXiv:2602.07594*. https://arxiv.org/abs/2602.07594.

Cobbe, K. et al. (2021). Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*. https://arxiv.org/abs/2110.14168.

DeepSeek-AI (2025). DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning. *arXiv preprint arXiv:2511.22570*. https://arxiv.org/abs/2511.22570.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. (2022). GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. *arXiv preprint arXiv:2210.17323*. https://arxiv.org/abs/2210.17323.

Gemini Team (2026). Gemini 3 Flash Preview: A Low-Latency Multimodal Model. https://blog.google/technology/ai/google-gemini-3-ai/.

Gemma Team (2026). Gemma 4: Open Models for Advanced Reasoning. https://blog.google/technology/developers/gemma-4-open-models/.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. In *International Conference on Machine Learning (ICML)*. http://proceedings.mlr.press/v70/guo17a.html.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2310.01798.

Kadavath, S. et al. (2022). Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*. https://arxiv.org/abs/2207.05221.

Kamoi, R. et al. (2024). When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. *Transactions of the Association for Computational Linguistics (TACL)*. https://arxiv.org/abs/2406.01297.

Kang, M. et al. (2025). T1: Tool-Integrated Self-Verification for Test-Time Compute Scaling in Small Language Models. *arXiv preprint arXiv:2504.04718*. https://arxiv.org/abs/2504.04718.

Kumar, A. et al. (2024). Training Language Models to Self-Correct via Reinforcement Learning. *arXiv preprint arXiv:2409.12917*. https://arxiv.org/abs/2409.12917.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*. https://arxiv.org/abs/2309.06180.

Lin, J., Tang, J., Tang, H., Yang, S., Deng, X., and Han, S. (2023). AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration. *arXiv preprint arXiv:2306.00978*. https://arxiv.org/abs/2306.00978.

Madaan, A. et al. (2023). Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2303.17651.

Panickssery, A., Bowman, S. R., and Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. *arXiv preprint arXiv:2404.13076*. https://arxiv.org/abs/2404.13076.

Qwen Team (2026). Qwen3.5 Technical Report. https://github.com/QwenLM/Qwen3.5.

Stechly, K., Marquez, M., and Kambhampati, S. (2023). GPT-4 Doesn't Know It's Wrong: An Analysis of Iterative Prompting for Reasoning Problems. *arXiv preprint arXiv:2310.12397*. https://arxiv.org/abs/2310.12397.

Wan, Y., Jia, X., and Li, X. L. (2025). Unveiling Confirmation Bias in Chain-of-Thought Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*. doi: 10.18653/v1/2025.findings-acl.195. https://aclanthology.org/2025.findings-acl.195/.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2023). Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2203.11171.

Zhang, Y. et al. (2024). Small Language Models Need Strong Verifiers to Self-Correct Reasoning. In *Findings of the Association for Computational Linguistics (ACL Findings)*. https://arxiv.org/abs/2404.17140.

## A. Verification Prompts

**Condition B (Naive Self-Check).** *"Check your previous answer. If there is an error, correct it. Otherwise, repeat the final answer."*

**Condition C (Role-Separated Self-Verification).** *"Analyze the candidate solution for correctness. Identify any logical errors or calculation mistakes. Respond with VERDICT: CORRECT or INCORRECT, followed by a brief justification."*

**Condition D.** Identical to Condition C, but the verifier call is routed to Gemini 3 Flash Preview instead of the target SLM.

## B. Bootstrap Confidence Intervals

We compute 95% bootstrap CIs over the question set with $n$=10,000 resamples. Full per-cell intervals appear in Table 5.

*Table 5.* 95% bootstrap CIs (per cell, accuracy in %).

| Cell | Acc | 95% CI |
|---|---|---|
| Gemma4-E2B GSM8K A | 91.1 | [89.5, 92.7] |
| Gemma4-E2B GSM8K B | 85.5 | [83.5, 87.4] |
| Gemma4-E2B GSM8K C | 91.0 | [89.3, 92.6] |
| Gemma4-E4B GSM8K A | 93.3 | [91.9, 94.5] |
| Gemma4-E4B GSM8K B | 92.6 | [91.1, 93.9] |
| Gemma4-E4B GSM8K C | 91.6 | [90.0, 93.0] |
| Gemma4-31B GSM8K A | 97.1 | [96.1, 98.0] |
| Gemma4-31B GSM8K B | 92.4 | [91.0, 93.8] |
| Qwen3.5-4B GSM8K A | 86.0 | [84.1, 87.9] |
| Qwen3.5-4B GSM8K B | 70.5 | [68.0, 73.0] |
| Qwen3.5-4B GSM8K C | 80.2 | [78.2, 82.3] |
| Qwen3.5-9B GSM8K A | 84.2 | [82.3, 86.3] |
| Qwen3.5-9B GSM8K B | 82.9 | [80.8, 84.8] |
| Qwen3.5-9B GSM8K C | 73.2 | [70.8, 75.7] |
| Qwen3.5-27B GSM8K A | 90.8 | [89.2, 92.4] |
| Qwen3.5-27B GSM8K B | 81.3 | [79.2, 83.3] |
| Qwen3.5-27B GSM8K C | 84.8 | [82.8, 86.7] |
| Gemma4-E2B HEval A | 81.7 | [75.6, 87.2] |
| Gemma4-E2B HEval B | 84.8 | [79.3, 90.2] |
| Gemma4-E2B HEval C | 88.4 | [82.9, 92.7] |
| Gemma4-E4B HEval A | 89.0 | [84.1, 93.3] |
| Gemma4-E4B HEval B | 91.5 | [86.6, 95.7] |
| Gemma4-E4B HEval C | 93.9 | [90.2, 97.6] |
| Qwen3.5-4B HEval A | 68.3 | [61.0, 75.0] |
| Qwen3.5-4B HEval B | 47.0 | [38.4, 54.9] |
| Qwen3.5-4B HEval C | 53.0 | [45.7, 61.0] |
| Qwen3.5-9B HEval A | 62.2 | [54.9, 69.5] |
| Qwen3.5-9B HEval B | 58.5 | [51.2, 66.5] |
| Qwen3.5-9B HEval C | 51.8 | [44.5, 59.8] |
| Qwen3.5-27B HEval A | 72.6 | [65.9, 79.3] |
| Qwen3.5-27B HEval B | 65.2 | [57.9, 72.6] |
| Qwen3.5-27B HEval C | 67.1 | [59.8, 74.4] |
| Gemma4-E2B GSM8K D | 93.7 | [92.4, 95.0] |
| Gemma4-E2B HEval D | 92.1 | [87.8, 95.7] |
| Gemma4-E4B GSM8K D | 96.7 | [95.8, 97.6] |
| Gemma4-E4B HEval D | 95.1 | [91.5, 98.2] |
| Gemma4-31B GSM8K D | 96.7 | [95.8, 97.6] |
| Qwen3.5-4B GSM8K D | 84.6 | [82.6, 86.5] |
| Qwen3.5-4B HEval D | 74.4 | [67.7, 81.1] |
| Qwen3.5-9B GSM8K D | 87.2 | [85.4, 88.9] |
| Qwen3.5-9B HEval D | 64.0 | [56.7, 71.3] |
| Qwen3.5-27B GSM8K D | 92.9 | [91.5, 94.3] |
| Qwen3.5-27B HEval D | 71.3 | [64.0, 78.0] |