QUANTIZATION CAN PRESERVE VERIFICATION WHILE DEGRADING SELF-GENERATED SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Can a quantized model retain its ability to select answers yet lose the learning value of its own solutions? We separate verification ranking, final-answer correctness and held-out learning gain. On Gemma 4 12B and Qwen3.6-35B-A3B, all nine registered 4- and 3-bit verifier contrasts are equivalent to full precision within ± 0.025 AUROC, even as generation degrades. Crossing sample source with learner precision reveals that SmolLM3-3B's full-precision solutions at 38% correctness teach both learners more than NF4 solutions at 74% (by 1.9 and 3.6 accuracy points) on a holdout selected for partial solvability. On the 12B, approximately matching sample count and problem coverage leaves an 18.9-point source advantage for the quantized learner on this holdout and 4.6 points on MATH-500. Completed, mostly correct NF4 solutions also reduce the full-precision learner's accuracy. The penalty increases along the GGUF precision ladder on both learning subjects, with format-dependent severity, and no tested selector shows an established recovery. Reciprocal continuations separately localize the 12B's excess truncation to the continuing model for the tested prefixes. Under this math supervised fine-tuning recipe, verification robustness and final-answer correctness are insufficient summaries of useful self-generated supervision.

1 INTRODUCTION

Self-training offers a way for a language model to learn from its own outputs: generate candidate solutions, select them with a verifier, and train on the accepted set (Zelikman et al., 2022; Singh et al., 2024; Yuan et al., 2024). Quantization reduces memory use by storing weights at lower numerical precision, changing the generator, verifier and learner. Its effect on generation accuracy is documented (Liu et al., 2025; Li et al., 2025b; Liang et al., 2026), but two further questions matter for self-training: does verification ranking survive, and does the selected data remain useful for learning?

These questions concern different properties of a solution set. A verifier ranks candidates by predicted correctness; a final-answer oracle measures how many accepted answers are correct; supervised fine-tuning (SFT) trains on the *whole retained solution*, including its intermediate text and its termination. Two sets can therefore have similar answer correctness yet provide different training signals. The generation–verification gap formalizes the benefit of reweighting generated answers, while realizing that benefit through training also requires a sufficiently faithful model update (Song et al., 2025; Amin et al., 2025). We test how quantization separates these quantities in an actual training recipe.

Our central finding is that **preserving verification ranking does not preserve the learning value of self-generated solutions**. Full-precision solutions can teach more even when fewer of their final answers are correct. Conversely, completed, mostly correct solutions from a quantized generator can damage a full-precision learner (Fig. 1). This comparison changes the generator's numerical representation while holding its underlying checkpoint fixed, and crosses that source change with the learner's representation. It therefore measures a source effect beyond the usual comparison between unrelated strong and weak teachers. The equivalence result concerns fixed-candidate ranking; the learning controls test whether changing the verifier or the source changes realized gains.

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.

Fixed-candidate verification tests (§3), source-by-learner training controls (§4) and reciprocal continuations (§5) separate these sensitivities. Together, they motivate evaluating generated solutions by realized learning value alongside correctness.

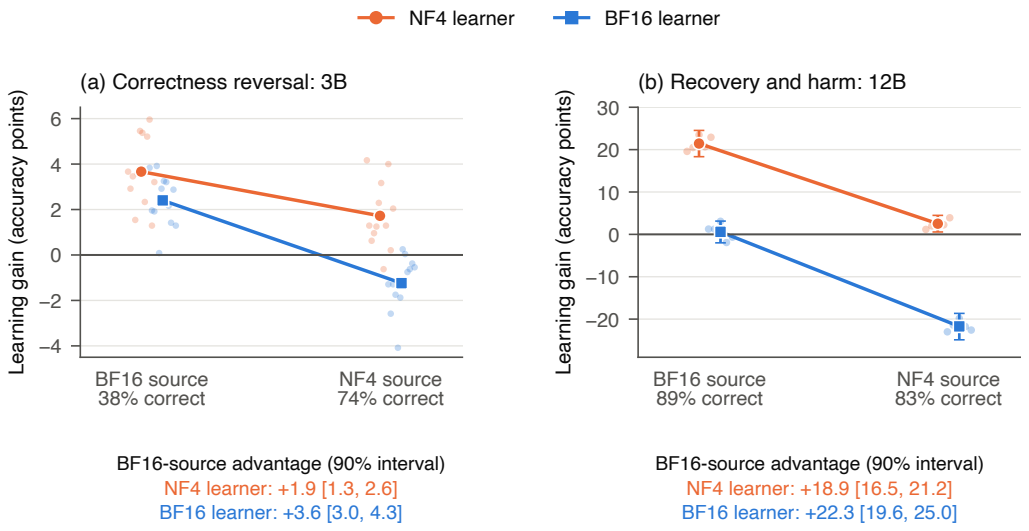

Figure 1: **The source changes what both learners gain from completed solutions. (a)** SmolLM3-3B: the least-correct BF16-source set (38%) outperforms the most-correct NF4-source set (74%) for both learners, with approximately equal count and coverage. **(b)** Gemma 4 12B: the count/coverage-matched BF16 source improves the NF4 learner, while the NF4 source harms the BF16 learner. Gains are accuracy changes in percentage points on 300 disjoint problems selected for partial solvability (§4); faint points are seeds. Source-advantage annotations use 90% intervals: paired-seed t in (a), crossed seed $\times$ problem in (b); bars on the cell means in (b) show crossed 95% intervals. The reversal is post hoc; the matched 12B source contrasts were pre-specified.

2 Study Design and Measurements

Ranking, correctness and learning gain. For model M and oracle correctness $c(x, y) \in \{0, 1\}$, generation accuracy is $G = \mathbb{E}_x \mathbb{E}_{y \sim M(\cdot|x)}[c]$ (pass@1). The model scores a supplied candidate on a fixed Yes/No prompt using $s = p_{\text{yes}}/(p_{\text{yes}} + p_{\text{no}})$ at the next output token. Verification ranking is $V = \text{AUROC}(s, c)$: the probability that a correct candidate outranks an incorrect one, with half credit for ties (alternative scores in Appendix J). Accepted-set correctness is the fraction of selected solutions whose final answers pass the oracle. Learning gain is $\Delta G_{\text{train}} = G_{\text{after}} - G_{\text{before}}$ on a disjoint holdout, relative to each learner's own starting checkpoint (inference in Appendices A and O).

Models and numerical representations. The primary subjects are Gemma 4 12B-it (dense) and Qwen3.6-35B-A3B (mixture of experts, MoE; 3 B active). Each uses one pinned 16-bit floating-point (BF16) checkpoint, one llama.cpp build and one importance matrix to construct a GGUF ladder: `Q8_0`, `Q4_K_M`, `Q3_K_M` and `IQ2_M` (nominally 8-, 4-, 3- and 2-bit; Gerganov & ggml contributors, 2026). Four-bit verifier comparisons also include bitsandbytes NormalFloat (NF4), used by QLoRA (Dettmers et al., 2023), in-repo AWQ (Lin et al., 2024), and the 12B vendor's QAT-Q4 checkpoint with its unquantized QAT-BF16 twin (Google, 2026). The learning experiments use the 12B and SmolLM3-3B, each with a BF16 or NF4 learner. Additional models test scope and held-out perturbation predictions (Appendix B); equal nominal bit widths need not induce equal perturbations.

Tasks and correctness. The static primary set is MATH-500 (Hendrycks et al., 2021; Lightman et al., 2024): all 500 problems, eight samples each, temperature 0.8, top-p 0.95, a 1,536-token cap, thinking off and deterministic per-sample seeds. The `math-verify` oracle checks the final

answer by mathematical equivalence (Hugging Face, 2025); it does not validate every reasoning step. A hard tier contains 153 AIME 2024–2026 and HMMT February 2025–2026 problems at a 4,096-token cap. Code uses HumanEval+/MBPP+ with EvalPlus execution oracles (Liu et al., 2023a) and a post-checkpoint AtCoder set (Appendix G). All generated responses enter generation accuracy: parse failures are incorrect, and truncated responses are graded on any answer already present. Thus truncation and correctness are separate, potentially overlapping outcomes.

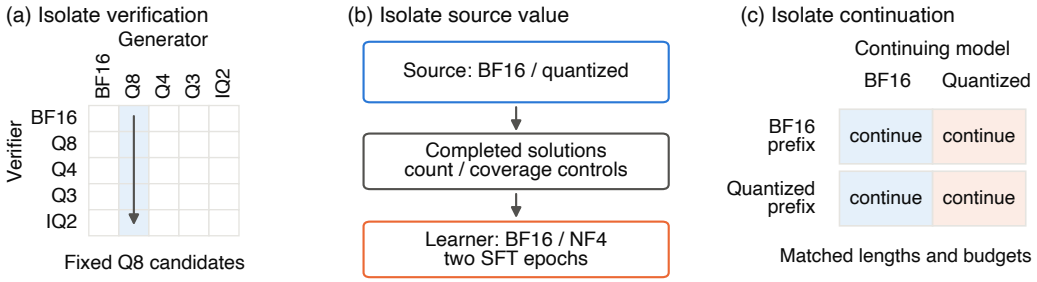

Figure 2: **Three interventions separate ranking, learning and continuation.** **(a)** Every verifier judges every generator's candidates; a fixed generator column isolates verifier quantization. **(b)** Completed samples from each source train both learners on identical rows; source comparisons control count and coverage, and separate experiments vary correctness or selection. **(c)** Each of two matched prefixes is continued by both models, with equal remaining token budgets. The source ladder uses random completed-sample selection; the other selectors are tested in the BF16/NF4 experiments.

Isolating verification. Solutions are sampled once per generator and judged by every verifier (Fig. 2a). The registered comparison holds the `Q8_0` generator's candidates fixed while varying the verifier. Since truncated candidates are usually wrong and easy to reject, we report AUROC both on all candidates and on *completed correct and completed incorrect candidates*. The latter removes truncation detection; its negative class comprises the *real errors*. Truncations account for 79% of the 12B's wrong Q8 candidates on MATH-500, 98% on the 35B and 62% on the 12B hard tier. We separately measure discrimination between completed attempts at the *same problem*. A median Yes/No probability mass below 0.5 disqualifies a comparison from equivalence claims under the declared interface rule.

Statistical interpretation. Static intervals use paired cluster bootstraps over problems (Efron & Tibshirani, 1993). The registered H1 family comprises nine Level-5 comparisons on fixed Q8 candidates: Q4, Q3, NF4 and AWQ on each primary subject, plus QAT-Q4 on the 12B. A contrast is *equivalent* if its 90% interval lies inside the margin (± 0.025 AUROC for this family; two one-sided tests, 10,000 draws; Schuirmann, 1987; Lakens, 2017), *different* if its Holm-adjusted p is below 0.05 within the declared family (Holm, 1979), and *inconclusive* if neither; a difference is *established* when the second flag holds. Two-bit comparisons are descriptive and outside this family. Training experiments use crossed seed $\times$ problem intervals with block-specific margins and Holm families. The two flags are separate decisions; an inconclusive contrast is never evidence of no effect. Protocols, amendments and seed/calibration diagnostics are in Appendix A.

3 Verification ranking survives where generation degrades

The registered comparisons preserve ranking. All nine registered verifier contrasts are equivalent to BF16 within ± 0.025 AUROC; every 90% interval lies within approximately ± 0.012 (Table 1). This is a comparison on identical candidates, so declining generator accuracy cannot itself explain the preserved ranking. Generation nevertheless loses 4.1 and 13.3 accuracy points from Q8 to Q4 and Q3 on the 12B hard tier. Completed-only and within-problem measurements test the strength of the preserved signal.

Table 1: **Verification changes on fixed Q8-generated candidates.** BF16 baseline AUROCs are shown above the differences (90% problem-bootstrap intervals). The registered test uses all Level-5 candidates; completed-only comparisons remove truncation detection. † equivalence within ± 0.025 not established; ‡ 2-bit, outside the registered family (Level-5 point estimate). Parentheses give completed-but-wrong counts.

Gemma 4 12B			
Verifier	Registered Level 5	Completed MATH (69)	Completed hard (250)
BF16 baseline	0.969	0.914	0.850
Q4_K_M	-0.005 $[-0.011, +0.001]$	-0.007 $[-0.020, +0.004]$	-0.015 $[-0.025, -0.006]$
Q3_K_M	-0.005 $[-0.012, +0.001]$	-0.015 $[-0.039, +0.012]^{\dagger}$	-0.016 $[-0.029, -0.004]^{\dagger}$
NF4	-0.002 $[-0.009, +0.006]$	$+0.005$ $[-0.014, +0.027]^{\dagger}$	–
AWQ-4	$+0.004$ $[-0.001, +0.009]$	-0.008 $[-0.023, +0.004]$	–
QAT-Q4	-0.001 $[-0.007, +0.005]$	-0.010 $[-0.021, +0.000]$	–
IQ2_M	$-0.137^{\ddagger}$	-0.494 $[-0.591, -0.399]^{\dagger}$	-0.155 $[-0.219, -0.094]^{\dagger}$

Qwen3.6-35B-A3B		
Verifier	Registered Level 5	Completed MATH, cap 4,096 (44)
BF16 baseline	0.967	0.879
Q4_K_M	-0.003 $[-0.006, +0.000]$	-0.007 $[-0.021, +0.008]$
Q3_K_M	$+0.002$ $[-0.001, +0.006]$	$+0.006$ $[-0.018, +0.040]^{\dagger}$
NF4	$+0.000$ $[-0.003, +0.003]$	–
AWQ-4	-0.002 $[-0.006, +0.002]$	–
IQ2_M	$+0.001^{\ddagger}$	-0.036 $[-0.060, -0.007]^{\dagger}$

What survives after removing easy signals. On the hard tier's 250 completed errors, the 12B Q4 verifier changes by -0.015 $[-0.025,\ -0.006]$ AUROC: a small loss, with its 90% interval still inside the equivalence margin. Q3 changes by -0.016 $[-0.029,\ -0.004]$ and does not establish equivalence (Fig. 3b). Pooled HumanEval+/MBPP+ comparisons, where $\geq$3-bit outputs rarely truncate, establish equivalence at 4-bit for both subjects and at 3-bit for the 12B; the 35B has a 3-bit loss of -0.017 $[-0.031,\ -0.003]$ (Appendix G). Thus the broad registered result has measurable limits on harder completed candidates.

Aggregate AUROC also reflects differences in problem difficulty. With 64 attempts per hard-tier problem, the BF16 verifier's within-problem AUROC is only 0.573 $[0.516,\ 0.629]$, versus 0.935 in aggregate (92 problems with both classes; 2,109 completed errors). Its changes at Q4 and Q3 are $+0.000$ $[-0.011,\ +0.011]$ and $+0.008$ $[-0.013,\ +0.029]$, equivalent within the separately prespecified ± 0.05 margin (Fig. 3c; rationale in Appendix K). This preserves an already weak signal within that tolerance; it does not establish a strong selector or preservation of every increment above chance. A pairwise read also preserves ranking through 3-bit; a linear hidden-state probe adds no detectable within-problem advantage over the verdict ($+0.001$ $[-0.032,\ +0.032]$, equivalent at ± 0.05; Appendix K).

Generation loss includes more than insufficient time. The 12B's MATH-500 loss is smaller than its hard-tier loss and grows with solution length within difficulty levels; it persists under greedy decoding. At Q3, the MATH-500 change separates into $+0.050$ truncation and -0.003 completed-candidate correctness (Appendix E). Doubling the hard-tier token budget to 8,192 closes the Q4 pass@8 deficit ($0.634 \rightarrow 0.732$, versus BF16's 0.719). Q3 remains at 0.556, losing 28 BF16-solvable problems while gaining 3; 46% of its truncations contain token loops (Appendix F). The Q4 budget effect and Q3 persistent failure should therefore be distinguished.

Boundaries of preservation. At 2-bit the dense 12B collapses on both sides ($G = 0.439$ with 58% truncation on MATH-500; completed-error AUROC 0.694 versus 0.850 on the hard tier), while the MoEs retain more: the 35B steps -0.050 in generation with its Level-5 column at -0.009 at the 4,096 cap, and the 26B-A4B keeps every column within 0.005 of BF16 (G $0.930 \rightarrow 0.829$). The dense 31B also retains generation ($0.962 \rightarrow 0.920$), so these comparisons do not isolate an architectural cause (Appendix H). Smaller models can lose ranking already at 4-bit (SmolLM3 -0.024 $[-0.037,\ -0.012]$); the 3–4 B panel's 3-bit Level-5 changes range from -0.020 to -0.056.

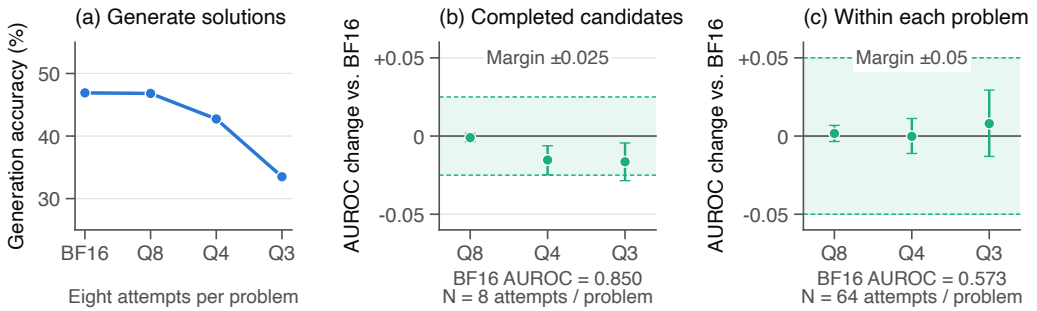

Figure 3: **Generation loss and verification changes on the 12B hard tier. (a)** Generation accuracy falls from Q8 to Q3 at the 4,096-token cap. **(b)** Completed-candidate AUROC changes relative to BF16 on the fixed Q8 set ($N = 8$; BF16 $V = 0.850$). **(c)** Within-problem AUROC changes on the separately sampled $N = 64$ Q8 set (BF16 $V = 0.573$). Bars in (b,c) are paired 90% problem-bootstrap intervals; shading marks the respective ± 0.025 and ± 0.05 equivalence margins. The margins were pre-specified separately; the wider within-problem tolerance reflects the design's expected precision. Q3 meets only the wider tolerance in (c). Full ladders and 2-bit boundaries are in Appendices D–H.

Ranking robustness also coexists with score drift: Q3 answers Yes on 60% of hard-tier completed errors, versus 27% at BF16. Thresholds must therefore be recalibrated or compared at matched acceptance rates (Fig. 4). The dissociation is not specific to weight quantization: with the 12B's weights at full precision, a 4-bit K/V cache leaves every verifier column equivalent to the f16-cache control (-0.002 $[-0.004, +0.000]$ on all levels) while generation truncates more (Δtruncation $+0.013$ $[+0.006, +0.020]$) at unchanged completed-candidate correctness (Appendix L).

4 Source controls the value of self-generated supervision

Crossing the source and the learner. We vary who generates the training solutions while holding the learner and training recipe fixed, then repeat on the other learner (Fig. 2b). The recipe is two epochs of low-rank adapter SFT: LoRA on BF16 or QLoRA on NF4, with loss on the retained solution tokens and prompt tokens masked; completion is measured at generation, and serialization applies a fixed 2,048-token sequence cap that can clip a few retained rows (Appendix O). One *training round* contains multiple optimizer steps. The fixed-pool experiments start afresh from the base checkpoint and use completed samples from stored pools of 449 MATH-train problems on the 12B and 1,000 on the 3B, with eight generated attempts per problem. Per-problem training caps are two and four, respectively; both learners receive identical rows within each source/selection condition. Separate adaptive experiments repeat generation, selection and training for four rounds.

The primary learning endpoint is a *partial-solvability holdout* ("pool-like" in the appendix): 300 disjoint MATH-train problems that the subject's BF16 pre-pass solved in one to three of four attempts. This targets problems with room for improvement; MATH-500 is the secondary endpoint. The source comparisons use five paired seeds on the 12B and twelve on the 3B.

Source reverses the ordering by correctness. To vary correctness without changing the intended count or problem coverage, a label-only selector ranges from prioritizing incorrect completed samples ($\lambda = 0$, anti-oracle) to prioritizing correct ones ($\lambda = 1$, oracle), filling the remaining slots from the other class. The attainable endpoints need not be 0% and 100% correct. On the 3B, even the least-correct BF16-source set (38%) teaches more than the most-correct NF4-source set (74%; Fig. 1a). The paired gain advantages are $+1.9$ $[+1.3, +2.6]$ points for the NF4 learner and $+3.6$ $[+3.0, +4.3]$ for the BF16 learner (seed-t 90% intervals). This post-hoc comparison of pre-registered cells shows that final-answer correctness does not order learning value across these sources, without fitting a correctness-response line. On this subject the NF4 verifier column is itself

eroded (-0.028 $[-0.046, -0.010]$ at Level 5; Appendix O.7), so the 3B tests correctness against source; the preserved-verifier premise is the 12B's.

The reversal is endpoint-dependent: on MATH-500 it is $+0.3$ $[-0.1, +0.7]$ points for the NF4 learner and $+0.7$ $[+0.4, +1.0]$ for BF16, with no established difference for NF4. The larger holdout effect is therefore a finding on partially solvable problems, not a general benchmark gain.

The 12B source effect survives matched coverage and transfers harm. The initial 12B source comparison also differed in sample count and problem coverage. We address this by drawing completed BF16-source solutions at the NF4 reference's per-problem counts (the matched-random set, mU). The resulting set has 477 rows on 272 problems, versus 483 on 275 for NF4, approximately a 1% difference. Correctness remains 0.886 versus 0.830 (Appendix O.2).

At these nearly matched counts, the NF4 learner gains $+21.4$ points on BF16-source solutions versus $+2.5$ on NF4-source solutions. The source advantage is $+18.9$ $[+16.5, +21.2]$ points, positive in 5/5 seeds; on MATH-500 it is $+4.6$ $[+3.5, +5.6]$. The BF16 learner shows the converse harm: its gain is $+0.6$ points on the matched BF16 source but -21.7 on the NF4 source (Fig. 1b). Its source advantage is $+22.3$ $[+19.6, +25.0]$ points on the primary holdout and $+4.9$ $[+3.8, +6.1]$ on MATH-500. Thus quantized-source solutions can impair a learner whose own base remains full precision, even after unfinished training samples have been removed. On the 12B this effect is mainly a change in termination; on the 3B it appears as completed-answer errors (below).

A second matched set prioritizes the NF4 oracle set's correct/incorrect counts per problem (mO); 5% of rows require filling from the other class. It gives a similar NF4-learner source advantage, $+19.3$ $[+16.7, +21.8]$ points. Neither comparison fixes correctness exactly. Applying the BF16 pool's measured correctness slope to mU's correctness gap predicts approximately $+0.9$ points of additional gain, versus the much larger measured source advantage. This linear sensitivity is not a bound or an identified direct effect; it supports reading the matched comparison alongside the independent 3B reversal.

The penalty extends across source precision. Replacing the generator with GGUF BF16, Q8, Q4 or Q3 holds the learner and recipe fixed while checking whether the NF4 source result extends to another format. On the 12B, the GGUF BF16 generator's learning outcomes are equivalent to the Hugging Face BF16 source within the declared margin (Appendix O.3). Table 2 displays the source penalties alongside the actual training-set compositions. Counts are matched per problem where completed samples exist; the 12B Q3 set is 7% short in rows and 6% in problems.

Table 2: **Source penalties and the controls behind them.** Penalty = gain on GGUF BF16-source solutions minus gain on each row's solutions, in accuracy percentage points on the partial-solvability holdout (crossed 90% intervals). Counts and correctness are seed means for the selected training sets, shared by both learners. BF16 is the reference; correctness is not matched. Five seeds on the 12B, twelve on the 3B; the NF4 row repeats the matched contrast of the preceding paragraphs on the GGUF path. † difference not established after Holm adjustment.

				Source penalty (points)	
Generator	Correct (%)	Rows	Problems	NF4 learner	BF16 learner
Gemma 4 12B					
BF16	87.8	476	273	reference	reference
Q8_0	87.7	476	272	$+0.2$ $[-1.8, +2.2]^\dagger$	-1.5 $[-3.4, +0.4]^\dagger$
Q4_K_M	86.0	474	271	$+2.7$ $[+0.6, +4.8]^\dagger$	$+4.4$ $[+2.5, +6.3]$
Q3_K_M	84.4	443	255	$+10.3$ $[+7.9, +12.6]$	$+11.8$ $[+9.5, +13.9]$
NF4	83.0	483	275	$+18.9$ $[+16.1, +21.7]$	$+22.9$ $[+20.3, +25.4]$
SmolLM3-3B					
BF16	58.7	3,994	1,000	reference	reference
Q4_K_M	55.4	3,992	1,000	$+1.4$ $[+0.3, +2.4]$	$+2.0$ $[+1.0, +3.0]$
Q3_K_M	48.6	3,993	1,000	$+4.2$ $[+3.1, +5.3]$	$+5.9$ $[+4.7, +7.1]$
NF4	52.3	3,996	1,000	$+1.8$ $[+0.6, +3.1]$	$+4.6$ $[+3.4, +5.9]$

The point penalties increase from Q8 through Q4 to Q3 on the 12B, and from Q4 to Q3 on the 3B. Q8 has no established penalty. The 12B Q4 NF4-learner contrast is equivalent within ± 5 points;

its difference from zero is not established after Holm ($p = 0.084$). Both Q3 learner contrasts on the 12B and the Q4/Q3 contrasts on the 3B establish differences. Nominal bit width does not order the formats: the 12B's Q3 solutions outperform its NF4 solutions by $+8.6$ $[+6.5, +10.7]$ points on the NF4 learner, whereas NF4 lies between Q4 and Q3 on the 3B. Across the tested cells, the supervision penalty tracks generator damage descriptively; this is not a transfer law (Appendix O.3).

What correctness-based selection contributes. Within a fixed source, more correct targets can help, but the measured effect is smaller and depends on the endpoint. On the 3B NF4 pool, the label-only selector spans correctness 0.308–0.739. The NF4 learner's gain slope is $+2.4$ $[+0.2, +4.4]$ points per unit correctness (crossed 90%); its 95% interval $[-0.2, +4.8]$ includes zero. The MATH-500 slope is $+1.8$ $[+0.8, +2.8]$. The narrower 12B range establishes no slope on its own pool. Consensus and verdict ranking add little beyond correctness-matched random selections on the quantized 3B learner: differences are $+0.1$ $[-0.8, +1.0]$ and -0.1 $[-0.9, +0.8]$ points, both equivalent within ± 1 point (Appendix O).

Changing the verifier does not produce a detected recovery. In the four-round loops, replacing the NF4 verifier with BF16 changes the quantized learner's gain by $+0.7$ $[-0.8, +2.1]$ points on the 12B and $+0.4$ $[-1.0, +1.8]$ on the 3B. Neither establishes a difference or a zero effect (12B minimum detectable effect ± 2.5 points). On the 12B, random completed-sample selection gains $+7.2$ $[+5.3, +9.1]$ points, compared with $+2.1$ $[+0.8, +3.5]$ for the deployed self-judge gate ($s \geq 0.5$). About a fifth of the NF4 gate's accepted samples are truncated, with a gain difference of -5.0 $[-6.6, -3.5]$ points (gate minus random, Holm-adjusted difference). Prioritizing correct completed samples at the same counts then raises correctness from 0.83 to 0.87 but changes gain by only -0.1 $[-1.4, +1.3]$ points, equivalent within ± 2.5 points. The 3B oracle increment, $+2.2$ $[+0.7, +3.6]$ points, is inconclusive after Holm (Table 44; Appendix O.8). These four-round effects and the one-round source contrasts are distinct estimands; retention of half the BF16 verifier's ranking increment over random selection remains inconclusive.

The outcome is clear; the textual mediator is not. The 12B's termination change is large: the NF4 learner's truncation changes by -0.294 on GGUF BF16-source solutions versus -0.037 on its own completed solutions, while a BF16 learner trained on the NF4 source increases truncation by $+0.247$. On the 3B, truncation changes by at most 0.012, and the source penalty appears as completed-answer errors. These outcomes do not identify which property of the text causes the learning difference.

A per-token likelihood fingerprint almost perfectly separates the two sources. Within NF4, selecting the least versus most quantization-typical completed solutions at matched correctness, count and coverage yields no Holm-established recovery. The selection spans only about 23% of the between-source fingerprint gap on the 3B and an eighth on the 12B (Appendix O.2). Its limited reach leaves other filters open.

5 Localizing excess generation failure

The source comparisons leave the textual mediator unidentified. We separately investigate generation failure: does excess truncation come from the prefix already written or from the model continuing it? For the 12B, we cross prefix source with continuation model (Fig. 2c). A *departure* is the first sampled token outside BF16's top-p nucleus at the quantized model's current prefix; it is a distributional event, not an annotated reasoning error. Both models continue the departed prefix and a BF16 prefix from the same problem and sample index, cut to the same length, under equal remaining token budgets (Table 3).

Table 3: **Continuation model dominates the tested prefix-source effect.** 995 2-bit and 1,628 3-bit documents; first departure at $\geq$ 16 tokens, two suffixes per arm. Document-mean truncation and correctness rates can overlap. Truncation contrasts hold the departed prefix fixed (continuation effect) or BF16 continuing model fixed (prefix effect), with paired problem-bootstrap 95% intervals (Appendix M).

		BF16 continues		Quantized model continues	
Precision	Prefix source	Truncated	Correct	Truncated	Correct
IQ2_M	BF16	0.066	0.924	0.563	0.473
	Quantized	0.063	0.918	0.557	0.472
Q3_K_M	BF16	0.068	0.913	0.123	0.861
	Quantized	0.076	0.909	0.126	0.857
		Continuation effect		Prefix effect	
IQ2_M		$+0.493$ [$+0.453$, $+0.537$]		-0.003 [-0.013, $+0.006$]	
Q3_K_M		$+0.050$ [$+0.037$, $+0.066$]		$+0.008$ [$+0.000$, $+0.017$]	

BF16 continuing a departed 2-bit prefix truncates at approximately its rate on its own matched prefixes; the prefix-effect interval lies inside the declared ± 0.05 margin. The 2-bit continuing model adds about 49 percentage points of truncation on the departed prefixes, and similarly fails more often from BF16 prefixes. At 3-bit the continuation effect is smaller, about five points, and the prefix effect remains within the same margin. Both decompositions of the 2×2 comparison agree. The result is local to the eligible first-departure prefixes, including only the later-departing half at 2-bit; it does not test recovery from established loops or show that every quantized prefix is harmless.

A limited perturbation account. Quantization changes the conditional distribution at each forward pass. A verifier produces one scored output after reading a fixed candidate; generation repeatedly samples outputs that become the inputs to subsequent steps. The difference is autoregressive feedback; quantization need not add independent fresh noise at each step. An empirical shifted/scaled score-plus-noise model approximates the 12B's completed-candidate AUROC losses: predicted -0.009 / -0.016 / -0.145 versus measured -0.015 / -0.016 / -0.155 at 4 / 3 / 2-bit (Fig. 6). Teacher-forced comparisons and seven pre-specified tests examine this account across formats, tensor locations and a quantized K/V cache (Table 19; Evidence Fig. E4 (companion); Appendix L). For example, quantizing attention to the 2-bit assignment on a Q4 anchor produces less verdict perturbation than its generation-distribution divergence predicts on the 12B; Qwen3.5-4B shows no such advantage.

On three held-out models, predictions committed before quantized verdict scoring have mean absolute errors of 0.0069 / 0.0072 AUROC at 4 / 3-bit, outperforming a no-loss baseline but not an input-matched baseline using the fitted models' median perturbation per rung. All three miss the registered band on the fitted constant (Appendix N). The evidence supports a local account of verification and generation sensitivity, not a transferable law or an identified mechanism of the learning penalty.

6 Related work

Verification and learning from synthetic data. The generation–verification framework separates improvement by reweighting from the fidelity of the subsequent model update (Song et al., 2025). Related analyses establish collapse resistance under assumptions on verification, data and learning (Fu et al., 2025; Amin et al., 2025; Yi et al., 2025; Sun et al., 2025); unreliable gates and limited verification have documented consequences (Song et al., 2026; Jin et al., 2026; Lin et al., 2026; Pan et al., 2026; Feng et al., 2025; Gureja et al., 2025). Our results do not contradict these conditional guarantees. Alemohammad et al. (2026) show source–student compatibility effects in unconditional self-training. Here the source intervention changes numerical representation within one checkpoint, testing verification and final-answer correctness alongside learning.

Distillation and low-precision learning. LLM-QAT already uses full-precision generations to train a quantized model (Liu et al., 2023b). We isolate the source change by crossing versions of

the same checkpoint with both learners, controlling count and coverage, and varying correctness. The reverse direction—a BF16 learner harmed by quantized-source solutions—and the correctness reversal expose more than a stronger teacher's benefit. On-policy distillation trains on student sequences with teacher feedback (Agarwal et al., 2024; Li et al., 2025a); that objective differs from imitating the selected solutions as hard targets here. Quantized training with ground-truth rewards (Huang et al., 2025; Kumar et al., 2025; Gu et al., 2026) or without a paired full-precision control (Thawakar et al., 2025) addresses different comparisons.

Quantization and verification. Reasoning degradation, longer outputs and failure to terminate under quantization are established phenomena (Liu et al., 2025; Li et al., 2025b; Oladri et al., 2026; Zhou et al., 2026a;b; Lotfi et al., 2026; Alimaskina et al., 2026; Lian et al., 2026; Ettifouri et al., 2026a;b). Confidence measurements under compression (Cacioli, 2026; Patarlapalli & Avvaru, 2026; Vipulanandan et al., 2026; Tong et al., 2026) and hidden-state probes (Aiersilan, 2026) motivate our fixed-candidate equivalence design and within-problem controls. Related work also distinguishes aggregate from within-question signals (Bahuguna, 2026); estimator choice matters, with mean-token probability showing a different quantization sensitivity (Ferdous, 2026).

7 Discussion and limitations

What the results change. A compressed model's ability to rank answers and the correctness of its selected answers are useful measurements, but they do not determine what its full solutions teach under SFT. The source-by-learner crossing shows both recovery and transferred harm, and the source ladder extends the effect beyond one NF4 implementation. For the tested recipe, generating with a less damaging representation and filtering for completion are supported interventions; a budget-matched optimum is not established. Correctness should remain an evaluation target, accompanied by the learning effect of the selected text. Static ranking comparisons should hold candidates fixed, inspect within-problem discrimination and distinguish ranking from threshold calibration.

Scope of the evidence. Verification here is a one-token Yes/No score with weak within-problem discrimination. NF4 loses its Yes/No interface on long candidates (median mass 0.06 versus 0.91 at BF16), and the 35B's math completed-error estimates have few negatives; code offers stronger evidence. The primary sub-4-bit ladder is one quantization family, four of eight fitted models are Gemma 4, and three held-out models do not establish population-wide generality. The learning results cover two subjects, one math adapter-SFT recipe, one-round fixed-pool experiments and four-round adaptive loops. They establish source-dependent learning value under these conditions, not the failure of all quantized self-training.

Count and coverage are approximately matched; correctness is not exact, and the 12B Q3 lists are short. The sensitivity calculations do not identify a correctness-fixed direct effect, and the harmful textual mediator remains unknown. The training bootstrap is anti-conservative under the reported Gaussian seed-only null: nominal 0.05 tests reject at 0.158 with five seeds and 0.09 with twelve. These diagnostics are not the actual false-positive rates of the source contrasts; they motivate reading the large paired effects and seed consistency separately from borderline selector and slope results (Appendix A).

The 12B pools near-duplicate 12 MATH-500 problems. Removing them leaves the majority-gate comparison similar ($+0.012$ $[+0.004, +0.021]$), but that check does not cover every source contrast (Appendix P). The larger effects on the selected partial-solvability holdout should not be interpreted as general benchmark improvements.

Under the tested recipe, source precision changes what completed solutions teach, and final-answer correctness alone does not order their learning value.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets; propose and refine hypotheses; develop conceptual framing; design and provide feedback on research methodology and experiments; implement methods; clean and reformat datasets; support qualitative and thematic data analysis; and interpret results. We did not use generative AI tools to develop formal theoretical models, formulate mathematical claims, or provide critical ingredients for proving mathematical claims. Proof writing and translation were not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures, create or edit software code, create artifacts, draft and edit parts of the paper, improve readability, and identify relevant literature.

We reviewed all AI-assisted work, including infrastructure code, analysis scripts and statistical methods, for results integrity. The numerical results supporting the paper's claims were independently checked against the frozen records, and the reported numbers, tables and figures were reproduced from those records. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The study uses public benchmarks and open-weight models, involves no human subjects, and its findings concern deployment guidance for quantized models.

REPRODUCIBILITY STATEMENT

Scripts generate the reported numbers, tables and figures from frozen metrics files; an independent checker verifies those files and the claim invariants. The pre-registration chain (twelve tags), the deviations ledger, the six committed prediction files of the held-out block with their external registration, the prompts, the per-arm manifests with artifact digests, and the recipes accompany the code and readable evidence as anonymized supplementary material (Appendices A–Q). Raw sample/verdict records, loop artifacts and teacher-forcing documents are archived separately in a versioned dataset for release after review.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.13649.

Aizierjiang Aiersilan. Hallucination is linearly decodable from mid-layer hidden states in quantized LLMs, 2026. arXiv:2606.02628.

Sina Alemohammad, Li Chen, Richard G. Baraniuk, and Zhangyang Wang. Not all synthetic data is yours to learn from, 2026. arXiv:2605.31126.

Ekaterina Alimaskina, Darya Rudas, Denis Shveykin, Gleb Molodtsov, Pavel Vasiliev, and Aleksandr Beznosikov. Extreme Low-Bit Inference in Reasoning Models: Failure Modes and Targeted Recovery, 2026. arXiv:2606.02011.

Kareem Amin, Sara Babakniya, Alex Bie, Weiwei Kong, Umar Syed, and Sergei Vassilvitskii. Escaping collapse: The strength of weak data for large language model training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.08924.

Utkarsh Bahuguna. When Self-Consistency Backfires: Majority Vote Hurts the Majority of Hard Science Problems for Small LLMs, 2026. arXiv:2608.11403; v1 accepted at the COLM 2026 Workshop on Efficient Reasoning; v2 additions not peer reviewed.

Jon-Paul Cacioli. Quantisation reshapes the metacognitive geometry of language models, 2026. arXiv:2604.08976.

Wenhua Cheng, Weiwei Zhang, Haihao Shen, Yiyang Cai, Xin He, Kaokao Lv, and Yi Liu. Optimize weight rounding via signed gradient descent for the quantization of LLMs, 2023. arXiv:2309.05516.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.14314.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

El Hassane Ettifouri, Ayoub Belfatmi, Mahaman Sanoussi Yahaya Alassan, and Walid Dahhane. Calibrated e-CUSUM Decoding for Quantized Reasoning Models: Why Token Log-Probability Is the Wrong Observable for Decoding Monitors, 2026a. arXiv:2607.11317.

El Hassane Ettifouri, Ayoub Belfatmi, Mahaman Sanoussi Yahaya Alassan, and Walid Dahhane. CUSUM-Shaped Inference-Time Monitoring and Targeted Re-Decoding for Quantized Small Language Model Reasoning, 2026b. arXiv:2607.20129.

Yunzhen Feng, Elvis Dohmatob, Pu Yang, François Charton, and Julia Kempe. Beyond model collapse: Scaling up with synthesized data requires verification. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.07515.

M M Asif Ferdous. Bigger or Cheaper? Scale and Quantization Effects on Uncertainty Signals in Vision-Language Models Under Image Degradation, 2026. arXiv:2607.24440.

Shi Fu, Yingjie Wang, Yuzhu Chen, Li Shen, and Dacheng Tao. Self-verification provably prevents model collapse in recursive synthetic training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Georgi Gerganov and ggml contributors. llama.cpp: LLM inference in C/C++. `https://github.com/ggml-org/llama.cpp`, 2026. release b10330, commit 687e7789.

Google. Gemma 4 with quantization-aware training. Google Developers Blog, `https://blog.google/innovation-and-ai/technology/developers-tools/quantization-aware-training-gemma-4/`, June 2026.

Hao Gu, Hao Wang, Jiacheng Liu, Lujun Li, Qiyuan Zhu, Bei Liu, Binxing Xu, Lei Wang, Xintong Yang, Sida Lin, Sirui Han, and Yike Guo. QaRL: Rollout-Aligned Quantization-Aware RL for Fast and Stable Training under Training–Inference Mismatch, 2026. arXiv:2604.07853.

Srishti Gureja, Elena Tommasone, Jingyi He, Sara Hooker, Matthias Gallé, and Marzieh Fadaee. Verification limits code LLM training, 2025. arXiv:2509.20837.

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 *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021. arXiv:2103.03874.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Wei Huang et al. QeRL: Beyond efficiency — quantization-enhanced reinforcement learning for LLMs, 2025. arXiv:2510.11696.

Hugging Face. Math-Verify: Mathematical expression evaluation and equivalence checking. `https://github.com/huggingface/Math-Verify`, 2025. version 0.9.0.

Ruijie Jin, Bingjie Zhao, Mingyu Kang, Qiuyu Zhang, and Xiang Li. Verification mirage: Mapping the reliability boundary of self-verification in medical VQA, 2026. arXiv:2605.10850.

Medha Kumar, Zeyuan Xu, Xue Wang, and Taylor Webb. The impact of quantization on large reasoning model reinforcement learning, 2025. arXiv:2511.15694.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners, 2025a. arXiv:2502.12143.

Zhen Li et al. Quantization meets reasoning: Exploring and mitigating degradation of low-bit LLMs in mathematical reasoning, 2025b. arXiv:2505.11574.

Xinyu Lian, Walid Krichene, Beichen Huang, Masahiro Tanaka, Olatunji Ruwase, Li Zhang, and Minjia Zhang. Quantization Inflates Reasoning: Token Inflation as a Hidden Cost of Low-Bit Reasoning Models, 2026. arXiv:2606.25519.

Yesheng Liang, Haisheng Chen, Zihan Zhang, Song Han, and Zhijian Liu. ParoQuant: Pairwise rotation quantization for efficient reasoning LLM inference. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2511.10645.

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 (ICLR)*, 2024. arXiv:2305.20050; the MATH-500 split.

Haoyang Lin, Zhipeng Kuai, Enze Xue, and Lei Wang. Detecting and mitigating the correct-answer extinction window in test-time reinforcement learning with majority voting, 2026. arXiv:2605.19444.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, 2024. arXiv:2306.00978.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? Rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a. arXiv:2305.01210.

Ruikang Liu et al. Quantization hurts reasoning? An empirical study on quantized reasoning models. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2504.04823.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models, 2023b. arXiv:2305.17888.

Sanae Lotfi, Polina Kirichenko, Steven Li, and Zechun Liu. Quantized Reasoning Models Think They Need to Think Longer, but They Do Not, 2026. arXiv:2606.00206.

Rohit Oladri, M. V. V. Priya, and Jiahao Wu. Silent failures in quantized LLM reasoning: A taxonomy-based analysis of hollow convergence and failure mode shifts, 2026. arXiv:2607.09999.

Tianyi Pan et al. CoVerRL: Breaking the consensus trap in label-free reasoning via generator-verifier co-evolution, 2026. arXiv:2603.17775.

Sri Bharath Patarlapalli and Sai Teja Avvaru. BitCal-TTS: Bit-calibrated test-time scaling for quantized reasoning models, 2026. arXiv:2605.05561.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Avi Singh et al. Beyond human data: Scaling self-training for problem-solving with language models, 2024. arXiv:2312.06585; Transactions on Machine Learning Research.

Xin Song, Ziyu Cai, and Le Zhao. When AI reviews its own code: Recursive self-training collapse in code LLMs, 2026. arXiv:2606.28438.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2412.02674.

Yifan Sun, Yushan Liang, Zhen Zhang, Xinyu Liu, and Jiaye Teng. Theoretical modeling of large language model self-improvement training dynamics through solver-verifier gap, 2025. arXiv:2507.00075.

Omkar Thawakar, Shravan Venkatraman, Ritesh Thawkar, Abdelrahman Shaker, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. EvoLMM: Self-Evolving Large Multimodal Models with Continuous Rewards, 2025. arXiv:2511.16672.

Yujia Tong, Yuxi Wang, Yunyang Wan, Tian Zhang, Junhao Dong, and Jingling Yuan. Does Compression Preserve Uncertainty? A Unified Benchmark for Quantized and Sparse LLMs via Conformal Prediction, 2026. arXiv:2606.01850.

Pragatheeswaran Vipulanandan, Kamal Premaratne, and Dilip Sarkar. Semantic Uncertainty Quantification of Hallucinations in LLMs: A Quantum Tensor Network Based Method, 2026. arXiv:2601.20026.

Baizhi Yi, Qiaoyu Liu, Yuan Cheng, and Han Xu. Escaping model collapse via synthetic data verification: Near-term improvements and long-term convergence, 2025. arXiv:2510.16657.

Weizhe Yuan et al. Self-rewarding language models. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2401.10020.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2203.14465.

Chenxi Zhou, Pengfei Cao, Jiang Li, Bohan Yu, Jinyu Ye, Jun Zhao, and Kang Liu. From Signal Degradation to Computation Collapse: Uncovering the Two Failure Modes of LLM Quantization, 2026a. arXiv:2604.19884; accepted to Findings of ACL 2026.

Chenxi Zhou et al. Quantization degradation in large language models: A signal-noise perspective, 2026b. arXiv:2608.08188.

GUIDE TO THE SUPPLEMENTARY EVIDENCE

The appendix contains the methods, primary comparisons and limits needed to assess the claims. The companion `evidence.pdf` retains the detailed diagnostics (tables and figures numbered E1, E2, ... within each kind); `evidence-index.md` maps each display to its generated table or figure and frozen experiment. These files are included with the anonymized supplementary material. For **verification robustness**, the full factorials and generation results are in Appendices D–E, the hard and code tasks in F–G, and the attempt-level decomposition in K. For **generation failure**, start with the reciprocal continuation intervention in Appendix M; the perturbation tests and held-out prediction are in L and N. For **learning value**, Appendix O groups the loop controls, fixed-pool factorials and source ladder, with a guide at its start. The protocol and inference rules are in Appendix A, model and prompt details in B–C, and contamination and reproduction in P–Q.

A PRE-REGISTRATION, AMENDMENTS AND DEVIATIONS

Hypotheses. The registered hypotheses (tag `v1`, 2026-08-09; `v1.2`, `v1.3` before the first counted sample) are: **H1** equivalence of the nine $\geq$3-bit verifier columns to the BF16 column on the Level-5 stratum of the `Q8_0` set (TOST ± 0.025, 90% CI, Holm sensitivity; all-levels columns and code columns secondary); **H2** the 35B's deployed operating points $(\beta, \alpha, \text{recall})$ at $\alpha^\star = 0.25$, estimation only; **H3** generation declines from `Q8_0` to `Q3_K_M` and collapses at `IQ2_M` (paired McNemar);

H4 mean-logprob AUROC below self-judge and majority agreement on every diagonal and degrading at 2-bit; **H5** the NF4 loop tracks its BF16 control with the static ladder as the $t = 0$ prediction; **H6** NF4/AWQ columns, execution-oracle code columns and a contamination-clean slice reproduce H1's flatness; **H7** QAT-Q4 equivalent to BF16 and its difference to `Q4_K_M`. Robustness recomputations (excluding truncated solutions; length terciles; alternate and label-swapped verdict prompts) and the interface-collapse rule (median Yes/No mass < 0.5 excludes a cell from the family) were fixed in the same document.

Amendment chain. The twelve-tag chain and its complete declarations accompany the code. After the static protocol (`v1`–`v1.3`), `v1.4` froze the training recipe and `v1.5` added a headroom control. The perturbation and trajectory tests were declared under `v1.6`–`v1.7`; the mechanism declarations were committed sequentially before each block's measurement, with a consolidated archival tag. `v1.8` added powered attempt-level verification, verifier isolation, a second training subject, hybrid-model mixes and longer-budget tests. `v1.9` added completed-sample selectors, the crossed inference rule, retention, held-out predictions, cache perturbation and probe/replay checks. `v1.10` declared the fixed-pool correctness/source factorial and reciprocal continuation; `v1.11` declared matched source lists and fingerprint selection; `v1.12` declared the source-precision ladder and its BF16 path control. Each block specified its estimands, predictions and decision rules before measurement. Declared exploratory additions (longer budgets, hard/code tasks, thinking-mode verdicts and the model panel) remain separate from the confirmatory family. Detailed declarations and the deviations ledger are included in the package.

Deviations affecting counted data. (D1) The thinking-mode verdict columns first ran at an effective ≈2,300-token budget because of a per-slot context limit; both subjects were re-run at the declared 4,096 budget and the measured-budget outputs archived. (D2) Qwen3.5-4B's panel row runs at a 4,096 cap instead of 1,536 (≈45% of its samples truncate at 1,536). (D3) Ministral 3 3B runs cap-bound at 1,536 and is labelled so. (D4) Three BF16 control chains of the first loop ran out of memory in round 2 (gradient checkpointing was not applied to a model already carrying an adapter); all ten chains restarted from per-batch checkpoints with identical per-sample seeds. (D5) The 26B-A4B's verdict prefix pins the thinking-off rendering its template does not emit. (D6) SmolLM3's HF-path loop runs eager attention: its SDPA path grew host memory with every new batch shape under the pinned transformers build, so the counted block was restarted from round 0 under one implementation before any readout (the smoke and pre-pass had run SDPA; round-0 metrics agree within ± 0.005). Incidents with no counted-data impact (a miscompiled 2-bit kernel found before the first counted sample, an oracle-watchdog bug whose recompute left every frozen number unchanged, a shared-volume outage, and in the last round a tool guard, a dead line in an analysis script and an analysis step that appended a derived column to four recorded prediction inputs—restored byte-identically and re-verified before the comparisons ran; in the final round a launch script whose guard ended it before any submission) are listed in the released ledger.

Inference for the loop blocks. Gains are round-4 minus round-0 holdout accuracy. The first loops use the problem cluster bootstrap with seeds pooled and, from `v1.8`, the seed-level t-interval beside it (both must resolve for "resolved"). The selection cells (`v1.9`) use a crossed seed × problem bootstrap—per draw one multinomial weight vector over the common seed bases and one over the holdout problems, product weights applied to every arm and every round, 5,000 draws—as the decision interval, with two flags without precedence: *equivalent* if the 90% interval lies inside ± 0.025; *different* if the Holm-adjusted two-sided bootstrap $p < 0.05$ within the block's fixed family (12B: Δ_{oracle}, Δ_{rank}, Δ_{sel}, $\Delta_{\text{ver,rank}}$; 3B: Δ_{oracle}, Δ_{rank}, Δ_{cons}, $\Delta_{\text{cons}-\text{rank}}$ and the retention test); *inconclusive* otherwise. The retention endpoint asks whether the quantized read keeps at least half of the full-precision ranking's increment over random selection—$b = g(\text{K}\times) - g(\text{U})$, $q = g(\text{K}) - g(\text{U})$, $H = q - b/2$—as an intersection–union test with both directions controlled through one p; it returned unresolved on both subjects. Under a Gaussian seed-only null, the bootstrap test's size at nominal 0.05 is 0.158 at five seeds and 0.09 at twelve; nominal 90% interval coverage is 0.789 and 0.85, respectively. These diagnostics show anti-conservative inference in that setting, not the actual false-positive rate of a given source contrast. Paired seed-level t-intervals are reported as additional diagnostics; they condition on the evaluation problems and do not replace the crossed intervals' treatment of problem variation. Secondary contrasts are unadjusted and labelled; a ratio is reported only over a resolved positive denominator. The fixed-pool factorial (`v1.10`) estimates, per

pool × learner, the within-seed least-squares slope of the one-update gain on realized precision over the mixture cells, refit in every crossed draw, with the same two flags at ± 0.01 for the composition contrasts and ± 0.03 per unit for the learner slope difference; the continuation intervention uses a problem-clustered bootstrap over documents (2,000 draws, 95%) with a ± 0.05 equivalence margin on truncation and both readings written before the data. The row-list blocks (`v1.11`, `v1.12`) use the same crossed estimator on one-update gains with the two flags at ± 0.05 (the 3B fingerprint contrasts at ± 0.01) and Holm within each declared family; the pre-written rule for "the source survives matched coverage" was a crossed 90% lower bound above $+0.05$, and the ladder's falsifier was equivalence of both K-quant rungs' penalties to zero on the quantized learner.

External timestamp. The tag chain and the six held-out prediction files are registered on the Open Science Framework, each prediction hash added before the measurement it predicts; the registration is reachable through an anonymized view-only link (OSF registration).

B SUBJECTS, LADDERS, BRIDGES

Table 4 lists the models with pinned revisions; Table 5 the measured bits per weight. All GGUF rungs are built with one llama.cpp build (b10330) from one BF16 checkpoint per model with one importance matrix computed on a committed calibration text; production images use the CUDA 12.8.1 toolchain after nvcc 13.2 was found to miscompile the 2-bit kernels for the B200 (caught by the acceptance replication before any counted sample). NF4 is bitsandbytes 4-bit with BF16 compute and double quantization disabled on the HF path; AWQ is an in-repo implementation of Lin et al. (2024) (activation-aware scales, α grid, group 128, asymmetric round-to-nearest, per-expert for the MoE) because the available exporter implemented a different algorithm (Cheng et al., 2023); QAT-Q4 is the vendor's Q4_0 GGUF and QAT-BF16 its unquantized twin. For the non-GGUF 3-bit block, fake-quantized checkpoints (AWQ 3-bit g128/g32; auto-round 3-bit g128) were exported as BF16 GGUFs and served by the registered engine, so sampler, prompts and read are identical; the export was accepted against a BF16 control at the numerical floor of the HF-vs-llama.cpp read.

Table 4: Subjects (Hugging Face revision prefixes; thinking off by default, except the ablation in Appendix I).

Model	Role	Params	Revision
Gemma 4 12B-it	dense subject; loop subject; QAT triplet	12B	`707f0a3b`; QAT-BF16 `b6ed8627`; QAT-Q4 `29d09777`
Qwen3.6-35B-A3B	MoE subject	35B / 3B active	`995ad96e`
Gemma 4 26B-A4B-it	panel (MoE)	26B / 4B active	`4d7ae498`
Gemma 4 31B-it	panel	31B	`842da379`
Gemma 4 E4B-it	panel	≈4B effective	`ee0ef602`
Qwen3.5-4B	panel (4,096 cap)	4B	`851bf6e8`
Ministral 3 3B-Instruct-2512	panel (cap-bound)	3B	`b35d4dfe`
SmolLM3-3B	panel	3B	`a07cc9a0`
Qwen3-0.6B	interface floor	0.6B	`c1899de2`
OLMo 3 7B Instruct	held-out (4,096 cap)	7B	`6e5971d9`
Granite 4.0-H-Micro	held-out (hybrid Mamba-2/attention)	≈3B	`d5f01a3e`
Phi-4-mini-instruct	held-out	3.8B	`cfbefacb`

Table 5: Bits per weight per rung (body only / overall incl. embeddings), from the build logs.

Model	Q8_0	Q4_K_M	Q3_K_M	IQ2_M
Gemma 4 12B	8.50 / 8.50	4.80 / 4.95	3.85 / 4.08	2.69 / 2.93
Qwen3.6-35B-A3B	8.52 / 8.52	4.87 / 4.89	3.83 / 3.87	2.78 / 2.83
Gemma 4 26B-A4B	8.51 / 8.51	5.28 / 5.32	4.14 / 4.21	3.22 / 3.29
Gemma 4 31B	8.50 / 8.50	4.78 / 4.87	3.86 / 3.98	2.71 / 2.84
Gemma 4 E4B	8.55 / 8.53	4.88 / 5.66	3.92 / 5.14	2.84 / 4.08
Qwen3.5-4B	8.51 / 8.51	4.88 / 5.13	3.87 / 4.27	3.08 / 3.44
Ministral 3 3B	8.50 / 8.50	4.78 / 4.99	3.85 / 4.17	2.74 / 3.06
SmolLM3-3B	8.50 / 8.50	4.81 / 4.96	3.83 / 4.07	2.67 / 2.91

C PROMPTS AND THE VERDICT INTERFACE

Generation prompt (chat template, user turn): *"Solve the following math problem step by step. Put your final answer in \boxed{}."* followed by the problem. Verdict prompt: *"You are grading a candidate solution to a math problem. Problem: {problem} Candidate solution: {solution} Is the candidate solution's final answer correct? Answer with exactly one word: Yes or No."* The verdict is the top-25 raw softmax at the first assistant position; p_{yes} and p_{no} sum the case/whitespace variants of Yes and No; candidate solutions have all special tokens stripped. The alternate phrasing ("student's solution ... Reply with a single word") moves the $\geq$3-bit 12B verifiers by at most 0.004 AUROC and the 2-bit verifier by $+0.063$ $[+0.035, +0.090]$ on the Q8_0 set—prompt sensitivity grows with quantization; the label-swapped prompt ("incorrect?", score read as p_{no}) costs the 12B's top rungs ≤ 0.013. Rebuilding the score as $\log p_{\text{yes}} - \log p_{\text{no}}$ moves no AUROC (max $|\Delta| = 1 \times 10^{-16}$): the ties are exact 0/1 reads with one of the two tokens outside the top-25. Code uses the analogous prompt pair ("would it pass every test of the task's specification?"); the 26B-A4B pins the assistant prefix "Answer:".

D FIXED-CANDIDATE RANKING AND GENERATION

Table 6 reports the fixed-candidate verifier columns. Generation accuracy, truncation and pass@8 are in Table 8; Table 1 gives the completed-error equivalence contrasts. The full diagonal metrics and graphical comparison are in Evidence Table E1 and Evidence Fig. E1.

The NF4 and AWQ columns on the BF16 solution set are the one place a $\geq$3-bit bridge misses equivalence at Level 5 (-0.029 / -0.027): the HF-path verdict read puts < 0.5 Yes/No mass on 16% / 3% of those rows, and on the rows with mass ≥ 0.5 the columns are equivalent ($+0.000$ / -0.001); the AWQ-4 column re-read through the registered engine is equivalent on both tiers (-0.001 $[-0.004, +0.002]$; hard tier $+0.004$ $[-0.003, +0.011]$, completed candidates $+0.012$ $[-0.007, +0.034]$). Seed-base replication of the 12B arms gives between-run SDs ≤ 0.0026 on G and ≤ 0.0049 on V.

Matched-acceptance calibration. Figure 4 compares accepted-set precision at the same acceptance rate, separating ranking robustness from score-threshold drift.

Table 6: Verifier columns on the `Q8_0` solution set (all samples; Level 5; and the BF16 solution set, Level 5), Yes/No mass, and the all-levels TOST contrast vs. BF16 (90% CI; † equivalence not established). Each block names its subject, task and generation cap.

Verifier	V all	V L5	V BF16-set L5	mass	Δ all levels
12B: MATH-500, cap 1,536					
BF16	0.979	0.969	0.986	0.991	–
Q8_0	0.978	0.969	0.987	0.990	–
Q4_K_M	0.978	0.964	0.984	0.990	$-0.001\,[-0.004, +0.001]$
Q3_K_M	0.975	0.964	0.975	0.980	$-0.004\,[-0.009, +0.002]$
IQ2_M	0.842	0.832	0.842	0.974	–
NF4	0.979	0.968	0.957	0.868	$+0.000\,[-0.004, +0.005]$
AWQ-4	0.976	0.973	0.959	0.991	$-0.003\,[-0.006, +0.000]$
QAT-Q4	0.977	0.968	0.983	0.937	$-0.003\,[-0.005, +0.000]$
QAT-BF16	0.977	0.969	0.986	0.978	–
12B: hard tier, cap 4,096					
BF16	0.940	–	–	0.800	–
Q8_0	0.939	–	–	0.788	–
Q4_K_M	0.934	–	–	0.784	$-0.006\,[-0.010, -0.002]$
Q3_K_M	0.933	–	–	0.820	$-0.006\,[-0.011, -0.001]$
IQ2_M	0.876	–	–	0.698	–
35B: MATH-500, cap 1,536					
BF16	0.975	0.967	0.961	1.000	–
Q8_0	0.976	0.967	0.961	1.000	–
Q4_K_M	0.973	0.964	0.959	1.000	$-0.003\,[-0.005, -0.001]$
Q3_K_M	0.979	0.969	0.963	1.000	$+0.003\,[+0.001, +0.006]$
IQ2_M	0.975	0.968	0.958	0.996	–
NF4	0.975	0.967	0.960	1.000	$+0.000\,[-0.002, +0.002]$
AWQ-4	0.972	0.966	0.954	1.000	$-0.003\,[-0.006, -0.001]$
35B: MATH-500, cap 4,096					
BF16	0.963	0.955	0.944	1.000	–
Q8_0	0.962	0.954	0.946	1.000	–
Q4_K_M	0.963	0.953	0.940	1.000	$+0.000\,[-0.004, +0.003]$
Q3_K_M	0.969	0.955	0.935	1.000	$+0.006\,[+0.000, +0.012]$
IQ2_M	0.958	0.946	0.932	0.995	–
35B: hard tier, cap 4,096					
BF16	0.980	–	–	1.000	–
Q8_0	0.981	–	–	1.000	–
Q4_K_M	0.981	–	–	1.000	$+0.001\,[-0.001, +0.002]$
Q3_K_M	0.978	–	–	1.000	$-0.002\,[-0.006, +0.002]$
IQ2_M	0.977	–	–	0.972	–

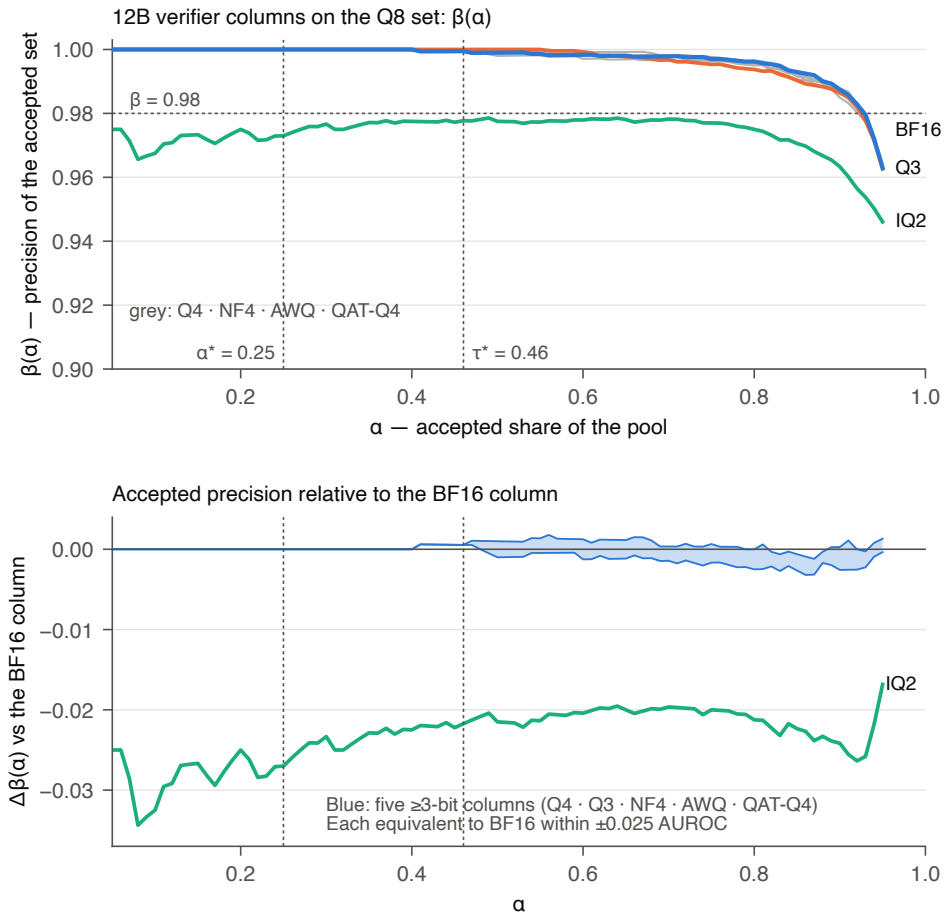

Figure 4: Accepted-set precision $\beta(\alpha)$ of the 12B verifier columns on the `Q8_0` set (upper) and $\Delta\beta(\alpha)$ vs. the BF16 column (lower): the five $\geq$3-bit columns, each TOST-equivalent within ± 0.025 AUROC, stay within a few thousandths of BF16 at matched acceptance rates; the 2-bit column sits 0.02–0.03 below.

E GENERATION: COMPOSITION, LENGTH, LOOPS

Repetitive-loop outcome. A response is labelled a repetitive loop when it is truncated and its word-8-gram repetition rate is at least 0.10, using whitespace-delimited words and the released detector. Loop shares below are conditional on truncation unless stated otherwise; the continuation intervention reports the fraction of all continuations meeting this joint criterion.

Table 7: Generation steps between adjacent rungs (paired ΔG, 95% CI).

Block	BF16→Q8	Q8→Q4	Q4→Q3	Q3→IQ2
12B MATH-500	$+0.001$ $[-0.005, +0.006]$	-0.011 $[-0.019, -0.004]$	-0.038 $[-0.051, -0.026]$	-0.429 $[-0.462, -0.397]$
12B hard tier	-0.001 $[-0.020, +0.019]$	-0.041 $[-0.062, -0.019]$	-0.092 $[-0.123, -0.065]$	-0.304 $[-0.364, -0.245]$
35B MATH-500	$+0.001$ $[-0.007, +0.009]$	$+0.010$ $[+0.002, +0.019]$	-0.021 $[-0.031, -0.009]$	-0.032 $[-0.046, -0.019]$
35B MATH-500 (4,096)	$+0.001$ $[-0.004, +0.006]$	$+0.002$ $[-0.006, +0.009]$	-0.009 $[-0.017, -0.001]$	-0.050 $[-0.064, -0.037]$
35B hard tier	-0.016 $[-0.035, +0.004]$	$+0.000$ $[-0.023, +0.022]$	-0.012 $[-0.038, +0.013]$	-0.097 $[-0.128, -0.069]$

Table 8: **Generation per precision**: pass@1 G, truncation share, pass@8.

	12B, MATH-500			12B, hard tier (4,096)			35B, MATH-500 (4,096)		
	G	trunc.	pass@8	G	trunc.	pass@8	G	trunc.	pass@8
BF16	0.917	0.07	0.96	0.469	0.33	0.69	0.943	0.05	0.98
Q8	0.918	0.07	0.97	0.468	0.33	0.70	0.944	0.05	0.98
Q4	0.906	0.08	0.97	0.427	0.34	0.63	0.946	0.04	0.98
Q3	0.868	0.12	0.96	0.335	0.54	0.56	0.937	0.05	0.98
IQ2	0.439	0.58	0.69	0.031	0.93	0.08	0.887	0.09	0.97

For the 12B on MATH-500, 79% of the `Q8_0` arm's wrong samples are truncations; the completed-candidate correctness is 0.979 / 0.981 / 0.979 / 0.977 / 0.906 down the ladder and pass@8 0.964 / 0.966 / 0.974 / 0.958 / 0.690. The 2-bit loss splits into $+0.517$ truncation and -0.073 completed correctness; by BF16-length tercile it is $-0.245/-0.551/-0.808$. The share of truncations that loop is $0.10 \to 0.13 \to 0.24 \to 0.86$ on the hard tier; SmolLM3 loops in ≈0.5 of its truncated responses at every rung including BF16 (a looping habit; its quantization loss is real errors), the E4B truncates more but loops little. A greedy ablation on the hard tier (the 153 problems, one greedy sample each, cap 4,096; declared before its data) reproduces the loss at temperature 0: paired ΔG vs. BF16 greedy -0.072 $[-0.137, -0.013]$ / -0.131 $[-0.203, -0.065]$ / -0.425 $[-0.503, -0.346]$ at 4 / 3 / 2-bit, with the same length inflation (mean $2,446 \to 3,982$ tokens). On the first loop's holdout (200 MATH-500 problems $\times$ 4) regenerated at a 4,096 cap, the NF4 generation deficit is -0.082 $[-0.113, -0.056]$ with truncation 0.124 vs. 0.019 and completed-candidate correctness no lower than BF16's (0.977 vs. 0.967). Bits-and-bytes NF4 costs the 12B far more generation than `Q4_K_M` ($G = 0.772$ with 0.193 truncation on the loop's $\tau^\star$ slice; paired loop-holdout deficit -0.097 $[-0.127, -0.072]$ at round 0) while its verifier column is equivalent ("4-bit $\neq$ 4-bit" for generation, "=" for verification).

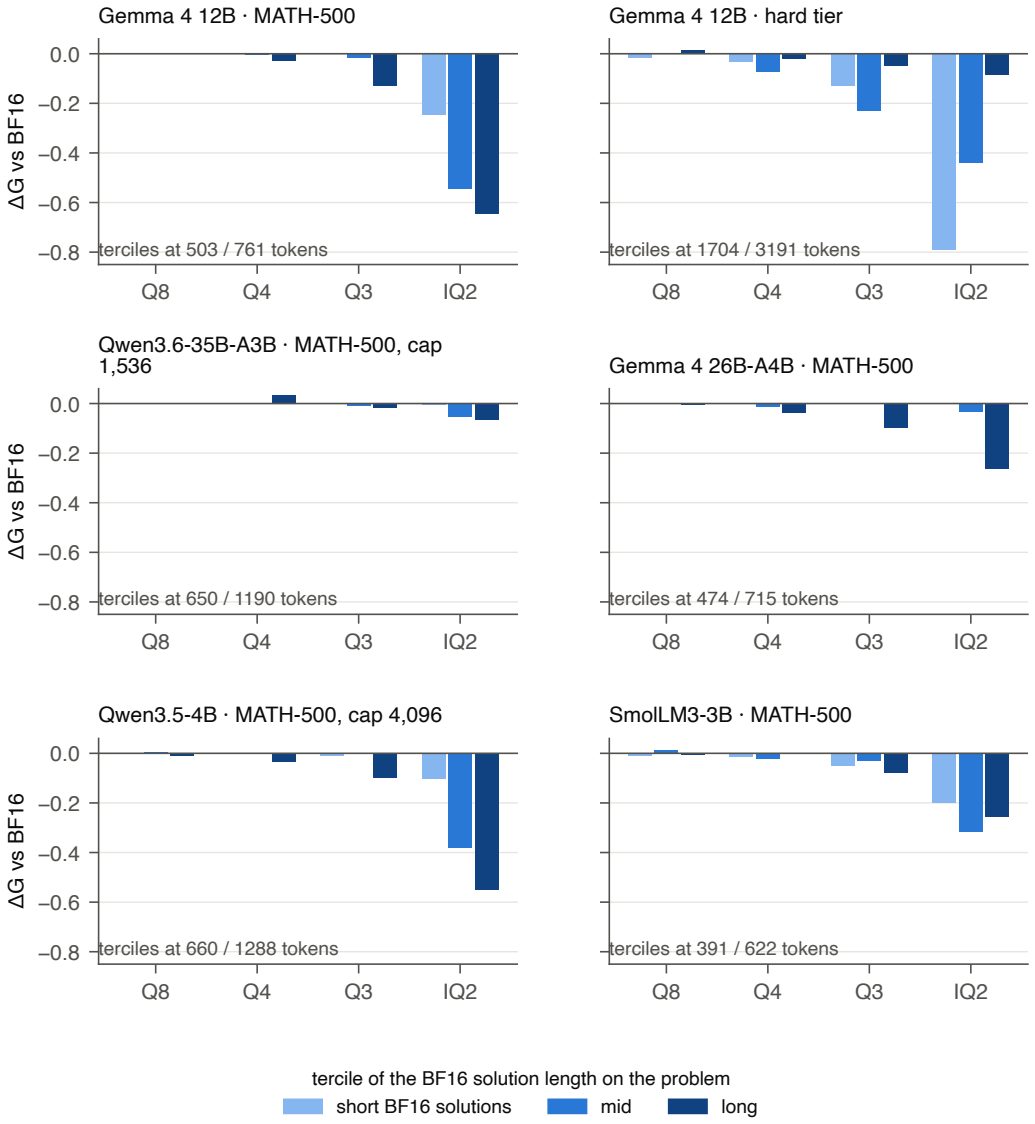

Figure 5: ΔG vs. BF16 by tercile of the BF16 solution length on the problem, per rung and model.

F HARD TIER

Table 9: 12B hard tier by competition: G per rung (BF16 / Q8 / Q4 / Q3 / IQ2), verifier AUROC on the `Q8_0` set on all samples and on completed candidates (BF16 / Q4 / Q3 / IQ2).

Competition	n	G	V all	V completed
AIME 2024	30	0.629 / 0.642 / 0.592 0.475 / 0.087	0.940 / 0.935 / 0.950 0.922	0.844 / 0.827 / 0.866 0.791
AIME 2025	30	0.487 / 0.475 / 0.467 0.371 / 0.025	0.964 / 0.956 / 0.956 0.896	0.909 / 0.887 / 0.886 0.743
AIME 2026	30	0.575 / 0.571 / 0.537 0.446 / 0.017	0.937 / 0.939 / 0.932 0.872	0.837 / 0.843 / 0.826 0.666
HMMT Feb 2025	30	0.329 / 0.329 / 0.296 0.179 / 0.013	0.942 / 0.933 / 0.936 0.892	0.867 / 0.844 / 0.850 0.755
HMMT Feb 2026	33	0.337 / 0.337 / 0.261 0.216 / 0.015	0.899 / 0.893 / 0.881 0.822	0.752 / 0.734 / 0.703 0.593

AIME 2024 is verbatim in NuminaMath-1.5 while AIME 2025–2026 and HMMT are not (Table 47); both subjects score highest on AIME 2024, and the column pattern holds inside every competition. The 35B remains cap-bound on the hard tier at 4,096 (truncation 0.616–0.723; 96%–98% of its errors are truncations; 17–36 real errors per set), so its columns there are truncation detection and carry no claim.

Table 10: 12B hard tier at caps 4,096 and 8,192 (declared under `v1.8`): pass@1, pass@8, truncation share, completed-candidate correctness, BF16-solvable problems lost, and the share of truncations that are token loops.

Arm	cap	pass@1	pass@8	trunc.	compl. corr.	lost	loops
BF16	4,096	0.469	0.686	0.33	0.698	–	10%
Q4_K_M	4,096	0.427	0.634	0.34	0.649	13	13%
Q3_K_M	4,096	0.335	0.556	0.54	0.707	22	24%
BF16	8,192	0.499	0.719	0.23	0.649	–	22%
Q4_K_M	8,192	0.470	0.732	0.29	0.658	8	32%
Q3_K_M	8,192	0.355	0.556	0.49	0.676	28	46%

Doubling the budget (Table 10) closes the 4-bit pass@8 deficit ($0.052 \rightarrow -0.013$) and widens the 3-bit one ($0.131 \rightarrow 0.163$; the pre-registered prediction that it would halve is falsified); the verifier columns at 8,192 stay flat to 3-bit on completed candidates (0.842 / 0.840 / 0.818 on 331 real errors; 2-bit 0.714).

G CODE DOMAIN

Table 11: Code sets: G and verifier AUROC on the `Q8_0` set per rung (BF16 / Q8 / Q4 / Q3 / IQ2), TOST contrasts vs. BF16 (90% CI), real errors on the BF16 set and the BF16 verifier's AUROC on them. † Equivalence not established within ± 0.025.

Per-rung generation and verification

Subject	Set	G	V (columns)
12B	HumanEval+	0.940 / 0.931 / 0.940 0.915 / 0.248	0.852 / 0.851 / 0.846 0.835 / 0.655
12B	MBPP+	0.777 / 0.773 / 0.773 0.757 / 0.352	0.727 / 0.724 / 0.725 0.738 / 0.676
12B	AtCoder slice (1,536)	0.470 / 0.466 / 0.447 0.354 / 0.091	0.940 / 0.941 / 0.931 0.939 / 0.656
12B	AtCoder slice (4,096)	0.570 / 0.576 / 0.534 0.451 / 0.098	0.896 / 0.896 / 0.887 0.888 / 0.515
35B	HumanEval+	0.908 / 0.908 / 0.899 0.891 / 0.820	0.869 / 0.873 / 0.877 0.875 / 0.881
35B	MBPP+	0.772 / 0.778 / 0.779 0.776 / 0.725	0.736 / 0.734 / 0.729 0.716 / 0.732
35B	AtCoder slice (1,536)	0.441 / 0.434 / 0.426 0.417 / 0.354	0.957 / 0.955 / 0.960 0.956 / 0.951
35B	AtCoder slice (4,096)	0.542 / 0.553 / 0.523 0.519 / 0.417	0.903 / 0.906 / 0.902 0.904 / 0.896

Contrasts and completed-error view

Subject	Set	Δ Q4	Δ Q3	n err.	V real
12B	HumanEval+	-0.006 $[-0.018, +0.007]$	-0.017 $[-0.043, +0.005]^{\dagger}$	86	0.845
12B	MBPP+	-0.002 $[-0.014, +0.010]$	$+0.012$ $[-0.009, +0.033]^{\dagger}$	637	0.706
12B	AtCoder slice (1,536)	-0.009 $[-0.022, -0.001]$	-0.001 $[-0.012, +0.010]$	54	0.737
12B	AtCoder slice (4,096)	-0.009 $[-0.022, +0.001]$	-0.008 $[-0.023, +0.008]$	71	0.713
35B	HumanEval+	$+0.008$ $[-0.005, +0.022]$	$+0.007$ $[-0.026, +0.043]^{\dagger}$	116	0.864
35B	MBPP+	-0.006 $[-0.016, +0.003]$	-0.020 $[-0.036, -0.004]^{\dagger}$	653	0.729
35B	AtCoder slice (1,536)	$+0.003$ $[-0.002, +0.009]$	-0.002 $[-0.007, +0.004]$	56	0.826
35B	AtCoder slice (4,096)	-0.001 $[-0.008, +0.007]$	$+0.000$ $[-0.009, +0.010]$	122	0.843

Table 12: Pooled HumanEval+ + MBPP+ TOST per subject (90% CI); Q8–IQ2 columns are contrasts vs. BF16. † Equivalence not established within ± 0.025.

Subject	n problems	V BF16	Q8	Q4	Q3	IQ2
	4,336		-0.001	-0.010	-0.007	-0.074
12B	542	0.728	$[-0.004, +0.001]$	$[-0.021, +0.000]$	$[-0.024, +0.010]$	$[-0.118, -0.030]^{\dagger}$
	4,336		$+0.000$	-0.005	-0.017	-0.038
35B	542	0.757	$[-0.003, +0.003]$	$[-0.013, +0.003]$	$[-0.031, -0.003]^{\dagger}$	$[-0.060, -0.017]^{\dagger}$

The AtCoder slice holds the 59 ABC/ARC problems of the 17 contests held after 2026-06-03 (the later of the two subjects' weight dates) with difficulty $\leq$ 1600; official test data being unavailable, each problem carries a reference solution, an independent brute-force solution and a seeded input generator, produced with LLM assistance and validated centrally (every public sample reproduced; 1,804 reference-vs-brute cross-checks with 0 mismatches over 2,497 tests; three problems with several valid outputs ship a checker). The slice stays hard un-capped ($G = 0.570$ for the 12B at 4,096) and is the one place a real 12B `Q8_0`→`Q4_K_M` generation step appears on code (-0.042 $[-0.083, -0.002]$, $p = 0.022$); its verifier columns stay equivalent. The 12B's 2-bit collapse on code is intrinsic non-termination (75% of HumanEval+ samples cut at 1,536; 89% on the

slice at 4,096); the 35B degrades gracefully (HumanEval+ $0.908 \rightarrow 0.820$ with its 2-bit column at 0.881).

H THE PANEL

Table 13: Panel: G per rung, Level-5 verifier columns on the `Q8_0` set (BF16 / Q8 / Q4 / Q3 / IQ2), TOST contrasts, real errors and the BF16 $\rightarrow$ 2-bit real-error AUROC. TOST contrasts use 90% CIs; † equivalence not established within ± 0.025. Qwen3.5-4B at a 4,096 cap; Ministral cap-bound at 1,536.

Per-rung generation and verification

Model	size	G	V L5
Qwen3.5-4B	4B	0.878 / 0.874 / 0.866 0.842 / 0.533	0.950 / 0.950 / 0.947 0.929 / 0.855
SmolLM3-3B	3B	0.718 / 0.716 / 0.705 0.665 / 0.462	0.794 / 0.796 / 0.770 0.750 / 0.722
Ministral 3 3B	3B	0.656 / 0.647 / 0.633 0.582 / 0.345	0.835 / 0.834 / 0.843 0.796 / 0.690
Gemma 4 E4B	≈4B	0.703 / 0.703 / 0.708 0.588 / 0.531	0.909 / 0.909 / 0.916 0.853 / 0.860
Gemma 4 26B-A4B	26B/4B	0.930 / 0.926 / 0.913 0.896 / 0.829	0.962 / 0.964 / 0.959 0.958 / 0.966
Gemma 4 31B	31B	0.962 / 0.960 / 0.961 0.961 / 0.920	0.896 / 0.905 / 0.895 0.888 / 0.939

Contrasts and completed-error view

Model	Δ Q4	Δ Q3	n err.	V real
Qwen3.5-4B	-0.002 $[-0.006, +0.001]$	-0.020 $[-0.030, -0.012]^{\dagger}$	107	$0.795 \rightarrow 0.670$
SmolLM3-3B	-0.024 $[-0.037, -0.012]^{\dagger}$	-0.044 $[-0.063, -0.027]^{\dagger}$	1079	$0.848 \rightarrow 0.752$
Ministral 3 3B	$+0.008$ $[+0.001, +0.015]$	-0.039 $[-0.058, -0.020]^{\dagger}$	227	$0.705 \rightarrow 0.628$
Gemma 4 E4B	$+0.007$ $[+0.003, +0.012]$	-0.056 $[-0.077, -0.035]^{\dagger}$	399	$0.565 \rightarrow 0.545$
Gemma 4 26B-A4B	-0.003 $[-0.008, +0.001]$	-0.004 $[-0.011, +0.001]$	28	$0.902 \rightarrow 0.885$
Gemma 4 31B	-0.001 $[-0.010, +0.008]$	-0.009 $[-0.026, +0.007]^{\dagger}$	59	$0.800 \rightarrow 0.902$

The corresponding trajectories are retained in Evidence Fig. E2.

Two notes. The 31B's 2-bit verifier column reads *above* its own BF16 column (0.939 vs. 0.896); this is a read-resolution artefact, not better 2-bit verification: up to 84% of the 31B's $\geq$3-bit verdict scores are exactly 1.0 (the No token outside the top-25 read), so those columns cannot rank inside the tied mass while the 2-bit column resolves them. The E4B's verifiers are near chance on completed candidates at every rung (0.565 at BF16) and its 3-bit loss is the one verdict whose noise jumps ($0.18 \rightarrow 0.91$); the 26B-A4B (MoE, 4 B active, 3.22 body bpw at 2-bit) reads like the 35B MoE, not like the dense 4 B models, keeping its real-error margin at 2-bit ($0.902 \rightarrow 0.885$, $n = 28$). Body bpw at the 2-bit rung is iso for the 12B, 31B and 35B (2.69 / 2.71 / 2.78) and milder for the 26B-A4B and Qwen3.5-4B (3.22 / 3.08).

I THINKING-MODE VERDICTS

Two regimes, neither an improvement over the one-token read. The 12B's thinking verdict equals the $t = 0$ read when its thinking block closes (within 0.02 AUROC at every $\geq$3-bit rung) and its cost is closure, falling 0.80 / 0.80 / 0.75 / 0.57 / 0.016 down the ladder; the 35B closes (0.89–0.96) but reads 0.65–0.72 against 0.97 at $t = 0$ on the same rows, largely because it says Yes to most truncated candidates; its real-error discrimination is unmeasurable at 10–12 negatives. A verifier that generates pays the generation exposure.

The closed-row and unconditional comparisons are retained in Evidence Table E2.

J ESTIMATORS

We compare the self-judge score defined in §2 with two generation-derived scores. Majority agreement is the fraction of the k samples for a problem in the candidate's answer-equivalence cluster (the candidate included; a missing extracted answer scores zero). Mean log-probability is the mean of the generated tokens' log-probabilities under the generating model. Each score is evaluated against oracle correctness on the same diagonal population.

Table 14: Estimator dissociation on the diagonal: self-judge, majority agreement (maj@k) and mean token log-probability AUROC, with paired contrasts (95% CI).

Block	Arm	self-judge	maj@k	logprob	self − logprob	maj@k − self
12B MATH-500	BF16	0.971	0.985	0.785	+0.186 [+0.134, +0.241]	+0.014 [−0.001, +0.034]
12B MATH-500	Q8	0.978	0.989	0.785	+0.193 [+0.141, +0.247]	+0.010 [+0.005, +0.018]
12B MATH-500	Q4	0.967	0.971	0.799	+0.168 [+0.121, +0.216]	+0.004 [−0.024, +0.028]
12B MATH-500	Q3	0.963	0.987	0.807	+0.156 [+0.118, +0.196]	+0.024 [+0.011, +0.038]
12B MATH-500	IQ2	0.929	0.952	0.647	+0.281 [+0.250, +0.314]	+0.023 [+0.010, +0.036]
12B hard tier	BF16	0.941	0.957	0.823	+0.118 [+0.071, +0.170]	+0.016 [−0.002, +0.036]
12B hard tier	Q8	0.939	0.950	0.829	+0.110 [+0.067, +0.155]	+0.011 [−0.008, +0.032]
12B hard tier	Q4	0.925	0.944	0.826	+0.099 [+0.055, +0.149]	+0.019 [−0.011, +0.047]
12B hard tier	Q3	0.948	0.947	0.829	+0.119 [+0.077, +0.167]	−0.001 [−0.023, +0.021]
12B hard tier	IQ2	0.963	0.848	0.590	+0.373 [+0.293, +0.452]	−0.115 [−0.269, −0.036]
35B MATH-500	BF16	0.974	0.992	0.751	+0.223 [+0.183, +0.268]	+0.019 [+0.011, +0.027]
35B MATH-500	Q8	0.976	0.989	0.747	+0.228 [+0.189, +0.273]	+0.014 [+0.006, +0.022]
35B MATH-500	Q4	0.971	0.993	0.772	+0.199 [+0.163, +0.242]	+0.021 [+0.012, +0.033]
35B MATH-500	Q3	0.977	0.991	0.748	+0.229 [+0.190, +0.272]	+0.014 [+0.008, +0.020]
35B MATH-500	IQ2	0.969	0.990	0.763	+0.205 [+0.171, +0.246]	+0.022 [+0.014, +0.031]
35B MATH-500 (4,096)	BF16	0.958	0.984	0.878	+0.080 [+0.047, +0.117]	+0.026 [+0.005, +0.048]
35B MATH-500 (4,096)	Q8	0.962	0.985	0.864	+0.098 [+0.056, +0.145]	+0.023 [+0.009, +0.036]
35B MATH-500 (4,096)	Q4	0.964	0.985	0.857	+0.107 [+0.061, +0.164]	+0.021 [+0.007, +0.034]
35B MATH-500 (4,096)	Q3	0.945	0.973	0.853	+0.092 [+0.054, +0.141]	+0.028 [−0.012, +0.060]
35B MATH-500 (4,096)	IQ2	0.926	0.992	0.805	+0.121 [+0.085, +0.162]	+0.066 [+0.047, +0.089]

Majority agreement usually matches or exceeds the self-judge, and both outperform mean log-probability (Table 14). The hard-tier 12B diagonal is an exception at 2-bit: majority AUROC is 0.848 versus self-judge 0.963 (majority minus self-judge -0.115 $[-0.269, -0.036]$, 95% CI); the small Q3 reversal is unresolved. Majority AUROC also falls from 0.957 at BF16. These diagonal candidate populations change with the generator, so high 2-bit self-judge AUROC here does not establish preserved discrimination on a fixed completed-candidate set. The majority estimator's advantage over the self-judge grows as models shrink (Ministral $+0.13$ to $+0.19$, SmolLM3 $+0.11$ to $+0.15$ across the rungs).

K DECOMPOSITION OF THE VERDICT

Units and weighting. For both $N = 8$ and $N = 64$, within-problem AUROC averages each eligible problem's AUROC with weight $n_{\text{correct}} n_{\text{wrong}}$, using completed candidates with finite scores and problems holding both classes. Intervals resample problems and recompute this weighted mean. Between-problem AUROC instead uses each problem's mean verdict score to classify whether its observed pass rate is at least 0.5. These statistics use different units and are not additive.

Table 15 gives a compact view of the original eight-attempt decomposition; the complete ladder and graphical comparison are in Evidence Table E3 and Evidence Fig. E3.

Table 15: Original eight-attempt diagnostics at BF16: aggregate and between-problem AUROC, within-problem completed-candidate AUROC and eligible problem count. The small counts motivate the separate $N = 64$ experiment below. These views are not additive components.

Block	All	Between	Within (n problems)
12B MATH-500	0.979	0.979	0.53 (20)
12B hard tier	0.940	0.956	0.63 (45)
35B MATH-500	0.975	0.981	0.83 (4)
35B MATH-500 (4,096)	0.963	0.980	0.67 (18)

No $\geq$3-bit rung moves the within-problem read resolvably on any $\geq$4 B model (SmolLM3 at 3-bit, -0.042 $[-0.066, -0.020]$, is the one resolved case); at 2-bit the 12B's within-problem loss is -0.103 $[-0.220, +0.015]$ on the hard tier and its between-problem loss -0.084 $[-0.126, -0.046]$ on MATH-500—what 2-bit and small-model 3-bit lose resolvably is difficulty recognition. The linear-probe block (Table 16) adds that the frozen BF16 probe direction already fails at 4-bit on the hard tier (0.477 vs. native 0.599; NF4 0.496 vs. 0.596) while 8-bit keeps it (0.572): quantization rotates the linear read before it removes the signal. On the pool-distribution holdout the re-fitted 2-bit probe beats the 2-bit self-judge by $+0.206$ $[+0.052, +0.362]$ (not reproduced on registered rows; below).

Table 16: Linear probe on the verdict prompt's last-token hidden state (logistic, layer and C by grouped CV on the loop pool): within-problem completed-candidate AUROC of the same-pass self-judge, the arm-native probe and the frozen BF16 probe on three test sets.

	MATH-500 (`Q8_0` set)			hard tier (`Q8_0` set)			pool-like holdout		
Arm	self	native	frozen	self	native	frozen	self	native	frozen
BF16	0.51	0.46	0.46	0.64	0.59	0.59	0.63	0.68	0.68
Q8_0	0.51	0.46	0.47	0.64	0.58	0.57	0.62	0.67	0.69
Q4_K_M	0.57	0.59	0.56	0.64	0.60	0.48	0.62	0.70	0.69
Q3_K_M	0.55	0.66	0.65	0.60	0.59	0.55	0.60	0.69	0.67
IQ2_M	0.63	0.54	0.60	0.53	0.56	0.53	0.50	0.71	0.49
NF4	0.42	0.55	0.58	0.63	0.60	0.50	0.50	0.61	0.59

An exploratory LLM-labelled taxonomy used a first pass, blind second pass and blind tie-break on disagreements ($\kappa = 0.66$ between full passes; human validation of 114 items remains pending). The BF16 verdict detects plausible method failures more readily than arithmetic slips or misreadings. These labels do not identify the mediator of the source effect; the labeling record and category-level rejection rates are in the supplementary data.

Attempt-level verification at $N = 64$ (`v1.8`). Sixty-four samples per hard-tier problem (9,792 rows; 92 problems holding both classes, 2,109 real errors) power the within-problem estimand at ± 0.05 (Table 17). The `v1.8` declaration chose ± 0.05 for an expected interval half-width of about 0.03, rather than inheriting the static ± 0.025 margin. It is a design tolerance, not a demonstrated negligible change in selector utility. As a sensitivity, Q3's interval $+0.008$ $[-0.013, +0.029]$ would miss ± 0.025 at its positive upper bound; this does not indicate a resolved loss. The one-token verdict's within-problem AUROC on completed candidates is 0.573 $[0.516, 0.629]$ at BF16 against majority agreement's 0.880; the 8-, 4- and 3-bit contrasts are equivalent and the 2-bit loss resolves on both the one-token and the pairwise read. The pairwise instrument samples at most 32 correct–wrong pairs per problem, presents both orders and gives half credit to ties. C11 evaluates the unary preference on those exact pairs with identical weights: at BF16, pairwise accuracy is 0.609 versus 0.605 for unary preference over 2,820 pairs from 92 problems. The matched difference $+0.003$ $[-0.049, +0.055]$ (95%) shows no detected advantage. The original difference between differently weighted summaries is retained in the generated data, separately from this instrument comparison. The NF4 one-token column is interface-collapsed on these long candidates (median Yes/No mass 0.06) and excluded from equivalence claims by rule.

The probes on the $N = 64$ set (`v1.9`). The frozen and arm-native linear probes of Table 16 applied unchanged to the same rows (Table 18): at BF16 native probe $-$ verdict is $+0.001$ $[-0.032, +0.032]$ (equivalent at ± 0.05); the arm-native probes retain the BF16 signal at every $\geq$3-bit rung and NF4 (above-chance retention 0.98 / 1.11 / 0.99) while the frozen BF16 direction erodes down the ladder; at 2-bit the native probe's within-problem AUROC is 0.534 $[0.488, 0.581]$ with retention 0.47 $[-0.12, +0.92]$, so the pool-holdout "readout collapse" reading of Table 16 does not reproduce on registered rows.

Table 17: Attempt-level verification on the hard tier at $N = 64$: Yes/No mass, AUROC on all rows and on completed candidates, the within-problem completed-candidate AUROC (95% CI at BF16) and its contrast vs. BF16 (90%; † = resolved loss), the pairwise read's accuracy and contrast (90%), and $\Delta_{\mathrm{read,matched}}$ = pairwise − unary preference on identical sampled pairs (95%; C11). ‡ = interface-collapsed.

One-token scores				
Verifier	mass	V all	V compl.	within
BF16	0.91	0.935	0.843	0.573 [0.516, 0.629]
Q8_0	0.91	0.935	0.842	0.574
Q4_K_M	0.92	0.931	0.832	0.572
Q3_K_M	0.95	0.929	0.831	0.581
IQ2_M	0.78	0.868	0.683	0.532
NF4 (HF path)‡	0.06	0.908	0.807	0.564

Paired readout contrasts				
Verifier	Δ within	pairwise	Δ pairwise	$\Delta_{\mathrm{read,matched}}$
BF16	–	0.609	–	+0.003 [−0.049, +0.055]
Q8_0	+0.002 [−0.003, +0.007]	0.609	+0.001 [−0.003, +0.004]	−0.001 [−0.053, +0.052]
Q4_K_M	+0.000 [−0.011, +0.011]	0.611	+0.003 [−0.005, +0.011]	+0.007 [−0.044, +0.059]
Q3_K_M	+0.008 [−0.013, +0.029]	0.602	−0.007 [−0.021, +0.008]	+0.014 [−0.038, +0.068]
IQ2_M	−0.040 [−0.068, −0.012]†	0.567	−0.042 [−0.066, −0.019]†	+0.009 [−0.034, +0.051]
NF4 (HF path)‡	−0.008 [−0.031, +0.014]	–	–	–

Table 18: The linear probes on the $N = 64$ hard set: within-problem completed-candidate AUROC of the arm-native probe, the same-pass self-judge (95% cluster CIs), the frozen BF16 probe (point estimates), native − self-judge (90%) with flag E for equivalence within ± 0.05, and the above-chance retention r_{ac} (90%).

Arm	native probe	self-judge	frozen probe	native − self	flag	r_{ac}
BF16	0.573 [0.516, 0.626]	0.572 [0.518, 0.624]	0.573	+0.001 [−0.032, +0.032]	E	1.00 [+1.00, +1.00]
Q8_0	0.571 [0.514, 0.628]	0.571 [0.517, 0.622]	0.565	+0.001 [−0.030, +0.031]	E	0.98 [+0.82, +1.09]
Q4_K_M	0.571 [0.518, 0.619]	0.580 [0.524, 0.632]	0.552	−0.009 [−0.036, +0.018]	E	0.98 [+0.57, +1.78]
Q3_K_M	0.581 [0.531, 0.628]	0.575 [0.527, 0.624]	0.562	+0.006 [−0.027, +0.039]	E	1.11 [+0.71, +2.02]
IQ2_M	0.534 [0.488, 0.581]	0.535 [0.493, 0.581]	0.525	−0.001 [−0.036, +0.036]	E	0.47 [−0.12, +0.92]
NF4	0.572 [0.527, 0.619]	0.573 [0.520, 0.623]	0.550	+0.000 [−0.035, +0.035]	E	0.99 [+0.60, +1.90]

L MECHANISM BLOCKS

Two perturbation statistics. With verdict log-odds $z = \log p_{\mathrm{yes}} - \log p_{\mathrm{no}}$, teacher-forced margin perturbation is $\rho_{\mathrm{shift}} = \mathrm{SD}(z_q - z_b)/\mathrm{SD}(z_b)$. The column regression fits $z_q = a + g z_b + \epsilon$ on jointly uncensored rows and reports $\rho_{\mathrm{reg}} = \mathrm{SD}(\epsilon)/(|g|\mathrm{SD}(z_b))$. The counterfactual applies the fitted shift and gain plus independently resampled residuals to BF16 scores, clipping log-odds to ± 20 before computing AUROC. The two normalizations and their evaluation rows differ; they are not interchangeable.

The generation perturbation and trajectory plots are retained in Evidence Fig. E4.

The per-column counterfactual values underlying Fig. 6 are in Evidence Table E4.

The full 33-pair teacher-forced ladder is in Evidence Table E5.

Additivity on the 12B: the KL of the two 2-bit mixes sums to $0.88\times$ the uniform 2-bit rung's and their noise variances to $0.65\times$ (sub-additive), at 3-bit $1.24\times$ and $1.27\times$ (super-additive). No mix touches the stop decision (end-of-turn flip rate 0). The 3.15-bpw non-GGUF arms collapse the Yes/No interface (mass 0.03 / 0.08), recorded before any counted sample; their column AUROCs are descriptive. The hard tier repeats the two 2-bit location calls; its 3-bit-level residuals are unreadable

Table 19: **Seven pre-specified tests of the mechanism** (predictions declared in committed amendments before each block's measurement; on the 12B unless stated): what held and what fell. Appendix A documents the consolidated archival tag chain; these exploratory tests are outside the confirmatory family.

Test	Prediction	Outcome
Hazard shape (8 models; within BF16-entropy quartiles)	The hazard does not rise with position, marginally or within entropy quartiles	**Held marginally, fell conditionally:** 0 of 33 and 0 of 9 pairs rising; the quartile test fired (8 rising of 344 cells, 275 undetermined)
Quantizer (AWQ 3-bit g128/g32, auto-round 3-bit, AWQ 4-bit)	Every arm on the ladder's KL→noise curve; nominal bits need not	**Held:** noise factors 0.79 / 0.83 / 0.69 / 1.10; "3-bit" g128 lands 1.4–$1.8\times$ beyond the 2-bit rung in KL
Tensor location (12B; Qwen3.5-4B)	Generation damage follows the KL regardless of location; the verdict may not	**Held on the 12B, fell on the 4B:** 12B mixes on the hazard curve (ϵ_T factors 1.05–1.09), attention @ 2-bit perturbs its verdict $0.62\times$ the prediction; the 4B's attention-path mix is on-curve for the verdict ($1.08\times$; column -0.049 $[-0.060,\ -0.037]$) and $1.76\times$ above the hazard curve
Hidden-state readout ($N = 64$ hard set)	At BF16 the probe reads no more than the verdict; at 2-bit a re-fitted probe recovers what the read lost	**Held at BF16, fell at 2-bit:** probe − verdict $+0.001$ $[-0.032,\ +0.032]$; the 2-bit-native probe not resolved above chance (0.534 $[0.488,\ 0.581]$; retention 0.47 $[-0.12,\ +0.92]$)
KV cache (4-bit K/V, full-precision weights)	Columns equivalent; generation truncates more; the 8-bit cache inert; the read's perturbation grows with context length	**Held in part:** -0.002 $[-0.004,\ +0.000]$ / $+0.000$ $[-0.004,\ +0.004]$; Δtrunc $+0.013$ $[+0.006,\ +0.020]$; 8-bit inert; the context prediction not supported (the read's KL by length tercile is non-monotone)
Continuation from matched prefixes (12B; 3- and 2-bit)	Two pre-written readings: the departed prefix carries the excess (prefix-driven) or the continuing model does (perturbation-driven); ± 0.05 margin	**Perturbation-driven at both rungs:** prefix effect under BF16 continuation -0.003 $[-0.013,\ +0.006]$ / $+0.008$ $[+0.000,\ +0.017]$ (negligible); continuation effect on the departed prefix $+0.493$ $[+0.453,\ +0.537]$ / $+0.050$ $[+0.037,\ +0.066]$; interaction negligible; both decompositions agree
Held-out prediction (3 models, registered before measurement)	A's per-rung error below the no-loss, rung-only-mean and input-matched baselines	**Not established:** MAE 0.0069 / 0.0072 at 4- / 3-bit, below no loss at both rungs and the rung-only mean at 4-bit, never below the input-matched baseline; the measured perturbation predicts within 0.0060 / 0.0062

Table 20: Mixed-precision arms of the 12B on the `Q4_K_M` anchor: per-token KL, verdict-margin noise and its factor against the uniform ladder's curve, generation, truncation, loop share, the truncation-hazard factor, the completed-candidate ΔV and the pre-specified call.

Arm	KL ρ	ρ factor	G trunc. loop	ϵ_T factor	ΔV compl.	call
			0.906			
	0.013		0.08			
Q4_K_M (anchor)	0.205	–	0.06	–	-0.007	–
			0.868			
	0.048		0.12			
Q3_K_M	0.402	–	0.13	–	-0.015	–
			0.439			
	0.219		0.58			
IQ2_M	0.882	–	0.50	–	-0.494	–
			0.670			
	0.142		0.33			
M1: FFN @ IQ2_M	0.664	0.94	0.42	1.05	-0.127	location-neutral
			0.856			
	0.051		0.13			verification-localised (below
M2: attention @ IQ2_M	0.257	0.62	0.10	0.90	-0.020	the curve)
			0.883			
	0.035		0.10			
M1′: FFN @ Q3_K_M	0.342	1.00	0.08	1.02	-0.018	location-neutral
			0.890			
	0.024		0.09			
M2′: attention @ Q3_K_M	0.295	1.04	0.07	1.09	-0.017	location-neutral
			0.906			
	0.015		0.08			
M3: embedding @ IQ3_S	0.202	0.92	0.04	0.81	-0.005	location-neutral

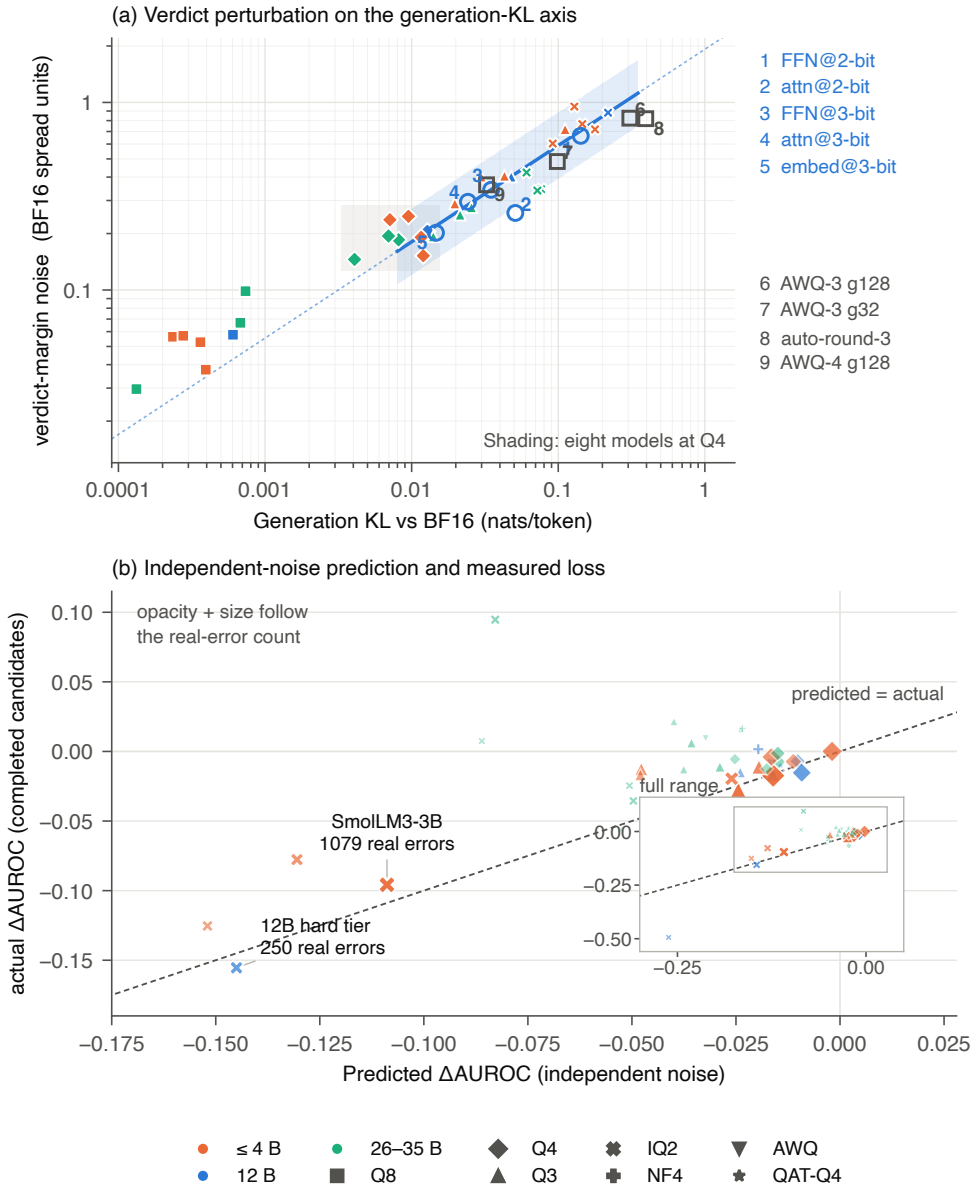

Figure 6: **The verdict side.** Upper panel: verdict-margin perturbation ρ_{shift} (SD of the shift in verdict log-odds, in units of the BF16 spread) against the per-token KL of the same (model, precision) pair under teacher forcing; the 12B's uniform rungs give $\rho_{\text{shift}} = 1.92 \cdot \text{KL}^{0.51}$ (factor-of-1.5 band). The non-GGUF arms and most mixed-precision arms lie inside the band; the 12B attention-at-2-bit mix lies below it ($0.62\times$). Lower panel: AUROC loss on completed candidates predicted from independent logit noise vs. measured, for every quantized verifier column of the eleven math blocks.

Table 21: Non-GGUF arms of the 12B through the registered engine: per-token KL, verdict noise and factor, Yes/No mass, generation, truncation, hazard factor and the all-samples column contrast vs. BF16.

Arm	KL	ρ_{shift}	factor	mass	G	trunc.	ϵ_T factor	ΔV all
AWQ 4-bit g128 (bridge; columns only)	0.032	0.364	1.10	1.00	–	–	–	-0.001 [-0.004, $+0.002$]
AWQ 3-bit g32	0.099	0.484	0.83	1.00	0.757	0.23	1.28	-0.023 [-0.033, -0.014]
AWQ 3-bit g128	0.308	0.824	0.79	0.03	0.155	0.90	2.03	-0.103 [-0.128, -0.080]
auto-round 3-bit g128	0.395	0.819	0.69	0.08	0.208	0.77	1.00	-0.024 [-0.035, -0.015]

because the uniform rungs' termination quantities at a 4,096 cap are not a power law in KL. The E4B's and Qwen3.5-4B's low-bit truncation is a destroyed stop decision (end-of-turn top-1 flips 0.57 and 0.86) whereas the 12B's is early derailment and loops (0.003); the QAT-Q4 checkpoint halves the 4-bit per-token perturbation (0.0059 vs. 0.0128 nats) with no verification consequence.

The KV cache as a second perturbation (`v1.9`). The 12B's BF16 weights under three server settings—the f16 cache under flash attention (the comparator; it matches the registered BF16 arm: G $+0.004$ [-0.002, $+0.009$], column $+0.000$ [$+0.000$, $+0.000$]), a `q8_0` and a `q4_0` K/V cache—with generation rows on MATH-500, columns on the registered set and the $N = 64$ set, and two teacher-forced pairs over the 12B's documents (Table 22). The 4-bit cache's ΔG by BF16-length tercile is -0.001 / -0.008 / -0.019; on the KL axis it is a 0.0053-nat perturbation per generation token with a verdict-margin perturbation of 0.194, between the `Q8_0` and `Q4_K_M` rungs, and it reaches the verdict read harder per unit of generation KL than a weight perturbation does (0.0120 nats per verdict token) with no AUROC consequence; the read's perturbation does not grow with the candidate's length (ρ_{shift} by length tercile 0.63 / 0.41 / 0.13), so the pre-specified context-growth prediction is not supported.

Table 22: **KV-cache quantization on the 12B at full-precision weights**, vs. the f16-cache control: TOST contrasts of the verifier column (all levels; Level 5; 90% CI), the within-problem contrast on the $N = 64$ hard set (95%), ΔG, Δtruncation and Δcompleted-candidate correctness on MATH-500 (95%), and the per-token KL of generation and verdict documents with the verdict-margin perturbation ρ_{shift}.

Verification

Cache	ΔV all	ΔV L5	Δ within
q8_0 K/V cache	$+0.000$ [$+0.000$, $+0.000$]	$+0.000$ [$+0.000$, $+0.001$]	$+0.002$ [-0.001, $+0.004$]
q4_0 K/V cache	-0.002 [-0.004, $+0.000$]	$+0.000$ [-0.004, $+0.004$]	-0.005 [-0.016, $+0.006$]

Generation

Cache	ΔG	Δtrunc.	Δcompl. corr.
q8_0 K/V cache	-0.002 [-0.008, $+0.004$]	$+0.004$ [-0.002, $+0.010$]	$+0.001$
q4_0 K/V cache	-0.010 [-0.016, -0.003]	$+0.013$ [$+0.006$, $+0.020$]	$+0.001$

Teacher-forced perturbation

Cache	KL gen.	KL verdict	ρ
q8_0 K/V cache	0.0000	0.0001	0.021
q4_0 K/V cache	0.0053	0.0120	0.194

Location on a hybrid 4B (`v1.8`). Qwen3.5-4B (8 full-attention and 24 linear-attention layers) with its attention path, its full-attention layers or its FFN at the 2-bit assignment on the `Q4_K_M` anchor, and its attention path at the 3-bit assignment (Table 23): the attention-path mix's verdict noise sits on its KL curve (1.08×; the 12B's sat at 0.62×), its column is not equivalent (-0.049 [-0.060, -0.037]), and it truncates 1.76× what the hazard curve predicts although its mean KL, top-1 flip rate, tail KL and end-of-turn flip rate all sit on the uniform curve; the FFN mix collapses the interface; the 3-bit-level mix is location-neutral. The allocation corollary is the 12B's.

Table 23: Qwen3.5-4B mixes on the Q4 anchor: perturbation factors and fixed-candidate column contrasts. Full uniform-rung comparisons, absolute perturbations and generation diagnostics are in Evidence Table E6. The attention advantage of the 12B does not generalize.

Arm	body bpw	noise factor	ΔV (90%)	hazard factor
attention path @ 2-bit	4.07	1.08	-0.049 $[-0.060, -0.037]$	1.76
full attention @ 2-bit	4.72	1.27	-0.042 $[-0.054, -0.030]$	1.23
FFN @ 2-bit	3.50	1.02	collapsed	2.33
attention path @ Q3	4.54	1.09	-0.003 $[-0.007, +0.001]$	0.88

The conditional hazard (`v1.8`). The first-divergence hazard within quartiles of the base model's entropy (Table 24): events concentrate in the top quartile, most cells are unresolved, and the declared test fires in 8 cells; the supplementary equal-event strata and a logistic hazard read flat-to-falling (3 and 1 rising), designed after the quartile table and reported as such.

Table 24: The conditional first-divergence hazard: cells and calls per instrument (constant / falling / rising / undetermined).

Instrument	cells	constant	falling	rising	undetermined
declared quartiles of at-risk positions	344	58	3	8	275
equal-event strata (supplement)	172	86	19	3	64
logistic hazard, 256–512 coefficient (supplement; pairs)	43	–	13	1	29

M DERAILMENT ON THE SAMPLED TRAJECTORY

The complete departure and replay diagnostics are in Evidence Tables E7 and E8.

A departure is a sampled token outside BF16's top-p nucleus at the rung's own prefix (the pinned sampler applies top-p before temperature, so the nucleus is a property of the raw softmax). The base-side floor of the test (BF16's own tokens outside its own nucleus) is 3.0×10^{-5} per token; rung-side floors are $\leq 10^{-3}$. Departure timing does not separate outcomes at $\geq$3-bit (survival without departure is the same for documents that end correct and documents that truncate), but a departure inside the first 16 tokens roughly doubles the truncation rate at 4-bit (0.148 vs. 0.059 for $T_1 \in [17, 64]$); at 2-bit every document departs within its first few dozen tokens (median T_1 15–37) and T_1 carries no information. Under every alternative departure definition (temperature-scaled nucleus, a 0.97 margin, nucleus $\wedge\, p_{\text{BF16}} < 10^{-2}$ or $< 10^{-3}$) the hazard calls are constant, falling or undetermined, never rising.

The attention-path mix of Qwen3.5-4B replayed on its own samples (`v1.9`; Table E8): post-departure amplification $\times 1.26$ over 1,673 documents against the uniform rungs' $\times 1.04$ / $\times 1.08$ / $\times 1.51$; departure rate 0.0135 per token; 133 looping documents with 0.99 of loop tokens inside BF16's nucleus.

Continuation from matched prefixes (`v1.10`). For every replayed 12B document at `Q3_K_M` and `IQ2_M` with a first departure at $16 \leq T_1 \leq 1{,}279$, the rung's prefix through the departing token and the BF16 base's own sample of the same problem and index cut at the same token count (documents whose BF16 sample is shorter were excluded outcome-blind) were each continued afresh by the rung and by BF16 under the registered sampler with $1{,}536 - (T_1 + 1)$ tokens of budget, two replicates per arm on fresh seed slots: 1,628 3-bit and 995 2-bit documents (the 2-bit population is the later-departing half, median T_1 34 vs. 112). The fresh rung-on-rung arm reproduces the originals (truncation 0.557 vs. 0.575 at 2-bit, 0.126 vs. 0.119 at 3-bit). Table 25 gives the declared estimands on truncation, repetitive loops (Appendix E) and correctness; the reading rule labels each effect positive, negative, negligible (inside ± 0.05) or unresolved and names a mechanism only when one decomposition's prefix effect is negligible and its continuation effect resolved, with the mirror decomposition required to agree—it did at both rungs.

Table 25: Matched-prefix continuation on the 12B: document-mean differences (95% problem-clustered bootstrap) on truncation (the primary, with its declared status), loop and correctness, for the prefix effect under each continuation model, the continuation effect on each prefix, and their interaction (r = rung, b = BF16; first letter the prefix, second the continuing model).

Rung effect	truncation	status	loop	correctness
IQ2_M	-0.003		-0.001	-0.006
prefix effect, BF16 continuation ($rb - bb$)	$[-0.013, +0.006]$	negligible	$[-0.005, +0.003]$	$[-0.017, +0.005]$
IQ2_M	$+0.493$		$+0.296$	-0.446
continuation effect, departed prefix ($rr - rb$)	$[+0.453, +0.537]$	positive	$[+0.261, +0.336]$	$[-0.488, -0.404]$
IQ2_M	-0.006		-0.003	-0.001
prefix effect, rung continuation ($rr - br$)	$[-0.026, +0.016]$	negligible	$[-0.024, +0.019]$	$[-0.021, +0.020]$
IQ2_M	$+0.496$		$+0.298$	-0.451
continuation effect, BF16 prefix ($br - bb$)	$[+0.455, +0.540]$	positive	$[+0.264, +0.336]$	$[-0.494, -0.411]$
IQ2_M	-0.003		-0.003	$+0.006$
interaction	$[-0.025, +0.020]$	negligible	$[-0.024, +0.020]$	$[-0.018, +0.029]$
Q3_K_M	$+0.008$		$+0.001$	-0.005
prefix effect, BF16 continuation ($rb - bb$)	$[+0.000, +0.017]$	negligible	$[-0.003, +0.004]$	$[-0.013, +0.004]$
Q3_K_M	$+0.050$		$+0.008$	-0.052
continuation effect, departed prefix ($rr - rb$)	$[+0.037, +0.066]$	positive	$[+0.003, +0.013]$	$[-0.068, -0.038]$
Q3_K_M	$+0.003$		$+0.002$	-0.004
prefix effect, rung continuation ($rr - br$)	$[-0.006, +0.013]$	negligible	$[-0.003, +0.006]$	$[-0.015, +0.007]$
Q3_K_M	$+0.055$		$+0.007$	-0.053
continuation effect, BF16 prefix ($br - bb$)	$[+0.042, +0.070]$	positive	$[+0.003, +0.012]$	$[-0.068, -0.039]$
Q3_K_M	-0.005		$+0.001$	$+0.001$
interaction	$[-0.016, +0.006]$	negligible	$[-0.005, +0.006]$	$[-0.013, +0.013]$

N THE RELATION ON HELD-OUT MODELS

Procedure. Eligibility criteria were pre-specified. Bring-up checked architecture support at the pinned llama.cpp build, an unpruned snapshot, the chat template, a clean BF16 Yes/No interface and truncation at the cap (E1–E5). Llama 3.1 8B was gated for the token and Phi-4-mini took its ordered fallback slot; OLMo 3 runs at a 4,096 cap by the panel's rule. The counted BF16 and `Q8_0` rows and the BF16 column on the `Q8_0` set then supplied E6: completed-candidate AUROC ≥ 0.60 with at least 30 completed errors. After E6, each model proceeded through generation-only teacher forcing at the four rungs; `predictions_A.json` committed (the loss per rung under the all-noise reading from $\hat{\rho}_{\text{shift}} = c\sqrt{\text{KL}}$ with the fitted models' leave-one-model-out band; B0, B1 and B2 in the same file; the input boundary admits no quantized verdict of the target); verdict-document teacher forcing; `predictions_B.json` committed (from the measured margin regression, a calibrated transfer to samples 1–7 of the calibration problems); then the quantized rows and the 5×5 columns. Every file's hash was registered externally before the measurement it predicts. The primary aggregate is the model-balanced MAE of ΔV on all rows over the 4- and 3-bit cells, per rung and pooled, every declared cell kept; incremental value is the paired problem-bootstrap 90% CI of MAE(A) $-$ MAE(B$_j$) with problems paired across models (Table 27); the 2-bit endpoint is reported by the point rule $\Delta V_{\text{completed}} \leq -0.05$ (Table 28; the declared classifier withdrew itself in development, its leave-one-model-out accuracy below the majority baseline). The constant is tested per rung against the median's leave-one-model-out band (Table E10); the generation half is a heuristic (order and band of the pooled truncation law, Table E11); consistency asks that the perturbation measured from the verdict documents and from the column regression agree within $\times 1.3$, which it does at every $\geq$3-bit rung. Table 29 is each model's panel row.

Table 26: **Held-out prediction of the verifier-column loss** on three models never fitted: model-balanced mean absolute error of ΔV (all rows, vs. the model's BF16 column) per rung, with the largest per-model error in parentheses, for forecast A from generation-only teacher forcing and one labeled full-precision evaluation, before any quantized target verdict, the measured-perturbation prediction B, and the baselines B0 (no loss), B1 (rung-only development mean) and B2 (A's pipeline with the fitted models' median perturbation). B is evaluated on samples 1–7 of the calibration problems, with sample 0 reserved for calibration; A and the baselines use all rows. Utility line 0.0125; the Q8 and 2-bit rows are reported, not scored.

Rung	A	B	B0	B1	B2
Q8_0 (reported)	0.0011 (0.0024)	0.0011 (0.0022)	0.0009 (0.0015)	0.0010 (0.0017)	0.0011 (0.0025)
Q4_K_M	0.0069 (0.0100)	0.0060 (0.0084)	0.0120 (0.0196)	0.0114 (0.0178)	0.0077 (0.0104)
Q3_K_M	0.0072 (0.0144)	0.0062 (0.0111)	0.0198 (0.0346)	0.0137 (0.0198)	0.0089 (0.0155)
IQ2_M (reported)	0.0426 (0.0571)	0.0250 (0.0303)	0.0839 (0.1494)	0.0648 (0.0921)	0.0571 (0.0816)
pooled Q4 + Q3	0.0070	0.0061	0.0159	0.0126	0.0083

The full per-model forecasts and errors are in Evidence Table E9. The fitted constant fails its registered transfer criterion on all three held-out models. Across the six 4-/3-bit cells, the empirical coverage of A's leave-one-model-out bands is 0.67, and of B's nested bands 0.17. Thus small mean prediction errors do not establish calibrated uncertainty or incremental value over B2. The per-rung constant/consistency and generation-heuristic diagnostics are in Evidence Tables E10 and E11.

Table 27: Incremental value: paired problem-bootstrap 90% CIs of the MAE differences (a negative interval favors the first method; in $\mathrm{A}_{\mathrm{transfer}} - \mathrm{B}$, it favors $\mathrm{A}_{\mathrm{transfer}}$).

Rung	comparison	point	90%	superior
Q4_K_M	$\mathrm{MAE}(\mathrm{A}_{\mathrm{transfer}}) - \mathrm{MAE}(\mathrm{B})$	$+0.0018$	$[+0.0009, +0.0018]$	no
Q4_K_M	MAE(A) − MAE(B0)	-0.0051	$[-0.0074, -0.0022]$	yes
Q4_K_M	MAE(A) − MAE(B1)	-0.0045	$[-0.0068, -0.0016]$	yes
Q4_K_M	MAE(A) − MAE(B2)	-0.0008	$[-0.0030, +0.0008]$	no
Q3_K_M	$\mathrm{MAE}(\mathrm{A}_{\mathrm{transfer}}) - \mathrm{MAE}(\mathrm{B})$	$+0.0004$	$[-0.0001, +0.0022]$	no
Q3_K_M	MAE(A) − MAE(B0)	-0.0127	$[-0.0189, -0.0039]$	yes
Q3_K_M	MAE(A) − MAE(B1)	-0.0065	$[-0.0113, +0.0023]$	no
Q3_K_M	MAE(A) − MAE(B2)	-0.0017	$[-0.0047, +0.0040]$	no
pooled	$\mathrm{MAE}(\mathrm{A}_{\mathrm{transfer}}) - \mathrm{MAE}(\mathrm{B})$	$+0.0011$	$[+0.0006, +0.0020]$	no
pooled	MAE(A) − MAE(B0)	-0.0089	$[-0.0122, -0.0043]$	yes
pooled	MAE(A) − MAE(B1)	-0.0055	$[-0.0080, -0.0008]$	yes
pooled	MAE(A) − MAE(B2)	-0.0013	$[-0.0035, +0.0019]$	no

Table 28: The 2-bit endpoint by the point rule (completed-candidate ΔV at `IQ2_M`, 95% CI, its position vs. -0.05, the label; the withdrawn classifier's call and B's reference call).

Model	ΔV compl.	95%	vs. -0.05	label	classifier	B
OLMo 3 7B	$+0.013$	$[-0.017, +0.043]$	above	dent	withdrawn	dent
Granite 4.0-H-Micro	-0.149	$[-0.176, -0.122]$	below	collapse	withdrawn	collapse
Phi-4-mini	-0.101	$[-0.120, -0.083]$	below	collapse	withdrawn	collapse

Table 29: Held-out models as panel rows: G, truncation share, verifier columns on the `Q8_0` set (all rows; Level 5), the diagonal, and the Level-5 TOST contrast vs. BF16 (90%; † equivalence not established).

Model	rung	G	trunc.	V all	V L5	V diag.	TOST L5
OLMo 3 7B	BF16	0.876	0.101	0.873	0.843	0.881	–
OLMo 3 7B	Q8_0	0.879	0.105	0.873	0.843	0.873	–
OLMo 3 7B	Q4_K_M	0.872	0.105	0.878	0.846	0.879	$+0.004\,[-0.003, +0.011]$
OLMo 3 7B	Q3_K_M	0.872	0.110	0.871	0.822	0.887	$-0.020\,[-0.034, -0.007]^{\dagger}$
OLMo 3 7B	IQ2_M	0.838	0.141	0.876	0.838	0.875	–
Granite 4.0-H-Micro	BF16	0.622	0.003	0.835	0.738	0.840	–
Granite 4.0-H-Micro	Q8_0	0.621	0.003	0.834	0.738	0.834	–
Granite 4.0-H-Micro	Q4_K_M	0.604	0.003	0.815	0.721	0.819	$-0.016\,[-0.036, +0.003]^{\dagger}$
Granite 4.0-H-Micro	Q3_K_M	0.556	0.001	0.800	0.708	0.796	$-0.029\,[-0.054, -0.006]^{\dagger}$
Granite 4.0-H-Micro	IQ2_M	0.341	0.001	0.685	0.572	0.737	–
Phi-4-mini	BF16	0.653	0.013	0.847	0.811	0.843	–
Phi-4-mini	Q8_0	0.652	0.011	0.848	0.812	0.848	–
Phi-4-mini	Q4_K_M	0.579	0.026	0.835	0.794	0.821	$-0.017\,[-0.024, -0.009]$
Phi-4-mini	Q3_K_M	0.600	0.014	0.824	0.783	0.822	$-0.028\,[-0.041, -0.014]^{\dagger}$
Phi-4-mini	IQ2_M	0.338	0.239	0.748	0.678	0.806	–

O SELF-TRAINING: CONTROLS, SOURCE EFFECTS AND THE PRECISION LADDER

Reading guide. The source/correctness factorial (§O.1), matched source and fingerprint controls (§O.2) and precision ladder (§O.3) come first. The initial loops (§O.4–O.5), verifier isolation (§O.6–O.7) and selectors (§O.8) follow as controls. All single-update blocks reuse stored pools and start from the base checkpoint; the adaptive loops continue their adapter across rounds. Protocol identifiers name the original frozen declarations; the reordered presentation does not change their timing.

Recipe (frozen under `v1.4`). LoRA rank 16, $\alpha = 32$, dropout 0.05 on q/k/v/o and gate/up/down (Hu et al., 2022); AdamW, learning rate 10^{-4}, cosine schedule with 3% warm-up, two epochs per round over the round's accepted samples, micro-batch 4 $\times$ accumulation 4, max length 2,048, gradient checkpointing; the adapter continues across rounds and each round trains on that round's accepted samples only; the NF4 arm is the bitsandbytes NF4 base with LoRA (QLoRA), generating and self-judging on the dequantized path; the BF16 control is the same recipe on the BF16 base. The single-seed smoke round met the pre-declared rule (loss falling, no holdout drop, round < 3 h), so no field changed.

O.1 CORRECTNESS AND SOURCE FACTORIAL (LOOP 7; V1.10)

A fixed-pool single-update factorial on both subjects from the selection cells' stored round-1 pools (the base model generated them; identical across cells within a seed) and each learner's stored round-0 evaluation: one round of the registered chain (the learner's own verdict read, the gate, one SFT update from the base, the holdouts), nothing regenerated. Factors: the pool's generator (NF4 for every seed; BF16 where a stored BF16 pool exists—twelve seeds on the 3B, five on the 12B), the learner (QLoRA on the NF4 base or LoRA on the BF16 base), and the policy: a label-only mixture gate that per problem takes $k_c = a + \lfloor \lambda(o - a) + u \rfloor$ correct completed samples (a the anti-oracle's count, o the oracle's, u a seeded rounding draw) and fills the rest with incorrect completed ones, so expected precision varies linearly with λ at the registered count and coverage; two composition controls at the consensus and verdict-rank selectors' realized precision (mS, mK); and the selectors themselves (U, S, K). Cap 4 on the 3B with 8 holdout samples per problem (a procedure simulation on the fitted seed/problem components put the slope's crossed half-width at 0.021 per unit with 8 samples against 0.029 with 4), cap 2 on the 12B (its pool holds 3.9 completed samples per problem at 82% correct, so the reachable precision range is 0.14 wide and no 12B slope was predicted). The estimand is the within-seed OLS slope of the one-update gain on realized precision over the λ cells, refit per crossed draw (5,000); policy cells sit beside as residuals from the fitted line. Tables 30–33.

Replication: the fresh U, S and K cells reproduce the historical round-1 cells on identical rows within $[-0.001, +0.015]$ (U), $[-0.010, +0.010]$ (S) and $[+0.001, +0.011]$ (K), the pipeline's reproducibility scale.

Post hoc and descriptive (seed-level 90% t-intervals): on the 3B's NF4 learner, $\lambda = 1$ minus U is -0.002 $[-0.007, +0.004]$ and $\lambda = 0$ minus U -0.012 $[-0.017, -0.006]$ on the pool-like holdout ($+0.005$ $[+0.002, +0.009]$ / -0.003 $[-0.006, +0.000]$ on MATH-500); on the BF16 learner $+0.006$ $[+0.001, +0.012]$ / -0.012 $[-0.019, -0.004]$; the two learners start from different levels (G_0 0.567 vs. 0.495 on the 3B, 0.581 vs. 0.269 on the 12B), so the identical-row learner contrasts of Table 30 are within-source learner differences on different baselines, not an erosion of what a correct sample is worth and not the factorial's source $\times$ learner interaction (which is the difference of the two learners' source effects); on the 12B the BF16-generated pool's own slope is positive on five seeds (Table 33), a descriptive reading only. The 12B's BF16-generated pool at cap 2 covers 426 problems against the NF4-generated pool's 276, so its `v1.10` source contrast differs in count and coverage as well as source; Loop 8 reduces these differences to approximately 1%.

Table 30: Loop 7 cells on SmolLM3-3B: pool $\rightarrow$ learner, policy, mixture λ or matched selector, complete seeds, realized precision (seed mean), accepted rows, the one-update gain on the pool-like holdout and on MATH-500, and the change in pool-like truncation (seed means).

Pool $\rightarrow$ learner	policy	λ / match	seeds	precision	rows	gain (pool-like)	gain (MATH-500)	Δtrunc.
NF4 $\rightarrow$ NF4	l0	λ 0.00	12	0.308	3,996	+0.007	+0.011	+0.000
NF4 $\rightarrow$ NF4	l25	λ 0.25	12	0.417	3,996	+0.013	+0.015	+0.002
NF4 $\rightarrow$ NF4	l50	λ 0.50	12	0.525	3,996	+0.013	+0.014	+0.003
NF4 $\rightarrow$ NF4	l75	λ 0.75	12	0.631	3,996	+0.019	+0.017	+0.003
NF4 $\rightarrow$ NF4	l100	λ 1.00	12	0.739	3,996	+0.017	+0.020	+0.006
NF4 $\rightarrow$ NF4	mS	consensus	12	0.678	3,996	+0.020	+0.018	+0.005
NF4 $\rightarrow$ NF4	mK	rank	12	0.584	3,996	+0.017	+0.017	+0.003
NF4 $\rightarrow$ NF4	U	all_completed	12	0.523	3,996	+0.019	+0.014	+0.002
NF4 $\rightarrow$ NF4	S	completed_consensus	12	0.678	3,996	+0.021	+0.019	+0.004
NF4 $\rightarrow$ NF4	K	completed_rank	12	0.584	3,996	+0.017	+0.019	+0.004
NF4 $\rightarrow$ BF16	l0	λ 0.00	12	0.308	3,996	−0.030	−0.008	+0.009
NF4 $\rightarrow$ BF16	l25	λ 0.25	12	0.417	3,996	−0.020	−0.007	+0.010
NF4 $\rightarrow$ BF16	l50	λ 0.50	12	0.525	3,996	−0.021	−0.004	+0.009
NF4 $\rightarrow$ BF16	l75	λ 0.75	12	0.631	3,996	−0.016	−0.002	+0.011
NF4 $\rightarrow$ BF16	l100	λ 1.00	12	0.739	3,996	−0.012	−0.002	+0.012
NF4 $\rightarrow$ BF16	U	all_completed	12	0.523	3,996	−0.019	−0.004	+0.008
NF4 $\rightarrow$ BF16	S	completed_consensus	12	0.678	3,996	−0.011	−0.002	+0.010
NF4 $\rightarrow$ BF16	mS	consensus	12	0.678	3,996	−0.015	−0.001	+0.010
BF16 $\rightarrow$ NF4	l0	λ 0.00	12	0.381	3,996	+0.037	+0.022	−0.004
BF16 $\rightarrow$ BF16	l0	λ 0.00	12	0.381	3,996	+0.024	+0.005	+0.000
BF16 $\rightarrow$ NF4	l50	λ 0.50	12	0.595	3,996	+0.042	+0.025	−0.006
BF16 $\rightarrow$ BF16	l50	λ 0.50	12	0.595	3,996	+0.025	+0.008	+0.001
BF16 $\rightarrow$ NF4	l100	λ 1.00	12	0.809	3,996	+0.046	+0.029	−0.004
BF16 $\rightarrow$ BF16	l100	λ 1.00	12	0.809	3,996	+0.033	+0.012	+0.002

Table 31: Loop 7 slopes on SmolLM3-3B: per pool × learner and holdout, the number of mixture cells and seeds, the within-seed slope of the one-update gain on realized precision with its crossed 90% interval, the crossed 95% interval, the seed-level 90% t-interval, and the pre-specified call (P-1 applies to the NF4 pool → NF4 learner group on the pool-like holdout).

Pool → learner (holdout)	cells / seeds	slope [90%]	95% seed t 90%	call
NF4 pool → NF4 learner (train-mixed)	5 / 12	$+0.024$ $[+0.002, +0.044]$	$[-0.002, +0.048]$ $[+0.011, +0.036]$	established: resolved positive, point inside the predicted band
NF4 pool → BF16 learner (train-mixed)	5 / 12	$+0.037$ $[+0.015, +0.059]$	$[+0.011, +0.064]$ $[+0.024, +0.051]$	–
BF16 pool → NF4 learner (train-mixed)	3 / 12	$+0.021$ $[-0.004, +0.046]$	$[-0.009, +0.051]$ $[+0.002, +0.040]$	–
BF16 pool → BF16 learner (train-mixed)	3 / 12	$+0.020$ $[-0.005, +0.046]$	$[-0.010, +0.051]$ $[+0.001, +0.039]$	–
NF4 pool → NF4 learner (MATH-500)	5 / 12	$+0.018$ $[+0.008, +0.028]$	$[+0.006, +0.031]$ $[+0.013, +0.023]$	gradient present: resolved positive, point outside the predicted band
NF4 pool → BF16 learner (MATH-500)	5 / 12	$+0.016$ $[+0.005, +0.027]$	$[+0.003, +0.030]$ $[+0.009, +0.023]$	–
BF16 pool → NF4 learner (MATH-500)	3 / 12	$+0.017$ $[+0.005, +0.029]$	$[+0.003, +0.031]$ $[+0.009, +0.025]$	–
BF16 pool → BF16 learner (MATH-500)	3 / 12	$+0.016$ $[+0.004, +0.027]$	$[+0.002, +0.029]$ $[+0.009, +0.022]$	–

Table 32: Loop 7 cells on Gemma 4 12B (cap 2), as Table 30.

Pool → learner	policy	λ / match	seeds	precision	rows	gain (pool-like)	gain (MATH-500)	Δtrunc.
NF4 → NF4	l0	λ 0.00	12	0.761	486	$+0.035$	$+0.012$	-0.050
NF4 → BF16	l0	λ 0.00	12	0.761	486	-0.214	-0.046	$+0.239$
NF4 → NF4	l50	λ 0.50	12	0.830	486	$+0.037$	$+0.013$	-0.049
NF4 → BF16	l50	λ 0.50	12	0.830	486	-0.208	-0.045	$+0.238$
NF4 → NF4	l100	λ 1.00	12	0.899	486	$+0.034$	$+0.013$	-0.041
NF4 → BF16	l100	λ 1.00	12	0.899	486	-0.205	-0.046	$+0.239$
NF4 → NF4	U	all_completed	12	0.833	486	$+0.032$	$+0.013$	-0.039
NF4 → BF16	U	all_completed	12	0.833	486	-0.209	-0.044	$+0.236$
BF16 → NF4	l0	λ 0.00	5	0.784	828	$+0.225$	$+0.062$	-0.314
BF16 → BF16	l0	λ 0.00	5	0.784	828	$+0.014$	$+0.002$	-0.028
BF16 → NF4	l50	λ 0.50	5	0.872	828	$+0.239$	$+0.065$	-0.325
BF16 → BF16	l50	λ 0.50	5	0.872	828	$+0.007$	$+0.002$	-0.019
BF16 → NF4	l100	λ 1.00	5	0.958	828	$+0.253$	$+0.066$	-0.327
BF16 → BF16	l100	λ 1.00	5	0.958	828	$+0.017$	$+0.004$	-0.008

Table 33: Loop 7 slopes on Gemma 4 12B (descriptive; no slope was predicted on its 0.14-wide reachable range), as Table 31.

Pool → learner (holdout)	cells / seeds	slope [90%]	95% seed t 90%	call
NF4 pool → NF4 learner (train-mixed)	3 / 12	-0.008 $[-0.106, +0.094]$	$[-0.124, +0.114]$ $[-0.092, +0.075]$	–
NF4 pool → BF16 learner (train-mixed)	3 / 12	$+0.059$ $[-0.036, +0.155]$	$[-0.053, +0.174]$ $[+0.001, +0.116]$	–
BF16 pool → NF4 learner (train-mixed)	3 / 5	$+0.162$ $[+0.039, +0.282]$	$[+0.013, +0.305]$ $[+0.079, +0.245]$	–
BF16 pool → BF16 learner (train-mixed)	3 / 5	$+0.014$ $[-0.096, +0.123]$	$[-0.120, +0.146]$ $[-0.044, +0.072]$	–
NF4 pool → NF4 learner (MATH-500)	3 / 12	$+0.009$ $[-0.031, +0.050]$	$[-0.039, +0.058]$ $[-0.019, +0.037]$	–
NF4 pool → BF16 learner (MATH-500)	3 / 12	-0.001 $[-0.034, +0.032]$	$[-0.041, +0.039]$ $[-0.022, +0.019]$	–
BF16 pool → NF4 learner (MATH-500)	3 / 5	$+0.017$ $[-0.019, +0.054]$	$[-0.027, +0.060]$ $[-0.005, +0.040]$	–
BF16 pool → BF16 learner (MATH-500)	3 / 5	$+0.012$ $[-0.023, +0.047]$	$[-0.033, +0.056]$ $[-0.016, +0.040]$	–

O.2 MATCHED SOURCE AND FINGERPRINT CONTROLS (LOOP 8; V1.11)

The row-list gate accepts exactly a listed set of completed rows (the list's hash enters the chain identity). E1 on the 12B (five seeds): the BF16-generated pool's completed rows restricted to the problems the NF4-generated pool completes and drawn per problem to the NF4 cells' counts at U's (mU) and at $\lambda = 1$'s (mO) composition, to both learners; against the NF4 cells of Loop 7, mU $-$ U is $+0.189$ $[+0.165, +0.212]$ (NF4 learner; Holm 0.001, 5/5 seeds) and $+0.223$ $[+0.196, +0.250]$ (BF16 learner), mO $-$ λ=1 $+0.193$ $[+0.167, +0.218]$ / $+0.226$ $[+0.200, +0.252]$ (the pre-written rule: the crossed 90% lower bound above $+0.05$); the unrestricted BF16 $\lambda = 1$ cell minus mO—the count-and-coverage component of Loop 7's 12B contrast—is $+0.028$ $[+0.006, +0.048]$ / -0.009 $[-0.029, +0.010]$; MATH-500 $+0.046$ $[+0.035, +0.056]$ / $+0.049$ $[+0.038, +0.061]$.

What the lists actually match, on the five common seeds: the NF4 U cell holds 483 rows on 275 problems at precision 0.830 against mU's 477 / 272 at 0.886—count and coverage to within 2–3 problems per seed without a completed BF16 row, correctness not (5.6 points); mO matches the correctness classes per problem where the BF16 pool holds them and fills 14–29 rows per seed (5% of the list) from the other class where it does not, as the protocol allows and the committed row-list files record. By the BF16-generated pool's own correctness slope on the 12B ($+0.162$ $[+0.039, +0.282]$ per unit, five seeds, descriptive) the mU correctness gap is worth $\approx +0.009$, a linear sensitivity.

E2 on both subjects (twelve seeds): every completed pool row scored under both models on the HF path, f = mean per-token $[\log p_{\text{BF16}} - \log p_{\text{NF4}}]$ on its own tokens (3B NF4-generated rows -0.0147 vs. BF16-generated $+0.0188$; 12B -0.0562 vs. $+0.0618$; $|\text{corr}(f, \log \text{length})| = +0.192$, so f itself was the pre-declared selection key); within the NF4 pool at the $\lambda = 1$ counts, the highest-f (fpLow, least quantization-typical) against the lowest-f (fpHigh, most quantization-typical) rows per problem, to both learners, with the mirror on the BF16 pool. fpLow $-$ fpHigh: 3B BF16 learner $+0.009$ $[+0.001, +0.018]$ (Holm 0.235; 11/12 seeds; the cells -0.003 $[-0.009, +0.003]$ / -0.012 $[-0.018, -0.007]$ against the $\lambda = 1$ reference -0.012 $[-0.019, -0.006]$), 3B NF4 learner $+0.001$ $[-0.009, +0.011]$, 12B $+0.007$ $[-0.004, +0.018]$ / $+0.008$ $[-0.008, +0.023]$ (equivalent within ± 0.05), the mirror -0.004 $[-0.014, +0.005]$ / $+0.001$ $[-0.008, +0.009]$; MATH-500 all equivalent. The selection moved the mean fingerprint by -0.0109 vs. -0.0186 on the 3B (about 23% of the between-source gap) and -0.0490 vs. -0.0639 on the 12B (an eighth), the dose the pool affords; the 3B response is proportional to it, the 12B's below proportional. Tables 34–35.

Table 34: Loop 8 cells on SmolLM3-3B: pool → learner, policy, complete seeds, realized precision, accepted rows and problems, the selected rows' mean fingerprint f and mean length, the one-update gain on the pool-like holdout and on MATH-500, and the change in pool-like truncation (seed means); the Loop 7 reference cells beside.

Pool → learner	policy	seeds	precision	rows	problems	f	tokens	gain (pool-like)	gain (MATH-500)	Δtrunc.
NF4 → NF4	fpLow	12	0.739	3,996	1,000	-0.0109	651	$+0.023$	$+0.018$	$+0.006$
NF4 → NF4	fpHigh	12	0.739	3,996	1,000	-0.0186	639	$+0.021$	$+0.016$	$+0.004$
BF16 → NF4	fpLow	12	0.809	3,996	1,000	$+0.0222$	655	$+0.043$	$+0.026$	-0.004
BF16 → NF4	fpHigh	12	0.809	3,996	1,000	$+0.0154$	664	$+0.047$	$+0.030$	-0.002
NF4 → BF16	fpLow	12	0.739	3,996	1,000	-0.0109	651	-0.003	$+0.002$	$+0.010$
NF4 → BF16	fpHigh	12	0.739	3,996	1,000	-0.0186	639	-0.012	-0.001	$+0.010$
BF16 → BF16	fpLow	12	0.809	3,996	1,000	$+0.0222$	655	$+0.032$	$+0.008$	$+0.001$
BF16 → BF16	fpHigh	12	0.809	3,996	1,000	$+0.0154$	664	$+0.031$	$+0.013$	$+0.003$
NF4 → NF4	l100 (e037 ref.)	12	0.739	3,996	1,000	–	–	$+0.017$	$+0.020$	$+0.006$
BF16 → NF4	l100 (e037 ref.)	12	0.809	3,996	1,000	–	–	$+0.046$	$+0.029$	-0.004
NF4 → BF16	l100 (e037 ref.)	12	0.739	3,996	1,000	–	–	-0.012	-0.002	$+0.012$
BF16 → BF16	l100 (e037 ref.)	12	0.809	3,996	1,000	–	–	$+0.033$	$+0.012$	$+0.002$

Table 35: Loop 8 cells on Gemma 4 12B (cap 2), as Table 34; mU / mO are the count- and coverage-matched BF16-generated row lists (mO class-matched, with recorded fills). Reference rows are twelve-seed (U, $\lambda = 1$) or five-seed means; the E1 contrasts pair the five common seeds, on which the U cell holds 483 rows / 275 problems at precision 0.830.

Pool → learner	policy	seeds	precision	rows	problems	f	tokens	gain (pool-like)	gain (MATH-500)	Δtrunc.
NF4 → NF4	fpLow	12	0.899	486	276	−0.0490	1,124	+0.042	+0.016	−0.054
NF4 → NF4	fpHigh	12	0.899	486	276	−0.0639	1,163	+0.036	+0.013	−0.042
BF16 → NF4	mU	5	0.886	477	272	+0.0587	1,041	+0.214	+0.060	−0.290
BF16 → NF4	mO	5	0.927	477	272	+0.0584	1,046	+0.225	+0.060	−0.296
NF4 → BF16	fpLow	12	0.899	486	276	−0.0490	1,124	−0.201	−0.041	+0.233
NF4 → BF16	fpHigh	12	0.899	486	276	−0.0639	1,163	−0.208	−0.045	+0.242
BF16 → BF16	mU	5	0.886	477	272	+0.0587	1,041	+0.006	+0.005	−0.011
BF16 → BF16	mO	5	0.927	477	272	+0.0584	1,046	+0.026	+0.005	−0.017
NF4 → NF4	l100 (e037 ref.)	12	0.899	486	276	–	–	+0.034	+0.013	−0.041
NF4 → NF4	U (e037 ref.)	12	0.833	486	276	–	–	+0.032	+0.013	−0.039
BF16 → NF4	l100 (e037 ref.)	5	0.958	828	426	–	–	+0.253	+0.066	−0.327
NF4 → BF16	l100 (e037 ref.)	12	0.899	486	276	–	–	−0.205	−0.046	+0.239
NF4 → BF16	U (e037 ref.)	12	0.833	486	276	–	–	−0.209	−0.044	+0.236
BF16 → BF16	l100 (e037 ref.)	5	0.958	828	426	–	–	+0.017	+0.004	−0.008

O.3 Source-precision ladder (Loop 9; v1.12)

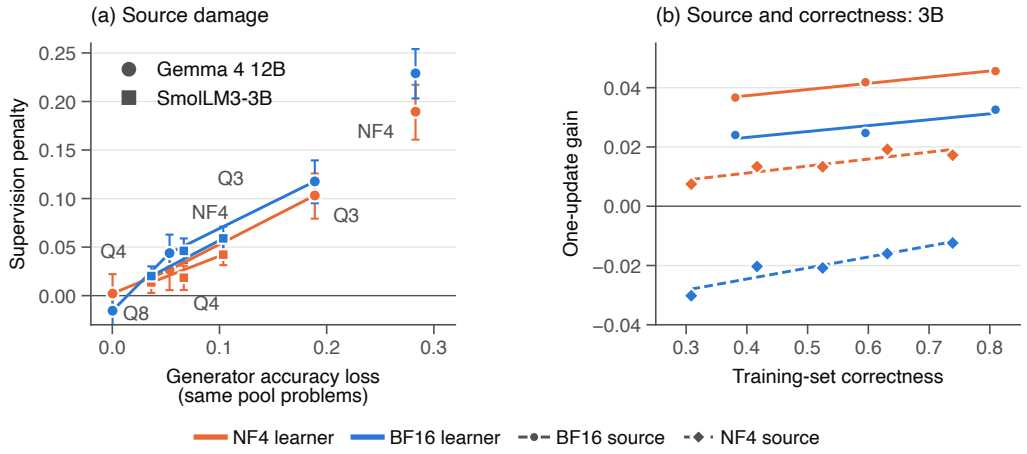

Figure 7: **Descriptive source relations.** Left: supervision penalty against the generator's accuracy loss on the same pool problems (crossed 90% intervals; five 12B and twelve 3B seeds). Right: SmolLM3-3B learning gain against selected-set correctness by source and learner (twelve seeds; least-squares lines). Both panels use the partial-solvability holdout; gain is measured after one two-epoch training round. Figure 1 shows the reversal contrast and absolute matched-source gains; these curves describe the broader cell pattern.

Training pools were generated by the GGUF rungs on the registered llama.cpp path with the chain's exact prompts (verified identical across the two container images), the registered sampler and $K = 8$, over the problems each seed's NF4 U cell completed (12B) or all 1,000 pool problems (3B), graded on the registered path and converted to the chain's pool format; row lists count-matched to the NF4 U cell per problem, to both learners; a GGUF BF16 arm as the path control. PATH: gguf-BF16 mU − hf-BF16 mU +0.000 [−0.021, +0.022] / +0.006 [−0.011, +0.022] (12B; equivalent), so every rung reads against the GGUF BF16 control and Loop 8's E1 reproduces on it (+0.190 [+0.161, +0.217] vs. +0.189 [+0.165, +0.212]); on the 3B the GGUF control against the NF4 U cell (+0.018 [+0.006, +0.031] / +0.046 [+0.034, +0.059]) reproduces Loop 7's HF-path source effect.

LADDER (Table 2; Holm within the family of rungs per learner): the 12B's `Q8_0` penalty is equivalent to zero, `Q4_K_M`'s +0.027 [+0.006, +0.048] (Holm 0.084; equivalent within ±0.05 with the interval above zero) / +0.044 [+0.025, +0.063] (different), `Q3_K_M`'s +0.103 [+0.079, +0.126]

/ $+0.118$ $[+0.095, +0.139]$ (different); adjacent steps $+0.024$ $[+0.003, +0.046]$ (Q4 − Q8) and $+0.077$ $[+0.053, +0.099]$ (Q3 − Q4). The 3B: `Q4_K_M` $+0.014$ $[+0.003, +0.024]$ / $+0.020$ $[+0.010, +0.030]$, `Q3_K_M` $+0.042$ $[+0.031, +0.053]$ / $+0.059$ $[+0.047, +0.071]$ (all resolved); MATH-500 `Q3_K_M` $+0.018$ $[+0.009, +0.026]$ (12B) / $+0.020$ $[+0.013, +0.027]$ (3B), the 4- and 8-bit rungs within ± 0.01. NF4: against the NF4 generator's own count-matched samples the rungs' samples are worth $+0.187$ $[+0.163, +0.211]$ (`Q8_0`), $+0.163$ $[+0.134, +0.190]$ (`Q4_K_M`) and $+0.086$ $[+0.065, +0.107]$ (`Q3_K_M`) more to the 12B's quantized learner, and $+0.005$ $[-0.007, +0.017]$ / -0.024 $[-0.036, -0.011]$ on the 3B—on the 12B every K-quant rung beats NF4's own samples, on the 3B NF4 sits between the 4- and the 3-bit K-quant.

Correctness was not matched on the rungs (the amendment declares no class-matched lists): the selected precision runs $0.878 \to 0.844$ on the 12B and $0.587 \to 0.486$ on the 3B against the U cell's 0.523; by each subject's own slope those gaps are worth $\approx +0.006$ (12B, the five-seed BF16-pool slope) and $\approx +0.002$ (3B) of the 3-bit penalty, and the 12B 3-bit lists' shortfall (443 rows / 255 problems against 476 / 273, 7%) scales the measured coverage component to $\approx +0.002$—linear sensitivities, not bounds. For Fig. 7 every generator's accuracy loss is measured on the drawn problems: the NF4 reference's accuracy there is 0.426 (truncation 0.521), a loss of $+0.283$ against the GGUF BF16 pool on the same problems, not its full-draw loss of $+0.435$. Tables 36–37.

Table 36: Loop 9 cells on Gemma 4 12B (cap 2): generator → learner, seeds, pool truncation and accuracy, selected-set composition, and one-update gains and truncation changes (seed means). Loop 7/8 references are shown. NF4 U pool statistics use the drawn problems, matching Fig. 7; no matched pool statistics are available for the HF BF16 references.

Generator → learner	seeds	pool trunc.	pool G	precision	rows	problems	tokens	gain (pool-like)	gain (MATH-500)	Δtrunc.
BF16 → NF4	5	0.211	0.709	0.878	476	273	1,034	+0.215	+0.056	−0.294
Q8_0 → NF4	5	0.217	0.709	0.877	476	272	1,037	+0.212	+0.058	−0.289
Q4_K_M → NF4	5	0.251	0.656	0.860	474	271	1,048	+0.188	+0.057	−0.277
Q3_K_M → NF4	5	0.410	0.520	0.844	443	255	1,077	+0.112	+0.039	−0.158
BF16 → BF16	5	0.211	0.709	0.878	476	273	1,034	+0.012	+0.003	−0.014
Q8_0 → BF16	5	0.217	0.709	0.877	476	272	1,037	+0.027	+0.004	−0.028
Q4_K_M → BF16	5	0.251	0.656	0.860	474	271	1,048	−0.032	−0.001	+0.017
Q3_K_M → BF16	5	0.410	0.520	0.844	443	255	1,077	−0.106	−0.024	+0.120
nf4U → NF4 (ref.)	5	0.521	0.426	0.830	483	275	–	+0.025	+0.014	−0.037
hfbf16mU → NF4 (ref.)	5	–	–	0.886	477	272	–	+0.214	+0.060	−0.290
nf4U → BF16 (ref.)	5	0.521	0.426	0.830	483	275	–	−0.217	−0.044	+0.247
hfbf16mU → BF16 (ref.)	5	–	–	0.886	477	272	–	+0.006	+0.005	−0.011

Table 37: Loop 9 cells on SmolLM3-3B, as Table 36.

Generator → learner	seeds	pool trunc.	pool G	precision	rows	problems	tokens	gain (pool-like)	gain (MATH-500)	Δtrunc.
BF16 → NF4	12	0.024	0.580	0.587	3,994	1,000	670	+0.038	+0.023	−0.006
Q4_K_M → NF4	12	0.030	0.543	0.554	3,992	1,000	673	+0.024	+0.022	+0.001
Q3_K_M → NF4	12	0.033	0.476	0.486	3,993	1,000	669	−0.005	+0.003	+0.002
BF16 → BF16	12	0.024	0.580	0.587	3,994	1,000	670	+0.028	+0.010	−0.001
Q4_K_M → BF16	12	0.030	0.543	0.554	3,992	1,000	673	+0.007	+0.004	+0.004
Q3_K_M → BF16	12	0.033	0.476	0.486	3,993	1,000	669	−0.031	−0.015	+0.007
nf4U → NF4 (ref.)	12	–	–	0.523	3,996	1,000	–	+0.019	+0.014	+0.002
nf4U → BF16 (ref.)	12	–	–	0.523	3,996	1,000	–	−0.019	−0.004	+0.008

O.4 Initial loop without headroom (Loop 1)

Pool 1,200 MATH-train problems, $K = 4$; gate = the top-$\tau^\star$ fraction of each arm's own samples by its own self-judge score, $\tau^\star = 0.460$ being the largest acceptance rate with $\hat{\beta} \geq 0.98$ on a static NF4 calibration slice; 5 seeds × {NF4, BF16} × 4 rounds; holdout 200 MATH-500 problems × 4. Evidence Table E12. The gate's score threshold sat inside float saturation and admitted almost only samples from problems solved 4/4; plateau fired in 10/10 chains by round 4.

O.5 Loop with a headroom control (Loop 2)

A BF16 pre-pass selected 449 pool problems; each round generated eight attempts and selected at most four samples per problem under the specified gate. The two gates were self-judge score ≥ 0.5

and majority agreement ≥ 0.5; each used five BF16 and five NF4 chains over four rounds. Holdouts were MATH-500 and 300 disjoint pool-like problems, with four attempts each. The positive-control gain and NF4-minus-BF16 gain were fixed readouts (Table 38). The original BF16 positive control was not established; seed-level intervals also left the gain contrasts unresolved. This limits what the experiment says about preserved self-improvement. Per-round trajectories, seed vectors and gate composition are in Evidence Tables E13–E14 and Evidence Fig. E5. A leakage sensitivity excluded 12 MATH-500 and 7 pool-like holdout problems near-duplicating the pool. The majority-gate contrast was unchanged ($+0.012$ $[+0.004, +0.021]$); the NF4 score-gate MATH-500 gain became $+0.007$ $[+0.000, +0.014]$, and the BF16 pool-like gain $+0.022$ $[+0.005, +0.041]$. This sensitivity does not replace the registered control criterion.

Table 38: Loop 2 readouts (NF4 vs. BF16 control, 5 seeds $\times$ 4 rounds; round-4 minus round-0 holdout accuracy, 95% cluster-bootstrap CI over problems with seeds pooled; last column: seeds with Δgain > 0).

Gate	Holdout	NF4 gain	BF16 gain	Δgain (NF4 $-$ BF16)	seeds
self-judge ≥ 0.5	train-mixed	$+0.021$ $[+0.008, +0.035]$	$+0.018$ $[+0.000, +0.036]$	$+0.003$ $[-0.019, +0.025]$	3/5
self-judge ≥ 0.5	MATH-500	$+0.008$ $[+0.001, +0.015]$	$+0.000$ $[-0.005, +0.006]$	$+0.007$ $[-0.001, +0.016]$	4/5
maj@$k \geq 0.5$	train-mixed	$+0.026$ $[+0.012, +0.041]$	$+0.012$ $[-0.006, +0.029]$	$+0.015$ $[-0.007, +0.036]$	4/5
maj@$k \geq 0.5$	MATH-500	$+0.011$ $[+0.005, +0.018]$	-0.001 $[-0.007, +0.004]$	$+0.012$ $[+0.004, +0.021]$	5/5

O.6 Verifier isolation on the 12B (Loop 3; v1.8)

Four cells against the headroom loop's chains, five seeds each: X1 (NF4 generator, BF16 verdict on the same adapter), U1 (every completed sample, a random draw of four per problem), X2 (BF16 generator, NF4 verdict) and U2 (every sample). Tables 39–40; individual seed deltas are in Evidence Table E15. At either verifier precision, roughly a fifth to a quarter of the accepted set is truncated. Across policies, gain is broadly negatively associated with that share, but not monotone: X1 has both higher gain and more accepted truncations than the NF4 self-judge gate. The NF4 self-judge's Yes/No mass median on the pool is 0.54 against 0.99 for the BF16 verdict on the same samples.

Table 39: Loop 3 readouts: Δ with the 95% problem-bootstrap CI, the seed-level 95% t-interval, positive seeds, the pre-specified call and the seed-level minimum detectable effect.

Readout	holdout	Δ	95% (problems)	95% (seeds)	positive	call	MDE
R1 Δ_{ver} = g(X1) $-$ g(NF4/NF4)	train-mixed	$+0.007$	$[-0.008, +0.021]$	$[-0.018, +0.031]$	3/5	unresolved	± 0.025
R1 Δ_{ver} = g(X1) $-$ g(NF4/NF4)	MATH-500	-0.003	$[-0.009, +0.003]$	$[-0.009, +0.004]$	1/5	consistent	± 0.006
R2 Δ_{gate1} = g(NF4/NF4) $-$ g(U1)	train-mixed	-0.050	$[-0.066, -0.035]$	$[-0.078, -0.022]$	0/5	resolved	± 0.028
R2 Δ_{gate1} = g(NF4/NF4) $-$ g(U1)	MATH-500	-0.012	$[-0.020, -0.005]$	$[-0.028, +0.004]$	0/5	consistent	± 0.016
R3 Δ_{gate2} = g(NF4/NF4) $-$ g(U2)	train-mixed	$+0.028$	$[+0.015, +0.042]$	$[+0.002, +0.055]$	5/5	resolved	± 0.026
R3 Δ_{gate2} = g(NF4/NF4) $-$ g(U2)	MATH-500	$+0.010$	$[+0.003, +0.017]$	$[+0.006, +0.013]$	5/5	resolved	± 0.003
R4 Δ'_{ver} = g(X2) $-$ g(BF16/BF16)	train-mixed	-0.015	$[-0.032, +0.002]$	$[-0.030, +0.000]$	0/5	consistent	± 0.015
R4 Δ'_{ver} = g(X2) $-$ g(BF16/BF16)	MATH-500	-0.005	$[-0.010, +0.001]$	$[-0.013, +0.004]$	1/5	consistent	± 0.008

Table 40: Loop 3 cells: gate, accepted samples per round (round 1 $\rightarrow$ 4), truncated share of the accepted set, oracle precision, own four-round gain on the pool-like holdout (95%) and the paired deficit vs. the BF16 control (round 0 $\rightarrow$ 4).

Cell gate	accepted	trunc. share	precision	gain	deficit
U1 every completed sample	780 $\rightarrow$ 1,092	0.00 $\rightarrow$ 0.00	0.83 $\rightarrow$ 0.83	$+0.072$ $[+0.053, +0.091]$	$-0.305 \rightarrow -0.252$
NF4/NF4 NF4 self-judge ≥ 0.5 (e020)	908 $\rightarrow$ 1,076	0.22 $\rightarrow$ 0.18	0.71 $\rightarrow$ 0.72	$+0.021$ $[+0.008, +0.035]$	$-0.305 \rightarrow -0.302$
X1 BF16 verdict ≥ 0.5	1,028 $\rightarrow$ 1,173	0.28 $\rightarrow$ 0.21	0.65 $\rightarrow$ 0.69	$+0.028$ $[+0.013, +0.043]$	$-0.305 \rightarrow -0.296$
U2 every sample	1,796 $\rightarrow$ 1,796	0.70 $\rightarrow$ 0.62	0.28 $\rightarrow$ 0.27	-0.007 $[-0.021, +0.008]$	$-0.305 \rightarrow -0.330$
BF16/BF16 BF16 self-judge ≥ 0.5 (e020)	1,510 $\rightarrow$ 1,620	0.03 $\rightarrow$ 0.02	0.89 $\rightarrow$ 0.89	$+0.018$ $[+0.000, +0.036]$	–
X2 NF4 verdict ≥ 0.5	1,353 $\rightarrow$ 1,369	0.02 $\rightarrow$ 0.02	0.90 $\rightarrow$ 0.90	$+0.003$ $[-0.016, +0.022]$	$+0.000 \rightarrow -0.015$

O.7 SECOND SUBJECT: SMOLLM3-3B (LOOP 4; V1.8)

The 12B's recipe applied to a 3 B subject whose NF4 verifier column erodes (precondition branch A, Table 41): a BF16 control, the NF4 arm and X1 (NF4 generator, BF16 verdict), five seeds, four rounds, a pool of 1,000 mixed MATH-train problems, $K = 8$, cap 4; eager attention on the HF path (D6). Table 42; composition and individual seed deltas are in Evidence Tables E16–E17. The score gate at 0.5 admits 0.97 of the pool before the cap, so the top-4-by-score cap is the selector; truncation stays at 0.03–0.04, so the gain is real-error repair; the saturation detector fired in 14/15 chains.

Table 41: Loop 4 precondition: SmolLM3-3B's NF4 verifier column vs. BF16, like-for-like on the HF path and against the engine, with the path floor (Level 5 and all levels; 90%; † equivalence not established).

Contrast	Δ L5	n	Δ all	n
NF4 − BF16 (HF path, like-for-like)	-0.028 $[-0.046, -0.010]^{\dagger}$	134	-0.020 $[-0.029, -0.012]^{\dagger}$	500
NF4 − BF16 (engine)	-0.030 $[-0.047, -0.013]^{\dagger}$	134	-0.022 $[-0.030, -0.014]^{\dagger}$	500
BF16 HF − BF16 engine (path floor)	-0.002 $[-0.005, +0.001]$	134	-0.001 $[-0.003, +0.000]$	500

Table 42: Loop 4 readouts (as Table 39).

Readout	holdout	Δ	95% (problems)	95% (seeds)	positive	call	MDE
(i) g_{BF16} (positive control)	train-mixed	$+0.029$	$[+0.011, +0.047]$	$[+0.006, +0.051]$	5/5	resolved	–
(i) g_{BF16} (positive control)	MATH-500	$+0.017$	$[+0.007, +0.027]$	$[+0.012, +0.021]$	5/5	resolved	–
(ii) Δgain = $g_{\mathrm{NF4}} - g_{\mathrm{BF16}}$	train-mixed	-0.007	$[-0.032, +0.018]$	$[-0.043, +0.029]$	2/5	unresolved	± 0.036
(ii) Δgain = $g_{\mathrm{NF4}} - g_{\mathrm{BF16}}$	MATH-500	$+0.007$	$[-0.006, +0.020]$	$[-0.003, +0.018]$	4/5	consistent	± 0.011
(iii) Δ_{ver} = g(X1) − g(NF4)	train-mixed	-0.005	$[-0.019, +0.011]$	$[-0.031, +0.021]$	2/5	unresolved	± 0.026
(iii) Δ_{ver} = g(X1) − g(NF4)	MATH-500	-0.003	$[-0.011, +0.004]$	$[-0.009, +0.002]$	1/5	consistent	± 0.006

Table 43: Loop 4 re-powered at twelve seeds (loop 4's five plus seven fresh), reported beside Table 42, never in its place.

Contrast holdout	Δ 90%	95% half-width	p Holm p	flags	seed-t 90% positive
(ii) Δgain = NF4+ − BF16+ (12 seeds) train-mixed	-0.013 $[-0.035, +0.009]$	$[-0.040, +0.013]$ 0.022	0.315 –		$[-0.025, -0.002]$ 2/12
(iii) Δ_{ver} = X1+ − NF4+ (12 seeds) train-mixed	$+0.004$ $[-0.010, +0.018]$	$[-0.013, +0.021]$ 0.014	0.601 –		$[-0.005, +0.013]$ 7/12
(i) g_{BF16} (12 seeds) train-mixed	$+0.034$ –	$[+0.014, +0.053]$ –	– –	– –	– –
(ii) Δgain = NF4+ − BF16+ (12 seeds) MATH-500	$+0.013$ $[+0.001, +0.025]$	$[-0.001, +0.028]$ 0.012	0.073 –		$[+0.007, +0.020]$ 10/12
(iii) Δ_{ver} = X1+ − NF4+ (12 seeds) MATH-500	-0.002 $[-0.009, +0.004]$	$[-0.010, +0.005]$ 0.006	0.543 –		$[-0.004, +0.000]$ 2/12
(i) g_{BF16} (12 seeds) MATH-500	$+0.011$ –	$[+0.000, +0.022]$ –	– –	– –	– –

O.8 COMPLETED-SAMPLE SELECTION (LOOPS 5–6; V1.9)

On the 12B (the headroom loop's protocol): U1 accept-all-completed with a random draw of four per problem (twelve seeds), O1 the correctness oracle (correct first, seeded random within ties), K1 the verdict's ranking (top four by the NF4 score, no threshold), V1 the deployed policy (score ≥ 0.5, top four by score), V1× and K1× the same two gates on the BF16 read (eight seeds), U1b accept-all-completed on the BF16 arm (five). On SmolLM3-3B (loop 4's protocol): U1n, O1n, K1n, K1n× and S1n (answer consensus: cluster size over completed count under the frozen answer equivalence, top four, no verdict, no label), twelve seeds each; U1b five; and loop 4's three cells at

seven fresh seeds, a twelve-seed re-power reported beside loop 4 (Table 43). Every contrast is paired by seed base on a shared round-1 pool; later rounds are adaptive. Tables 44–46 summarize the policies, primary contrasts and retention; full secondary contrasts, cell bookkeeping and seed diagnostics are in Evidence Tables E18–E25. The crossed-bootstrap rule and families are in Appendix A. The fresh-seeds-only sensitivity gives the same picture (12B: Δ_{oracle} -0.009 $[-0.025, +0.006]$, Δ_{rank} -0.018 $[-0.037, +0.001]$; 3B: $+0.019$ $[+0.000, +0.038]$, $+0.004$ $[-0.013, +0.020]$, Δ_{cons} $+0.011$ $[-0.006, +0.030]$).

Table 44: **Selection among completed samples** (twelve paired seeds per cell, four rounds; the pool-like holdout): the accepted set's oracle precision at round 1, the cell's four-round gain (95% crossed-bootstrap CI), its difference to the random completed draw (90%) and the pre-specified label; the full-precision-read cells and secondary contrasts are in Evidence Tables E18–E25. The registered self-judge gate of the headroom loop (five seeds) is shown for the 12B, with its difference to the deployed policy V.

Policy (top 4 completed / problem)	precision	4-round gain	Δ vs. random	outcome
Gemma 4 12B				
U: random draw	0.83	$+0.082$ $[+0.062, +0.102]$	—	
K: verdict ranking, NF4 read	0.84	$+0.073$ $[+0.055, +0.090]$	-0.010 $[-0.024, +0.005]$	equivalent
V: score ≥ 0.5 + ranking (deployed)	0.87	$+0.065$ $[+0.046, +0.083]$	-0.017 $[-0.031, -0.004]$	inconclusive – intervals only
O: correctness oracle	0.87	$+0.081$ $[+0.062, +0.100]$	-0.001 $[-0.014, +0.013]$	equivalent
registered gate, admits truncations (5 seeds)	0.71	$+0.021$ $[+0.003, +0.039]$	V − gate: $+0.034$ $[+0.016, +0.053]$	different (+)
SmolLM3-3B				
U: random draw	0.52	$+0.014$ $[-0.004, +0.032]$	—	
K: verdict ranking, NF4 read	0.58	$+0.023$ $[+0.002, +0.044]$	$+0.009$ $[-0.005, +0.022]$	equivalent
S: answer consensus (label-free)	0.68	$+0.026$ $[+0.007, +0.044]$	$+0.012$ $[-0.001, +0.024]$	equivalent
O: correctness oracle	0.74	$+0.036$ $[+0.016, +0.056]$	$+0.022$ $[+0.007, +0.036]$	inconclusive – intervals only

Table 45: Primary selector contrasts on both holdouts. Each row reports the estimate with crossed 90% and 95% intervals, the paired-seed t 90% interval, Holm-adjusted p, E/D flags and positive seeds. E uses ± 0.025; D uses the declared Holm family. A dash means neither flag, not no effect.

Contrast holdout	Δ crossed 90%	crossed 95% seed-t 90%	Holm p	flags	positive
Gemma 4 12B					
Δ_{oracle}	-0.001	$[-0.017, +0.016]$			
train-mixed	$[-0.014, +0.013]$	$[-0.011, +0.009]$	1.000	E	6/12
Δ_{rank}	-0.010	$[-0.027, +0.008]$			
train-mixed	$[-0.024, +0.005]$	$[-0.021, +0.001]$	0.845	E	5/12
Δ_{sel}	-0.017	$[-0.034, -0.001]$			
train-mixed	$[-0.031, -0.004]$	$[-0.027, -0.008]$	0.166	–	2/12
$\Delta_{\text{ver,rank}}$	-0.004	$[-0.023, +0.015]$			
train-mixed	$[-0.020, +0.012]$	$[-0.014, +0.007]$	1.000	E	4/8
Δ_{oracle}	$+0.002$	$[-0.006, +0.011]$			
MATH-500	$[-0.005, +0.010]$	$[-0.004, +0.008]$	1.000	E	6/12
Δ_{rank}	$+0.000$	$[-0.009, +0.008]$			
MATH-500	$[-0.007, +0.007]$	$[-0.006, +0.006]$	1.000	E	6/12
Δ_{sel}	$+0.000$	$[-0.008, +0.008]$			
MATH-500	$[-0.006, +0.006]$	$[-0.005, +0.005]$	1.000	E	5/12
$\Delta_{\text{ver,rank}}$	$+0.000$	$[-0.007, +0.008]$			
MATH-500	$[-0.006, +0.007]$	$[-0.004, +0.005]$	1.000	E	4/8
SmolLM3-3B					
Δ_{oracle}	$+0.022$	$[+0.004, +0.039]$			
train-mixed	$[+0.007, +0.036]$	$[+0.012, +0.032]$	0.088	–	10/12
Δ_{rank}	$+0.009$	$[-0.007, +0.025]$			
train-mixed	$[-0.005, +0.022]$	$[+0.003, +0.015]$	0.804	E	9/12
Δ_{cons}	$+0.012$	$[-0.004, +0.027]$			
train-mixed	$[-0.001, +0.024]$	$[+0.006, +0.018]$	0.533	E	9/12
$\Delta_{\text{cons}-\text{rank}}$	$+0.003$	$[-0.015, +0.020]$			
train-mixed	$[-0.012, +0.017]$	$[-0.006, +0.011]$	1.000	E	7/12
Δ_{oracle}	$+0.004$	$[-0.006, +0.013]$			
MATH-500	$[-0.004, +0.011]$	$[-0.001, +0.009]$	1.000	E	9/12
Δ_{rank}	-0.001	$[-0.010, +0.008]$			
MATH-500	$[-0.009, +0.007]$	$[-0.006, +0.004]$	1.000	E	5/12
Δ_{cons}	$+0.000$	$[-0.010, +0.010]$			
MATH-500	$[-0.008, +0.008]$	$[-0.006, +0.006]$	1.000	E	6/12
$\Delta_{\text{cons}-\text{rank}}$	$+0.001$	$[-0.009, +0.010]$			
MATH-500	$[-0.007, +0.009]$	$[-0.004, +0.006]$	1.000	E	8/12

Table 46: Retention on both subjects and holdouts: b is the BF16-read ranking increment, q the NF4-read increment, $H = q - b/2$ (crossed 90% intervals). Both subjects return unresolved; the denominator is not resolved positive. Decision p is unadjusted secondary on the 12B and Holm-adjusted on the 3B. Full 95% intervals and the mirror quantity $q - b$ are in Evidence Tables E20 and E24.

Subject holdout	b 90%	q 90%	$H = q - b/2$ 90%	Decision p outcome
12B	-0.012	-0.008	-0.002	0.869
train-mixed	$[-0.029, +0.006]$	$[-0.026, +0.010]$	$[-0.016, +0.012]$	unresolved
12B	$+0.002$	$+0.002$	$+0.001$	0.835
MATH-500	$[-0.006, +0.010]$	$[-0.006, +0.010]$	$[-0.005, +0.007]$	unresolved
3B	$+0.011$	$+0.009$	$+0.004$	1.000
train-mixed	$[-0.002, +0.024]$	$[-0.005, +0.022]$	$[-0.008, +0.015]$	unresolved
3B	$+0.002$	-0.001	-0.002	1.000
MATH-500	$[-0.006, +0.010]$	$[-0.009, +0.007]$	$[-0.009, +0.004]$	unresolved

P CONTAMINATION AND MEMORIZATION

The subjects' pretraining corpora are not public, so the scan uses public proxies (MATH train, NuminaMath-1.5) and the loops' own pools. The memorization probe (greedy continuation of the first 60% of each MATH-500 statement; ROUGE-L ≥ 0.7 on the held-back 40%) fires 0% of the time for the 12B and 0% for the 35B; 3.0% of 12B solutions share a 13-gram with the reference solution. Hard-tier problems predate both subjects' weights (a difficulty instrument, not a clean slice); the AtCoder slice is clean by date.

Table 47: Near-duplicates ($\geq 50\%$ of a problem's distinct 8-grams in one corpus problem) / verbatim matches per query set.

Corpus	MATH-500 (500)	hard tier (153)	AtCoder slice (59)
MATH train	27 / 0	0 / 0	0 / 0
NuminaMath-1.5	53 / 14	31 / 31	0 / 0
loop-v2 pool	4 / 0	0 / 0	0 / 0
loop-v1 pool	6 / 0	0 / 0	0 / 0

Q COMPUTE, REPLICATION AND REPRODUCTION

Counted runs used ≈5,800 NVIDIA B200 GPU-hours under Slurm, one arm per GPU, in two pinned images: llama.cpp b10330/CUDA 12.8.1 and PyTorch (HF columns, loops and probes). Cluster results replicated Apple Metal and RTX 4090 probes within their intervals: 12B `Q8_0` $G = 0.906$ and `IQ2_M` 0.375 on 120 problems, versus 0.900 / 0.368 on the 4090. The 12B ladder rebuilt byte-identically. Seeds derive from problem id and sample index; resumed runs reuse ledgered samples. Re-generation matches within the disclosed replay tolerance, including cached-prefix effects and the HF/engine verdict-read floor. Arm manifests record digests, environments and deviations. The anonymized review package contains these manifests, code, frozen metrics, assertion and figure/number scripts, registration documents, the deviations ledger and both evidence PDFs. Raw sample/verdict records, loop artifacts and teacher-forcing documents are archived in a separate versioned dataset for release after review; they are not included in the review package.