What Agentic SFT Buys: Rapid Behavioral Convergence, Base-Conditional Capability Gains

Anonymous ACL submission

Abstract

Open coding agents are increasingly trained on the trajectories of a frontier agent, and the recipes are compared by resolve rate, which at 731 and 500 tasks cannot separate "the trajectories did not transfer" from "they transferred and bought no score". We curate 8,137 Claude Opus 4.8/Claude Code trajectories from 32,830 captured runs with an explicit rule set, fine-tune two open Qwen families on them for one to three epochs from bare and post-trained bases, and evaluate twelve arms in one harness on SWE-bench Pro and SWE-bench Verified; one extractor fingerprints all 46,542 teacher and student trajectories, and two same-pool anchors put teacher and students on the same tasks. On SWE-bench Pro, one epoch produces most of the dense model's convergence in command vocabulary, tool composition and protocol-error recovery, and on the same tasks cuts the distance to the teacher's own trajectories to about one tenth of the native model's; supervised fine-tuning (SFT) cuts median cumulative input tokens by 45–53 % on every dense checkpoint, whatever its score. Fingerprint proximity changes whether or not pass@1 changes: the corpus lifts a bare base by $+8.9$ pp on Pro ($+3.4$ pp on Verified), every comparison against a post-trained base (-2.2 to $+2.8$ pp) stays below the paired design's detectable effect, and the next Qwen generation gains $+5.6$ to $+5.8$ pp over the native dense model at $1.35\times$ its input. Where paired pass@1 at 731 and 500 tasks cannot separate the arms, the behavioral rulers separate them by seven to thirty times their sampling floors.

1 Introduction

Open coding agents are increasingly trained on the trajectories of a frontier agent (Pan et al., 2025; Yang et al., 2025b; Ahmad et al., 2026), and the recipes are compared by resolve rate. Resolve rate hides the process that produced it (Shu et al., 2026). At 731 and 500 tasks a paired comparison cannot separate "the trajectories did not transfer" from "they transferred and bought no score", and two runs of the same weights flip more than a tenth of the tasks.

Imitation SFT is known to copy a model's style before its capability (Gudibande et al., 2024), and agent behaviors can be fingerprinted and made more teacher-like (Oderinwale, 2026; Yang et al., 2026a,b); we show where that convergence saturates, what operational habits and input-token footprint it transfers, and why it can be sharply dissociated from paired solve-rate gains on post-trained coding models. We vary the dose (epochs) and the starting model (bare or post-trained) together, pair every behavioral measurement with a capability measurement on the same arm, add two same-pool anchors (held-out training-pool tasks; the teacher run on the benchmark) and the next Qwen generation as a control, all in one coding harness, and document the curation rule by rule.

The corpus is 8,137 Claude Opus 4.8/Claude Code trajectories (v12) selected from 32,830 captured runs by an explicit, fully reported rule set (§3). Two open Qwen families are fine-tuned on it from a post-trained dense base, a post-trained mixture-of-experts (MoE) base and a bare MoE base, and twelve arms run in one harness on SWE-bench Pro and SWE-bench Verified (§4). One extractor measures all 46,542 teacher and student trajectories on four rulers, 345 held-out and 359 benchmark tasks anchor teacher and students on the same tasks, and 180 endings and 68 rationalization windows are audited (§5–§7).

On the dense line on SWE-bench Pro, one epoch produces most of the observed convergence in command vocabulary, tool composition and protocol-error recovery; the second epoch adds a small vocabulary step and the third nothing above the sampling floor (Figure 1a), on the evaluation pools and on the same tasks the teacher ran (Figure 4). Supervised fine-tuning (SFT) cuts median cumu-

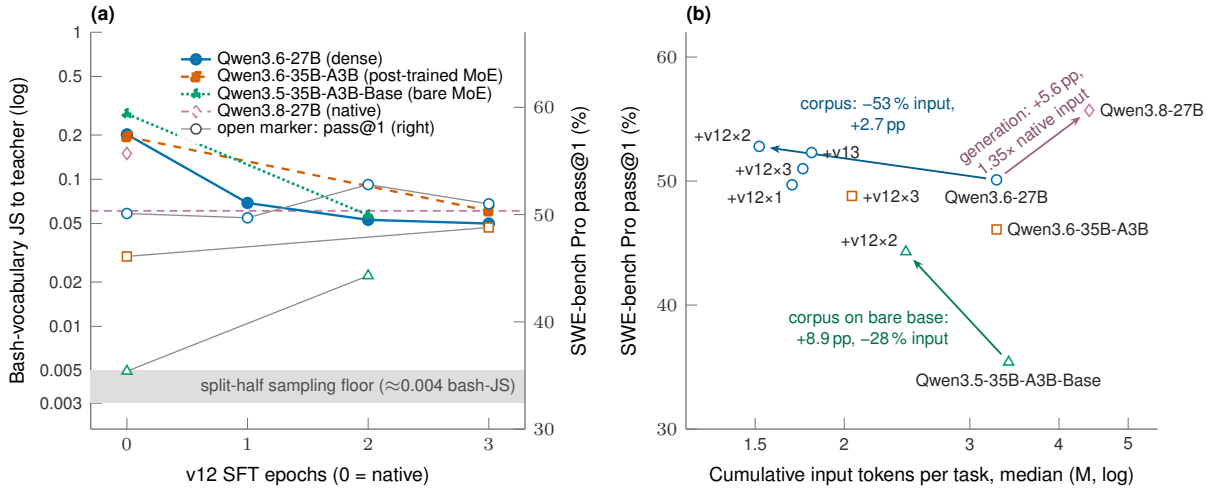


Figure 1: One epoch produces most of the dense model's behavioral convergence on SWE-bench Pro (731 tasks); SFT cuts median input by up to 53 %, while the capability response depends on the starting model. **(a)** Bash-vocabulary Jensen–Shannon distance to the pooled teacher reference (left, log; filled markers) and pass@1 (right; open markers) over v12 epochs for a post-trained dense model (Qwen3.6-27B), a post-trained MoE (Qwen3.6-35B-A3B) and a bare MoE base (Qwen3.5-35B-A3B-Base); dashed line = the natively post-trained Qwen3.8-27B. **(b)** pass@1 against median cumulative input tokens per task (log): the dense model moves from 3.27 M to 1.52 M at +2.7 pp, the bare base from 3.40 M to 2.44 M at +8.9 pp, and the generation step from 3.27 M to 4.41 M at +5.6 pp.

lative input tokens by 45–53 %, from 3.27 M to 1.52–1.80 M per task, on every dense checkpoint whatever its score. Fingerprint proximity changes whether or not pass@1 changes: the same corpus lifts a bare base by +8.9 pp on Pro (+3.4 pp on Verified) and moves post-trained bases by −2.2 to +2.8 pp, below the detectable effect of the paired design, while the next Qwen generation gains +5.6 to +5.8 pp over the native dense model at 1.35× its input (Figure 1b). Where paired pass@1 cannot separate the arms, the behavioral rulers separate them by seven to thirty times their sampling floors (§7).

Our contributions are:

1. a documented curation pipeline for frontier-agent trajectories with a rule-by-rule impact analysis (§3);
2. a dose-response fingerprint analysis: a one-epoch result on the dense line, MoE-line convergence, within-language replication, and two same-pool anchors on which the student's reads and edits overlap the teacher's on tasks it never trained on (§5);
3. the dissociation of fingerprint transfer from capability across starting points, with the input-token footprint as a transferred trait (§6);
4. a paired detectable-effect analysis for SWE-bench Pro (Deng et al., 2025) and SWE-bench Verified (Jimenez et al., 2024; OpenAI, 2024), and an audit showing that a success-claim regex tracks completion phrasing rather than outcome (§7).

2 Related Work

Trajectory SFT for coding agents. Fine-tuning on agent trajectories began with search and tool-use agents (Chen et al., 2023; Zeng et al., 2024), where format following and reasoning were already seen to be learned at different speeds (Chen et al., 2024). Issue-resolution agents are now routinely trained on trajectories collected at scale (Pan et al., 2025; Yang et al., 2025b; Zeng et al., 2025; Wang et al., 2025a,b), distilled from a stronger agent (Ahmad et al., 2026), curated down to a few hundred examples (CodeArts Model Team et al., 2026), or combined with skill priors and inference-time scaling (Yang et al., 2025c; Rastogi et al., 2025; Tao et al., 2026); reinforcement learning on software history is the alternative (Wei et al., 2025). These recipes, and the study that asks which trajectories teach well (Yang et al., 2026b), are compared by resolve rate; we keep the recipe fixed and ask what the student inherits, at what dose, and whether it shows in the score at all.

How trajectory corpora are curated. Trajectory- and segment-level selection with loss masks is used as an accuracy lever (Tao

et al., 2026; Zheng et al., 2026a; Yang et al., 2026b), failed trajectories can be kept under a signal (Wang et al., 2024), the student is known to inherit the flaws of every retained step (Ma et al., 2026), and leakage of the upstream fix is a known contamination path (Prathifkumar et al., 2025; Liang et al., 2025). Our object is the rule system itself and the behavioral profile of what each rule removes.

Behavioral fingerprints and distillation-induced similarity. Trajectories and pull requests identify the agent that produced them (Oderinwale, 2026; Gupta et al., 2026; Ghaleb, 2026), models carry detectable idiosyncrasies (Sun et al., 2025), distillation leaves detectable traces (Rawat et al., 2026) and moves a student's tool-use behavior toward its teacher (Yang et al., 2026a), sometimes through signals invisible in the data (Cloud et al., 2025; Dang et al., 2026; König et al., 2026). Our extractor is in this family; we add the epoch dose, capability measured on the same arms, a same-task anchor and a vendor next generation.

Style versus capability in supervised fine-tuning. Imitating a proprietary model copies its style before its capability (Gudibande et al., 2024); small instruction sets change format more than knowledge (Zhou et al., 2023; Lin et al., 2024; Raghavendra et al., 2024); whether SFT generalizes depends on the starting model and on repetition (Chu et al., 2025; Ren et al., 2026; Kopiczko et al., 2026); long-chain-of-thought distillation transfers a reasoning *form* (Yeo et al., 2025; Muennighoff et al., 2025; Ye et al., 2025); and trajectory SFT can raise an agent benchmark by memorizing its interface rather than its tools (Gu et al., 2026). The dissociation we report is the agentic, teacher-relative instance of this pattern.

Evaluation validity. Resolve rate conceals trajectory structure (Shu et al., 2026; Mehtiyev and Assunção, 2026; Majgaonkar et al., 2025), passes wrong patches through weak tests (Yu et al., 2026), is flaky across runs (Wang et al., 2026; Yagubyan, 2026) and needs paired error bars (Miller, 2024); confident endings mask silent failure (Advani, 2026), fewer tokens need not mean a smaller bill under prompt caching (Weinberger and Hozez, 2026), and SWE-bench results are exposed to memorization (Prathifkumar et al., 2025; Liang et al., 2025).

3 Corpus: generation and curation

3.1 Generation and verification

The teacher is Claude Opus 4.8 driven by Claude Code (Anthropic, 2026b,a). A vendor captured 32,830 of its runs on 16,328 issue-resolution instances from 296 open-source repositories at fixed commits, each with a hidden test patch. A run passes only if, with its candidate patch applied (test files stripped) and the gold test patch on top, every required fail-to-pass and pass-to-pass test passes. The captures fall into 12,257 high-precision passes, 9,843 by-product passes and 10,483 by-product failures; 247 errored before verification.

3.2 Curation funnel and rules

Figure 2a draws the v12 funnel; Table A17 (Appendix C) tabulates it with the loss masks. A vendor flag for access to the upstream fix removes 221 of the 12,257 high-precision passes. Ten drop rules (the P0 tier) then remove 3,435 of the remaining 12,036 (28.5 %): a raw-thinking coverage floor of 0.98, vendor sourcing marks, six content-leak rules, sandbox escapes and multi-user turns (thresholds in Table A18). A demote tier flags but keeps rows, a same-task cap of two removes 39 more, and a task-level split leaves 8,137 training and 425 validation rows, of which 8,120 and 424 fit the 196,608-token training length. Nine message-level loss masks touch 5,528 of the 8,137 rows and remove none, the two largest covering harness-rejected tool calls and calls from an ignored working directory; masked segments stay in context (2.77 % of learnable characters; Table A17).

3.3 The v13 variant

v13 starts from all 12,257 high-precision rows (the 221 vendor-flagged rows face its leak rules rather than up-front removal), relaxes the thinking floor to 0.90, narrows the vendor-mark rule, adds the by-product population (7,503 passes and 6,578 failures) (cf. Wang et al., 2024), drops the per-task cap and trains for one epoch on the 23,146 of its 23,238 training rows that fit the training length (Table A17). On 3,104 of the failures (47.2 %) the final message asserts completion in the strong form and receives no loss. The same regex fires on 46.4 % of the teacher's *successful* endings and on 44.6 % of its failed ones, so the mask edits phrasing rather than detecting failure (Advani, 2026). Filters, composition, cap and dose change together, so v12 versus v13 is a composite comparison (§6).

3.4 What the rules remove

Figure 2 compares what each active rule family removes with the 8,601 rows it keeps (numbers in Table A19). The thinking-coverage floor accounts for 2,450 of the 3,435 drops (71 %) and shifts the tool mix least of any family (JS 0.005 to the kept set, above the 0.0001 size-matched random floor), but it removes the longest and most think-heavy trajectories (median 68 steps against 44, thinking share 0.897 against 0.838). The sourcing rules remove 1–14 % each, and their rows are Bash-heavier and recover from protocol errors far less often (0.24 for the leak family against 0.60): a small, behaviorally distinct slice. Language and repository shares barely move along the funnel. The trained corpus therefore carries a thinking share of 0.881, the teacher's harness-error recoveries and truncation habits, and one habit no rule targets, editing a file without reading it (0.110 of edits).

4 Students, evaluation and fingerprint extractor

Students and arms. Table 1 lists the twelve arms, named by base model and corpus dose. The dense line starts from Qwen3.6-27B (Yang et al., 2025a; Qwen Team, 2026b) and adds v12 for one, two or three epochs (128, 256 and 381 updates) or v13 for one epoch (362 updates). The MoE line starts from a pretrained base without post-training, Qwen3.5-35B-A3B-Base (Qwen Team, 2026a) (it ships a chat template), trained for two epochs, and from the post-trained Qwen3.6-35B-A3B (Qwen Team, 2026c), the same architecture one release later, trained for three; the two post-trained MoE arms are also run a second time on Pro (repeat samples). The next Qwen generation, Qwen3.8-27B (Qwen Team, 2026d), is evaluated natively. All students are full-parameter fine-tunes with a cosine schedule (peak 8×10^{-6} on the post-trained bases, 3×10^{-5} on the bare base), a global batch of 64 sequences of up to 196,608 tokens, and loss on assistant text, thinking and tool calls except masked segments (Appendix D).

Evaluation. Every arm runs inside the same Claude Code harness (Liu et al., 2026) (version 2.1.202, 262,144-token context with automatic compaction) on SWE-bench Pro (731 tasks in Go, Python, TypeScript and JavaScript) and SWE-bench Verified (500 Python tasks): one rollout per task, temperature 0.8, a 5,400-second wall-time limit per task, and an added instruction that forbids looking up the fix over the network or in git history (serving in Appendix D). A task counts as solved only if its single trial passes the hidden tests; timeouts, crashes and tasks an arm did not run (0–14 per arm and pool) count as unsolved (Appendix A). No evaluation instance appears in the corpus and the Verified repositories are disjoint from it; one Pro repository, `teleport`, contributes 1,622 teacher trajectories and 76 Pro tasks, so Table 1 also reports differences on the other 655. With one rollout per task the paired design (Miller, 2024) resolves a median effect of about 4 pp on Pro and 5 pp on Verified; smaller differences are reported as below the detectable effect (Appendix A).

Fingerprint extractor. One extractor maps the teacher's 32,830 captures and the students' 13,712 parsed evaluation trajectories (from 13,733 trials) to a common step sequence and measures each arm's Jensen–Shannon divergence ($\log_2$) from the pooled high-precision teacher reference (12,257 passes) on four rulers: tool mix over eleven buckets, tool bigrams, Bash command-head vocabulary, and thinking position in deciles of the trajectory. Arm-level distributions are equal-weight means over trajectories; the composite is the mean of mix and bigram. Differences are read against two floors: the split-half floor (random halves of one arm; about 0.001 composite, 0.002–0.005 Bash) and the two repeat samples of the same weights. Two same-pool anchors put teacher and students on the same tasks: the 345 oracle-passed validation tasks of the v12 split (teacher trajectories exist, no task overlap with training), run once by eight arms; and the teacher itself, same harness version, instruction and wall-time limit at effort `high`, on 359 SWE-bench Pro tasks drawn in stratified batches without teleport (Appendix E).

5 What the student inherits

5.1 Dose response

On SWE-bench Pro, one v12 epoch produces most of the dense model's convergence toward the teacher reference. Bash-vocabulary distance to the teacher falls from 0.202 to 0.069 after one epoch, then to 0.053 and 0.050 (Figure 1a); the composite and thinking-position rulers fall the same way (Figure A2, Appendix B). Against the floors of §4, the first epoch moves the Bash, bigram, thinking-position and composite rulers well clear of noise (tool mix alone does not), the second epoch moves

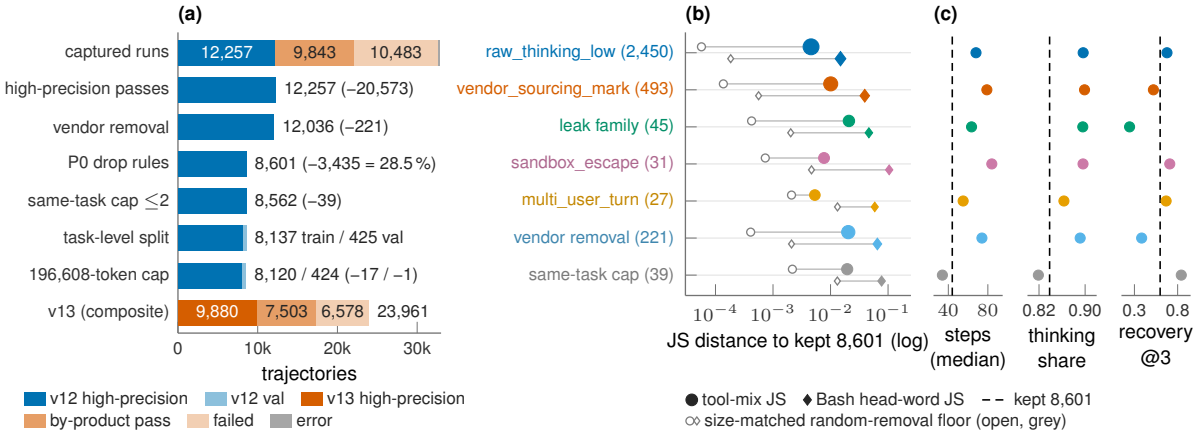


Figure 2: The v12 curation funnel and what each rule removes. **(a)** Rows surviving each stage, from 32,830 captured runs to 8,120 training rows at the 196,608-token cap, with the composite v13 corpus (9,880 high-precision rows plus 7,503 by-product passes and 6,578 failures) for comparison. **(b)** Jensen–Shannon distance between the trajectories a rule family removes and the 8,601 it keeps, on tool mix (circles) and Bash head-words (diamonds), against the size-matched random-removal floor (open grey markers): the thinking-coverage floor removes most rows yet has the smallest tool-mix distance of any family, while the sourcing and leak rules remove a small, distinct slice. **(c)** Median steps, thinking share and protocol-error recovery of the removed rows against the kept set (dashed).

Arm (base model + corpus dose)	pass@1 (%)		Δ vs. native (pp)		ΔPro−tp	Input	Bash-JS
	Pro	Ver.	Pro	Ver.	(pp)	(M tok)	to teacher
Qwen3.6-27B	50.1	68.4	–	–	–	3.27	0.202
Qwen3.6-27B+v12×1	49.7	67.8	−0.41	−0.60	−1.22	1.69	0.069
Qwen3.6-27B+v12×2	52.8	68.6	+2.74	+0.20	+2.60	1.52	0.053
Qwen3.6-27B+v12×3	51.0	71.2	+0.96	+2.80	−0.15	1.75	0.050
Qwen3.6-27B+v13×1	52.3	66.2	+2.19	−2.20	+1.98	1.80	0.055
Qwen3.5-35B-A3B-Base	35.4	52.0	–	–	–	3.40	0.278
Qwen3.5-35B-A3B-Base+v12×2	44.3	55.4	+8.89	+3.40	+9.62	2.44	0.057
Qwen3.6-35B-A3B	46.1	63.0	–	–	–	3.27	0.195
Qwen3.6-35B-A3B+v12×3	48.8	63.0	+2.74	+0.00	+3.05	2.05	0.061
Qwen3.6-35B-A3B (repeat)	47.1	–	–	–	–	3.13	0.200
Qwen3.6-35B-A3B+v12×3 (repeat)	49.5	–	+2.46	–	+1.83	2.01	0.069
Qwen3.8-27B	55.7	74.2	+5.61	+5.80	+5.04	4.41	0.061

Table 1: The twelve evaluation arms, named by base model and corpus dose (+v12×2 = two epochs of v12; +v13×1 = one epoch of v13, 362 updates). pass@1 = solved tasks over the pool (731 SWE-bench Pro, 500 SWE-bench Verified; anomalous trials count as unsolved), one rollout. Δ: percentage points against the line's native arm (Qwen3.8-27B: against Qwen3.6-27B, the generation jump). ΔPro−tp: Pro without the 76 teleport tasks ($N = 655$). Input: median cumulative input tokens per task on Pro, millions. Bash-JS: Jensen–Shannon distance ($\log_2$) of the Bash head-word distribution to the teacher on Pro. The two T=0.6 rows are the same weights as the rows above them, run a second time (the run-to-run reference, §7). "–": not defined for a native arm.

only the Bash ruler, and the third moves none (Appendix B). The MoE line converges on either starting base, and the drop holds inside each language stratum (Figure A6).

Off-pool, tool composition follows task structure: on SWE-bench Verified the composite distance rises from 0.046 to 0.065–0.068 while the Bash and thinking-position rulers still fall (Figure A2, Appendix B). On the same tasks the pool confound disappears (Figure 4): on 345 held-out validation tasks the dense line's composite distance to the teacher's own trajectories falls from 0.051 to 0.006 after one epoch and stays (Bash 0.128 to 0.026, 0.014 and 0.015; Table A29); on the 359 SWE-bench Pro tasks the teacher itself ran, every v12 and v13 arm sits at 0.005–0.007 (tool mix) and 0.005–0.010 (Bash, at its split-half floor) from the teacher, against 0.045–0.061 and 0.11–0.13 for the post-trained natives of the same families (Qwen3.8-27B: 0.061 and 0.020; Table A26); and the drop

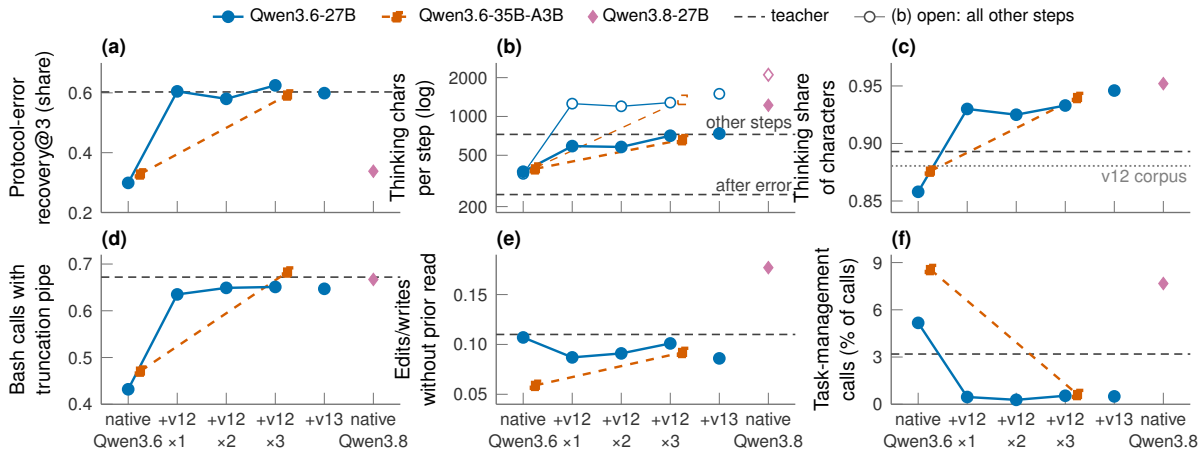


Figure 3: What the student inherits (SWE-bench Pro; dashed = pooled teacher reference) along the dense line (native, +v12 ×1/2/3, +v13) and the post-trained MoE pair (native, +v12 ×3). Protocol-error recovery (a) and truncation-pipe use (d) reach the teacher's level at the first trained checkpoint; write-without-read (e) sits at the teacher's rate on the dense line and rises to it on the post-trained MoE line; the post-error pattern (b), less thinking after an error (filled) than elsewhere (open), is copied while both levels rise past the teacher; thinking share (c) overshoots corpus and teacher; task-management calls (f) collapse below the corpus rate (the teacher makes none on the benchmark itself, Appendix E). Qwen3.8-27B matches the distilled arms on (c) and (d) only.

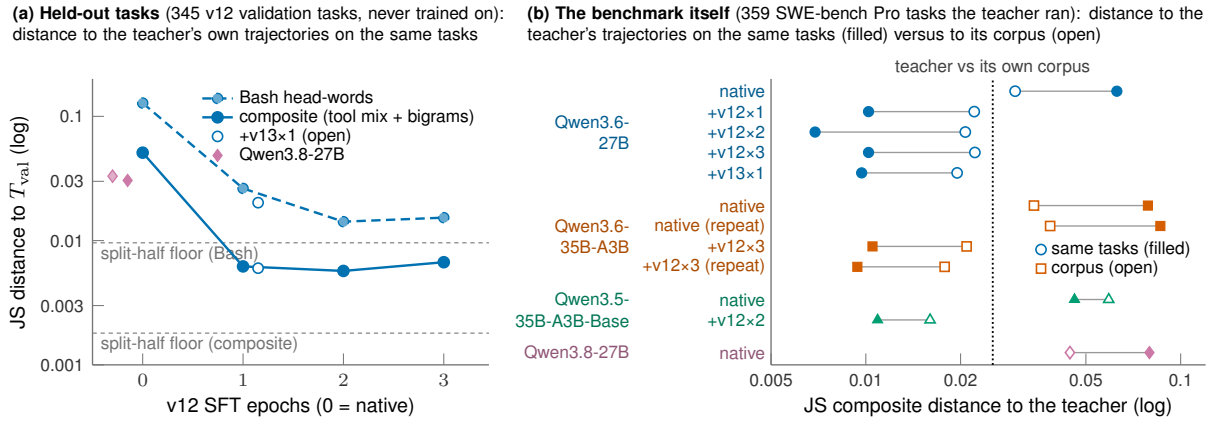


Figure 4: Teacher and students on the same tasks. **(a)** 345 held-out validation tasks (teacher trajectories exist, never trained on): Jensen–Shannon distance from the dense line to the teacher's own trajectories on these tasks against v12 epochs, composite and Bash head-words, with the split-half sampling floors (dashed); open circle = Qwen3.6-27B+v13×1, diamond = Qwen3.8-27B. **(b)** 359 SWE-bench Pro tasks run by the teacher itself: each arm's composite distance to the teacher's trajectories on the same tasks (filled) and to the corpus reference (open); dotted line = the teacher's own benchmark trajectories against its corpus. Every v12 and v13 arm is closer to the teacher on the benchmark than the teacher's corpus is; every native arm is farther.

holds within each language stratum. On the held-out tasks thinking-position distance is 0.003–0.005 for every dense arm (floor 0.002), so the native model is already close there.

5.2 Anchors

On tasks it never trained on, the student reads fewer files and more of the teacher's: read-set precision against the teacher rises from 0.56 to 0.78 on the 345 held-out tasks and from 0.50 to 0.75 on the 359 benchmark tasks (Jaccard 0.50 to 0.64 and 0.50 to 0.60), a gain of the same order as on 88 trained tasks replayed three times (0.50 to 0.80, 0.44 to 0.60; Figure A3), against negative controls of 0.001 and 0.05 on the held-out tasks (0.003 and 0.02 on the benchmark) and repeat-run differences below 0.01 (Tables A27 and A30). The overlap measures imitation of a reading style rather than recall of trained tasks (the held-out tasks share repositories, not tasks, with the corpus); imitation gain is uncorrelated with solve gain in every anchor (Spearman ρ from -0.12 to 0.03). On the benchmark the

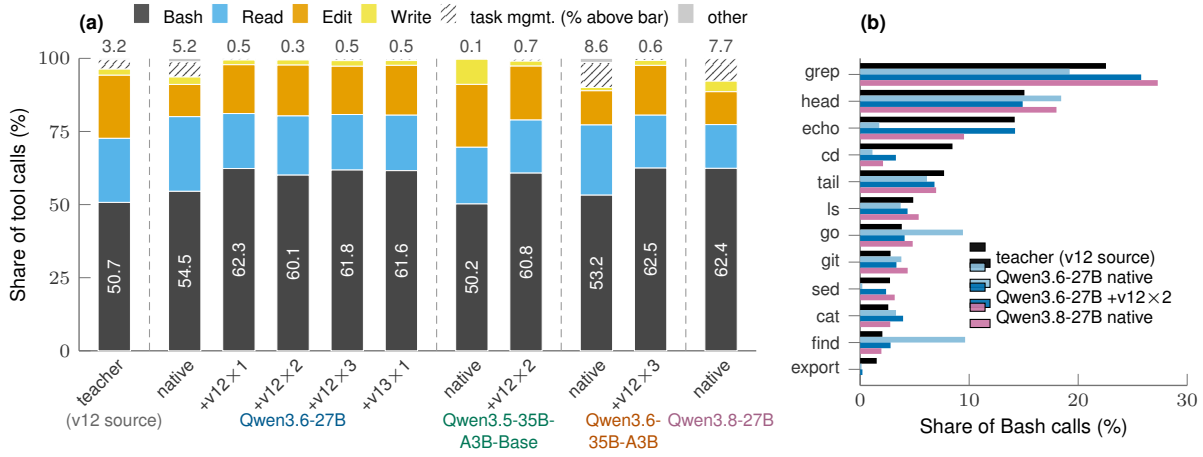


Figure 5: Tool composition and Bash vocabulary on SWE-bench Pro. **(a)** Share of tool calls per arm (inside the bar: Bash share; above it: task-management share). v12 and v13 raise the Bash share from 50–55 % to 60–62 % and collapse task management on the two post-trained natives from 5–9 % to below 1 % (teacher 50.7 and 3.2 %). **(b)** The teacher's twelve most frequent Bash command heads as a share of each group's Bash calls: native Qwen3.6-27B barely uses `echo` (1.8 against 14.2 %); two v12 epochs restore the teacher's profile, which native Qwen3.8-27B matches only in part.

teacher itself stays within 0.01–0.08 of its corpus values for Bash share, `echo`, truncation pipes and write-without-read, recovers from protocol errors more often (0.74) and issues no task-management calls (Table A28).

5.3 The inherited bundle

Protocol-error recovery, the share of harness-rejected calls after which the same tool succeeds on the same target within three calls, rises on SWE-bench Pro from 0.30 to 0.58–0.63 in the dense line and from 0.33 to 0.59 in the post-trained MoE line, level with the teacher's 0.60 (Figure 3a); Qwen3.8-27B stays at 0.34. On SWE-bench Verified four of the five fine-tuned arms rise to 0.57–0.71 and the one-epoch checkpoint falls to 0.27 (Table A22). The student copies the teacher's habit of thinking less after an error than on other steps, while both levels rise past the teacher (Figure 3b; Appendix B). Thinking share rises from 0.858 to 0.925–0.946, above the corpus (0.881) and the teacher (0.893), and thinking position moves toward the teacher in both pools except for the post-trained MoE pair on Verified (Figure 3c). Command habits move after one epoch and stay: the `echo` share of Bash heads reaches the teacher's 14.2 % from 1.8 % (Figure 5b), and commands with a truncation pipe go from 0.43 to 0.64–0.65 (teacher 0.67; Figure 3d). Editing a file without reading it settles at the corpus rate of 0.11 from either side: the post-trained MoE line rises from 0.059 to 0.092, the bare base falls from 0.213 to 0.065, and the dense line stays at 0.09–0.11 (Figure 3e; Table A22). Task-management calls collapse from 5.2 % of calls to 0.3–0.5 %, below the corpus rate of 3.2 % and toward the teacher's own rate on the benchmark, which is zero (Figure 3f; Table A28).

Qwen3.8-27B, a single model, matches the distilled arms on Bash vocabulary (distance 0.061 against 0.050–0.069) and on the `echo` profile of Figure 5b but not on tool composition, reading habits or error recovery, which is consistent with its developer having trained on similar traces; it therefore bounds rather than controls the generation comparison of §6.

6 What it costs and what it buys

6.1 Token footprint

The input-token footprint is part of the fingerprint. Median cumulative input per SWE-bench Pro task falls from 3.27 M tokens (Qwen3.6-27B) to 1.52 M (Qwen3.6-27B+v12×2), −53 %: 32 % fewer steps times 29 % less context per step, with whole-file reads carrying 74 % of the drop in tool output (Figure A8, Appendix C); rewriting file reading in the harness reaches 9–50 % (Zheng et al., 2026b; Zhang et al., 2026). The post-trained MoE line moves the same way (3.27 M to 2.05 M) and so does the bare base (3.40 M to 2.44 M); every v12 and v13 checkpoint of the dense line lands at 1.52–1.80 M regardless of its score. Output tokens rise by 16–49 %, reflecting the thinking overshoot of §5 (Table A25). These are input-token counts, not

bills: under prompt caching the two diverge (Weinberger and Hozez, 2026).

6.2 pass@1

The corpus on a bare base improves SWE-bench Pro by $+8.9$ pp (Qwen3.5-35B-A3B-Base+v12×2 against Qwen3.5-35B-A3B-Base; Table 1); its $+3.4$ pp on SWE-bench Verified is below the detectable effect. The bare-base gain exceeds the post-trained MoE gain by 6.2 pp (standard error 2.3 pp), a post-hoc contrast that also crosses generation, dose and learning rate (§4), so it resolves that the gains differ, not why. Against post-trained bases the differences span -2.2 to $+2.8$ pp in both pools and both families, all below the detectable effect; excluding the 76 teleport tasks changes no decision (Appendix A). The generation jump (Qwen3.8-27B against Qwen3.6-27B) gives $+5.6$ pp on Pro and $+5.8$ pp on Verified, more than any checkpoint on the post-trained dense line, at 4.41 M input tokens per task ($1.35\times$ the native model); against the two-epoch student the Pro margin is $+2.9$ pp at $2.9\times$ the input. The bare-base gain is consistent with the corpus supplying part of the agentic post-training transition on this base, the diminishing return on stronger starting models that curated-trajectory SFT also shows (CodeArts Model Team et al., 2026). The composite v13 arm gives $+1.3$ pp on Pro and -5.0 pp on Verified against Qwen3.6-27B+v12×3, the one composition signal at the design's detectable effect (Appendix A); as a composite change it is not attributable, and the arm's fingerprint and footprint stay within the v12 range.

7 What the score tables can resolve

Three properties of the evaluation bound the score tables. First, single-rollout outcomes are noisy: two runs of the same weights flip 13 % of SWE-bench Pro tasks for Qwen3.6-35B-A3B and 11 % for Qwen3.6-35B-A3B+v12×3, for net differences below 1 pp, the run-to-run reference for every paired comparison (Wang et al., 2026; Yagubyan, 2026). Second, the paired design resolves differences of about 4 pp on Pro and 5 pp on Verified (§4), so the 0–3 pp differences between post-trained bases and their students are read as unresolved, whereas the behavioral rulers separate the same arms by seven to thirty times their sampling floors (Appendix B): at this scale the fingerprint is the sensitive readout, as interface perturbation is for agent benchmarks (Gu et al., 2026). Third, the success-claim regex

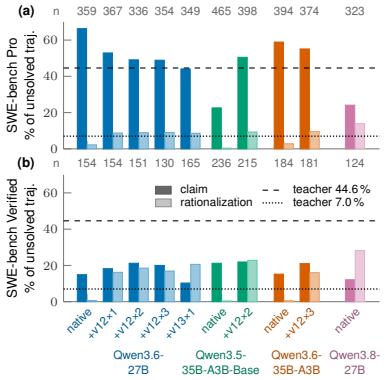


Figure 6: Endings of unsolved trajectories: share whose final message matches the strong-form completion regex (solid) and the rationalization regex (light) per arm on SWE-bench Pro (top) and SWE-bench Verified (bottom); dashed and dotted lines = the teacher's failed by-product trajectories (44.6 and 7.0 %, measured on the teacher's own task pool). v12 moves the dense line's strong-form share toward the teacher on Pro; Verified is lower and not monotone; audited endings that escape the regex assert completion in variant wording in every arm.

tracks completion phrasing rather than outcome (§3; Advani, 2026). On SWE-bench Pro the share of unsolved trajectories whose final message asserts completion in the strong form falls from 66 % (Qwen3.6-27B) to 49–53 % after v12 and 44 % after v13, level with the teacher's 45 %, so distillation moves this phrasing toward the teacher; on Verified the shares are lower and not monotone (Figure 6; Table A7). The 110 audited endings that escape the regex assert completion in variant wording in every arm, teacher included (Appendix A).

8 Conclusion

In this single teacher–corpus study, v12 SFT on 8,137 trajectories moved command vocabulary, tool composition and protocol-error recovery toward the teacher in two Qwen families, on the evaluation pools and on the same tasks the teacher ran, with most dense-model change after one epoch, and cut median cumulative input tokens by 45–53 %. pass@1 depended on base and pool: $+8.9$ pp on Pro and $+3.4$ pp on Verified for a bare base, and -2.2 to $+2.8$ pp against post-trained bases, the latter two below what 731 and 500 paired tasks can resolve. The composite v13 variant stayed inside the v12 fingerprint and footprint range. One epoch buys the habits and half the input tokens; transfer is verified on the fingerprint, not on pass@1.

Limitations

One teacher, one corpus, one harness. Every student learns from trajectories of a single frontier agent captured in a single harness, and every arm is evaluated in that same harness. Two student families and three starting points (bare MoE, post-trained MoE, post-trained dense) vary the receiver, and the native dense arm stays far from the teacher inside the same harness, so the harness alone does not create the fingerprint; but the transfer of a different teacher's habits, or of these habits into a different scaffold, is not measured.

Anchors cover Pro, not Verified. The teacher's benchmark trajectories cover 359 of the 731 SWE-bench Pro tasks and none of SWE-bench Verified, and its thinking is returned encrypted there, so same-task comparisons on the benchmark use tool rulers only; the held-out anchor shares repositories, not tasks, with the corpus, and its two MoE arms start from different bases.

Model-annotated endings. The 180 endings (110 regex-negative, 70 regex-positive) and 68 rationalization windows were labeled by two model annotators, which agree on the two quantities the body uses (0 of 110 admissions of failure; regex precision 69 of 70) and differ on the rationalization split (48 versus 49 verified attributions, 12 versus 6 bare dismissals); no human pass was run, and the body states only the agreed quantities.

Repository-level exposure. No SWE-bench Pro or SWE-bench Verified instance appears in the corpus (zero title matches, description Jaccard at most 0.21, zero commit matches), and the Verified repositories are disjoint from it; but the corpus contains 1,622 teacher trajectories from the `teleport` repository, which supplies 76 of the 731 Pro tasks. Teleport-excluded differences are reported next to the full-pool ones (Table 1); every v12 and v13 checkpoint gains more on those 76 tasks than on the other 655.

No single-rule ablation. The rule-impact analysis is offline: it measures what each rule removed, not what training without it would have changed. The one trained variation, v13, changes the thinking floor, the by-product inclusion, the task cap and the epoch count together, and is reported as a composite. The bare-base result likewise comes from one base and one corpus; the design does not isolate a general base-to-instruct mechanism.

Sparse dose curve and one rollout. The MoE dose curve has two points per base and every arm has one rollout per task, so the behavioral rulers are read against split-half floors and the two repeat-sample arms.

Ethics statement

The corpus consists of trajectories of a commercial frontier agent working on 296 public open-source repositories at fixed commits, each used under its own license; the runs were produced and verified by a data vendor, every trajectory was audited for answer-source access, and trajectories flagged for fetching the upstream fix were removed from v12 and subjected to the content-leak rules in v13. The trajectories are outputs of a commercial model used here for research on open models; they are not redistributed, and Appendix D states what this submission releases. The evaluation benchmarks are the public SWE-bench Pro and SWE-bench Verified task sets; §4 and the Limitations section declare the one repository-level overlap between corpus and benchmark. No human subjects were involved; the failure-ending annotators are models.

References

Laksh Advani. 2026. From confident closing to silent failure: Characterizing false success in LLM agents. *Preprint*, arXiv:2606.09863.

Wasi Uddin Ahmad, Nikolai Ludwig, Somshubra Majumdar, and Boris Ginsburg. 2026. Open-SWE-Traces: Advancing Dual-Mode multilingual distillation for software engineering agents. *Preprint*, arXiv:2606.16038.

Anthropic. 2026a. Claude Code overview. `https://docs.claude.com/en/docs/claude-code/overview`. Documentation of the harness used for the teacher and every student arm (version 2.1.202); accessed 6 September 2026.

Anthropic. 2026b. Introducing Claude Opus 4.8. `https://www.anthropic.com/news/claude-opus-4-8`. Announcement of the teacher model (28 May 2026); system card at `https://www.anthropic.com/claude-opus-4-8-system-card`.

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. 2023. FireAct: Toward language agent fine-tuning. *Preprint*, arXiv:2310.05915.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Wenwei Zhang, Jiangning Liu, Dahua Lin, Kai Chen, and Feng Zhao. 2024. Agent-FLAN: Designing data and

methods of effective agent tuning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*. Findings of ACL 2024; arXiv:2403.12881.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning*. ICML 2025; arXiv:2501.17161.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. 2025. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *Preprint*, arXiv:2507.14805.

CodeArts Model Team, Yang Ye, Jingyuan Tan, Tianyue Jiang, Ruizhe Ye, Qiankun He, Jiarui Yang, Jian Dong, Sicong Liang, Chongjian Yue, Peibai Xu, Lufan Lu, Shiguan Pang, Taotao Qian, Junbao Hu, Yuechan Hao, Ensheng Shi, Qi Zhang, Yi Hao, and 8 others. 2026. Yet even less is even better for agentic, reasoning, and coding LLMs. *Preprint*, arXiv:2604.00824.

Jacob Dang, Brian Y. Xie, and Omar G. Younis. 2026. Subliminal transfer of unsafe behaviors in AI agent distillation. *Preprint*, arXiv:2604.15559.

Xiang Deng, Jeff Da, Edwin Pan, Yannis Yiming He, Charles Ide, Kanak Garg, Niklas Lauffer, Andrew Park, Nitin Pasari, Chetan Rane, Karmini Sampath, Maya Krishnan, Srivatsa Kundurthy, Sean Hendryx, Zifan Wang, Vijay Bharadwaj, Jeff Holm, Raja Aluri, Chen Bo Calvin Zhang, and 3 others. 2025. SWE-Bench pro: Can AI agents solve Long-Horizon software engineering tasks? *Preprint*, arXiv:2509.16941.

Taher A. Ghaleb. 2026. Fingerprinting AI coding agents on GitHub. *Preprint*, arXiv:2601.17406.

Weizheng Gu, Chengze Li, Zhuohao Yu, Mengyuan Sun, Zhibang Yang, Wei Wang, Hongrui Jia, Shikun Zhang, and Wei Ye. 2026. What do agents learn from Trajectory-SFT: Semantics or interfaces? *Preprint*, arXiv:2602.01611.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2024. The false promise of imitating proprietary LLMs. In *International Conference on Learning Representations*. ICLR 2024; arXiv:2305.15717.

Gaurav Gupta, Vatshank Chaturvedi, Jun Huan, and Anoop Deoras. 2026. Dissecting model behavior through agent trajectories. *Preprint*, arXiv:2606.17454.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. SWE-bench: Can language models resolve Real-World GitHub issues? In *International Conference on Learning Representations*. ICLR 2024; arXiv:2310.06770.

Dawid J. Kopiczko, Sagar Vaze, Tijmen Blankevoort, and Yuki M. Asano. 2026. Data repetition beats data scaling in Long-CoT supervised Fine-Tuning. *Preprint*, arXiv:2602.11149.

Uwe König, Hamza Kazmi, Ruizhe Li, and Maheep Chaudhary. 2026. Quantifying subliminal behavioral transfer ratios in language model distillation. *Preprint*, arXiv:2606.11270.

Shanchao Liang, Spandan Garg, and Roshanak Zilouchian Moghaddam. 2025. The SWE-Bench illusion: When State-of-the-Art LLMs remember instead of reason. *Preprint*, arXiv:2506.12286.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base LLMs: Rethinking alignment via In-Context learning. In *International Conference on Learning Representations*. ICLR 2024; arXiv:2312.01552.

Jiacheng Liu, Xiaohan Zhao, Xinyi Shang, and Zhiqiang Shen. 2026. Dive into claude code: The design space of today's and future AI agent systems. *Preprint*, arXiv:2604.14228.

Murong Ma, Tianyu Chen, Yun Lin, Shuai Lu, Qinglin Zhu, Yeyun Gong, Zhiyong Huang, Peng Cheng, Yan Lu, and Jin Song Dong. 2026. From patches to trajectories: Privileged process supervision for Software-Engineering agents. *Preprint*, arXiv:2605.21996.

Oorja Majgaonkar, Zhiwei Fei, Xiang Li, Federica Sarro, and He Ye. 2025. Understanding code agent behaviour: An empirical study of success and failure trajectories. *Preprint*, arXiv:2511.00197.

Tural Mehtiyev and Wesley Assunção. 2026. Beyond resolution rates: Behavioral drivers of coding agent success and failure. *Preprint*, arXiv:2604.02547.

Evan Miller. 2024. Adding error bars to evals: A statistical approach to language model evaluations. *Preprint*, arXiv:2411.00640.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of EMNLP 2025*. EMNLP 2025 (main); arXiv:2501.19393.

Hamidah Oderinwale. 2026. Agent trajectories as programs: fingerprinting and programming coding-agent behavior. *Preprint*, arXiv:2606.16988.

OpenAI. 2024. Introducing SWE-bench Verified. `https://openai.com/index/introducing-swe-bench-verified/`. Human-validated 500-task subset of SWE-bench.

Jiayi Pan, Xingyao Wang, Graham Neubig, Navdeep Jaitly, Heng Ji, Alane Suhr, and Yizhe Zhang. 2025. Training software engineering agents and verifiers with SWE-Gym. In *Proceedings of the 42nd International Conference on Machine Learning*. ICML 2025; arXiv:2412.21139.

Thanosan Prathifkumar, Noble Saji Mathews, and Meiyappan Nagappan. 2025. Does SWE-Bench-Verified test agent ability or model memory? *Preprint*, arXiv:2512.10218.

Qwen Team. 2026a. Qwen3.5-35B-A3B-Base (model card). `https://huggingface.co/Qwen/Qwen3.5-35B-A3B-Base`. Bare MoE base of the bare-MoE arms; repository created 24 February 2026.

Qwen Team. 2026b. Qwen3.6-27B (model card). `https://huggingface.co/Qwen/Qwen3.6-27B`. Post-trained dense base of the dense line; repository created 21 April 2026.

Qwen Team. 2026c. Qwen3.6-35B-A3B (model card). `https://huggingface.co/Qwen/Qwen3.6-35B-A3B`. Post-trained MoE base of the post-trained MoE arms; repository created 15 April 2026.

Qwen Team. 2026d. Qwen3.8-27B (model card). `https://huggingface.co/Qwen/Qwen3.8-27B`. The next Qwen generation, evaluated natively; repository created 5 August 2026.

Mohit Raghavendra, Vaskar Nath, and Sean Hendryx. 2024. Revisiting the superficial alignment hypothesis. *Preprint*, arXiv:2410.03717.

Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexander H. Liu, Alexandre Sablayrolles, Amélie Héliou, Amélie Martin, Anmol Agarwal, Andy Ehrenberg, Andy Lo, Antoine Roux, Arthur Darcet, Arthur Mensch, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Chris Bamford, Christian Wallenwein, Christophe Renaudin, and 85 others. 2025. Devstral: Fine-tuning language models for coding agent applications. *Preprint*, arXiv:2509.25193.

Rajat Rawat, Sizhe Chen, Akshay Anand, Michael Duan, Bob Rotsted, and Sewon Min. 2026. Reference-Based distillation detection in LLMs. *Preprint*, arXiv:2607.09692.

Qihan Ren, Peng Wang, Ruikun Cai, Shuai Shao, Dadi Guo, Yuejin Xie, Yafu Li, Quanshi Zhang, Xia Hu, Jing Shao, and Dongrui Liu. 2026. Rethinking generalization in reasoning SFT: A conditional analysis on optimization, data, and model capability. *Preprint*, arXiv:2604.06628.

Rui Shu, Chun Yong Chong, Xin Zhou, Yun Peng, Zihan Wu, Xu Han, Zeyang Zhuang, Guowen Yuan, and Yuan Wang. 2026. What resolve rate hides: Trajectory structure diagnostics for coding agents. *Preprint*, arXiv:2607.06184.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J. Zico Kolter, and Zhuang Liu. 2025. Idiosyncrasies in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*. ICML 2025; arXiv:2502.12150.

Chaofan Tao, Jierun Chen, Yuxin Jiang, Kaiqi Kou, Shaowei Wang, Ruoyu Wang, Xiaohui Li, Sidi Yang, Yiming Du, Jianbo Dai, Zhiming Mao, Xinyu Wang, Lifeng Shang, and Haoli Bai. 2026. SWE-Lego: Pushing the limits of supervised fine-tuning for software issue resolving. *Preprint*, arXiv:2601.01426.

Haoran Wang, Zhenyu Hou, Yao Wei, Jie Tang, and Yuxiao Dong. 2025a. SWE-Dev: Building software engineering agents with training and inference scaling. *Preprint*, arXiv:2506.07636.

Jicheng Wang, Yifeng He, Zili Wang, Hanwen Xing, Arkaprava De, and Hao Chen. 2026. ContractBench: Can LLM agents preserve observation contracts? *Preprint*, arXiv:2605.17281.

Junhao Wang, Daoguang Zan, Shulin Xin, Siyao Liu, Yurong Wu, and Kai Shen. 2025b. SWE-Mirror: Scaling Issue-Resolving datasets by mirroring issues across repositories. *Preprint*, arXiv:2509.08724.

Renxi Wang, Haonan Li, Xudong Han, Yixuan Zhang, and Timothy Baldwin. 2024. Learning from failure: Integrating negative examples when fine-tuning large language models as agents. *Preprint*, arXiv:2402.11651.

Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida I. Wang. 2025. SWE-RL: Advancing LLM reasoning via reinforcement learning on open software evolution. *Preprint*, arXiv:2502.18449.

Sarel Weinberger and Amir Hozez. 2026. Token reduction is not cost reduction. *Preprint*, arXiv:2607.12161.

Abel Yagubyan. 2026. How consistent are LLM agents? measuring behavioral reproducibility in Multi-Step Tool-Calling pipelines. *Preprint*, arXiv:2605.28840.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Chenghao Yang, Yuning Zhang, Zhoufutu Wen, Tao Gong, Jiaheng Liu, Qi Chu, and Nenghai Yu. 2026a. When agents look the same: Quantifying Distillation-Induced similarity in Tool-Use behaviors. In *Proceedings of ACL 2026*. ACL 2026 (main); arXiv:2604.21255.

John Yang, Kilian Lieret, Carlos E. Jimenez, Alexander Wettig, Kabir Khandpur, Yanzhe Zhang, Binyuan Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang. 2025b. SWE-smith: Scaling data for software engineering agents. In *Advances in Neural Information Processing Systems 38*. NeurIPS 2025; arXiv:2504.21798.

Sidi Yang, Chaofan Tao, Jierun Chen, Tiezheng Yu, Ruoyu Wang, Yuxin Jiang, Yiming Du, Wendong Xu, Jing Xiong, Taiqiang Wu, Lifeng Shang, Xiaohui Li, Ngai Wong, and Haoli Bai. 2026b. What makes interaction trajectories effective for training terminal agents? *Preprint*, arXiv:2606.03461.

Zonghan Yang, Shengjie Wang, Kelin Fu, Wenyang He, Weimin Xiong, Yibo Liu, Yibo Miao, Bofei Gao, Yejie Wang, Yingwei Ma, Yanhao Li, Yue Liu, Zhenxing Hu, Kaitai Zhang, Shuyi Wang, Huarong Chen, Flood Sung, Yang Liu, Yang Gao, and 2 others. 2025c. Kimi-Dev: Agentless training as skill prior for SWE-Agents. *Preprint*, arXiv:2509.23045.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. LIMO: Less is more for reasoning. *Preprint*, arXiv:2502.03387.

Edward Yeo, Yuxuan Tong, Morry Niu, Graham Neubig, and Xiang Yue. 2025. Demystifying long Chain-of-Thought reasoning in LLMs. *Preprint*, arXiv:2502.03373.

Boxi Yu, Yang Cao, Yuzhong Zhang, Liting Lin, Junjielong Xu, Zhiqing Zhong, Qinghua Xu, Guancheng Wang, Jialun Cao, Shing-Chi Cheung, Pinjia He, and Lionel Briand. 2026. SWE-ABS: Adversarial benchmark strengthening exposes inflated success rates on test-based benchmark. *Preprint*, arXiv:2603.00520.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. 2024. AgentTuning: Enabling generalized agent abilities for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*. Findings of ACL 2024; arXiv:2310.12823.

Liang Zeng, Yongcong Li, Yuzhen Xiao, Changshi Li, Chris Yuhao Liu, Rui Yan, Tianwen Wei, Jujie He, Xuchen Song, Yang Liu, and Yahui Zhou. 2025. Skywork-SWE: Unveiling data scaling laws for software engineering in LLMs. *Preprint*, arXiv:2506.19290.

Shaoqiu Zhang, Maoquan Wang, Yuling Shi, Yuhang Wang, Xiaodong Gu, Yongqiang Yao, Tori Gong, Sheng Chen, Rao Fu, Anisha Agarwal, Spandan Grag, Gabriel Ryan, Colin Merkel, Yufan Huang, and Shengyu Fu. 2026. FastContext: Training efficient repository explorer for coding agents. *Preprint*, arXiv:2606.14066.

Dewu Zheng, Ruizhe Ye, Yanlin Wang, Yang Ye, Hongyu Zhang, Ensheng Shi, Xilin Liu, Yuchi Ma, Jianxing Yu, and Zibin Zheng. 2026a. SWE-Prime: Fewer trajectories, better performance. *Preprint*, arXiv:2608.27449.

Mingwei Zheng, David OBrien, Siwei Cui, Pardis Pashakhanloo, Rajdeep Mukherjee, Myeongsoo Kim, and Sachit Kuhar. 2026b. CORVUS: Context optimization and reduction via underlying synchronization for LLM coding agents. *Preprint*, arXiv:2607.22711.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems 36*. NeurIPS 2023; arXiv:2305.11206.

Zexin Zhuang, Yanhang Li, and Zhichao Fan. 2026. Pre-Registering the detectable effect: A Paired-MDE budget for 4-bit quantization benchmarks, with a pilot audit. *Preprint*, arXiv:2605.28873.

A Evaluation noise

Conventions. Every score-level statistic in this appendix is recomputed with the standard library from `data/report.json` (`analysis/stats/stats_fixes.py`). The pool is the union of task ids over the arms of that pool (731 on SWE-bench Pro, 500 on SWE-bench Verified) and is the denominator of every rate and difference; a task missing from an arm counts as unsolved (0–14 missing per arm and pool, Table A21; scoring each arm on the tasks it ran raises pass@1 by 0.0–2.1 pp, most for Qwen3.6-27B+v12×3 on Verified, and leaves every comparison on the same side of its detectable effect). For a pair A vs B, b is the number of tasks solved by A only, c by B only, $\Delta = (b - c)/N$, and p is the two-sided exact binomial McNemar test on the $b + c$ discordant tasks. Of the 132 pairs in `report.json`, 21 involve the repeat-sample arms on Verified, which were never run, and are void; the remaining 111 reproduce `report.json` exactly in b, c, Δ and p.

Pre-declared paired tests. Table A1 lists the two pre-declared families with Holm correction as the primary decision rule: the 13 baseline-relative corpus tests (each fine-tuned arm against its own starting model) and the 6 generation/architecture tests. Exactly one corpus test survives Holm, Qwen3.5-35B-A3B-Base+v12×2 − Qwen3.5-35B-A3B-Base on Pro (+8.89 pp, Holm $p = 5.87 \times 10^{-6}$); every post-trained-base comparison is null after correction, including Qwen3.6-27B+v12×2 − Qwen3.6-27B on Pro (exact $p = 0.054$, Holm $p = 0.653$), whose Wald interval $[+0.09, +5.38]$ excludes zero only through the

normal approximation. Five of the six generation/architecture tests survive Holm; Qwen3.6-27B+v12×3 − Qwen3.6-35B-A3B+v12×3 on Pro (+2.19 pp, $p = 0.137$) does not. Over all 111 valid pairs, 74 are significant before correction, 49 after Holm and 71 after Benjamini–Hochberg; the two nominal checkpoint-versus-checkpoint exceptions, Qwen3.6-27B+v13×1 vs Qwen3.6-27B+v12×3 on Verified (−5.00 pp, $p = 0.005$) and Qwen3.6-27B+v12×2 vs Qwen3.6-27B+v12×1 on Pro (+3.15 pp, $p = 0.014$), survive Benjamini–Hochberg only (Holm-111 $p = 0.250$ and 0.658) and are not pre-declared endpoints.

Detectable effect. Table A2 derives the minimum detectable effect of the one-rollout paired design (Miller, 2024; Zhuang et al., 2026) from the observed discordance of the baseline-relative pairs: the median is 4.03 pp on Pro (range 3.77–4.83) and 5.04 pp on Verified (4.87–5.94), and about 1,322 Pro tasks would be needed to resolve 3 pp at one rollout. Table A3 projects the effect of more rollouts per task from a variance decomposition estimated on the 237-task stratified rollout×3 subsample: at $N = 731$ it gives 5.99 / 4.07 / 3.25 / 2.63 pp for $k = 1$ / 3 / 8 / ∞. Because the subsample over-states the pools' single-rollout discordance by about a factor of two in variance, the $k = 1$ projection is an upper bound and Table A2 is the figure to quote for the actual design.

Run-to-run noise. Figure A1 draws the flip rates and the teleport slice; Table A4 gives the two same-weights pairs on Pro: a second run of Qwen3.6-35B-A3B flips 97 of 731 task verdicts (13.3 %) for a net +0.96 pp ($p = 0.543$), and Qwen3.6-35B-A3B+v12×3 flips 83 (11.4 %) for +0.68 pp ($p = 0.661$). The flip rate ranges from Go (10.4 / 8.6 %) to JavaScript (25.0 / 20.5 %) and, by repository, from `flipt` (4.7 / 1.2 %) to `NodeBB` (25.0 / 20.5 %), so the noise is not uniform across strata.

Three rollouts. On the 237-task stratified subsample replayed three times, 36–42 % of Pro tasks receive inconsistent verdicts across the three rollouts, and the stratified re-estimate of Qwen3.6-27B+v12×2 − Qwen3.6-27B is +1.40 / +2.75 pp, not distinguishable from zero.

Teleport slice. The corpus contains 1,622 teacher trajectories from the `teleport` repository, which supplies 76 of the 731 Pro tasks (Verified has none). Table A5 gives every arm's solved counts on the 76 teleport tasks and on the other 655, and Table A6 repeats the seven Pro baseline-relative tests on each slice. Excluding teleport changes no decision: nothing becomes significant and nothing significant becomes null. It removes Qwen3.6-27B+v12×3's apparent Pro gain (+0.96 → −0.15 pp) and slightly strengthens Qwen3.6-27B+v12×2 (+2.74 → +2.60 pp) and Qwen3.6-35B-A3B+v12×3 (+2.74 → +3.05 pp), so the post-trained-base corpus effect is not a teleport artefact but is below the detectable effect everywhere. On the 76 teleport tasks alone every v12/v13 checkpoint beats Qwen3.6-27B, by +3.9 to +10.5 pp (Qwen3.6-27B+v12×3: 8 vs 0 discordant, $p = 0.008$, Holm-7 $p = 0.055$), a repository-level lift consistent with corpus exposure although instance-level matches are zero (Appendix C); a 76-task slice has a standard error of about 4 pp and the repeat runs flip 9 of its tasks, so this is a one-rollout observation rather than a measured effect.

Endings audit. Figure 6 draws the regex rates per arm and Table A7 gives the regex rates behind §7. Two model annotators read the same 180 endings (110 regex-negative, 70 regex-positive) and the 68 rationalization hits, the second (a different model family) blind to the first's labels (Table A8). They agree on the two quantities the body uses: 0 of 110 regex-negative failure endings admit failure, and the strong-form regex precision is 69 of 70. They differ on one Qwen3.8-27B sample (claim versus truncated) and on the split of the rationalization hits (48 versus 49 verified attributions, 12 versus 6 bare dismissals, 8 versus 11 false positives); the verified-attribution share on which they agree (Qwen3.6-27B 2/8, Qwen3.6-27B+v12×2 12–13/15, teacher 10–11/15) is the only rationalization statistic the paper reports. No human pass was run.

Test ($A - B$)	Pool	b	c	Δ (pp)	p	p_{Holm}
Corpus family: 13 baseline-relative tests, Holm within $m = 13$						
Qwen3.6-27B+v12×1 − Qwen3.6-27B	Pro	47	50	-0.41	0.839	1.000
Qwen3.6-27B+v12×2 − Qwen3.6-27B	Pro	59	39	+2.74	0.054	0.653
Qwen3.6-27B+v12×3 − Qwen3.6-27B	Pro	59	52	+0.96	0.569	1.000
Qwen3.6-27B+v13×1 − Qwen3.6-27B	Pro	61	45	+2.19	0.145	1.000
Qwen3.6-27B+v12×1 − Qwen3.6-27B	Verified	38	41	-0.60	0.822	1.000
Qwen3.6-27B+v12×2 − Qwen3.6-27B	Verified	42	41	+0.20	1.000	1.000
Qwen3.6-27B+v12×3 − Qwen3.6-27B	Verified	45	31	+2.80	0.135	1.000
Qwen3.6-27B+v13×1 − Qwen3.6-27B	Verified	34	45	-2.20	0.260	1.000
Qwen3.5-35B-A3B-Base+v12×2 − Qwen3.5-35B-A3B-Base	Pro	115	50	+8.89	4.51×10^{-7}	5.87×10^{-6}
Qwen3.5-35B-A3B-Base+v12×2 − Qwen3.5-35B-A3B-Base	Verified	65	48	+3.40	0.132	1.000
Qwen3.6-35B-A3B+v12×3 − Qwen3.6-35B-A3B	Pro	67	47	+2.74	0.075	0.822
Qwen3.6-35B-A3B+v12×3 − Qwen3.6-35B-A3B	Verified	45	45	+0.00	1.000	1.000
Qwen3.6-35B-A3B+v12×3 (repeat) − Qwen3.6-35B-A3B (repeat)	Pro	70	52	+2.46	0.123	1.000
Generation / architecture family: 6 tests, Holm within $m = 6$						
Qwen3.8-27B − Qwen3.6-27B	Pro	75	34	+5.61	1.07×10^{-4}	5.37×10^{-4}
Qwen3.8-27B − Qwen3.6-27B	Verified	57	28	+5.80	0.002	0.009
Qwen3.6-27B − Qwen3.6-35B-A3B	Pro	70	41	+3.97	0.008	0.023
Qwen3.6-27B − Qwen3.6-35B-A3B	Verified	64	37	+5.40	0.009	0.023
Qwen3.6-27B+v12×3 − Qwen3.6-35B-A3B+v12×3	Pro	59	43	+2.19	0.137	0.137
Qwen3.6-27B+v12×3 − Qwen3.6-35B-A3B+v12×3	Verified	64	23	+8.20	1.27×10^{-5}	7.59×10^{-5}

Table A1: Pre-declared paired tests (`analysis/stats/stats_fixes.md`, §b). b = tasks solved by A only, c = by B only, $\Delta = (b - c)/N$ with $N = 731$ (Pro) or 500 (Verified) and a task missing from an arm counted as unsolved; p = two-sided exact binomial McNemar test on the discordant pairs; p_{Holm} = Holm-adjusted within the family shown. Positive Δ = the corpus (first family) or the first-named arm (second family) solved more. Over all 111 valid same-pool pairs, 74 are significant at $p < 0.05$ before correction, 49 survive Holm and 71 survive Benjamini–Hochberg at 0.05.

Test ($A - B$)	b+c	SE (pp)	MDE_{80} (pp)	N for 3 pp
Pro ($N = 731$)				
Qwen3.6-27B+v12×1 − Qwen3.6-27B	97	1.35	3.77	1,156
Qwen3.6-27B+v12×2 − Qwen3.6-27B	98	1.35	3.78	1,162
Qwen3.6-27B+v12×3 − Qwen3.6-27B	111	1.44	4.03	1,322
Qwen3.6-27B+v13×1 − Qwen3.6-27B	106	1.41	3.94	1,259
Qwen3.5-35B-A3B-Base+v12×2 − Qwen3.5-35B-A3B-Base	165	1.73	4.83	1,898
Qwen3.6-35B-A3B+v12×3 − Qwen3.6-35B-A3B	114	1.46	4.08	1,352
Qwen3.6-35B-A3B+v12×3 (repeat) − Qwen3.6-35B-A3B (repeat)	122	1.51	4.22	1,449
Verified ($N = 500$)				
Qwen3.6-27B+v12×1 − Qwen3.6-27B	79	1.78	4.98	1,377
Qwen3.6-27B+v12×2 − Qwen3.6-27B	83	1.82	5.10	1,447
Qwen3.6-27B+v12×3 − Qwen3.6-27B	76	1.74	4.87	1,318
Qwen3.6-27B+v13×1 − Qwen3.6-27B	79	1.77	4.97	1,373
Qwen3.5-35B-A3B-Base+v12×2 − Qwen3.5-35B-A3B-Base	113	2.12	5.94	1,959
Qwen3.6-35B-A3B+v12×3 − Qwen3.6-35B-A3B	90	1.90	5.31	1,568

Table A2: Detectable effect of the one-rollout paired design from the observed discordance (`stats_fixes.md`, §c1). $\text{SE} = \sqrt{(b + c) - (b - c)^2/N}/N$, $\text{MDE}_{80} = 2.8\,\text{SE}$ (80 % power, two-sided 0.05); the last column is the pool size at which the same discordance would give $\text{MDE}_{80} = 3$ pp with one rollout. Median MDE_{80} over the baseline-relative pairs: 4.03 pp on Pro and 5.04 pp on Verified; median N for 3 pp: 1,322 and 1,412 tasks.

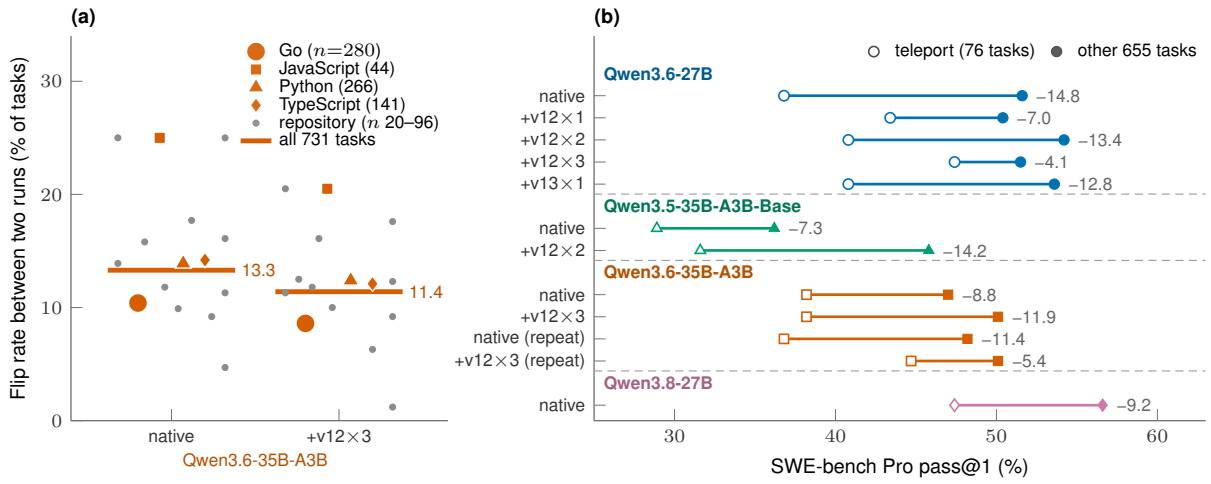


Figure A1: Robustness of the score tables on SWE-bench Pro. **(a)** Repeat samples: share of tasks solved in exactly one of two runs of the same weights, overall (bars), per language (symbols sized by n) and per repository (grey dots). **(b)** Teleport slice: pass@1 on the 76 teleport tasks (open) against the other 655 (filled) for the twelve arms, labelled with the difference in pp. Neither a second run nor the teleport repository reorders the arms.

	MDE_{80} (pp) at k rollouts per task			
N	$k=1$	$k=3$	$k=8$	$k\to\infty$
237	10.52	7.15	5.71	4.63
500	7.24	4.93	3.93	3.19
731	5.99	4.07	3.25	2.63
N for 3 pp	2,916	1,348	858	564

Table A3: Variance-decomposition projection of the detectable effect (`stats_fixes.md`, §c2): $\mathrm{Var}(\hat{\Delta}) = [\sigma^2_{\text{between}} + (\sigma^2_{\text{within},A} + \sigma^2_{\text{within},B})/k]/N$ with $\sigma^2_{\text{between}} = 0.0647$ and $\sigma^2_{\text{within},A} + \sigma^2_{\text{within},B} = 0.27$, all three estimated on the 237-task stratified rollout$\times$3 subsample, not on the evaluation pools. The decomposition implies a single-rollout discordance of 0.335, whereas the pools show 0.13–0.23 for the baseline-relative pairs (Table A2), so the $k=1$ row is an upper bound.

Stratum	n	flips	rate (%)	run 2 only	run 1 only	net Δ (pp)	p
Qwen3.6-35B-A3B (repeat) vs Qwen3.6-35B-A3B (same weights, Pro)							
all tasks	731	97	13.3	52	45	+0.96	0.543
Go	280	29	10.4	17	12	+1.79	–
JavaScript	44	11	25.0	7	4	+6.82	–
Python	266	37	13.9	18	19	-0.38	–
TypeScript	141	20	14.2	10	10	+0.00	–
repo range			4.7–25.0	(flipt $n = 85$ to NodeBB $n = 44$)			
Qwen3.6-35B-A3B+v12×3 (repeat) vs Qwen3.6-35B-A3B+v12×3 (same weights, Pro)							
all tasks	731	83	11.4	44	39	+0.68	0.661
Go	280	24	8.6	14	10	+1.43	–
JavaScript	44	9	20.5	4	5	-2.27	–
Python	266	33	12.4	15	18	-1.13	–
TypeScript	141	17	12.1	11	6	+3.55	–
repo range			1.2–20.5	(flipt $n = 85$ to NodeBB $n = 44$)			

Table A4: Run-to-run reference on SWE-bench Pro (`stats_fixes.md`, §d): the same weights run twice. A flip is a task solved in exactly one of the two runs; net Δ = solved(run 2) $-$ solved(run 1) over $N = 731$; p = exact McNemar. Restricting to tasks present in both runs gives the same rates (97/730 and 82/722). Per-language and per-repository rows show that the noise is not uniform across strata.

	teleport (76)		rest (655)		all (731)	
Arm	solved	%	solved	%	solved	%
Qwen3.6-27B						
native	28	36.8	338	51.6	366	50.1
+v12×1	33	43.4	330	50.4	363	49.7
+v12×2	31	40.8	355	54.2	386	52.8
+v12×3	36	47.4	337	51.5	373	51.0
+v13×1	31	40.8	351	53.6	382	52.3
Qwen3.5-35B-A3B-Base						
native	22	28.9	237	36.2	259	35.4
+v12×2	24	31.6	300	45.8	324	44.3
Qwen3.6-35B-A3B						
native	29	38.2	308	47.0	337	46.1
+v12×3	29	38.2	328	50.1	357	48.8
native (repeat)	28	36.8	316	48.2	344	47.1
+v12×3 (repeat)	34	44.7	328	50.1	362	49.5
Qwen3.8-27B						
native	36	47.4	371	56.6	407	55.7

Table A5: Solved counts on the 76 SWE-bench Pro tasks from the `teleport` repository (the one Pro repository that also appears in the corpus: 1,622 teacher trajectories) and on the remaining 655 tasks (`stats_fixes.md`, §e1). Column groups give the subset size N; denominator = subset size, a missing task counts as unsolved. SWE-bench Verified contains no teleport task.

Test ($A - B$)	b	c	Δ (pp)	p	p_{Holm}
Pro without teleport ($N = 655$), Holm within $m = 7$					
Qwen3.6-27B+v12×1 − Qwen3.6-27B	40	48	-1.22	0.456	0.911
Qwen3.6-27B+v12×2 − Qwen3.6-27B	52	35	+2.60	0.086	0.428
Qwen3.6-27B+v12×3 − Qwen3.6-27B	51	52	-0.15	1.000	1.000
Qwen3.6-27B+v13×1 − Qwen3.6-27B	54	41	+1.98	0.218	0.872
Qwen3.5-35B-A3B-Base+v12×2 − Qwen3.5-35B-A3B-Base	104	41	+9.62	1.70×10^{-7}	1.19×10^{-6}
Qwen3.6-35B-A3B+v12×3 − Qwen3.6-35B-A3B	59	39	+3.05	0.054	0.326
Qwen3.6-35B-A3B+v12×3 (repeat) − Qwen3.6-35B-A3B (repeat)	61	49	+1.83	0.294	0.883
teleport only ($N = 76$), Holm within $m = 7$					
Qwen3.6-27B+v12×1 − Qwen3.6-27B	7	2	+6.58	0.180	0.898
Qwen3.6-27B+v12×2 − Qwen3.6-27B	7	4	+3.95	0.549	1.000
Qwen3.6-27B+v12×3 − Qwen3.6-27B	8	0	+10.53	0.008	0.055
Qwen3.6-27B+v13×1 − Qwen3.6-27B	7	4	+3.95	0.549	1.000
Qwen3.5-35B-A3B-Base+v12×2 − Qwen3.5-35B-A3B-Base	11	9	+2.63	0.824	1.000
Qwen3.6-35B-A3B+v12×3 − Qwen3.6-35B-A3B	8	8	+0.00	1.000	1.000
Qwen3.6-35B-A3B+v12×3 (repeat) − Qwen3.6-35B-A3B (repeat)	9	3	+7.89	0.146	0.876

Table A6: The seven Pro baseline-relative tests recomputed without the teleport repository and on teleport alone (`stats_fixes.md`, §e2, e4). Conventions as in Table A1. Re-running the 13-test corpus family with the seven Pro tests replaced by their teleport-excluded versions (§e3) leaves the same single Holm survivor (Qwen3.5-35B-A3B-Base+v12×2 − Qwen3.5-35B-A3B-Base).

	SWE-bench Pro			SWE-bench Verified		
Arm	n	claim	rat.	n	claim	rat.
Teacher (failed)	10,483	44.6	7.0	–	–	–
Qwen3.6-27B						
native	359	66.3	2.2	154	14.9	0.6
+v12×1	367	52.9	8.7	154	18.2	16.2
+v12×2	336	49.1	8.9	151	21.2	18.5
+v12×3	354	48.9	9.0	130	20.0	16.9
+v13×1	349	44.1	8.6	165	10.3	20.6
Qwen3.5-35B-A3B-Base						
native	465	22.6	0.4	236	21.2	0.4
+v12×2	398	50.5	9.3	215	21.9	22.8
Qwen3.6-35B-A3B						
native	394	58.9	2.8	184	15.2	0.5
+v12×3	374	55.1	9.6	181	21.0	16.0
native (repeat)	386	63.2	1.0	–	–	–
+v12×3 (repeat)	360	58.3	8.6	–	–	–
Qwen3.8-27B						
native	323	24.1	13.9	124	12.1	28.2

Table A7: Endings audit. For each arm, n = unsolved trajectories; "claim" = share (%) whose final assistant message matches the strong-form success-claim regex; "rat." = share (%) matching the rationalization regex. The teacher row ("Teacher (failed)") is the corpus by-product failed population, 10,483 trajectories with no pool. Regex precision is 69/70 on the audited positives. Two model annotators, working blind, find 0 of 110 regex-negative failure endings that admit failure (the remainder claim completion in variant wording or stop mid-sentence at the context wall); of 68 rationalization hits, 48 carry baseline-verification evidence (re-running on a clean checkout) and 12 are bare dismissals.

Quantity	Annot. 1	Annot. 2
Regex-negative endings that admit failure	0/110	0/110
claim / truncated / admit: teacher	20/0/0	20/0/0
Qwen3.6-27B	20/0/0	20/0/0
Qwen3.6-27B+v12×2	20/0/0	20/0/0
Qwen3.6-27B+v13×1	13/2/0	13/2/0
Qwen3.8-27B	16/4/0	17/3/0
Qwen3.6-35B-A3B+v12×3	15/0/0	15/0/0
Strong-form regex precision (true / hits)	69/70	69/70
Rationalization hits (68): verified attribution	48	49
bare dismissal	12	6
false positive	8	11
grader-will-replace (Qwen3.6-35B-A3B+v12×3)	2‡	2
Verified-attribution share: teacher	10/15	11/15
Qwen3.6-27B	2/8	2/8
Qwen3.6-27B+v12×2	12/15	13/15
Qwen3.8-27B	13/15	13/15
Qwen3.6-35B-A3B+v12×3	11/15	10/15

Table A8: Agreement between the two model annotators of the endings audit (`second_annotator_agreement.md`); annotator 2 worked blind on the same 180 + 68 windows. Annotator 1's per-arm tallies sum to 110. ‡Counted inside the three categories for annotator 1 ($48 + 12 + 8 = 68$), separately for annotator 2 ($49 + 6 + 11 + 2 = 68$). The two agree on the admission and precision counts the body uses and differ on one Qwen3.8-27B sample and on the rationalization split.

B Fingerprint robustness

Setup. Every distance in this appendix is the trajectory-weighted divergence between an arm's step distribution and the teacher's high-precision population (12,257 trajectories) over the four rulers of §5: tool mix (11 buckets), tool bigrams, Bash command heads and thinking-position deciles, with the composite the mean of mix and bigram. Uncertainty is a trajectory-level bootstrap ($B = 1000$, seed 20260902; percentile 2.5/97.5, teacher fixed), reported as a bootstrap 95 % interval; the point estimates agree with `fp_summary.json` to 10^{-15}.

Consecutive-arm differences. Figure A4 draws the intervals; Table A9 gives the change in JS distance between arms with its bootstrap interval. On Pro, Qwen3.6-27B →Qwen3.6-27B+v12×1 is above noise on Bash (-0.133), bigram (-0.010), thinking position (-0.008) and the composite (-0.006) but not on mix; Qwen3.6-27B+v12×1 →Qwen3.6-27B+v12×2 moves only Bash (-0.016, about seven times the same-weights floor); Qwen3.6-27B+v12×2 →Qwen3.6-27B+v12×3 moves nothing. "One epoch does most of it" is therefore a difference statement, and the residual second-epoch effect exists only in the Bash vocabulary. On Verified the mix and bigram rulers move away from the teacher for every v12 checkpoint while Bash converges.

Alternative distances. Table A10 repeats the epoch ordering under Hellinger and total-variation distance. Hellinger preserves every JS ordering and the habit/strategy split on Verified; total variation does not preserve the Pro tool-mix ordering, because the SFT arms overshoot the teacher's Bash share (Table A11: teacher 50.7 %, Qwen3.6-27B 54.5 %, Qwen3.6-27B+v12×1/Qwen3.6-27B+v12×2/Qwen3.6-27B+v12×3 62.3 / 60.1 / 61.8 %), which a linear distance registers while JS is dominated by the collapsed small buckets. Tool-mix "convergence" is thus ruler-dependent; Bash vocabulary and thinking position are not.

Noise floors. Table A12 gives the two references against which the body quotes "above noise": the same-weights repeat samples (composite 0.0008 / 0.0003, Bash 0.0021 / 0.0011 at $n \approx 730$) and split-half distances within one arm (composite ≈0.0008, Bash 0.004–0.005; Qwen3.6-35B-A3B Bash 0.015 from its heterogeneous vocabulary). The sampling floor at student sample size is therefore about 0.001 on the composite and 0.002–0.005 on Bash; the "teacher-internal 0.002–0.003" figure is the distance between teacher sub-populations, not sampling noise.

Pool confound. Distances are comparable only within a pool: on the 88-task anchor Qwen3.6-27B+v12×2's composite is 0.0090 against 0.0215 on Pro, and Qwen3.6-27B's 0.0493 against 0.0285; the pool effect is larger than the residual dose effect and is not additive.

Post-error thinking. Figure A5 draws the two levels per arm; Table A13 reports absolute thinking characters per step after a tool error and on all other steps. The teacher thinks less after an error than on ordinary steps (248 vs 726 characters, ratio 0.34); Qwen3.6-27B does not (373 vs 359, 1.04); after SFT the student thinks more after an error in absolute terms (Qwen3.6-27B+v12×2 580) and its ratio falls to 0.48 because ordinary-step thinking triples (1,199). What transfers is the teacher's pattern, not a reduction in post-error thinking; the paper reports both numbers.

Language strata. Figure A6 draws the strata; Table A14 restricts the Pro distance to one language at a time and, for Python, adds a Python-only teacher reference ($n = 574$). Bash-vocabulary convergence holds inside every language (Qwen3.6-27B →Qwen3.6-27B+v12×2: Python 0.341→0.133, Go 0.217→0.066, JS/TS 0.237→0.105); with the Python reference the composite falls on the Pro Python slice (0.0267→0.0195→0.0178 for Qwen3.6-27B, Qwen3.6-27B+v12×2, Qwen3.6-27B+v12×3) but rises on Verified (0.0369→0.0476→0.0465), the strategy-layer divergence noted in §5.

Anchor diagnostics. Tables A15 and A16 document the 88-task anchor of Figure A3. Both jobs ran exactly three rollouts per task (263 and 296 parsed trajectories: one void Qwen3.6-27B trial; the Qwen3.6-27B+v12×2 job also covers the 12 tasks later removed by the oracle golden-patch filter). The student reads fewer files (median 11→7; teacher 6) at higher precision against the teacher's read set (0.50→0.80, 75:8 tasks) and similar recall (0.86→0.80, 18:30, n.s.), so its read-set Jaccard rises (0.44→0.60, 66:15) while the negative controls stay at 0.001 (another task's teacher set) and 0.03 (same repository, different task; 45:38, $p = 0.51$); the gain survives on the 15 tasks where Qwen3.6-27B+v12×2 reads as many or more files

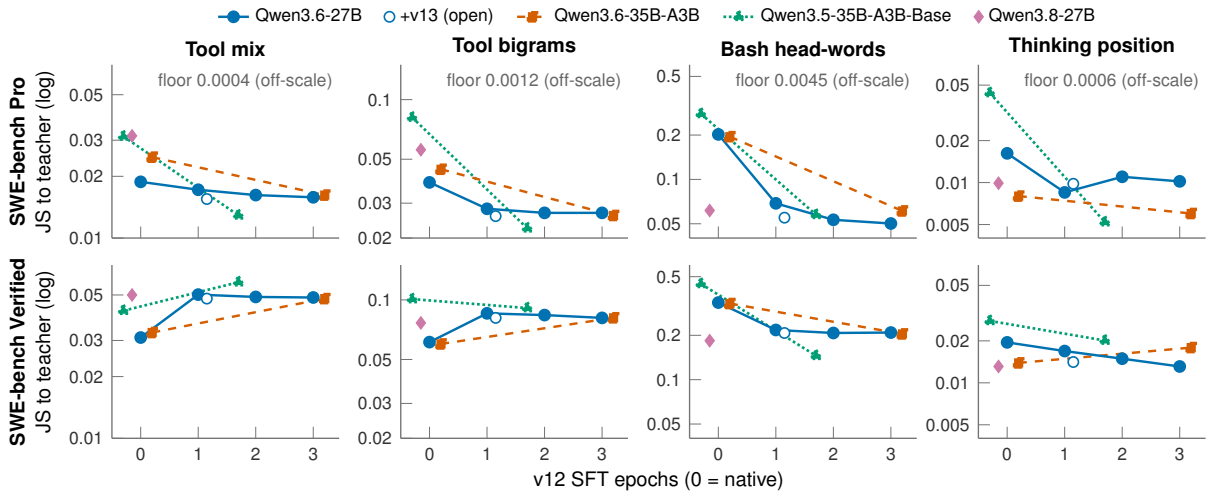


Figure A2: Jensen–Shannon distance to the pooled teacher reference on four rulers (columns) and two pools (rows) against v12 epochs: dense line (circles; open circle at epoch 1 = Qwen3.6-27B+v13×1), post-trained MoE (squares), bare MoE (triangles), Qwen3.8-27B (diamond); the split-half sampling floors (0.0004–0.0045, noted in each Pro panel) lie one to two decades below every plotted point. On SWE-bench Pro every ruler falls at the first epoch and then flattens; on SWE-bench Verified the Bash and thinking-position rulers fall while tool mix and bigrams move away from the teacher, the task-sensitive composition layer.

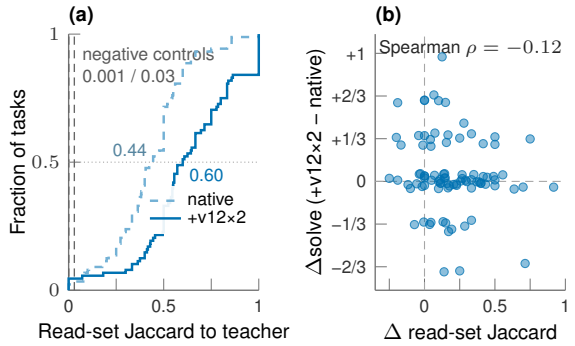


Figure A3: Same-task anchor: 88 v12 training tasks replayed three times by Qwen3.6-27B and Qwen3.6-27B+v12×2, per-task medians. (a) Distribution over tasks of read-set Jaccard to the teacher's read set (median 0.44 to 0.60; Qwen3.6-27B+v12×2 closer on 66 of 88 tasks); dashed: negative controls, another task's teacher set (0.001) and a same-repository different-task set (0.03). The student reads fewer files (median 11 to 7; teacher 6) at higher precision (0.50 to 0.80) and similar recall (0.86 to 0.80). (b) Per-task change in solve fraction against change in read-set Jaccard (y jittered): imitation gain does not track solve gain (Spearman $\rho = -0.12$; mean solve 55.3 to 61.6 %).

(+0.054, 8:3). Table A16 shows that the two published solve-rate p-values (0.073 and 0.099) differ only in whether a void trial counts as unsolved, a gap of one task; the interval is a plain task bootstrap, not a stratified estimate. Per-task solve gain does not track imitation gain (Spearman $\rho = -0.12$, Figure A3b).

Thinking-position profiles. Figure A7 draws the decile profiles behind the thinking-position ruler: the native models spread thinking evenly or load it into the last decile, whereas the fine-tuned arms and Qwen3.8-27B reproduce the teacher's early hump on Pro; on Verified the fine-tuned profiles flatten, which is why the thinking-position distance moves less there.

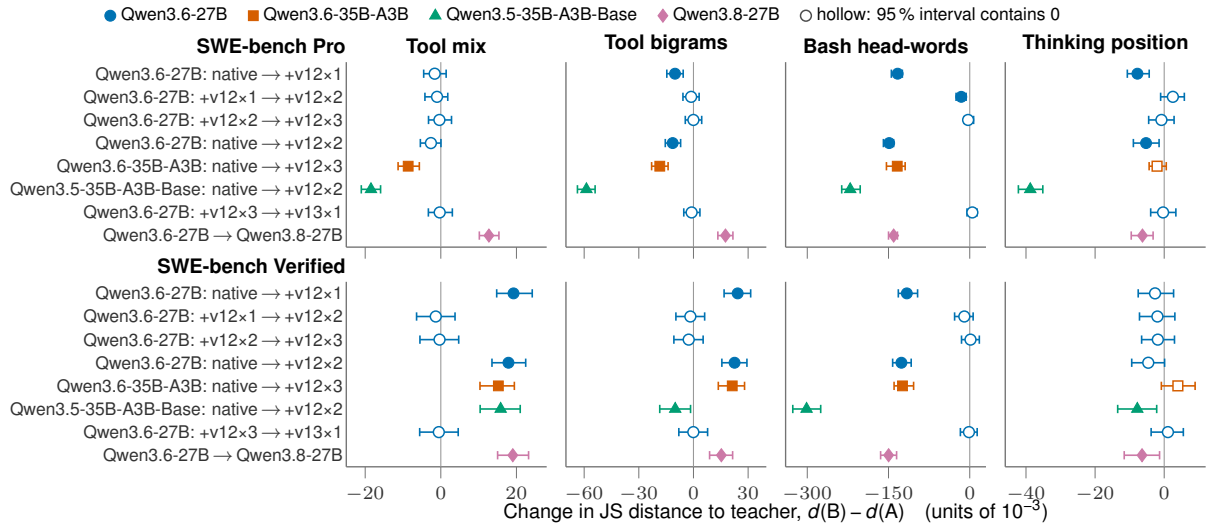


Figure A4: Change in JS distance to the teacher between arms, $d(B) - d(A)$, in units of 10^{-3}, with trajectory-level bootstrap 95 % intervals (teacher fixed; Table A9); filled markers = interval excludes zero. Every native→v12 pair moves Bash head-words toward the teacher in both pools; the second and third dense epochs and the Qwen3.6-27B+v13×1 swap are indistinguishable from zero on every ruler; on SWE-bench Verified tool mix and bigrams move away from the teacher.

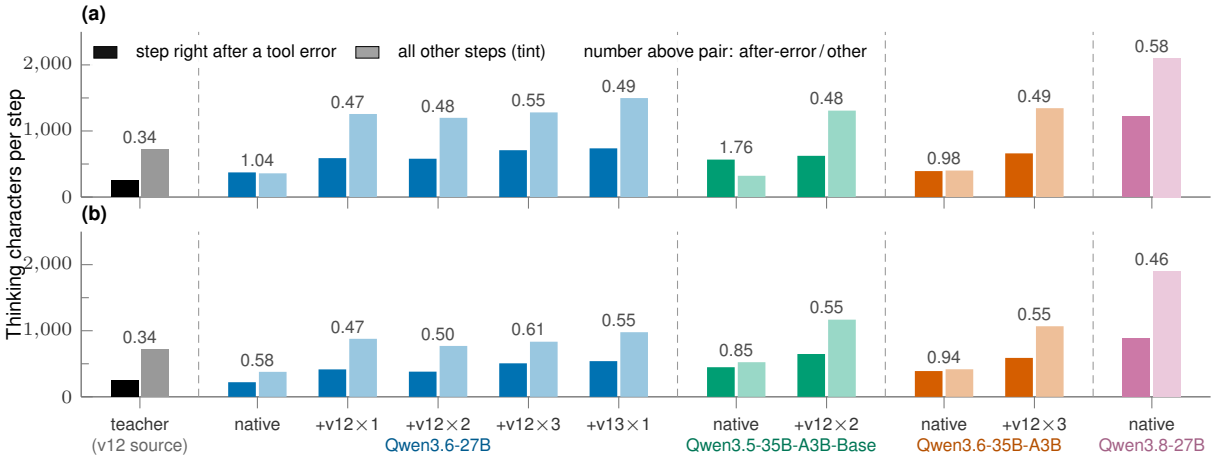


Figure A5: Mean thinking characters per assistant step on the step after a tool error (filled) and on all other steps (tint), with their ratio, on SWE-bench Pro (a) and SWE-bench Verified (b). The teacher and every fine-tuned arm think about half as much after an error (ratios 0.34 and 0.47–0.61); the native models think as much or more (1.04, 1.76 and 0.98 on Pro).

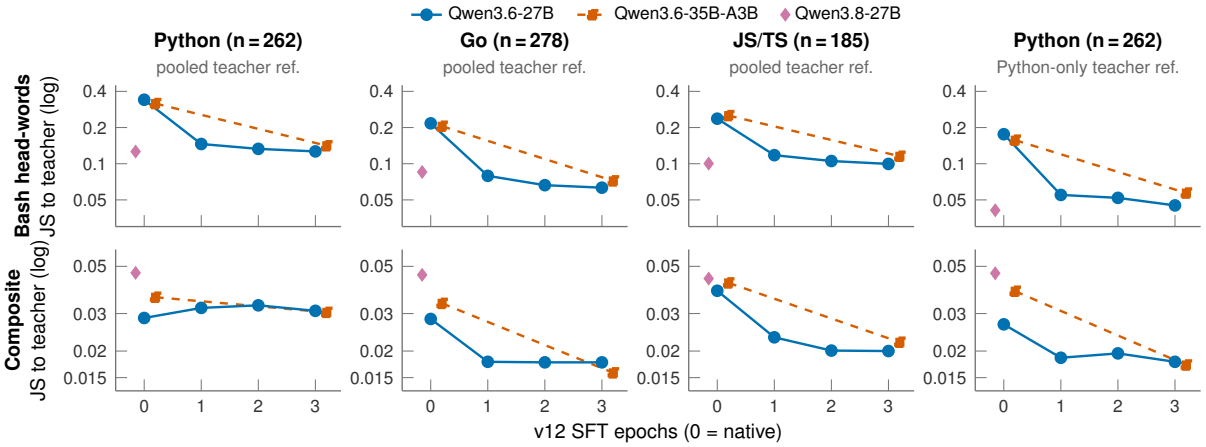


Figure A6: Within-language replication on SWE-bench Pro: Bash head-word (top) and composite (bottom) JS distance to the teacher inside the Python, Go and JS/TS strata against v12 epochs, and the Python stratum measured against a Python-only teacher reference (right column). Bash vocabulary converges in every stratum; the composite is flat in Python only against the pooled reference and falls against the Python-only one, the pool effect of §5.

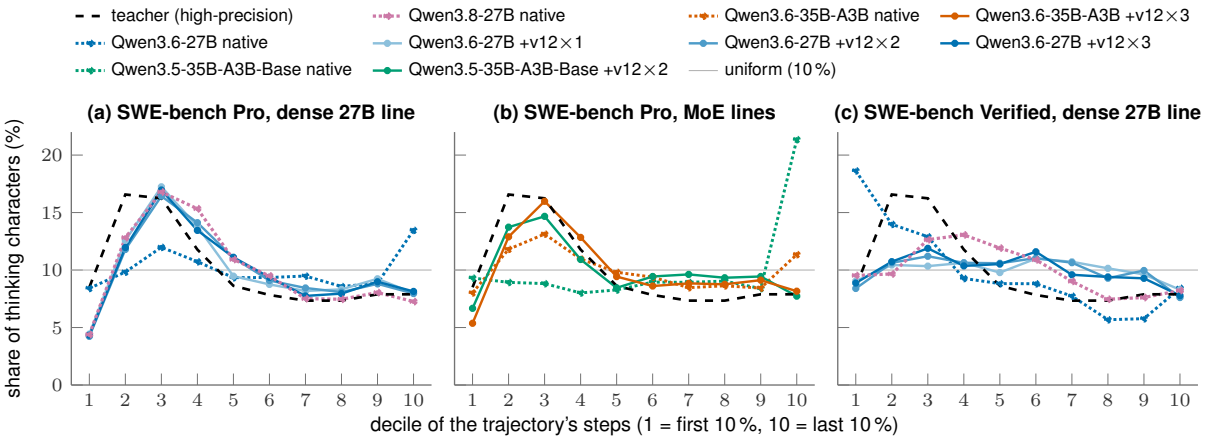


Figure A7: Share of thinking characters per decile of the trajectory, the distribution behind the thinking-position ruler. **(a)** SWE-bench Pro, dense line; **(b)** SWE-bench Pro, MoE lines; **(c)** SWE-bench Verified, dense line. Native models think flat across the trajectory (the bare base loads 21 % into the last decile); every fine-tuned arm and Qwen3.8-27B reproduce the teacher's early hump at deciles 2–4; on Verified the fine-tuned curves flatten toward uniform.

Pair ($A \to B$)	mix	bigram	Bash	think-pos.
Pro				
Qwen3.6-27B: native → +v12×1	$-1.6\ [-4.5, 1.4]$	$-10.1\ [-14.6, -5.6]^*$	$-133.1\ [-144.7, -124.4]^*$	$-7.8\ [-10.7, -4.3]^*$
Qwen3.6-27B: +v12×1 → +v12×2	$-1.0\ [-4.2, 1.8]$	$-1.3\ [-5.8, 3.1]$	$-15.7\ [-25.2, -6.7]^*$	$2.5\ [-1.0, 5.8]$
Qwen3.6-27B: +v12×2 → +v12×3	$-0.4\ [-3.2, 2.8]$	$-0.0\ [-4.5, 4.5]$	$-3.0\ [-10.4, 7.2]$	$-0.8\ [-4.5, 2.9]$
Qwen3.6-27B: native → +v12×2	$-2.6\ [-5.4, 0.0]$	$-11.4\ [-15.5, -7.0]^*$	$-148.8\ [-159.4, -140.6]^*$	$-5.2\ [-8.9, -1.5]^*$
Qwen3.6-35B-A3B: native → +v12×3	$-8.6\ [-11.3, -5.7]^*$	$-18.5\ [-23.0, -13.9]^*$	$-133.9\ [-153.8, -119.3]^*$	$-2.1\ [-4.4, 0.6]$
Qwen3.5-35B-A3B-Base: native → +v12×2	$-18.5\ [-21.0, -15.9]^*$	$-58.8\ [-63.7, -54.0]^*$	$-220.7\ [-236.4, -202.3]^*$	$-38.8\ [-42.2, -35.1]^*$
Qwen3.6-27B: +v12×3 → +v13×1	$-0.3\ [-3.3, 3.1]$	$-1.0\ [-5.3, 3.5]$	$4.8\ [-5.1, 13.4]$	$-0.3\ [-3.9, 3.4]$
Qwen3.6-27B → Qwen3.8-27B	$12.7\ [10.2, 15.4]^*$	$17.6\ [13.4, 21.7]^*$	$-140.5\ [-150.0, -133.5]^*$	$-6.3\ [-9.5, -3.2]^*$
Verified				
Qwen3.6-27B: native → +v12×1	$19.2\ [14.8, 24.2]^*$	$24.3\ [16.8, 31.4]^*$	$-116.2\ [-131.9, -96.3]^*$	$-2.7\ [-7.5, 2.7]$
Qwen3.6-27B: +v12×1 → +v12×2	$-1.3\ [-6.4, 3.8]$	$-1.7\ [-9.6, 6.1]$	$-10.0\ [-28.0, 6.3]$	$-2.0\ [-7.1, 3.1]$
Qwen3.6-27B: +v12×2 → +v12×3	$-0.3\ [-5.5, 4.7]$	$-2.6\ [-10.8, 5.3]$	$1.4\ [-15.2, 17.7]$	$-1.8\ [-6.5, 3.0]$
Qwen3.6-27B: native → +v12×2	$17.9\ [13.5, 22.4]^*$	$22.6\ [15.6, 29.4]^*$	$-126.2\ [-142.3, -108.1]^*$	$-4.6\ [-9.4, 0.1]$
Qwen3.6-35B-A3B: native → +v12×3	$15.2\ [10.4, 19.4]^*$	$21.3\ [13.6, 28.0]^*$	$-124.2\ [-139.7, -103.7]^*$	$4.0\ [-0.8, 9.0]$
Qwen3.5-35B-A3B-Base: native → +v12×2	$15.8\ [10.4, 21.0]^*$	$-10.0\ [-18.5, -1.6]^*$	$-301.3\ [-326.7, -275.2]^*$	$-7.8\ [-13.5, -2.2]^*$
Qwen3.6-27B: +v12×3 → +v13×1	$-0.5\ [-5.6, 4.6]$	$-0.1\ [-8.0, 7.8]$	$-1.5\ [-17.3, 13.7]$	$1.1\ [-3.8, 5.5]$
Qwen3.6-27B → Qwen3.8-27B	$19.0\ [15.0, 23.2]^*$	$15.3\ [8.8, 21.5]^*$	$-149.8\ [-164.6, -135.1]^*$	$-6.4\ [-11.6, -1.3]^*$

Table A9: Change in JS distance to the teacher between arms, $d(B) - d(A)$, in units of 10^{-3}, with the bootstrap 95 % interval (`analysis/fp/fp_ci.md`, §b; $B = 1000$ trajectory-level replicates with independent resampling of both arms, teacher fixed). Negative = B is closer to the teacher. * = the interval excludes 0. The composite (mean of mix and bigram) is omitted; its intervals are in the source file. On Pro the first epoch moves bigram, Bash, thinking position and the composite, the second only Bash, the third nothing; on Verified only Bash and (for two of four checkpoints) thinking position move toward the teacher.

			Qwen3.6-27B line						
Pool	Distance	Ruler	native	+v12×1	+v12×2	+v12×3	monotone	epoch-1 share	steps ≠ 0
Pro	JS	mix	0.0188	0.0172	0.0162	0.0158	yes	0.54	nnn
Pro	JS	bigram	0.0382	0.0281	0.0268	0.0268	yes	0.89	Ynn
Pro	JS	Bash	0.2020	0.0688	0.0532	0.0502	yes	0.88	YYn
Pro	JS	think-pos.	0.0162	0.0085	0.0110	0.0102	no	1.28	Ynn
Pro	JS	composite	0.0285	0.0226	0.0215	0.0213	yes	0.81	Ynn
Pro	Hellinger	mix	0.1154	0.1123	0.1107	0.1077	yes	0.41	nnn
Pro	Hellinger	bigram	0.1665	0.1444	0.1428	0.1408	yes	0.86	Ynn
Pro	Hellinger	Bash	0.3985	0.2397	0.2090	0.2038	yes	0.82	YYn
Pro	Hellinger	think-pos.	0.1062	0.0767	0.0874	0.0841	no	1.33	Ynn
Pro	Hellinger	composite	0.1410	0.1283	0.1267	0.1242	yes	0.76	Ynn
Pro	total variation	mix	0.1074	0.1159	0.0934	0.1107	no	2.55	nYY
Pro	total variation	bigram	0.1524	0.1452	0.1265	0.1434	no	0.80	nYY
Pro	total variation	Bash	0.3663	0.1711	0.1350	0.1284	yes	0.82	YYn
Pro	total variation	think-pos.	0.1214	0.0806	0.0905	0.0878	no	1.21	Ynn
Pro	total variation	composite	0.1299	0.1305	0.1099	0.1271	no	-0.23	nYY
Verified	JS	mix	0.0310	0.0502	0.0489	0.0486	no	1.10	Ynn
Verified	JS	bigram	0.0612	0.0855	0.0838	0.0811	no	1.22	Ynn
Verified	JS	Bash	0.3336	0.2174	0.2074	0.2088	no	0.93	Ynn
Verified	JS	think-pos.	0.0195	0.0169	0.0149	0.0131	yes	0.41	nnn
Verified	JS	composite	0.0461	0.0679	0.0663	0.0649	no	1.16	Ynn
Verified	Hellinger	mix	0.1480	0.1932	0.1899	0.1884	no	1.12	Ynn
Verified	Hellinger	bigram	0.2139	0.2553	0.2524	0.2479	no	1.22	Ynn
Verified	Hellinger	Bash	0.5281	0.4236	0.4147	0.4145	yes	0.92	Ynn
Verified	Hellinger	think-pos.	0.1168	0.1083	0.1018	0.0953	yes	0.40	nnn
Verified	Hellinger	composite	0.1810	0.2242	0.2212	0.2182	no	1.16	Ynn
Verified	total variation	mix	0.1609	0.2130	0.2127	0.2118	no	1.02	Ynn
Verified	total variation	bigram	0.1969	0.2458	0.2439	0.2341	no	1.31	Ynn
Verified	total variation	Bash	0.4868	0.3747	0.3727	0.3847	no	1.10	Ynn
Verified	total variation	think-pos.	0.1222	0.1318	0.1250	0.1172	no	-1.94	nnn
Verified	total variation	composite	0.1789	0.2294	0.2283	0.2230	no	1.15	Ynn

Table A10: Epoch ordering along the Qwen3.6-27B line (native, +v12×1, +v12×2, +v12×3) under three distances to the teacher (`fp_ci.md`, §c). "monotone" = non-increasing along the epochs; "epoch-1 share" = share of the native→+v12×3 drop realised by the first epoch; "steps" = whether the bootstrap 95 % interval of each adjacent step (1→2, 2→3, 3→4) excludes 0 (Y/n). Hellinger preserves every JS ordering; total variation does not preserve the Pro tool-mix ordering because the SFT arms overshoot the teacher's Bash share (Table A11).

Pool	Arm	Bash	Read	Edit	Write
corpus	Teacher	50.7	21.9	21.6	2.1
Pro	Qwen3.6-27B	54.5	25.5	11.0	2.5
Pro	Qwen3.6-27B+v12×1	62.3	18.8	16.7	1.6
Pro	Qwen3.6-27B+v12×2	60.1	20.2	17.4	1.7
Pro	Qwen3.6-27B+v12×3	61.8	19.0	16.5	1.9
Pro	Qwen3.6-27B+v13×1	61.6	19.0	17.0	1.6
Pro	Qwen3.8-27B	62.4	15.0	11.2	3.6
Pro	Qwen3.5-35B-A3B-Base	50.2	19.4	21.5	8.6
Pro	Qwen3.5-35B-A3B-Base+v12×2	60.8	18.2	18.4	1.7
Pro	Qwen3.6-35B-A3B	53.2	24.0	11.7	1.1
Pro	Qwen3.6-35B-A3B+v12×3	62.5	18.1	17.0	1.7
Verified	Qwen3.6-27B	66.1	20.6	8.9	1.5
Verified	Qwen3.6-27B+v12×1	72.0	16.6	11.0	0.2
Verified	Qwen3.6-27B+v12×2	72.0	16.6	11.0	0.3
Verified	Qwen3.6-27B+v12×3	71.9	17.0	10.6	0.2
Verified	Qwen3.6-27B+v13×1	71.8	16.6	11.1	0.2
Verified	Qwen3.8-27B	73.6	15.8	7.1	1.5
Verified	Qwen3.5-35B-A3B-Base	61.7	17.9	11.6	8.2
Verified	Qwen3.5-35B-A3B-Base+v12×2	75.0	14.8	9.8	0.2
Verified	Qwen3.6-35B-A3B	65.0	20.3	9.5	0.6
Verified	Qwen3.6-35B-A3B+v12×3	71.7	16.4	11.5	0.2

Table A11: Share (%) of tool calls in the four largest tool-mix buckets, trajectory-weighted (`bench/scripts/paper/out/fp_summary.json`). The SFT arms overshoot the teacher's Bash share on Pro (teacher 50.7 %, Qwen3.6-27B 54.5 %, its +v12×1/+v12×2/+v12×3 checkpoints 62.3/60.1/61.8 %), which a linear distance registers and JS, dominated by the collapsed small buckets, does not.

Reference	n	mix	bigram	Bash	think-pos.	composite
Same weights run twice (Pro): JS between the two runs, bootstrap 95 % interval						
Qwen3.6-35B-A3B, Pro: T=0.8 vs 0.6	731	0.4 [0.2, 1.4]	1.1 [1.0, 3.0]	2.1 [2.3, 23.9]	0.3 [0.2, 1.3]	0.8 [0.6, 2.2]
Qwen3.6-35B-A3B+v12×3, Pro: T=0.8 vs 0.6	731	0.1 [0.1, 0.7]	0.5 [0.4, 1.7]	1.1 [1.4, 6.8]	0.3 [0.3, 1.7]	0.3 [0.2, 1.2]
Split-half within one group: JS between two random halves, mean [2.5 %, 97.5 %] over 200 splits						
Qwen3.6-27B (Pro)	362×2	0.4 [0.1, 1.0]	1.2 [0.6, 2.3]	4.5 [1.8, 7.6]	0.6 [0.2, 1.5]	0.8 [0.4, 1.7]
Qwen3.6-27B+v12×2 (Pro)	361×2	0.3 [0.0, 0.8]	1.3 [0.7, 2.1]	3.9 [1.6, 7.5]	1.2 [0.3, 2.8]	0.8 [0.4, 1.4]
Qwen3.8-27B (Pro)	365×2	0.2 [0.0, 0.8]	0.7 [0.3, 1.4]	1.3 [0.6, 2.8]	0.8 [0.2, 1.8]	0.4 [0.2, 1.2]
Qwen3.6-35B-A3B (Pro)	365×2	0.4 [0.1, 0.9]	1.4 [0.8, 2.1]	14.9 [10.7, 20.5]	0.7 [0.2, 1.4]	0.9 [0.5, 1.5]
Teacher, 725-trajectory subsample	362×2	0.5 [0.1, 1.1]	1.7 [0.9, 2.8]	2.4 [0.7, 6.0]	1.5 [0.4, 2.9]	1.1 [0.5, 1.9]
Teacher, high-precision (v12 source)	6128×2	0.0 [0.0, 0.1]	0.1 [0.1, 0.2]	0.1 [0.0, 0.2]	0.1 [0.0, 0.2]	0.1 [0.0, 0.1]

Table A12: Noise references for the fingerprint distances, JS in units of 10^{-3} (`fp_ci.md`, §d). The same-weights repeat runs give the run-to-run floor at the arm sample size ($n \approx 730$); the split-half rows give the sampling floor at half the arm size, for two student arms, the vendor model, the MoE base (whose heterogeneous Bash vocabulary raises its Bash floor) and the teacher, once subsampled to 725 trajectories and once in full. The "teacher-internal 0.002–0.003" figure quoted elsewhere is the distance between teacher sub-populations, not sampling noise.

	trajectories		steps after an error			all other steps			mean ratio	
Group	all	w/ error	n	mean	med.>0	n	mean	med.>0	err/other	err/ok
Teacher, high-precision (v12 source)	12,257	9,399	28,164	248	145	711,171	726	487	0.34	0.34
Teacher by-product, passed	9,843	5,787	12,349	315	210	447,281	727	444	0.43	0.43
Teacher by-product, failed	10,483	7,950	23,970	297	197	735,848	897	553	0.33	0.33
Pro										
Qwen3.6-27B	725	604	1,819	373	86	38,585	359	111	1.04	1.04
Qwen3.6-27B+v12×1	730	514	1,413	589	145	32,449	1257	253	0.47	0.46
Qwen3.6-27B+v12×2	722	450	996	580	133	28,881	1199	253	0.48	0.47
Qwen3.6-27B+v12×3	727	453	971	709	134	31,928	1281	262	0.55	0.54
Qwen3.6-27B+v13×1	731	452	1,153	737	139	33,734	1498	286	0.49	0.48
Qwen3.8-27B	730	550	1,414	1222	196	48,230	2104	344	0.58	0.58
Qwen3.5-35B-A3B-Base	731	719	7,994	567	107	43,945	322	95	1.76	1.84
Qwen3.5-35B-A3B-Base+v12×2	722	601	3,303	624	188	42,459	1308	311	0.48	0.47
Qwen3.6-35B-A3B	731	599	2,012	393	96	40,012	401	113	0.98	0.98
Qwen3.6-35B-A3B+v12×3	731	528	1,554	661	142	38,892	1345	286	0.49	0.48
Verified										
Qwen3.6-27B	499	476	2,220	221	65	20,069	378	95	0.58	0.59
Qwen3.6-27B+v12×1	500	445	1,917	415	121	19,779	878	185	0.47	0.46
Qwen3.6-27B+v12×2	494	423	1,394	381	104	18,274	770	200	0.50	0.48
Qwen3.6-27B+v12×3	488	433	1,376	507	104	18,776	834	202	0.61	0.60
Qwen3.6-27B+v13×1	500	436	1,597	540	105	21,112	978	212	0.55	0.55
Qwen3.8-27B	499	470	2,006	886	170	27,817	1906	333	0.46	0.47
Qwen3.5-35B-A3B-Base	500	499	6,005	448	92	26,419	524	100	0.85	0.87
Qwen3.5-35B-A3B-Base+v12×2	500	458	3,126	648	176	31,918	1168	289	0.55	0.55
Qwen3.6-35B-A3B	500	487	2,804	390	72	23,222	417	96	0.94	0.94
Qwen3.6-35B-A3B+v12×3	500	455	1,944	589	108	28,372	1068	232	0.55	0.54

Table A13: Absolute post-error thinking (`fp_ci.md`, §e): thinking characters per assistant step when the previous step returned at least one tool error versus every other step. "med.>0" = median over steps that carry a thinking block; "err/ok" divides by steps that follow a clean tool step (the `react_ratio` of `fp_pooled.json`). After SFT the student thinks more after an error in absolute terms (Qwen3.6-27B 373 → Qwen3.6-27B+v12×2 580 on Pro) and the ratio falls because ordinary-step thinking rises more; what transfers is the teacher's pattern of less thinking after errors than on ordinary steps.

	Python			Go			JS/TS			Python ref., Pro-Python		Python ref., Verified	
Arm	comp.	Bash	n	comp.	Bash	n	comp.	Bash	n	comp.	Bash	comp.	Bash
Qwen3.6-27B	0.0286	0.341	262	0.0283	0.217	278	0.0384	0.237	185	0.0267	0.176	0.0369	0.136
Qwen3.6-27B+v12×1	0.0319	0.146	265	0.0178	0.079	280	0.0232	0.118	185	0.0186	0.055	0.0488	0.083
Qwen3.6-27B+v12×2	0.0328	0.133	262	0.0177	0.066	278	0.0201	0.105	182	0.0195	0.052	0.0476	0.073
Qwen3.6-27B+v12×3	0.0309	0.127	263	0.0177	0.063	279	0.0200	0.100	185	0.0178	0.045	0.0465	0.071
Qwen3.6-35B-A3B	0.0358	0.317	266	0.0335	0.205	280	0.0418	0.251	185	0.0382	0.157	0.0388	0.134
Qwen3.6-35B-A3B+v12×3	0.0304	0.141	266	0.0158	0.072	280	0.0220	0.115	185	0.0172	0.057	0.0460	0.082
Qwen3.8-27B	0.0466	0.126	266	0.0456	0.085	280	0.0438	0.100	184	0.0464	0.041	0.0525	0.076

Table A14: Per-language strata on SWE-bench Pro (`bench/scripts/paper/out/lang_robustness.json`): composite and Bash-vocabulary JS to the pooled teacher within each language slice (n = trajectories in the slice). The last four columns use a Python-only teacher reference ($n = 574$ teacher trajectories) for the Pro Python slice and for SWE-bench Verified. Bash vocabulary converges inside every language (Qwen3.6-27B →Qwen3.6-27B+v12×2: Python 0.341→0.133, Go 0.217→0.066, JS/TS 0.237→0.105); with the Python reference the composite falls on Pro-Python (0.0267→0.0195→0.0178) but rises on Verified.

	Qwen3.6-27B			
Metric	native	+v12×2	+v12×2:native	sign p
files read	11.000	7.000	7:73	< 0.001
read precision	0.500	0.800	75:8	< 0.001
read recall	0.857	0.800	18:30	0.111
read Jaccard	0.444	0.604	66:15	< 0.001
control: other task	0.000	0.000	4:9	0.267
control: same repo	0.029	0.032	45:38	0.510
files edited	6.000	5.000	26:34	0.366
edit precision	0.809	0.857	42:21	0.011
edit recall	0.833	1.000	30:13	0.014
edit Jaccard	0.652	0.750	44:24	0.021
control: other task	0.000	0.000	6:7	1.000
control: same repo	0.021	0.030	49:29	0.031

Table A15: Same-task anchor, read and edit file sets against the teacher's on the 88 v12 training tasks (`analysis/fp/anchor_diag.md`, §b; per-rollout metric → per-task median → median over tasks; means are in the source file; three rollouts per task). Teacher medians: 6 files read, 5 edited. "+v12×2:native" counts tasks on which each arm is higher; sign p = two-sided sign test. Controls: Jaccard against another task's teacher set and against a same-repository different-task teacher set. The student reads fewer files and they are the teacher's files; the Jaccard gain survives on the 15 tasks where Qwen3.6-27B+v12×2 reads as many or more files ($+0.054$, 8:3).

	k/n		mean solve (%)				sign test W:L:T (p)	
Rule	native	+v12×2	native	+v12×2	Δ (pp)	95 % int.	task means	task majority
all rollouts, void dropped	145/263	161/259	55.3	61.6	+6.2	$[-0.0,\ +12.3]$	24:13:51 (0.099)	14:5:69 (0.064)
first 3, void = unsolved	145/264	161/264	54.9	61.0	+6.1	$[-0.0,\ +12.1]$	25:13:50 (0.073)	13:5:70 (0.096)
first 3, void dropped	145/263	161/259	55.3	61.6	+6.2	$[-0.0,\ +12.3]$	24:13:51 (0.099)	14:5:69 (0.064)

Table A16: Same-task anchor (Qwen3.6-27B native versus +v12×2), solve rate under the three accounting rules that produced the two published p-values (`anchor_diag.md`, §c). "void" = a trial with no session files and no verdict (one Qwen3.6-27B trial; five Qwen3.6-27B+v12×2 trials without a verdict, four of them without session files). The interval is a plain task bootstrap (5000 draws) of the mean per-task difference, not a stratified estimate. Sign tests are exact and two-sided (wins:losses:ties). Per-task Δsolve against Δread-Jaccard: Spearman $\rho = -0.12$ (task-bootstrap 95 % interval $[-0.31,\ 0.08]$).

C Additional analyses

Funnel, drop rules and rule impact. Table A17 tabulates the v12 funnel of §3 with the v13 column and the message-level loss masks; Table A18 lists the ten v12 P0 drop rules with their thresholds, hits and sole hits, and Table A19 gives the numbers behind Figure 2: for each active v12 rule family, the rows it touches and removes alone, the median step count and thinking share of the removed rows, their JS distance to the kept set on tool mix and Bash head-words next to the size-matched random-removal floor, and their protocol-error recovery.

Leak audit and decontamination. A golden-patch leak audit with three negative-control channels (network, local git history, package manager) leaves every conclusion unchanged on the clean subsets. At the instance level the corpus shares no task with either pool: 0 title matches, a maximum task-description Jaccard of 0.21 and 0 commit-SHA hits; the SWE-bench Verified repositories have 0 overlap with the corpus. At the repository level one SWE-bench Pro repository, `teleport`, appears in the corpus with 1,622 teacher trajectories (1,309 pull requests) and supplies 76 of the 731 Pro tasks; the teleport-excluded tests (Tables A5, A6) change no decision.

Cost decomposition. Figure A8 bridges the median cumulative input of §6 through its two factors and splits tool-output characters by source. On the MoE line, replicated in both runs, whole-file reads fall from 7,293 to 2,578 at unchanged size per read (8,975→8,516 characters), reads with a line limit rise from 37 to 64 % and commands with truncation pipes from 47 to 68 % (Table A22), observations of ≥20k characters fall by 72 %, and whole-file reads account for 78 % of the drop (74 % on the dense line).

Regression and compaction. Table A20 gives the SWE-bench Verified regression rates: among unsolved trials the share that breaks a previously passing test follows capability (Qwen3.5-35B-A3B-Base 37.7 % and 31.0 % after v12, Qwen3.6-35B-A3B 28.3 % and 27.1 % after v12, Qwen3.6-27B 24.0 %, Qwen3.8-27B 16.8 %); the solved-group rate is 0 by construction and is not a merit. Table A21 and Figure A9b give context compaction and peak context: it fired in 52 of 13,733 trials (0.4 %; 57 events), half of the trials peak below 65k tokens and 99 % below 181k against a trigger near 218k, so compaction is not a confound of the token-footprint results.

Protocol errors and edit rework. Figure A9a draws the error taxonomy; Table A22 pools the error and editing rates, and Table A23 the next action after an error. Recovery within three calls is teacher 0.603, Qwen3.6-27B 0.299, Qwen3.6-35B-A3B 0.329, the distilled arms 0.580–0.625 and Qwen3.8-27B 0.338; after an error the teacher retries verbatim 1 % of the time, changes parameters 62 % and switches tool 38 % (Qwen3.6-27B+v12×2 63/37, Qwen3.6-27B 53/46), the same measurement seen from another angle. Edit rework falls from Qwen3.6-27B 0.036 to Qwen3.6-27B+v12×2 0.025 (teacher 0.017) and from Qwen3.6-35B-A3B 0.050 to Qwen3.6-35B-A3B+v12×3 0.039; Qwen3.8-27B's 0.012 is discounted because it prefers whole-file `Write` (22 % of its writes; teacher 8.5 %). Write-without-read rates (Table A22): the bare base falls from 21.3 to 6.5 % after v12, the post-trained MoE rises from 5.9 to 9.2 %, and the dense line stays at 8.6–10.7 % against the teacher's 11.0 %, so the habit converges on the teacher's rate from both directions on the MoE lines and is already there on the dense line. Reading local git history, absent from the corpus and forbidden by the task instruction, drops from 140 to 92 trajectories (−34 %) on the MoE line in both runs. The 1,106 `Edit` parameter-name errors of the bare base are not corpus content: the offending names occur 0 times in the corpus' 474,417 tool calls.

Dense versus sparse. Table A24 compares the dense and the sparse student at equal generation and after the same corpus: native, Qwen3.6-27B beats Qwen3.6-35B-A3B by +3.97 pp on Pro and +5.40 pp on Verified; after the same corpus Qwen3.6-27B+v12×3 − Qwen3.6-35B-A3B+v12×3 is +2.19 pp on Pro ($p = 0.137$) and +8.20 pp on Verified. The difference has the same sign on both pools, is established on Verified and unresolved on Pro; behaviour follows the recipe (Appendix B), the score follows the architecture where it is resolved.

Teacher failures. The teacher's failed trajectories differ from its successes in quantity, not habit: about twice the rounds (median 128 vs 76–102) with correspondingly more output and thinking tokens, while the by-product passed and failed populations sit at composite JS 0.0021 and 0.0026 from

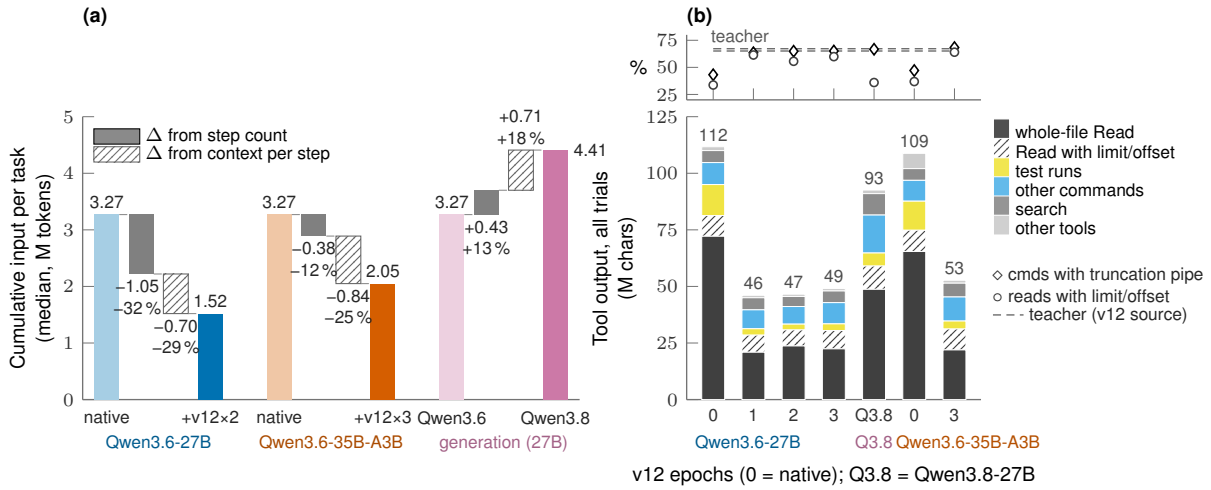


Figure A8: Where the input tokens go (SWE-bench Pro, 731 tasks). (a) Median cumulative input per task bridged through its two factors, step count and context per step: the corpus cuts steps by 32 % and context per step by 29 % on the dense line (3.27 to 1.52 M) and moves the MoE line the same way (3.27 to 2.05 M); the generation step moves the opposite way (3.27 to 4.41 M). (b) Tool-output characters by source over all trials: whole-file reads carry 74 % of the dense-line drop, while limited reads and truncation pipes (strip: diamonds = pipes, circles = limited reads; dashed = teacher) rise to the teacher's level after one epoch.

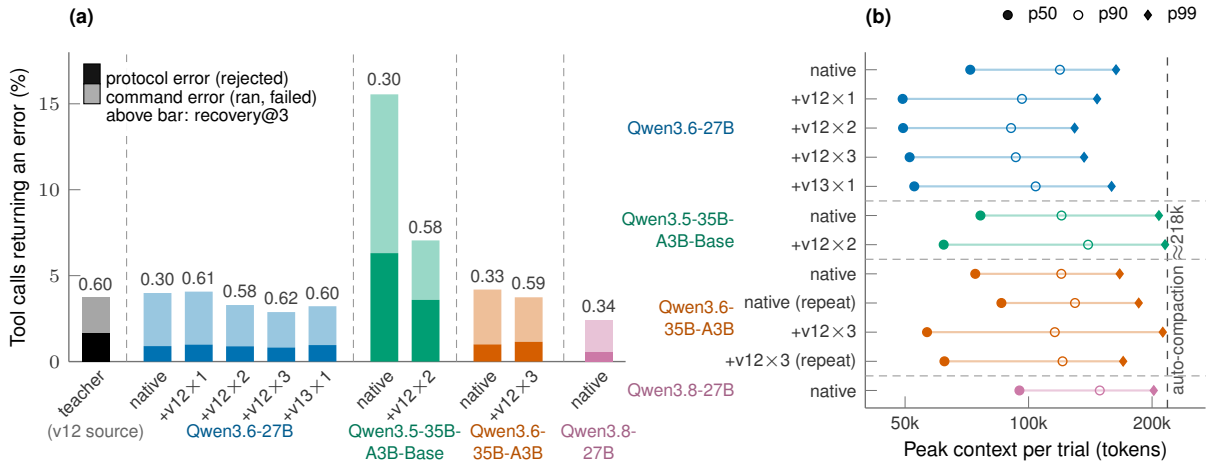


Figure A9: **(a)** Tool calls returning an error on SWE-bench Pro, split into protocol errors the harness rejects and commands that ran and failed, with protocol-error recovery@3 above each bar: fine-tuning leaves the error rate roughly in place and doubles recovery (natives 0.30–0.34, fine-tuned arms 0.58–0.62, teacher 0.60). **(b)** Peak context per trial (p50, p90, p99; both pools) against the ≈218k auto-compaction trigger.

the high-precision population (Bash 0.0070 and 0.0024), the teacher-internal floor of Table A12.

D Reproducibility

Training. Every student is trained with ms-swift on a Megatron backend with tensor parallelism 4, context parallelism 2 and micro-batch 1 (global batch 64), sequence length 196,608 tokens, optimizer state offloaded to CPU, on nodes of eight A100-80GB GPUs. Sequences are pre-paired for the two context-parallel ranks (seed 42) and each epoch is an independent shuffle of the pairs, so one epoch of v12 is 127 updates (8,120 rows at a global batch of 64, the last batch partial); the checkpoints were saved at updates 128 and 256, one update past the first and second epoch boundaries, and at 381, the end of training; Qwen3.6-27B+v13×1 runs 362 updates. Learning-rate schedules are cosine with 5 % warm-up: $8 \times 10^{-6} \rightarrow 2.4 \times 10^{-6}$ for the post-trained bases, $8 \times 10^{-6} \rightarrow 8 \times 10^{-7}$ for Qwen3.6-27B+v13×1, and $3 \times 10^{-5} \rightarrow 3 \times 10^{-6}$ for the bare base; weight decay is 0.01. Rows are rendered with the Qwen3.5 chat template; the teacher's raw thinking is kept as the thinking block

Funnel stage	v12	removed	v13	removed	Loss mask (message level)	rows
High-precision passes	12,257	–	12,257	–	`harness_error`	4,513
Vendor removal (answer fetches)	12,036	221	–	–	`ignored_cwd`	2,543
P0 drop rules	8,601	3,435	9,880	2,377	`assert_removed`	582
Same-task cap ≤ 2	8,562	39	none	–	`harness_internal`	502
By-product rows added	–	–	+14,081	–	`silence`	479
Train / val split (task level)	8,137 / 425	–	23,238 / 723	–	`summary_think`	258
262,144-token cap	8,137 / 425	0 / 0	23,236 / 723	2 / 0	`dup_repeat`	155
196,608-token cap	8,120 / 424	17 / 1	23,146 / –	90 / –	`grader_config`	39
					`network`	7
					rows with ≥ 1 mask (of 8,137)	5,528

Table A17: The v12 curation funnel with the v13 column (left; the 3,435 P0 drops are 28.5 % of the scanned rows; v13 adds 7,503 by-product passes and 6,578 failures) and the message-level loss masks among the v12 survivors (right). The ten P0 drop rules with thresholds and hit counts are listed in Table A18 (Appendix C). Masks remove loss from 11,713 of 437,589 assistant text segments and 16,856 of 474,417 tool-call segments; context is kept. The v13 validation count at the 196,608-token cap is not recorded.

Drop rule (P0)	Threshold	Hits	Sole
`raw_thinking_low`	raw-thinking length coverage < 0.98, or a vendor thinking code	2,681	2,173
`vendor_sourcing_mark`	any of three vendor sourcing codes	823	438
`content_leak_same_task`	a sibling trajectory of the same task leaked (186 tasks)	169	18
`answer_source_fetch`	answer-source URL of the task's own repository or fork	128	0
`sandbox_escape`	MITM CA variables, `-trusted-host`, `-no-hashes`, TLS off	99	27
`integrity_policy_violation`	policy clause $\wedge$ (answer fetch $\vee$ own-repo clone)	80	0
`content_leak_tool_return`	run of golden lines $\geq \max(8, \lceil 0.5 \cdot \text{golden} \rceil)$ inside one tool return	55	11
`multi_user_turn`	more than one real user turn	38	25
`upstream_clone`	`git clone/fetch/ls-remote` of the own repository	29	0
`gold_test_fetch`	≥ 10 gold-test lines returned by an external call	1	0
`skip_injected` / `broken_trace` / `subagent_dangling`	skip on a F2P test / unpaired call id / agent without result	0	0
Demote (kept)	`huge_patch` 478, `tiny_patch` 389, `memorized_upstream` 291, `git_all_refs` 219, `meta_grader_shrunk_scope` 214, `leak_in_think` 122, `unverified_success_claim` 28, `tool_protocol_unclosed` 1	1,050	–

Table A18: The v12 P0 drop rules with thresholds, hits over the 12,036 scanned rows and sole hits. "Sole" counts rows on which no rule of any tier fires besides this one (the build report's definition); Table A19 reports the larger "alone" count, which allows demote flags to co-occur. The demote tier flags 1,050 rows and keeps them.

(rows whose thinking survives only in summarized form are removed by the thinking-coverage rule), and loss falls on assistant text, thinking and tool calls except masked segments (§3).

Serving and evaluation. Every arm runs inside Claude Code 2.1.202 driven by a task runner that gives each trial a fresh container, a 262,144-token context with automatic compaction and a 5,400-second wall-time limit; students are served by SGLang behind an adapter that exposes the Anthropic Messages API and passes thinking blocks through unchanged, and the next Qwen generation is served the same way. Sampling is at temperature 0.8, one rollout per task, and the two repeat-sample arms are second runs of the same weights; the added instruction forbids looking up the fix over the network or in the repository's git history, and network access is otherwise unrestricted. Outcomes are the benchmark's hidden tests; a trial that times out or crashes counts as unsolved.

Release. This submission releases the fingerprint extractor, the per-arm behavioral distributions, the curation rule definitions with their thresholds and hit counts, the rule-impact analysis, the evaluation ledgers with per-task outcomes for all twelve arms, and the scripts that generate every table and figure in this paper. Release of the corpus and the checkpoints is not part of this submission.

Rule family	touched	alone	% drops	steps	think	JS mix / floor	JS heads / floor	recovery@3
`raw_thinking_low`	2,681	2,450	71.3	68	0.897	0.005 / 0.000	0.015 / 0.000	0.68
`vendor_sourcing_mark`	823	493	14.4	79	0.899	0.010 / 0.000	0.039 / 0.001	0.52
leak family (6 rules)	224	45	1.3	64	0.896	0.021 / 0.000	0.046 / 0.002	0.24
`sandbox_escape`	99	31	0.9	84	0.896	0.008 / 0.001	0.105 / 0.005	0.71
`multi_user_turn`	38	27	0.8	55	0.863	0.005 / 0.002	0.059 / 0.013	0.67

Table A19: What each v12 drop rule removes (12,036 scanned rows, 3,435 dropped). "touched": dropped rows the family flags; "alone": dropped rows on which no other drop-tier rule fires (demote flags may co-occur, so it exceeds the "sole" count of Table A18; for the leak family the six rules overlap, so its alone count exceeds the sum of the rules' sole counts); steps and think = median step count and thinking share of the removed rows (kept set: 44 steps, 0.838); JS = Jensen–Shannon distance ($\log_2$) of the removed rows' tool mix and Bash head-word distribution to the kept set, next to the size-matched random-removal floor; recovery@3 = protocol-error recovery within three calls (kept 0.60). The thinking-coverage floor removes volume with the smallest tool-mix shift of any family (still above its random-removal floor) but takes the longest, most think-heavy trajectories; the sourcing rules remove a Bash-heavier profile with poor recovery; the demote tier removes nothing and the cap removes 39 rows.

		regressed		solved	unsolved	
Arm	applied	n	%	%	n	%
Qwen3.6-27B						
native	496	37	7.5	0.0	154	24.0
+v12×1	491	41	8.4	0.0	151	27.2
+v12×2	494	42	8.5	0.0	151	27.8
+v12×3	486	30	6.2	0.0	130	23.1
+v13×1	494	43	8.7	0.0	163	26.4
Qwen3.5-35B-A3B-Base						
native	491	87	17.7	0.0	231	37.7
+v12×2	477	62	13.0	0.0	200	31.0
Qwen3.6-35B-A3B						
native	488	49	10.0	0.0	173	28.3
+v12×3	496	49	9.9	0.0	181	27.1
Qwen3.8-27B						
native	484	19	3.9	0.0	113	16.8

Table A20: Regression on SWE-bench Verified (`regression_summary.md`): among trials whose patch applied, the share whose verifier report lists at least one failing PASS_TO_PASS test, overall and within the solved and unsolved groups. Every produced patch applied (0 missing, 0 non-applying in 4,897 graded trials); trials without a verifier report (grader failure or trial exception; 0–23 per arm) are excluded. The solved-group rate is 0 by construction of the verdict; the unsolved-group rate follows capability. The harness never fills PASS_TO_FAIL, so this is the only regression signal.

E Same-pool anchors

Teacher on the benchmark. Claude Opus 4.8, driven by the same Claude Code version, effort `high`, with the same added instruction and wall-time limit as the students, ran once on 359 SWE-bench Pro tasks (four stratified batches and a fifth block of 92 tasks drawn the same way, teleport excluded and every arm valid; 149 easy, 64 medium and 146 hard by the number of arms that solved them, against 305/133/293 in the full pool). Its thinking is returned encrypted by the harness, so only tool-based rulers and scalars are defined for this reference; every trial has a verdict. Table A26 gives each arm's distance to the teacher's own trajectories on the same tasks next to its distance to the corpus reference, Table A27 the read-set and edit-set overlap on the same tasks, and Table A28 pass@1 and the bundle scalars on the 359. The teacher's pass@1 on these public tasks is contaminated and difficulty-stratified and is reported only to place the arms; the dense line does not move on this subset ($+3.1$ pp for Qwen3.6-27B+v12×2 against Qwen3.6-27B, 26:15 discordant) while both MoE pairs do ($+7.2$ and $+4.5$ pp), and per-task read-overlap gains do not track per-task solve gains (Spearman $\rho = 0.03$).

Held-out validation tasks. The 345 tasks are the validation split of the v12 corpus that passed the oracle gate (Go 166, TypeScript 52, Python 32, other 95); the teacher's trajectory on each is a high-precision pass that no student trained on, and the reference for the same-pool distances is the teacher's own 345 trajectories, which are 0.0012 (composite) from the pooled corpus. Table A29 gives the distances and floors and Table A30 pass@1, overlap and the bundle scalars; the composite distance falls from 0.051 to 0.006 at one epoch and moves by ± 0.001 afterwards, and the drop reproduces in every language stratum at the first epoch (native to Qwen3.6-27B+v12×1: Go 0.049 to 0.009, TypeScript 0.088 to 0.007, Python 0.071 to 0.005, other 0.047 to 0.010). pass@1 on the 345 is 50.7 % for Qwen3.6-27B and 52.2–56.5 % for the v12 and Qwen3.6-27B+v13×1 checkpoints, not ordered by dose

		auto-compaction				peak context per trial (tokens)				
Arm	trials	trials	%	events	med. dropped	p50	p90	p99	max	≥200k
Qwen3.8-27B	1,231	7	0.6	8	207,101	94,990	149,116	201,890	217,745	13
Qwen3.6-27B	1,224	1	0.1	1	206,815	72,090	119,223	163,372	217,464	2
Qwen3.6-27B+v12×1	1,230	1	0.1	1	212,475	49,319	96,347	146,883	217,375	2
Qwen3.6-27B+v12×2	1,216	0	0.0	0	–	49,441	90,633	129,461	160,210	0
Qwen3.6-27B+v12×3	1,215	1	0.1	1	215,781	51,288	93,070	136,563	217,737	1
Qwen3.6-27B+v13×1	1,231	1	0.1	1	215,251	52,648	104,049	159,340	217,710	4
Qwen3.5-35B-A3B-Base	1,231	9	0.7	9	209,962	76,324	120,304	207,789	217,809	14
Qwen3.5-35B-A3B-Base+v12×2	1,231	13	1.1	15	210,696	62,097	139,647	215,130	217,875	23
Qwen3.6-35B-A3B	1,231	2	0.2	2	206,787	74,149	120,167	166,750	217,121	4
Qwen3.6-35B-A3B (repeat)	731	3	0.4	3	207,060	85,852	129,751	185,556	217,145	3
Qwen3.6-35B-A3B+v12×3	1,231	13	1.1	15	213,152	56,557	115,852	212,383	217,713	16
Qwen3.6-35B-A3B+v12×3 (repeat)	731	1	0.1	1	211,761	62,366	121,007	170,054	217,492	1
All arms	13,733	52	0.4	57	210,222	–	–	–	–	–

Table A21: Context compaction across the twelve arms and both pools (`compaction_summary.md`): `compact_boundary` records in the Claude Code session files. All 57 events were automatic (none manual); one event drops a median 210k tokens (`preTokens` median 218,365 → `postTokens` ≈8k). Peak context = maximum over a trial's requests of input + cache-creation + cache-read tokens; the observed trigger is ≈218k, so compaction is excluded as a confound rather than compared as a performance dimension. Trials that compacted solve less often (15.7 % vs 55.1 %), a selection effect of long-context tasks.

	tool errors (% of calls)				editing (% of edits)						
Group	any	protocol	command	recovery@3	re-edit	rework	revert	no-read	Write share	trunc. pipes	read w/ limit
Teacher, high-precision	3.7	1.6	2.1	0.603	55.1	1.7	0.5	11.0	8.5	67.2	65.1
Teacher by-product, passed	2.6	1.0	1.6	0.727	55.1	0.8	0.3	10.0	8.0	69.1	62.4
Teacher by-product, failed	3.1	1.3	1.7	0.692	56.9	1.5	0.5	13.0	11.6	68.8	62.2
Pro											
Qwen3.6-27B	4.0	0.9	3.1	0.299	57.6	3.6	0.8	10.7	17.5	43.2	33.7
Qwen3.6-27B+v12×1	4.1	1.0	3.1	0.605	63.5	3.0	0.9	8.7	9.1	63.5	61.4
Qwen3.6-27B+v12×2	3.3	0.9	2.4	0.580	60.9	2.5	1.0	9.1	9.2	64.9	55.6
Qwen3.6-27B+v12×3	2.9	0.8	2.1	0.625	61.4	2.8	0.8	10.1	10.3	65.1	59.9
Qwen3.6-27B+v13×1	3.2	1.0	2.3	0.599	63.5	2.8	1.3	8.6	8.7	64.7	59.7
Qwen3.8-27B	2.4	0.5	1.9	0.338	59.7	1.2	0.3	17.7	22.0	66.7	36.0
Qwen3.5-35B-A3B-Base	15.7	6.3	9.2	0.305	75.2	15.6	9.0	21.3	24.1	20.6	51.1
Qwen3.5-35B-A3B-Base+v12×2	7.0	3.6	3.5	0.577	70.6	3.8	1.3	6.5	9.0	66.9	69.1
Qwen3.6-35B-A3B	4.2	1.0	3.2	0.329	63.0	5.0	1.0	5.9	9.2	47.1	36.9
Qwen3.6-35B-A3B+v12×3	3.8	1.2	2.6	0.593	67.5	3.9	1.7	9.2	9.3	68.3	64.0
Qwen3.6-35B-A3B (repeat)	3.8	1.0	2.8	0.361	63.7	5.0	1.2	5.3	9.2	47.5	35.6
Qwen3.6-35B-A3B+v12×3 (repeat)	3.4	1.1	2.4	0.580	66.4	3.8	1.7	8.5	8.8	68.6	62.7
Verified											
Qwen3.6-27B	9.8	0.4	9.4	0.448	49.5	10.3	1.8	10.6	18.4	44.3	75.1
Qwen3.6-27B+v12×1	8.9	1.6	7.4	0.273	60.9	17.9	11.8	3.1	2.7	62.6	93.4
Qwen3.6-27B+v12×2	7.2	0.7	6.5	0.709	54.5	9.3	2.4	3.0	3.1	64.1	91.5
Qwen3.6-27B+v12×3	6.9	0.7	6.2	0.578	52.6	7.5	2.8	3.6	3.0	62.8	91.8
Qwen3.6-27B+v13×1	7.1	0.8	6.3	0.567	56.0	9.5	3.7	2.9	2.7	63.4	93.6
Qwen3.8-27B	6.4	0.5	5.9	0.345	47.0	3.3	1.1	18.5	20.3	57.9	79.6
Qwen3.5-35B-A3B-Base	18.9	1.6	17.3	0.389	63.5	27.6	9.8	35.8	40.8	17.3	69.9
Qwen3.5-35B-A3B-Base+v12×2	9.0	2.4	6.5	0.629	66.1	8.8	3.2	4.1	4.9	64.0	93.9
Qwen3.6-35B-A3B	10.5	1.2	9.3	0.232	60.1	20.7	10.3	6.2	7.9	46.1	81.8
Qwen3.6-35B-A3B+v12×3	6.5	1.0	5.5	0.673	65.5	12.2	4.3	3.5	3.2	63.9	94.9

Table A22: Protocol-error taxonomy and edit rework, pooled over all trajectories of a group (`fp_pooled.json`). "any" = tool calls returning an error; "protocol" = calls the harness rejects (malformed or missing parameters); "command" = commands that ran and failed; recovery@3 = share of protocol errors followed within three calls by a successful call of the same tool on the same target. "re-edit" = edits touching a file already edited in the trajectory; "rework" = editing what one just wrote; "revert" = an edit undone by a later edit; "no-read" = edits without a prior read of the file; "Write share" = whole-file writes among edits; "trunc. pipes" = commands with a truncation pipe (`head`, `tail`, ...); "read w/ limit" = reads with a line limit.

(Qwen3.6-27B+v12×1 highest); the two MoE arms, Qwen3.6-35B-A3B and Qwen3.5-35B-A3B-Base+v12×2, start from different bases and are reported without a pairing.

Group	traj.	verbatim	changed	switch	no tool	end
Teacher, high-precision	9,399	0.5	61.6	37.7	0.1	0.0
Teacher by-product, passed	5,787	0.2	62.5	37.1	0.2	0.0
Teacher by-product, failed	7,950	0.3	61.6	38.0	0.2	0.0
Pro						
Qwen3.6-27B	604	0.2	53.3	46.3	0.2	0.0
Qwen3.6-27B+v12×1	514	0.2	69.1	30.5	0.2	0.0
Qwen3.6-27B+v12×2	450	0.2	62.5	37.0	0.3	0.0
Qwen3.6-27B+v12×3	453	0.1	65.0	34.8	0.1	0.0
Qwen3.6-27B+v13×1	452	0.8	67.5	31.0	0.6	0.1
Qwen3.8-27B	550	0.0	59.7	40.2	0.0	0.1
Qwen3.5-35B-A3B-Base	719	1.3	31.9	66.0	0.8	0.0
Qwen3.5-35B-A3B-Base+v12×2	601	1.7	69.2	29.1	0.1	0.0
Qwen3.6-35B-A3B	599	0.2	57.3	42.4	0.0	0.0
Qwen3.6-35B-A3B+v12×3	528	0.6	63.3	35.4	0.7	0.0
Verified						
Qwen3.6-27B	476	0.6	80.4	18.9	0.0	0.0
Qwen3.6-27B+v12×1	445	0.8	88.7	10.2	0.3	0.0
Qwen3.6-27B+v12×2	423	0.4	88.4	10.7	0.5	0.0
Qwen3.6-27B+v12×3	433	0.2	90.0	9.6	0.3	0.0
Qwen3.6-27B+v13×1	436	1.2	89.3	9.4	0.1	0.0
Qwen3.8-27B	470	0.1	85.3	14.4	0.1	0.0
Qwen3.5-35B-A3B-Base	499	0.8	62.2	36.6	0.3	0.0
Qwen3.5-35B-A3B-Base+v12×2	458	2.2	84.4	12.7	0.6	0.0
Qwen3.6-35B-A3B	487	0.3	83.1	16.6	0.0	0.0
Qwen3.6-35B-A3B+v12×3	455	0.8	83.0	15.5	0.6	0.0

Table A23: Next action after a tool error, as a share (%) of post-error steps averaged over trajectories with at least one error (`fp_summary.json`, `err_next`): the same call repeated verbatim, the same tool with changed parameters, a different tool, no tool call, or the trajectory ends. This is the recovery measure of §5 seen from another angle.

Comparison ($A - B$)	Pool	pass@1 A	pass@1 B	b	c	Δ (pp)	p
native, same generation							
Qwen3.6-27B − Qwen3.6-35B-A3B	Pro	50.1	46.1	70	41	+3.97	0.008
Qwen3.6-27B − Qwen3.6-35B-A3B	Verified	68.4	63.0	64	37	+5.40	0.009
after the same corpus							
Qwen3.6-27B+v12×3 − Qwen3.6-35B-A3B+v12×3	Pro	51.0	48.8	59	43	+2.19	0.137
Qwen3.6-27B+v12×3 − Qwen3.6-35B-A3B+v12×3	Verified	71.2	63.0	64	23	+8.20	1.27×10^{-5}
next generation vs distilled MoE							
Qwen3.8-27B − Qwen3.6-35B-A3B+v12×3	Pro	55.7	48.8	80	30	+6.84	2.02×10^{-6}
Qwen3.8-27B − Qwen3.6-35B-A3B+v12×3	Verified	74.2	63.0	72	16	+11.20	1.19×10^{-9}
generation jump, dense							
Qwen3.8-27B − Qwen3.6-27B	Pro	55.7	50.1	75	34	+5.61	1.07×10^{-4}
Qwen3.8-27B − Qwen3.6-27B	Verified	74.2	68.4	57	28	+5.80	0.002

Table A24: Dense versus sparse at equal generation and after the same corpus, and the vendor generation jump (`data/report_excerpt.json`; exact McNemar, pool denominators 731 / 500, missing = unsolved). The four generation/architecture rows of Table A1 give the Holm-adjusted values. Behaviour follows the recipe (Appendix B); the score difference between the dense and the sparse student has the same sign on both pools, is established on Verified and unresolved on Pro.

Arm	steps	ctx (k)	input (M)	output (k)
Qwen3.6-27B				
native	53	85	3.27	13.7
+v12×1	40	57	1.69	17.7
+v12×2	36	57	1.52	15.9
+v12×3	39	59	1.75	17.3
+v13×1	39	61	1.80	19.0
Qwen3.5-35B-A3B-Base				
native	62	79	3.40	18.4
+v12×2	50	66	2.44	23.3
Qwen3.6-35B-A3B				
native	51	87	3.27	14.1
+v12×3	45	63	2.05	21.1
native (repeat)	51	86	3.13	13.6
+v12×3 (repeat)	44	62	2.01	20.0
Qwen3.8-27B				
native	60	109	4.41	46.8

Table A25: Per-task medians on SWE-bench Pro: assistant steps, context of the final request (thousand tokens), cumulative input tokens (million; the ledger of §6) and output tokens (thousand, including thinking). Output rises by 16–49 % on the fine-tuned arms while input falls by 28–53 %; Qwen3.8-27B emits $3.4\times$ the native dense model's output.

	tool mix			tool bigrams			Bash heads			composite	
Arm	same	off	floor	same	off	floor	same	off	floor	same	off
Qwen3.6-27B											
native	0.0451	0.0190	0.0007	0.0803	0.0406	0.0021	0.1319	0.2163	0.0098	0.0627	0.0298
+v12×1	0.0069	0.0161	0.0007	0.0134	0.0281	0.0021	0.0066	0.0722	0.0074	0.0102	0.0221
+v12×2	0.0046	0.0152	0.0006	0.0093	0.0262	0.0021	0.0050	0.0501	0.0105	0.0069	0.0207
+v12×3	0.0069	0.0160	0.0006	0.0135	0.0284	0.0021	0.0048	0.0547	0.0066	0.0102	0.0222
+v13×1	0.0066	0.0141	0.0011	0.0129	0.0249	0.0028	0.0072	0.0576	0.0117	0.0097	0.0195
Qwen3.6-35B-A3B											
native	0.0589	0.0244	0.0007	0.0984	0.0441	0.0027	0.1075	0.2050	0.0116	0.0787	0.0342
native (repeat)	0.0656	0.0277	0.0008	0.1068	0.0492	0.0031	0.1129	0.2076	0.0132	0.0862	0.0385
+v12×3	0.0066	0.0158	0.0007	0.0143	0.0261	0.0020	0.0074	0.0575	0.0085	0.0105	0.0209
+v12×3 (repeat)	0.0060	0.0132	0.0007	0.0127	0.0225	0.0021	0.0113	0.0663	0.0138	0.0094	0.0178
Qwen3.5-35B-A3B-Base											
native	0.0222	0.0323	0.0005	0.0698	0.0860	0.0019	0.2023	0.3031	0.0196	0.0460	0.0591
+v12×2	0.0071	0.0113	0.0007	0.0146	0.0207	0.0023	0.0097	0.0594	0.0134	0.0109	0.0160
Qwen3.8-27B											
native	0.0609	0.0320	0.0004	0.0982	0.0570	0.0012	0.0199	0.0657	0.0031	0.0796	0.0445

Table A26: Same-pool anchor on SWE-bench Pro: Jensen–Shannon distance ($\log_2$) from each arm's 359 trajectories to the teacher's own trajectories on the same 359 tasks (same) and to the pooled high-precision corpus reference (off), with the arm's split-half floor (200 random halves). The teacher on the benchmark is at 0.0184 / 0.0321 / 0.0545 (mix / bigram / Bash) from its own corpus. Thinking rulers are undefined for this reference (the teacher's thinking is not returned by the harness).

	files read						files edited		
Arm	n	precision	recall	Jaccard	mean	control	n	Jaccard	mean
Qwen3.6-27B									
native	8	0.50	1.00	0.50	0.478	0.019	3	0.60	0.620
+v12×1	5	0.75	0.89	0.63	0.611	0.016	3	1.00	0.762
+v12×2	5	0.75	0.88	0.60	0.606	0.018	3	0.83	0.755
+v12×3	5	0.75	0.86	0.60	0.578	0.019	3	0.80	0.740
+v13×1	5	0.73	0.88	0.60	0.598	0.018	3	0.82	0.743
Qwen3.6-35B-A3B									
native	7	0.50	1.00	0.50	0.484	0.018	3	0.73	0.702
native (repeat)	7	0.50	1.00	0.50	0.483	0.020	3	0.75	0.700
+v12×3	5	0.67	0.88	0.57	0.579	0.018	3	0.80	0.741
+v12×3 (repeat)	5	0.67	0.89	0.56	0.570	0.018	3	0.83	0.738
Qwen3.5-35B-A3B-Base									
native	6	0.56	0.75	0.43	0.455	0.018	7	0.25	0.285
+v12×2	5	0.71	0.83	0.56	0.571	0.019	3	0.75	0.718
Qwen3.8-27B									
native	8	0.50	1.00	0.46	0.482	0.018	4	0.50	0.473

Table A27: Read-set and edit-set overlap with the teacher on the same 359 SWE-bench Pro tasks (one rollout per arm; medians over tasks, Jaccard also as a mean). Teacher medians: 4 files read, 2 edited. Control = mean Jaccard against the teacher's read set of another task of the same repository; against another random task it is 0.003–0.004 for every arm. Paired tests are in the analysis record (`analysis/anchors/pro359`).

Arm	pass@1	recovery@3	truncation pipe	write w/o read	task mgmt	echo share	Bash share	steps
Teacher on the 359 (contaminated)	56.8	0.736	0.594	0.098	0.000	0.130	0.546	24
Teacher corpus reference	–	0.603	0.672	0.110	0.032	0.142	0.507	51
Qwen3.6-27B								
native	49.6	0.309	0.416	0.123	0.056	0.020	0.533	49
+v12×1	49.0	0.551	0.616	0.091	0.005	0.129	0.618	37
+v12×2	52.6	0.550	0.631	0.086	0.003	0.145	0.596	35
+v12×3	49.3	0.667	0.634	0.106	0.004	0.136	0.612	38
+v13×1	51.3	0.560	0.609	0.083	0.005	0.136	0.606	38
Qwen3.6-35B-A3B								
native	44.8	0.358	0.425	0.062	0.088	0.024	0.522	50
native (repeat)	46.5	0.380	0.452	0.058	0.100	0.022	0.510	49
+v12×3	49.3	0.598	0.674	0.084	0.005	0.129	0.616	42
+v12×3 (repeat)	47.1	0.588	0.675	0.084	0.006	0.126	0.603	43
Qwen3.5-35B-A3B-Base								
native	37.0	0.293	0.202	0.215	0.001	0.016	0.497	59
+v12×2	44.3	0.581	0.657	0.066	0.007	0.151	0.598	46
Qwen3.8-27B								
native	52.4	0.344	0.655	0.195	0.078	0.092	0.620	57

Table A28: The same 359 tasks: pass@1 (per cent; the teacher's value is contaminated, the tasks are public and the sample is stratified by student difficulty) and the pooled bundle scalars of §5 for the teacher on the benchmark, the teacher's corpus, and every arm. Median steps per trajectory in the last column.

	tool mix		tool bigrams		Bash heads		thinking position		composite	
Arm	same	floor	same	floor	same	floor	same	floor	same	corpus
Qwen3.6-27B										
native	0.0390	0.0011	0.0629	0.0024	0.1279	0.0096	0.0046	0.0020	0.0510	0.0552
+v12×1	0.0043	0.0008	0.0080	0.0024	0.0263	0.0081	0.0038	0.0017	0.0062	0.0044
+v12×2	0.0035	0.0006	0.0079	0.0021	0.0142	0.0100	0.0031	0.0018	0.0057	0.0046
+v12×3	0.0048	0.0005	0.0087	0.0017	0.0153	0.0069	0.0034	0.0019	0.0067	0.0051
+v13×1	0.0038	0.0007	0.0082	0.0020	0.0202	0.0073	0.0054	0.0016	0.0060	0.0044
Qwen3.6-35B-A3B										
native	0.0347	0.0007	0.0598	0.0021	0.1402	0.0317	0.0082	0.0023	0.0472	0.0512
Qwen3.5-35B-A3B-Base										
+v12×2	0.0068	0.0006	0.0129	0.0016	0.0249	0.0099	0.0100	0.0015	0.0099	0.0076
Qwen3.8-27B										
native	0.0214	0.0005	0.0396	0.0015	0.0330	0.0048	0.0104	0.0016	0.0305	0.0297

Table A29: Held-out anchor: Jensen–Shannon distance ($\log_2$) from each arm's trajectories on the 345 validation tasks (never trained on; one rollout each) to the teacher's own trajectories on the same tasks (same), with the arm's split-half floor (200 random halves), and the composite distance of the same trajectories to the pooled corpus reference (corpus). The teacher's 345 validation trajectories are 0.0012 (composite) from its corpus. The two MoE arms start from different bases and are not a pair.

			overlap with the teacher					bundle scalars				
Arm	pass@1	void	n read	precision	recall	Jaccard	edit J.	recovery@3	trunc.	w/o read	task mgmt	echo
Teacher on the 345	100	0	8	1	1	1	1	0.62	0.67	0.107	0.036	0.136
Qwen3.6-27B												
native	50.7	2	13	0.56	0.86	0.50	0.69	0.38	0.57	0.061	0.150	0.017
+v12×1	56.5	0	8	0.80	0.80	0.62	0.75	0.64	0.67	0.076	0.023	0.127
+v12×2	53.3	3	9	0.78	0.80	0.64	0.78	0.60	0.68	0.072	0.020	0.136
+v12×3	52.2	0	8	0.78	0.82	0.63	0.75	0.65	0.69	0.086	0.017	0.131
+v13×1	55.1	0	9	0.77	0.83	0.62	0.77	0.57	0.69	0.077	0.023	0.135
Qwen3.6-35B-A3B												
native	47.2	2	13	0.53	0.80	0.46	0.62	0.36	0.52	0.041	0.139	0.016
Qwen3.5-35B-A3B-Base												
+v12×2	44.1	2	9	0.71	0.80	0.58	0.71	0.49	0.69	0.056	0.016	0.121
Qwen3.8-27B												
native	53.0	1	12	0.58	0.87	0.50	0.67	0.46	0.67	0.117	0.103	0.103

Table A30: Held-out anchor: pass@1 on the 345 validation tasks (per cent; trials without a verdict count as unsolved, their number in the void column), read-set and edit-set overlap with the teacher on the same task (medians over tasks), and the pooled bundle scalars of §5. The teacher passed every validation task by construction, so its pass@1 is not a comparison.