The Protocol Is the Agent: Decomposing RL, SFT, and Scaffold Effects in Tool-Use Agents from 0.8B to 27B

Anonymous authors
Paper under double-blind review

Abstract

On a single frozen instrument (the same 100 financial tool-agent tasks, execution harness, and analyzer, with three evaluation seeds per arm and pooled paired exact McNemar tests), we decompose the measured task-completion capability of one Qwen3.x family (0.8B, 2B, 9B, 27B, plus a 4B case study) into protocol, supervised-fine-tuning (SFT), and reinforcement-learning (RL) contributions. Three findings. **(1) A consistent RL null:** the same GRPO recipe with verifiable structural rewards changes nothing at any scale (+0.7 at 0.8B, p=.75; +2.0 at 2B, p=.27; 0.0 at 9B, p=1.0; −0.7 at 27B, p=.64), including the scale with the largest SFT headroom. We further document how a shared-but-wrong evaluation template alone manufactured p=.024 in our own 2B cell before correction. **(2) The protocol dominates:** switching the surrounding stack from a text-protocol scaffold to native tool-calling is worth +59 points on the same 2B SFT weights, versus +4.3 for SFT itself and +2.0 for RL. **(3) SFT's contribution is non-monotone and can be catastrophic:** the identical recipe adds +18 points at 0.8B and +11 at 9B but destroys 45 points at 4B by inducing temperature-dependent format fragility in an otherwise format-clean base, after which RL repairs the format while capturing a rewarded shortcut. Corollaries: measure the protocol before the weights, the headroom before the RL budget, and never compare training arms without tokenizer-exact evaluation.

1 Introduction

Practitioners budget for three ways to make a tool-use agent better: change the model's weights (post-training), change what surrounds the model (the protocol: prompts, tool-call conventions, execution harness), or leave both fixed and sample. A large literature treats the first as the frontier; RL with verifiable rewards (RLVR) on top of supervised fine-tuning (SFT) is the default recipe at every frontier lab. The second is treated as scaffolding: necessary engineering, rarely measured.

This paper asks the attribution question directly: *how much of measured agent capability comes from each source?* We hold everything fixed except one factor at a time, on a single instrument (a frozen 100-task financial tool-agent suite with a deterministic analyzer and a paired-statistics protocol: 3 evaluation seeds per arm, pooled exact McNemar tests over all seed pairings), across one model family at 0.8B, 2B, 4B, 9B, and 27B parameters (Qwen3.x: 3.5 at 0.8B/2B/4B/9B/27B, 3.8 at 27B), trained with an identical recipe (LoRA SFT on 1,624 distilled trajectories, then GRPO with a six-term verifiable structural reward).

Three results, none of which we set out to find:

1. **A consistent RL null** (Section 4). Across every seeded, paired cell on this instrument, GRPO adds no detectable capability: +0.7 points at 0.8B (p=.75), +2.0 at 2B (p=.27), 0.0 at 9B (p=1.0, discordant pairs perfectly symmetric at 87/87), −0.7 at 27B (p=.64). This includes the cell with the largest SFT headroom (0.8B: an +18-point SFT increment still furthest below the protocol ceiling) as well as cells

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

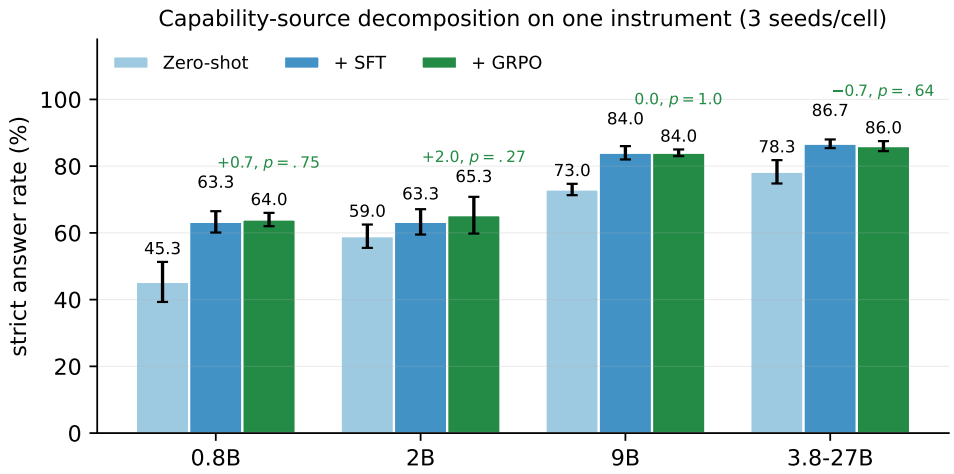


Figure 1: The capability-source decomposition. Per-scale strict answer rates for the base model (zero-shot on the native protocol), after SFT, and after GRPO, with pooled paired statistics for each increment. No RL increment is significant at any scale; SFT increments are scale-dependent; the base models already solve most tasks the instrument poses.

with measured headroom below the harness ceiling, and it replicates across four additional program cells (full-stack, long-horizon; Appendix A).

2. **The protocol dominates the weights** (Section 4). The same 2B SFT weights score 4.3% under a text-protocol scaffold and 63.3% under the native tool-calling protocol: +59 points from the surrounding stack, against +4.3 from SFT itself and +2.0 from RL. Zero-shot base models tell the same story from below: the 2B *base* scores 59.0% on the native protocol, 55 points above the SFT'd model on the text protocol.

3. **Measurement artifacts are the same size as the effects practitioners chase** (Section 5). Our own 2B cell showed a "significant" RL effect (+4.0, p=.024) that evaporated (p=.268) when we discovered both arms had been evaluated under the wrong chat template: the template default of a different scale's tokenizer. The inflation was *differential*: the template mismatch helped the RL arm 5.4 points, the SFT arm 3.4. We report the episode as a first-class result because it is the field's default failure mode in miniature.

The 4B scale contributes a fourth finding we did not predict (Section 6): the identical SFT recipe whose increment is +18.0 at 0.8B, +4.3 at 2B, and +11.0 at 9B *destroys* 45 points at 4B. The base is format-clean; the LoRA induces temperature-dependent JSON/XML hybrid emission; RL then repairs exactly what its reward penalizes (malformed calls drop to zero) while capturing a rewarded shortcut (an ever-succeeding documentation-loading tool consumed 81% of rollouts' redundancy budget). The triad of repair, capture, and null is the mechanistic face of the same conclusion: *RL optimizes the reward, and on a strong-SFT instrument the reward has almost nothing left to say.*

We state the scope carefully: these are boundary conditions of one recipe stack (GRPO-family objectives, structural verifiable rewards, LoRA SFT inits) on one domain and one model family, at bounded horizons. The claim is not that RL never helps agents: our own text-protocol lineage (Appendix A) contains a +54-point RL repair where the reward could finally see the loop. The claim is that *on the instrument where agents are actually fielded (native protocols, strong SFT anchors), the measured effect of RL was indistinguishable from zero at every scale we tested, while the protocol and the evaluation setup moved results by tens of points.* Budget accordingly.

2 Related work

RL with verifiable rewards for tool-use agents. GRPO-style RLVR (Shao et al., 2024; DeepSeek-AI, 2025) is the dominant post-SFT recipe, with many objective refinements (Yu et al., 2025; Zheng et al., 2025; Zhang et al., 2026; Liu et al., 2026). Reported agent gains typically compare a trained arm against a base or SFT initialization under a fixed scaffold, often with single evaluation runs. Our contribution to this line is a *negative* result with unusually tight instrumentation: across three scales and seven objective formulations (Appendix A), no RL increment survives paired statistics on a strong SFT anchor; the one apparent exception was a measurement artifact (Section 5).

Integrated agent stacks. Recent systems work integrates data engines, execution harnesses, and RLVR training into end-to-end agent stacks (Wang et al., 2026; Raoof et al., 2026; Zuo et al., 2026). These report aggregate capabilities but hold no factor fixed, so weights, protocol, and harness contributions cannot be separated. Our study is the complementary attribution experiment: one factor moves at a time against a frozen instrument, with the harness lineage itself documented separately (Xiao et al., 2026). The finding that the protocol term dominates the weights term at every scale is, to our knowledge, the first controlled measurement of this decomposition.

SFT, distillation, and interference. Distilled multi-teacher SFT is standard for small agentic models (Qwen Team, 2026); see also the Qwen3 technical report (Qwen Team, 2025). Catastrophic interference from fine-tuning is well documented in classification and instruction tuning. We contribute a clean, fully-instrumented agent case: a format-clean base rendered format-fragile at temperature by the very LoRA meant to teach it the protocol, invisible to greedy evaluation and to aggregate training loss, and then partially repaired by RL for reward-shaped rather than task reasons (Section 6).

Evaluation integrity. Contamination and leakage in agent benchmarks (Kapoor & Narayanan, 2023; Zuo et al., 2026) and variance from stochastic trajectories are recognized hazards; tool-use protocols shape what models can express (Schick et al., 2023; Yao et al., 2023). We extend this line with a new, minimal artifact class: the *shared-but-wrong chat template*, harmless-looking because it is constant across arms yet capable of manufacturing a paired-significant training effect through differential inflation (Section 5), alongside two infrastructure masquerades ($\pm$10–14 points) and a single-seed +6.0 that fresh seeds retired (Section 5; Appendix A).

3 Setting: one lineage, one instrument

Task domain and instrument. Tasks are drawn from a realistic financial-analysis domain: portfolio computations over price histories, technical indicators, company fundamentals, forecasting, and web research. The frozen evaluation suite is n=100 held-out tasks stratified into four difficulty tiers (atomic lookups; multi-ticker compositions; platform orchestration; adversarial market-shock recovery), never trained on, executed against deterministic mock services or live services (mode recorded per row, never mixed within a comparison). The analyzer is frozen and deterministic: it classifies every trajectory into a strict outcome taxonomy from raw transcripts. The primary metric throughout is the **strict answer rate**: the fraction of tasks whose final assistant message is a well-formed, non-empty answer with no tool-call opening. It is deliberately *format-level*: it does not grade answer *correctness*, and we label it as such wherever it appears; Section 7 discusses the correctness layer.

The native protocol. The instrument's scaffold enforces a native XML tool-calling convention, rendered by each model's own chat template: a reasoning block may precede a structured call, which the environment executes against the local skill registry (market data, technical analysis, quantitative computation, forecasting, web search, deep research) before returning an observation, until a final answer or the bounded turn budget (8 turns) is reached. A trajectory dies at the turn cap if the loop never terminates, which is the dominant

Table 1: Capability-source decomposition on one instrument (frozen 100-task suite, native protocol, 3 seeds/arm, strict format-level answer rate, pooled paired exact McNemar over 9 seed pairings per increment). Zero-shot rows are the untrained hub snapshots evaluated on the same instrument.

Scale	Zero-shot	SFT	GRPO	Δ(SFT−Zero-shot), p	Δ(RL−SFT), p
0.8B	45.3 ± 6.0	63.3 ± 3.2	64.0 ± 2.0	+18.0, $< 10^{-16}$ (b=273/c=111)	+0.7, .752 (b=128/c=122)
2B	59.0 ± 3.5	63.3 ± 3.8	65.3 ± 5.5	+4.3, .065 (b=231/c=192)	+2.0, .268 (b=127/c=109)
9B	73.0 ± 1.7	84.0 ± 2.0	84.0 ± 1.0	+11.0, $< 10^{-4}$ (b=157/c=58)	0.0, 1.00 (b=87/c=87)
3.8-27B	78.3 ± 3.5	86.7 ± 1.3	86.0 ± 1.5	+8.4, $< 10^{-4}$ (b=134/c=59)	−0.7, .64 (b=55/c=61)

failure mode below. All evaluation commands pin the tokenizer explicitly (Section 5 explains why this sentence earns its place in the protocol).

One lineage, one recipe. All scales are Qwen3.x models (Qwen3.5 at 0.8B, 2B, 4B, 9B, and 27B; Qwen3.8 at 27B) trained identically: LoRA (r=32, α=64) on 1,624 sanitized multi-teacher trajectories (after a 32K render-budget filter; median ∼13.6K tokens, max ∼31K; chunked cross-entropy at 32,768 context, one epoch, 203 steps, identical target-module set). The RL arm anchors on the frozen SFT checkpoint and runs GRPO (Shao et al., 2024; von Werra et al., 2024) via a multi-turn verl pipeline (Sheng et al., 2024) with a six-term structural reward (format validity, tool-registry and payload-schema checks, numeric hygiene, a process heuristic, integrity penalties for unclosed calls and empty turns, and a redundant-reload penalty), group size 4, 300-prompt training set, 37 steps (one epoch, so the RL arm also sees each training prompt exactly once). Training telemetry for every RL run is healthy (reward curves rising or stable, entropy stable, ratios on-policy); the nulls below are not training failures.

Seed discipline and paired statistics. Every headline cell has ≥3 evaluation seeds per arm (identical weights, re-sampled trajectories, temperature 0.2). The single decision test is pooled paired exact McNemar (McNemar, 1947) over *all* 3×3=9 seed pairings on matched task IDs, with discordant counts (b= arm-1-only successes, c= arm-2-only) reported alongside. Measured single-run swings in this program reach ±6–12 points; no single-seed delta enters any table. We report mean ± sd over seeds.

Zero-shot base rows. Base models are served from their unmodified hub snapshots under the *same* native instrument, same suite, same analyzer, 3 seeds each, with their own tokenizers pinned. These rows anchor the decomposition: they measure what the protocol stack alone extracts from each base model before any training.

4 Results I: the capability-source decomposition

A caveat before any reading, and its resolution. Every number in this section is the strict, format-level answer rate of Section 3: it certifies that a trajectory ended in a well-formed final answer, not that the answer is correct. The paired comparisons below are unaffected (both arms share the metric), but the zero-shot (zs) levels are the layer where hallucinated-but-well-formed answers would hide. We therefore ran a numeric-grounding pass over every cell: each well-formed final is graded *grounded* when every number it asserts is traceable to retrieved tool payloads, the task statement, or the skill registry (scale-invariant echoes and simple arithmetic derivations at 1% tolerance; *rambling* finals flagged by a repetition heuristic). Table 2 reports both layers. Grounding rates are 80–94% everywhere, with zero-shot cells at the top of that range: the zs rows are not standing on invented numbers, and the decomposition survives the stricter metric with its shape intact and two of its features sharpened.

Four readings. (i) The SFT increment at 9B *strengthens* on the strict layer (+11.0 → +13.3): once invented numbers are excluded, supervised training is worth even more. (ii) The 4B catastrophe *deepens* (−45.3 → −56.0; only 32% of the 4B SFT arm's few finals are

Table 2: The two metric layers per cell (3 seeds/arm). Zero-shot: the untrained hub snapshot on the instrument of Table 1. Format level: well-formed final answer. Strict level: well-formed final whose asserted numbers are all traceable (grounding rate = strict/format). Paired pooled exact McNemar per increment.

	Zero-shot		SFT		GRPO	
Cell	format	strict	format	strict	format	strict
0.8B	45.3	35.7	63.3	50.3	64.0	50.7
2B	59.0	49.7	63.3	51.3	65.3	53.3
4B	67.3	63.0	22.0	7.0	21.7	9.7
9B	73.0	66.3	84.0	79.7	84.0	79.7
3.8-27B	78.3	64.0	86.7	80.0	86.0	78.7

Paired strict increments:

SFT−Zero-shot: 0.8B +14.7 ($p<10^{-16}$, b=288/c=114); 2B +1.7 (p=.50); 4B −56.0 ($p<10^{-4}$); 9B +13.3 ($p<10^{-4}$); 27B +16.0 ($p<10^{-4}$)

RL−SFT: 0.8B +0.3 (p=.63, b=129/c=138); 2B +2.0 (p=.28); 4B +2.7 (p=.026; repair-consistent, 5 cells tested); 9B 0.0 (p=1.0, b=c=104); 27B −1.3 (p=.40)

grounded): its answers are not merely rare, they are the least numerically reliable in the program. (iii) The zs curve is monotone on the format layer but flattens on the strict layer (35.7/49.7/63.0/66.3/64.0): the 27B base asserts more aggregate quantities, and its manual-queue share reflects complexity, not fabrication. (iv) The RL nulls hold on both layers at every scale (0.8B strict +0.3, p=.63); the single exception, the 4B strict increment (+2.7, p=.026), is the repair mechanism of Section 6 extending to numeric grounding, and we report it with the same multiplicity caveat we demand elsewhere. Grounding is necessary but not sufficient for correctness (direction and interpretation remain ungraded); the residual manual-review queue covers 44 task IDs (34 from the R6 cells, 10 from the 0.8B row), bounded at ±1–2 points per cell.

The RL column is flat. Table 1 is the paper's centerpiece, and its rightmost column is its headline: the identical GRPO recipe, applied to the SFT anchor of each scale, produces no detectable change anywhere: +0.7 at 0.8B (p=.752), +2.0 at 2B (p=.268), 0.0 at 9B (p=1.00, with the 87 discordant pairs splitting perfectly symmetrically across 900 paired readings), −0.7 at 27B (p=.64). The null is not for want of headroom on this instrument: the 0.8B cell is the largest-headroom cell in the program (its SFT increment is +18.0, leaving it furthest below the protocol ceiling), and its RL arm is the cleanest null of the four; at 2B the SFT anchor sits 17–30 points below the harness-measurable ceiling (the full-stack protocol of Appendix A extracts 96.7 from the same 9B SFT weights that score 84.0 here), and the RL arm was given exactly the recipe that produced a +54-point repair under the text protocol (Appendix A). Headroom below the ceiling, in other words, is not by itself an RL claim ticket: the reward must be able to see, and pay for, the thing the headroom is made of. Training telemetry is healthy at every scale. The null, in other words, is not that RL cannot move these policies (Section 6 shows it moving the 4B sharply, and the text-protocol lineage moved a 2B by +54 points); it is that on a strong-SFT native anchor the reward has nothing left to say about the task itself.

The protocol term dominates. The decomposition's largest single number is not in the table; it is the gap between protocol regimes. The same 2B SFT weights that score 63.3±3.8 on the native instrument score 4.3 (3-seed mean; seeds 4/3/6) under a text-protocol scaffold that describes tools in prose and parses free-form calls: +59 **points from the surrounding stack, against** +4.3 **from SFT and** +2.0 **from RL**. We stress what this comparison is: a *regime* effect, not a single-variable manipulation. Moving text-to-native changes the wire format, template enforcement, harness execution semantics, prompt structure, and tool-registry representation at once; the +59 measures the whole stack, and only the instrument-internal increments below it (zs → SFT → GRPO on one fixed protocol) are single-variable. The same holds, with an additional confound, at 9B: the 55.0 → 84.0

comparison (+29) crosses a pipeline generation (the text-era anchor predates two service fixes), which makes it a lower bound on the protocol effect, not an upper one. The hierarchy *protocol* $\gg$ *SFT* > *RL* holds at both scales where SFT is well-behaved; Section 6 exhibits a scale where SFT itself is the hazard, and the hierarchy's first term still dominates.

The base models already solve most of the instrument. The zs column is the decomposition's most uncomfortable row: an untrained 2B scores 59.0 on the same suite where the text-protocol SFT'd model scored 4.3. The native protocol (template-enforced call structure, a bounded and typed skill registry, deterministic execution) supplies most of what we had attributed to training. What the bases cannot do is visible in the behavioral columns, not the score: the 2B base burns 33–42% of its calls on redundant reloads and 35–42% of trajectories on turn-cap death, while the 9B base is nearly clean (0.6% redundancy). The 0.8B base sharpens the point to its limit: it emits 100% well-formed native calls zero-shot: the wire-format syntax costs nothing even at 0.8B parameters, yet burns ∼26% of calls on redundant reloads and 49–61% of trajectories on turn-caps. Protocol *syntax* is free; protocol *economy* is what scales. Task knowledge is the other thing that scales: the zs curve is monotone across the family, 45.3 (0.8B) < 59.0 (2B) < 67.3 (4B) < 73.0 (9B) < 78.3 (3.8-27B).

What SFT actually buys. At 0.8B, SFT is worth +18.0 points ($p<10^{-16}$): the largest increment in the program, earned entirely by loop discipline: redundancy collapses from ∼26% to 2–6%, turn-caps from 49–61% to 33–39%, while the tier breakdown shows the knowledge floor unmoved (the orchestration tier sits at ∼30% before and after). At 2B the increment is +4.3 and marginal (p=.065; directionally consistent across pairings but short of α=.05), and the behavioral columns show what it consists of: redundancy collapses from 33–42% to 2–6%, turn-caps from 35–42% to 19–21%, mean turns from 5.7–6.0 to 4.4–4.6. Two regularities fall out. First, at small scale SFT is *loop discipline*, not knowledge: the base already retrieves and computes; the training teaches it to stop reloading what it already loaded. Second, the SFT *level* is scale-flat at the bottom of the ladder: 0.8B and 2B both land at 63.3, while the SFT *increment* anti-correlates with scale (+18.0 at 0.8B vs +4.3 at 2B): the corpus buys exactly the discipline the base lacks, and no more. At 9B, both discipline and procedure arrive (+11.0). At 4B, Section 6 shows the same recipe destroying the base's own format compliance. The tier breakdown (Appendix Table 6) sharpens this: SFT's gains concentrate in the composition and recovery tiers (the 0.8B composition tier jumps 43 → 83, the single largest tier movement in the program), while at 2B it *hurts* the orchestration tier (52→27), and GRPO gives back a few points on the adversarial tier at 9B/27B.

5 Results II: how measurement manufactures effects

A significant result that was not. Our first 2B native cell produced the program's only significant RL effect: SFT 66.7 ± 4.6 vs. GRPO 70.7 ± 4.9, pooled p=.024 (b=138/c=102). Before freezing it, a hygiene audit found the cause: the evaluation client's `tokenizer-path` flag silently defaulted to the 9B composite's tokenizer, so *both* 2B arms were rendered through the 9B's thinking-default chat template, while the 2B's own template is non-thinking-default. The re-run with explicit tokenizer paths moved SFT to 63.3 and GRPO to 65.3: the corrected cell is parity (+2.0, p=.268). The template mismatch had inflated GRPO by 5.4 points and SFT by only 3.4, a *differential* inflation, which is precisely what a paired test converts into significance. We report both cells, with artifacts retained (Figure 2), for three reasons: the artifact is the field's default failure mode (template defaults are invisible in results tables and constant across arms, so nothing in a standard ablation exposes them); the inflation direction favored the trained arm, suggesting RL'd policies are *more* template-sensitive than their SFT anchors; and the episode is what upgraded our protocol to pin tokenizers explicitly in every evaluation command.

The +6.0 that was not. The same discipline retired our program's previous best RL result. A single-seed reading of the 9B bounded cell (SFT 84.0 vs. GRPO 90.0) motivated an entire 27B replication cycle; fresh 3+3 seeds put both arms at exactly 84.0 (b=87/c=87, p=1.0),

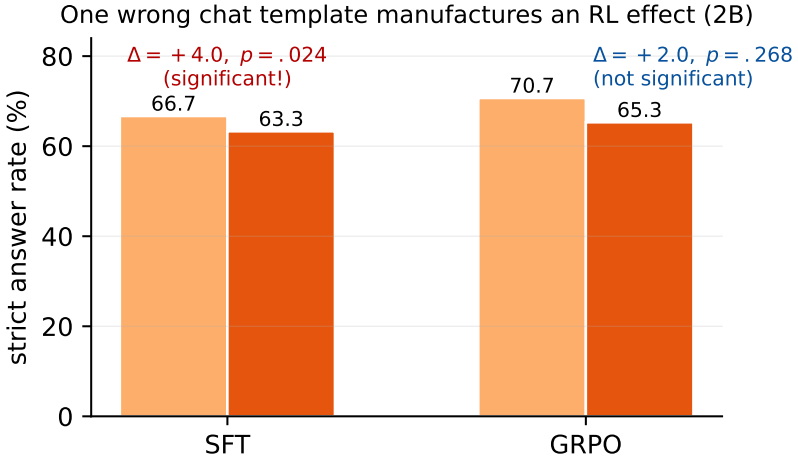


Figure 2: The template exhibit. Same 2B weights, same 100 tasks, same analyzer; only the chat template differs. Light bars: original runs under a shared 9B thinking-default template (both arms); dark bars: the tokenizer-correct re-run under the 2B's own non-thinking-default template. Under the shared-but-wrong template GRPO appears +4.0 over SFT at p=.024; under the correct tokenizer the delta is +2.0 at p=.268.

Table 3: The 4B case study (3 seeds/arm, native instrument, paired exact McNemar vs. the preceding stage). Behavioral forensics from the frozen analyzer over all 300 trajectories per arm.

Stage	strict answer	hybrid calls	skill-load failures	redundant loads
Zero-shot	67.3 ± 0.6 (67/68/67)	0/300 tasks	0/300 tasks	0–2%
+ SFT	22.0 ± 3.6 ($p<10^{-4}$, b=51/c=459)	60–90% of calls	dominant failure	loops
+ GRPO	21.7 ± 3.5 (p=.895, b=113/c=116)	0	0	81%

Strict (grounded) level: base 63.0, *SFT* 7.0, *GRPO* 9.7*; grounding* 92%/32%/45%

no GRPO seed approaches 90, and the original anchor files no longer exist in archivable form. Single-seed agent evaluation manufactures effects at exactly the scale practitioners care about.

Two infrastructure masquerades. Earlier in the program, two silent backend faults (a degraded statement service; a proxy polling fault) moved cells by +14.4 and ∼10 points respectively ($p \approx 3 \times 10^{-6}$ under pooling), each larger than every algorithmic effect we have ever measured on this instrument. Both were found by environment fingerprinting (per-row backend-error counts), not by score inspection; both masqueraded as training effects for weeks. The general lesson is architectural: in tool-agent evaluation, the environment is inside the causal path of every score, and unlogged environment state is indistinguishable from model capability.

Implications for reported agent results. Combining this section's artifacts (template, seeds, backend) with the decomposition of Section 4 yields an uncomfortable arithmetic: the plausible measurement variance between two honest evaluations of the *same weights* (template + seeds + backend) exceeds the total contribution of RL post-training by roughly an order of magnitude. Any RL-vs-SFT agent claim that does not pin tokenizers, pool paired seeds, and fingerprint the environment is, on this instrument's evidence, unfalsifiable.

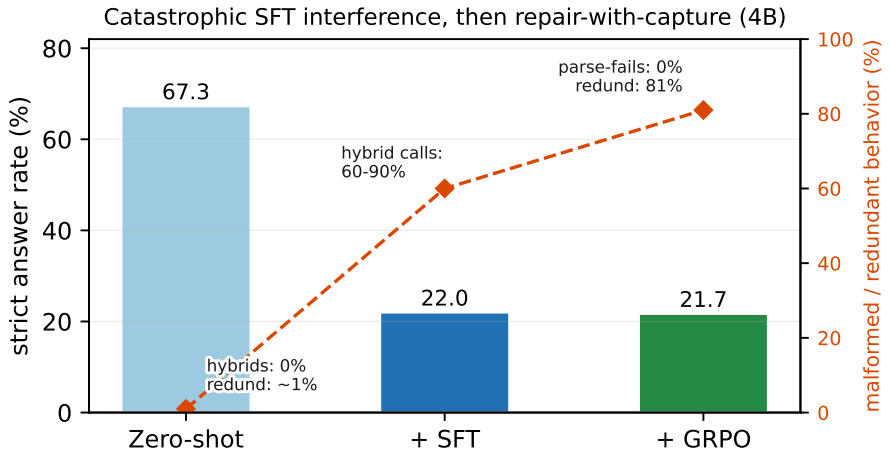


Figure 3: The 4B arc. The base is format-clean (zero hybrid emissions across 300 runs at temperature 0.2); the SFT LoRA induces hybrid derailment (60–90% of calls) and the score collapses by 45 points; GRPO eliminates the malformation (parse failures to zero) while a rewarded shortcut absorbs the recovered turns (redundancy 81%).

6 Results III: catastrophic SFT interference at 4B

The anomaly. The 4B scale breaks every trend in Table 1: its *base* outscores its SFT derivative by 45 points (67.3 ± 0.6 vs. 22.0 ± 3.6; pooled $p<10^{-4}$, $b=51/c=459$). The failure mode is a format pathology: at sampling temperature the SFT'd model emits JSON/XML *hybrid* calls (native call syntax wrapping JSON-argument payloads) on 60–90% of tool invocations; the registry rejects them, the model retries, and the trajectory dies at the turn cap. Greedy decoding is clean. Aggregate training metrics are not just unhelpful but *inverted*: the 4B SFT run has better loss and token accuracy than the healthy 2B run on the same corpus; format robustness lives in the ∼1–2% of tokens that encode call structure, and aggregate likelihood averages it away.

Falsifying our own hypothesis. We initially attributed the fragility to a base-model prior: the 4B had thrived under the text protocol's JSON-argument style (53.7 in the earlier regime), so we predicted, and pre-registered, that the 4B base would emit JSON-hybrid calls natively under the XML protocol. The zero-shot rows falsified it: across 300 base trajectories at temperature 0.2, *zero* hybrid emissions and *zero* skill-load failures, with 0–2% redundancy (Table 3). The fragility is induced by the SFT LoRA itself: the same 1,624-trajectory corpus, LoRA configuration, and single epoch that add +18.0 at 0.8B, +4.3 at 2B, and +11.0 at 9B. The 4B run is the program's second-cleanest by training metrics (final loss 0.210, token accuracy 93.8%, no instability), so the damage is invisible to aggregate likelihood. To our knowledge this is the cleanest documented case of *catastrophic SFT interference* in a tool-use agent: a training procedure that improves aggregate likelihood while destroying temperature-robust protocol compliance. A pre-registered generation control at 27B shows the hazard is not confined to this scale: transplanting the recipe, byte-identical, onto same-scale Qwen3.5-27B replicates the base arm (76.3 vs 78.3 zero-shot, inside seed noise, so the ladder's top is a scale result) yet *inverts* the SFT arm (50.0, a -26.3 regression at $p<10^{-36}$) with the mirror-image signature: format compliance survives while loop discipline collapses (redundant calls ∼48%, the atomic tier falls 91 → 49). With hazards at 4B and 3.5-27B and clean outcomes at 0.8B, 2B, and 9B, neither scale nor generation alone predicts where this recipe is safe (Appendix A).

RL repairs the format and captures a shortcut. GRPO on the 4B SFT anchor returns a perfect null on the score (21.7 vs. 22.0, $p=.895$), but the behavioral forensics show it did not stand still. Hybrid calls drop to *zero*: the reward's format terms penalize

malformation, and the policy concedes. The recovered turns flow not toward task completion but toward `load_skill`, an always-succeeding documentation tool whose successful execution the composite reward *pays for*: redundancy climbs to 81%, trajectories end in doc-loading loops at the turn cap, and the outcome term of the reward never rises (training score flat at 6.3–7.1 throughout, vs. a rising curve at 2B). The triad is complete: **RL repairs what the reward penalizes, captures what it rewards, and does nothing where SFT has saturated the task**. The 4B arc exhibits all three faces on one checkpoint pair.

The eval gate this motivates. Nothing in a standard evaluation pipeline would have caught the interference before deployment: greedy smoke tests pass, aggregate training metrics improve, and single-seed score variance (±13 points on the text-era 4B anchor) could have hidden the level shift for another cycle. The minimal gate that catches it: a *temperature-0.2 malformed-call probe* on a held-out suite before any SFT anchor is trusted: count hybrid emissions and registry rejections per trajectory, not just final answers. We have added it to our own pipeline; Section 7 folds it into the reporting protocol.

7 Synthesis, limitations, and conclusion

An attribution frame for agent capability. The decomposition suggests a budget rule that inverts current practice. Before spending on RL post-training, measure: (i) the protocol gap, measuring how far the current scaffold sits below the native instrument (worth +59 points at 2B in this study); (ii) the SFT increment over the zero-shot base on the native protocol (+4.3 to +11.0 here, and negative at one scale); (iii) the harness-measurable ceiling (the full-stack protocol extracts 96.7 from the same 9B weights that score 84.0 bounded). Only the residue justifies RL spend, and on this instrument the residue was statistically indistinguishable from zero at every scale, while the reward's real leverage showed up behaviorally: format repair and shortcut capture, not task gains.

Limitations. (i) One model family (Qwen3.x: 3.5 at 0.8B/2B/4B/9B/27B, 3.8 at 27B) and one domain (finance); the interference scales and the zs curve are family properties until replicated, and a 27B generation control shows the SFT hazard itself is generation-dependent in *form* (Appendix A). (ii) LoRA-only SFT/RL at one recipe point; full-parameter runs or longer schedules could shift the increments, though the text-protocol lineage (Appendix A) shows this reward stack can move a policy by +54 points when headroom and reward visibility both exist. (iii) The headline metric is format-level; the numeric-grounding layer of Table 2 bounds the hallucination exposure (grounding 80–94%, zs cells highest) and preserves every paired conclusion, but full semantic grading (direction and interpretation of answers) remains open and is future work. (iv) The zs curve's strict-layer flattening at 27B is queue-driven (aggregate quantities) and bounded by the manual-review margin. (v) Bounded horizons (8 turns); the long-horizon twin (200 tasks, 16–24 turns) replicates the null (p=.81) but at one seed pair.

Scope of the null. The null should be read as a property of the reward-on-anchor pair, not of RL per se: where the reward could see failure it repaired it (format), where it paid for something cheaper it captured that instead (the documentation loop), and where the SFT anchor left no observable deficit it moved nothing (0.8B/2B/9B/27B task outcomes). The positive control is our own text-protocol lineage, where the same reward family repaired a collapsed 2B by +54 points once the loop became visible to it (Appendix A).

Conclusion. We ran the attribution experiment the agent literature has largely skipped: one factor at a time, one instrument, paired statistics, five scales. The weights term, the term the field budgets for, was the smallest measurable contributor at every scale. The protocol term was the largest. The evaluation pipeline itself could manufacture or retire every "effect" we observed. And the one training result that survived scrutiny was a warning: the same SFT that teaches a small agent its loop discipline can, at one scale, destroy the format compliance the base already had, while RL, given that broken anchor, spent its optimization budget repairing the reward's complaints and exploiting its payouts. *The protocol is the agent; the weights are the last mile.*

References

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 2023.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-decoupled normalization policy optimization for multi-reward RL optimization, 2026. URL `https://arxiv.org/abs/2601.05242`.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Negin Raoof, Richard Zhuang, Marianna Nezhurina, Etash Guha, Atula Tejaswi, Ryan Marten, Charlie F. Ruan, Tyler Griggs, Alexander Glenn Shaw, Hritik Bansal, E. Kelly Buchanan, Artem Gazizov, Reinhard Heckel, Chinmay Hegde, Sankalp Jajee, Daanish Khazi, Emmanouil Koukoumidis, Xiangyi Li, Hange Liu, Shlok Natarajan, Harsh Raj, Nicholas Roberts, Ethan Shen, Nishad Singhi, Michael Siu, Ashima Suvarna, Hanwen Xing, Patrick Yubeaton, Robert Zhang, Leon Liangyu Chen, Xiaokun Chen, Steven Dillmann, Saadia Gabriel, Xunyi Jiang, Anurag Kashyap, Boxuan Li, Yein Park, Minh Pham, Sujay Sanghavi, Lin Shi, Ke Sun, Yixin Wang, Zhiwei Xu, Erica Zhang, Siyan Zhao, Wanjia Zhao, Jenia Jitsev, Alex Dimakis, Benjamin Feuer, and Ludwig Schmidt. Openthoughts-agent: Data recipes for agentic models, 2026. URL `https://arxiv.org/abs/2606.24855`.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024.

Leandro von Werra, Edward Beeching, Lewis Tunstall, et al. TRL: Transformer reinforcement learning. `https://github.com/huggingface/trl`, 2024.

Kun Wang, Gavin Zhang, Yaze Geng, Lei Tang, Yaoyang Yi, Zonghan Wu, Yifan Hu, Qingsong Wen, and Yilei Shao. Mint-agent: Introducing finance-native agentic foundation models, 2026. URL `https://arxiv.org/abs/2608.16386`.

Yijia Xiao, Rujun Han, Yanfei Chen, Zifeng Wang, Ke Jiang, Zhongying CuiZhu, Vishy Tirumalashetty, Wei Wang, Burak Gokturk, Tomas Pfister, and Chen-Yu Lee. FinanceHarness: Autonomous financial deep research with execution scaffolding and PIT evaluation, 2026. arXiv preprint arXiv:2607.27853.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Han Zhang, Ruibin Zheng, Zexuan Yi, Zhuo Zhang, Hanyang Peng, Hui Wang, Zike Yuan, Cai Ke, Shiwei Chen, Jiacheng Yang, Yangning Li, Xiang Li, Jiangyue Yan, Yaoqi Liu, Liwen Jing, Jiayin Qi, Ruifeng Xu, Binxing Fang, and Yue Yu. GEPO: Group expectation policy optimization for stable heterogeneous reinforcement learning, 2026. URL https://arxiv.org/abs/2508.17850.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Yuxin Zuo, Zikai Xiao, Li Sheng, Fei Huang, Jianhong Tu, Yuxuan Liu, Tianyi Tang, Xiaomeng Hu, Yang Su, Qingfeng Lan, Yantao Liu, Qin Zhu, Yinger Zhang, Bowen Yu, Haiquan Zhao, Haiyang Xu, Jianxin Yang, Jiayang Cheng, Junyang Wang, Lianghao Deng, Mingfeng Xue, Tianyi Bai, Yang Fan, Yubo Ma, Yucheng Li, Zeyu Cui, Zhihai Wang, Zhihui Xie, Zhuorui Ye, An Yang, Dayiheng Liu, Jingren Zhou, and Ning Ding. Qwen-agentworld: Language world models for general agents, 2026. URL https://arxiv.org/abs/2606.24597.

AI-use statement

All experimental code, analysis scripts, figure generation, and manuscript drafting for this paper were produced with AI assistance under direct human supervision. All experimental decisions (design, seeding, pre-registrations, scoping, and retirements) were made by the authors, and every number in the paper is transcribed from logged program artifacts rather than generated text.

Reproducibility statement

All results use a frozen 100-task suite, a deterministic transcript analyzer, and a fixed paired-statistics protocol (3 evaluation seeds per arm, pooled exact McNemar over all seed pairings); per-seed cells are reported in Appendix B. Training uses identical LoRA SFT recipes per scale (one epoch on 1,624 distilled trajectories) and a fixed GRPO configuration (300-prompt set, 37 steps = one epoch, group size 4, six-term structural reward). Evaluation commands pin tokenizer paths explicitly (Section 5). Model identities (Qwen3.5 at 0.8B/2B/4B/9B/27B, Qwen3.8 at 27B), environment fingerprints, and per-run logs are retained; we will release the task suite, analyzer, training configurations, and per-trajectory outputs upon publication.

A Text-protocol lineage results

The native-instrument results of the main text sit on a completed earlier program run under a *text-protocol* scaffold: tools described in prose, JSON-argument calls parsed from free-form output. That regime produced the program's positive RL result and the mechanism evidence for reward visibility; we summarize it here because it bounds the claims of the main text. All rows are regime-flagged and never share a table with native-protocol cells.

Seven formulations, one null. On the 2B SFT anchor (strict answer rate 4.3), seven single-turn RLVR objective formulations (token-level GRPO and DAPO-style defaults (Yu et al., 2025), GSPO (Zheng et al., 2025), chunk-level importance sampling, GEPO (Zhang et al., 2026), length-normalized LUSPO, Ω-Pure, and GDPO (Liu et al., 2026; DeepSeek-AI,

2025)) shared prompts, rewards, anchor, optimizer, and hardware. All landed in $[0.15, 1.0]$ after correction: the single-turn reward never observes the tool loop, and its failure modes are invisible in the rollout text it grades.

Closing the loop: the +54 repair. Executing the tool environment during RL rollouts (multi-turn GRPO, outcome reward visible) repaired the same 2B policy from 4.3 to 58.7 ($p \approx 10^{-129}$ pooled), matching the 2×-larger 4B SFT model's 53.7. Step-wise checkpoints locate the gains inside the pre-saturation window: $4 \to 52$ by step 25 ($p < 10^{-15}$), peak 64 at step 50, flat to step 100 ($p = .54$ for the 25→100 contrast); advantages collapse near step 25. A two-epoch SFT control holds at 4.0 ($p = 1.0$), so the delta is RL's, not more gradient steps. Credit-placement variants (turn-level vs. trajectory) land inside the repair band (53.7–57.3), pairwise indistinguishable.

Scale-side program cells. Four seeded paired cells on the 9B/27B lineage all null: 9B full-stack 96.7/95.3 ($p = .33$); 9B bounded 84.0/84.0 ($p = 1.0$); 27B bounded 86.7/86.0 ($p = .64$); the 200-task long-horizon twin 59.0/60.5 ($p = .81$, one seed pair). 27B GRPO training telemetry is healthy (reward 8.5–9.5 of 11, entropy 0.44–0.74, ratios on-policy). A pre-registered saturation rule skipped the 27B full-stack RL run: the fixed harness leaves no clean-task headroom (94.0, 100% completion on environment-clean tasks).

The harness ladder. On frozen 9B SFT weights, five cumulative harness interventions (service fixes, statement repair, retry policy, compaction, full-stack dispatch) raise the full-stack 50-task score $50.0 \to 56.0 \to 70.0 \to 84.0 \to 96.0$ (+46 total). Rungs are single-run and labeled as such; the final state re-measured at 3 seeds is 96.7 ± 2.5. This is the ceiling measurement the main text's headroom claims reference.

A 27B generation control (pre-registered). To separate scale from generation at the top of the ladder, we transplanted the anchor recipe onto Qwen3.5-27B, byte-identical to the 3.8-27B recipe except the base checkpoint. The base arm replicates across generations: 76.3 ± 2.3 vs 78.3 ± 3.5 zero-shot, a -2.0 format delta inside seed noise, so the ladder's top reading is a scale result. The SFT arm does not replicate:

Table 4: Generation control at 27B (3 seeds/arm, format-level metric, pooled paired exact McNemar).

Generation	Zero-shot	SFT	Δ(SFT−Zero-shot), p
3.5-27B	76.3 ± 2.3	50.0 ± 2.6	-26.3, $< 10^{-36}$ (b=65/c=302)
3.8-27B	78.3 ± 3.5	86.7 ± 1.3	$+8.4$, $< 10^{-4}$

The 3.5-27B failure mode is the mirror image of the 4B case (Section 6): format compliance survives (98.9–99.9% well-formed calls) while loop discipline collapses: redundant-call share ∼48% against ∼0% zero-shot, turn-caps 43–51, and the atomic tier falls $91 \to 49$. Numeric grounding stays high (86–96%), so the pathology is behavioral, not fabrication. The training run itself is healthy (final loss 0.215, token accuracy 92.9%, no instability; the corpus is imitated as well as at any scale), which rules out a diverged-run explanation: the recipe converges *and* damages. SFT interference in this stack therefore takes at least two forms: format fragility (4B) and discipline regression (3.5-27B). At fixed scale, generation decides the sign of the treatment; across scales no rule emerges: both failures are 3.5-generation, yet the same generation trains cleanly at 0.8B, 2B, and 9B. An initial 37-step GRPO run on the regressed anchor (the standard recipe of Section 3) improves 50.0 to 55.0 (+5.0, $p = .011$, pooled) without reducing the redundant-loop share (43–49%): RL polishes the metric but does not heal the anchor, echoing the repair-and-capture pattern of Section 6. Full forensics (greedy probes, optimizer-state analysis) is beyond this paper's scope and treated elsewhere.

B Per-seed tables

Per-seed strict answer rates underlying every mean in Table 1 and Table 3 (frozen analyzer, format-level metric; artifacts retained with environment fingerprints).

Table 5: Per-seed cells (100 tasks each). The two 2B training-arm pairs are the template exhibit of Section 5 and Fig. 2: the (2B template) rows are the tokenizer-correct frozen-protocol evaluations used everywhere in the main text, and the (9B template) rows are the original evaluations under the shared-but-wrong 9B thinking-default template, retained as the exhibit. The 2B base was always evaluated under its own template, hence the single Zero-shot row.

Scale	Arm	seed 1	seed 2	seed 3
0.8B	Zero-shot	51	39	46
0.8B	SFT	61	62	67
0.8B	GRPO	62	64	66
2B	Zero-shot	57	57	63
2B	SFT (2B template)	65	66	59
2B	GRPO (2B template)	59	69	68
2B	SFT (9B template)	72	64	64
2B	GRPO (9B template)	74	65	73
4B	Zero-shot	67	68	67
4B	SFT	27	19	20
4B	GRPO	17	25	23
9B	Zero-shot	74	71	74
9B	SFT	86	82	84
9B	GRPO	85	84	83
3.8-27B	Zero-shot	82	75	78
3.8-27B	SFT	87	88	85
3.8-27B	GRPO	84	87	87

Table 6: Per-tier strict answer rates (%, pooled over 3 seeds; 25 tasks/tier/seed). Tier A: atomic lookups; B: multi-ticker compositions; C: platform orchestration; D: adversarial market-shock recovery.

Cell	A	B	C	D
0.8B Zero-shot	52	43	23	47
0.8B SFT	79	83	31	61
0.8B GRPO	80	77	35	64
2B Zero-shot	83	49	52	52
2B SFT	75	85	27	67
2B GRPO	80	91	37	53
9B Zero-shot	87	89	47	69
9B SFT	87	92	65	92
9B GRPO	93	95	64	84
3.8-27B Zero-shot	93	85	51	84
3.8-27B SFT	97	91	79	80
3.8-27B GRPO	100	92	77	75

Reading the tier breakdown. Three patterns are invisible in the aggregate table. (i) SFT's gains concentrate in the discipline-heavy tiers (B and D): loop discipline is exactly what those tiers demand. (ii) At 2B, SFT *hurts* the orchestration tier (52→27): orchestration

exceeds the trained 2B's capacity, and training replaced the base's conservative fallbacks with confident incomplete loops, a within-scale echo of the 4B interference of Section 6. (iii) GRPO gives back a few points on the adversarial tier at 9B/27B (92→84, 80→75) while gaining on compositions at 2B (85→91): the consistent aggregate null conceals small opposing tier-level movements, none of which survives correction for 25-task cells; we report them as direction, not effect.

Table 7: Per-seed strict (grounded) answer rates underlying Table 2.

Scale	Arm	seed 1	seed 2	seed 3
0.8B	Zero-shot	36	32	39
0.8B	SFT	50	50	51
0.8B	GRPO	49	52	51
2B	Zero-shot	45	48	56
2B	SFT	54	55	45
2B	GRPO	47	55	58
4B	Zero-shot	62	67	60
4B	SFT	10	5	6
4B	GRPO	9	8	12
9B	Zero-shot	66	63	70
9B	SFT	82	78	79
9B	GRPO	81	77	81
3.8-27B	Zero-shot	69	61	62
3.8-27B	SFT	82	80	78
3.8-27B	GRPO	75	80	81