DOES POST-TRAINING TAX EXACT ARITHMETIC? IT DEPENDS ON THE ROUTE

Anonymous authors
Paper under double-blind review

ABSTRACT

Does post-training reduce accuracy on exact arithmetic, or can the evaluation interface change that conclusion? We study 16 released checkpoint pairs on six integer-arithmetic tasks, including seven documented post-training comparisons from three pretrained bases. We first compare checkpoints using the same completion input, without a chat template. We then compare completion and chat inputs for the same post-trained checkpoint. The common comparison between base completion and post-trained chat changes both checkpoint and interface. In a pre-registered replication that constrains outputs to integers, the common and fixed-completion comparisons give opposite conclusions in 24 of 96 pair–task combinations: one indicates significantly higher accuracy and the other significantly lower accuracy after Holm correction. In tokenizer-compatible pairs, the size of the checkpoint effect depends on the input interface. Two selected problem subsets show reversals under both constrained decoding and free generation with first-integer scoring. This does not establish independence from generation and scoring conditions. Some deficits are no longer detected after changing the prompt, whereas others persist. For OLMo-3, we ask subtraction questions without a chat template and allow free generation, scoring the first integer. On 14-digit problems, accuracy is 52.4% for the base checkpoint and 1.4% for the final checkpoint. Among answers that are correct at the base checkpoint but wrong at the final checkpoint, most first become wrong at the supervised fine-tuning stage. For Llama-3.1/Tulu, gains measured with the same completion input become negative estimates under the question form. A secondary transplant of base-checkpoint states corrects some selected expression-form errors but none of ten tested 14-digit subtractions. These findings concern accuracy under specific evaluation conditions. Comparisons of post-training should distinguish checkpoint, interface, and answer-extraction conditions before attributing score differences to changes in arithmetic ability.

1 INTRODUCTION

Post-training is meant to add capabilities, but it can also take them away. Instruction-tuned models lose completion-style infilling ability their bases retain—an "instruction-tuning tax" (Chang et al., 2026); whether they outperform their bases on mathematical benchmarks varies with prompt condition and grader (Munjal et al., 2026); instruction tuning can erode context awareness (Wang et al., 2025); and alignment fine-tuning regressed public NLP datasets—the "alignment tax" (Ouyang et al., 2022).

Does such a tax extend to exact arithmetic? Every item has one determined answer, the operation is explicit, and digit length is an adjustable, locally calibrated difficulty variable, allowing us to measure exact-answer emission rather than a broad score with numeracy as one ingredient (Yang et al., 2025; Mahendra et al., 2025). The common base-completion-to-chat evaluation path (bundled comparison) changes both checkpoint and route, and prompt construction alone moves scores even under meaning-preserving changes (Sclar et al., 2024; Mizrahi et al., 2024). Few-shot completion is also a common base-evaluation practice; the bundled comparison studied here is specifically the zero-shot completion-to-chat path.

These comparisons can support opposite conclusions. On the same 2,000 fresh-grid items involving multiplication by a single digit, the Gemma 4 E2B pair has a fixed-completion checkpoint difference

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.

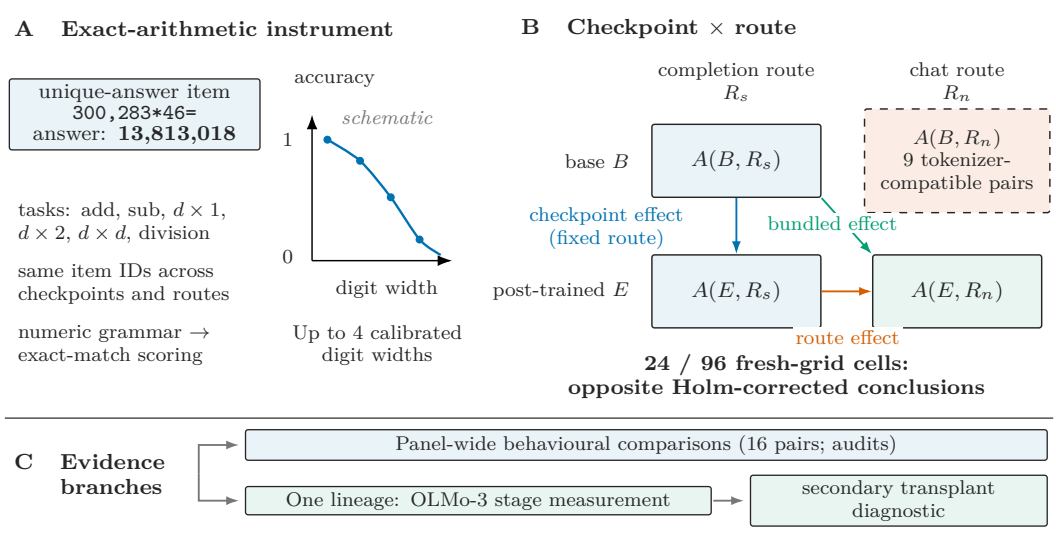

Figure 1: Exact arithmetic exposes opposite corrected conclusions from different evaluation paths. *A:* paired unique-answer tasks and a schematic accuracy curve over calibrated digit widths. *B:* the three labelled comparisons; 24 of 96 fresh-grid cells have opposite Holm-corrected checkpoint and bundled conclusions, and the dashed fourth condition completes the factorial for nine tokenizer-compatible pairs. *C:* panel-wide comparisons and selected-cell audits, separate from OLMo-3 stage measurement and its secondary transplant diagnostic.

of -43.70 pp versus a bundled difference of $+20.65$ pp, both Holm-significant: two commonly used comparisons support opposite conclusions about the same released checkpoint pair because they estimate different quantities. Neither interface is an a priori neutral assay of capability: checkpoint identity and input format can interact, and the measured outcome depends on the generation and scoring contract.

We therefore make the measurement route part of the estimand. For the same paired items, we report the *fixed-route checkpoint contrast*, the *route contrast at a fixed post-trained checkpoint*, and the *bundled contrast* in which both change. Let B be the base checkpoint, E its post-trained release, R_s the shared-completion route, and R_n the official chat template under the registered system configuration. With strict exact-answer accuracy A, the three contrasts are $\Delta_{\text{checkpoint}} = A(E, R_s) - A(B, R_s)$, $\Delta_{\text{route}} = A(E, R_n) - A(E, R_s)$, and $\Delta_{\text{bundle}} = A(E, R_n) - A(B, R_s)$. This ordered three-condition decomposition (Figure 1) answers three different questions; it does not uniquely attribute the bundled difference to checkpoint and route. Adding $A(B, R_n)$ on tokenizer-compatible pairs completes the 2×2 factorial and measures the interaction and checkpoint contrast through chat (Section 4.2). The three comparisons can support opposite family-wise-corrected conclusions, as the pre-registered, item-disjoint replication demonstrates. Exact arithmetic makes this dependence measurable with unique answers and paired items.

Operational definitions. A Holm-significant negative fixed-route checkpoint contrast is a *checkpoint-package deficit*; the significant positive case is a *gain*. We reserve *post-training tax* and *post-training gain* for the seven *documented contrasts (T1)* with pretrained bases (Appendix A); the other pairs remain checkpoint-package comparisons. Only OLMo-2 1B belongs to this subset in the factorial; SmolLM3 uses a mid-training reference. "Checkpoint" denotes the pinned package including tokenizer and stop convention. Tax/gain therefore names exact-answer emission under a controlled decoding contract, not a weight-only causal effect or a route-independent change in arithmetic ability. A *cell* is one checkpoint pair (contrast) crossed with one task. The *fresh grids* are the newly generated replication items, distinct from the discovery grids. A *reversal* is a cell where the bundled and fixed-completion comparisons reach opposite Holm-significant conclusions with a significant route contrast. A cell is at the *floor* when base chat accuracy is at or below 1%.

Research questions. **RQ1:** Under a shared route, are base–post-trained checkpoint-package effects uniformly negative? **RQ2:** Does route interact with checkpoint identity, changing the magnitude or direction of the effect? **RQ3:** How often do bundled and route-held-fixed comparisons support

opposite corrected conclusions? **RQ4:** Which deficits persist under a question form, at which stage do answers first fail, and which selected losses recover under base-state transfer?

Contributions. (i) We measure how often the direction of a post-training comparison flips: the three-condition comparison and pre-registered replication yield 24 reversals among 96 fresh-grid cells, robust to a global correction. In nine tokenizer-compatible pairs, checkpoint–route interactions are Holm-significant in 12 of 14 cells above the floor; the factorial excludes OLMo-3 and Tulu, so it adds breadth, not the basis for the documented tax claims. (ii) Free-generation and prompt-form audits separate deficits no longer detected from deficits that persist. Under the question form, OLMo-3's three-digit subtraction loss is no longer detected, while its 14-digit loss persists from SFT. Tulu's fixed-completion gains become negative question-form estimates. Two same-slice reversals persist under free generation, which does not establish contract independence. A secondary transplant of OLMo-3 base states recovers selected lost answers, without identifying the deficit's natural mechanism. Table 1 gives the evidence boundaries.

2 Related Work

Post-training taxes in other domains. Degradation from post-training is documented outside arithmetic. The alignment tax was reported as performance regressions on public NLP datasets (Ouyang et al., 2022); Chang et al. (2026) document an instruction-tuning tax in code infilling; Wang et al. (2025) trace degraded context awareness under chat-template wrapping to role bias in instruction-tuning data; instruction-tuned helpfulness can collapse under trivial lexical constraints (Baghaei Potraghloo et al., 2026); and Munjal et al. (2026) report that base-versus-instruct rankings on math benchmarks vary with prompt condition and grader. Post-training forgetting is tracked at the item level by Harmon et al. (2026), who distinguish $1 \to 0$ forgetting from $0 \to 1$ backward transfer—our paired four-state records apply the same logic—and across training stages by CapTrack (Thede et al., 2026); our same-item measurement follows the released OLMo-3 lineage (Section 4.5). We compare checkpoint, route, and bundled effects on exact-answer emission.

Alignment and elicitation. Zhou et al. (2023) propose the superficial-alignment hypothesis, showing that limited supervised data can elicit behaviour largely learned in pretraining. Lin et al. (2024) show that in-context examples can align base models without parameter updates, making elicitation a substantive part of base–post-trained comparisons.

Measurement-side sensitivity. Meaning-preserving formatting changes affect scores (Sclar et al., 2024), motivating multi-prompt evaluation (Mizrahi et al., 2024); prompt optimization can change rankings (Sadjoli et al., 2025). BOSE (Luan et al., 2025) and soft-prompt tuning (Erkan et al., 2026) optimize benchmark construction for bases; we instead compare fixed registered routes. Heuristic grading and rigid extraction can create apparent prompt sensitivity (Hua et al., 2025); our unique-answer items retain strict, tolerant, and parseability views, with identical strict/tolerant decisions in every cell (Appendix D).

The lm-evaluation-harness supports completion evaluation and optional chat-template application (Sutawika et al., 2024), making the base-completion-to-chat path an implemented evaluation practice.

Identification-strategy lineage. Wang et al. (2025) compare inputs with and without chat templates; Sanz-Guerrero et al. (2026) compare unwrapped bases with wrapped and unwrapped instruction-tuned checkpoints to isolate template-induced overconfidence. We apply this three-condition logic to multi-lineage exact-answer contrasts with familywise error control, then estimate checkpoint–route interactions and corrected-conclusion reversals. The contribution is the controlled comparison these constructions enable; they do not identify weight-only causal effects.

Mechanistic analyses of post-training. Cross-model patching reveals enhanced entity tracking after fine-tuning (Prakash et al., 2024) and tests retained knowledge and behavioural directions across checkpoints (Du et al., 2025). Interventions also recover post-training losses: attention steering (Wang et al., 2025), function vectors (Jiang et al., 2025), and conjugate prompting (Kotha et al., 2024). Crosscoder feature ablation across matched supervised and reinforcement-learning branches (Shi et al., 2026), mediation and patching for instruction vectors (Bigoulaeva et al., 2026), and edge-attribution

patching (Zhang et al., 2026) localize post-training effects; Zhou (2026) cross-patches upstream states across cumulative stages without isolating training increments. Our secondary diagnostic locates where each answer first fails, then transfers same-item states along a cumulative released lineage to test exact-answer restoration; it likewise does not isolate individual training increments.

Interventional interpretability. The Jacobian lens reads residual-stream states through corpus-averaged Jacobian transports (Gurnee et al., 2026); causal tracing and activation patching localize behavior-relevant states by substitution (Meng et al., 2022); representation-level steering moves behavior along chosen directions (Zou et al., 2023). Because subspace interventions can create an interpretability illusion (Makelov et al., 2024), our OLMo-3 transplant study is reported as sufficiency on selected items, with matched controls (Section 4.5).

Arithmetic and numeracy benchmarks. Number Cookbook introduces the NUPA benchmark of number understanding across representations and tasks (Yang et al., 2025); broader numeracy evaluations frame numeracy as natural language inference beyond exact arithmetic (Mahendra et al., 2025); and length-generalization studies ask how far learned arithmetic extends beyond trained operand lengths and counts (Jelassi et al., 2023; Cho et al., 2024; 2025). We instead use exact arithmetic as a controlled instrument for a post-training question: the narrow battery allows item pairing and a controlled digit sweep. Stolfo et al. (2023) use causal mediation to locate components involved in arithmetic reasoning. Nikankin et al. (2025) identify heuristic arithmetic circuits, motivating caution about inferring algorithms from correct answers. Reasoning evaluations show why a token budget is part of the measurement condition (Huang et al., 2025). The inferential layer uses paired McNemar tests (McNemar, 1947) with Holm's familywise step-down correction (Holm, 1979) and Newcombe paired-score intervals (Newcombe, 1998).

Our contribution is not the algebraic decomposition or activation transfer itself. We establish when route-explicit comparisons yield opposite multiplicity-controlled conclusions on exact-answer tasks, and distinguish deficits that are no longer detected under changed elicitation from those that persist and can be localized along a released training lineage.

3 METHOD

3.1 CONTROLLED EXACT-ARITHMETIC INSTRUMENT

The six tasks are addition (add), non-negative subtraction (sub), multiplication by one digit ($d \times 1$), by two digits ($d \times 2$), and balanced $d \times d$ multiplication, plus exact division constructed from $a = q \times b$. The multiplication variants locate effects within an operation: a single-digit multiplier exercises memorized products and a short carry chain; two digits add a partial-product combination; balanced multiplication scales digit interactions quadratically. Every item has a unique integer answer. For each base checkpoint and task, calibration selects a four-point digit-width window (Appendix B.4); contrasts use the same frozen item IDs. Capacity-aware redistribution leaves 39 of 96 fresh cells with three observed points and 78 without a digit-2 row (Appendix D).

The visible prompt is `Output ONLY a number.`, a newline, and {a}{op}{b}=, with three-digit grouping commas (for example, `300,283*46=`). A fixed character-level numeric grammar constrains the permitted answer emissions—the standard grammar-constrained decoding device (Geng et al., 2023; Park et al., 2024), used here to hold the output format fixed across checkpoints and routes. It permits unsigned decimal integers, including zero, either ungrouped or with valid three-digit comma grouping, without leading zeros in multi-digit numerals, whitespace, or signs, up to the task's registered digit cap; after any syntactically complete integer (including one shorter than the answer) the decoder also admits the registered stop tokens (end-of-sequence and tokens whose decoded piece contains a newline or is entirely whitespace), and emission ends when one is selected. Because the grammar forbids a leading space, a question-form prompt whose natural continuation starts with a standalone space after `Answer:` is an instrument mismatch (Appendix F.5). Strict exact-match scoring requires the last non-empty line to be one valid integer; tolerant extraction, parseability, and failure categories are retained separately. The registered contract measures exact-answer emission when a numeric grammar forces a numeral and a stop set ends the answer. Under free generation, some rows emit no numeral despite being registered-correct (OLMo-3 post-trained: 58/470 no-

Table 1: Evidence roles and inference boundaries.

Analysis	What it supports	What it does not support
Three-condition comparison, replication, and compatible-pair factorial	Opposite corrected conclusions and checkpoint–route interactions	A privileged route or weight-only attribution
Free-generation and question-form audits	Deficits no longer detected versus persistent deficits; two same-slice reversals persist under free generation	Route-independent arithmetic ability
Same-item stage tracking	The released stage at which each answer first fails	Isolation of a training increment's causal mechanism
Secondary base-content transplant	Selected-item recovery; Holm-supported dictionary-over-random advantage only for subtraction	The deficit's natural cause or recovery on all lost items

numeral subtraction rows and 239/651 multiplication rows; Appendix F.4), so strict accuracy is a property of the model under the contract, not a route-independent ability measure.

3.2 Checkpoint–route design

We evaluate base B by direct completion and each released post-trained stage E (SFT, DPO, or final) through shared completion R_s and its native chat template R_n. At a fixed checkpoint, the route contrast changes the rendered interface (template, system turn, special tokens) while holding the stop set fixed; base–post-trained contrasts may additionally differ in tokenizer and termination conventions (five pairs; Appendix H). Adding $A(B, R_n)$ yields the interaction and checkpoint contrast through chat (Figure 1B). Native-route analyses use *canonical* tokenization: each row's executed input IDs are verified against the tokenizer's template application after correcting a duplicated beginning-of-sequence token for six post-trained checkpoints (Appendix H). Shared-completion cells are unchanged. Registered templates, stop conventions, tokenizer compatibility, and qualification are detailed in Appendix H.

3.3 Panel, replication, and inference

Each contrast–task cell contains 2,000 paired items at base-calibrated digit widths. The primary analysis comprises three 96-cell contrast families, each corrected separately; a global 288-test correction is reported as a sensitivity analysis. The 96 cells are repeated task-level measurements, not 96 independent model replications; the seven documented post-training contrasts derive from three pretrained bases. The three families are not independent, because $\Delta_{\text{bundle}} = \Delta_{\text{checkpoint}} + \Delta_{\text{route}}$ identically in the means. A reversal therefore needs three corrected decisions at once: opposite Holm-significant checkpoint and bundled directions, plus a significant route contrast. Decisions use exact conditional McNemar tests; Newcombe intervals are descriptive and can in principle disagree with the test near a boundary.

The registered selection rule allocates ten pairs to quotas over model family, parameter scale, and recipe openness; two to the same-family size comparison (Llama-3.2 1B/3B), separating scale from family; and four to SFT/DPO stages of OLMo-3 and Llama-3.1/Tulu. SmolLM3 uses a mid-training reference; other pairs have less complete lineage documentation. Tax claims retain the seven documented contrasts, although public recipe descriptions also exist outside this subset. Tokenizer compatibility permits the fourth condition for nine pairs.

For each contrast-task cell, paired exact conditional McNemar tests are Holm-corrected within the registered family, with Newcombe paired-score intervals. The factorial completion's interaction test instead uses an item-level paired bootstrap (Appendix D), an approximate resampling procedure rather than a finite-sample exact test like McNemar's. The three decomposition families are discovery analyses run after their outcomes were observed, and every cell is reported. The three-condition fresh-grid replication was pre-registered outcome-blind. The fourth-condition completion was registered after the other three condition outcomes were observed, its own outcome still unobserved. Families, registration status, strict and tolerant views, and chronology appear in Appendix Table 8, Appendix D, and Appendix H. Table 1 summarises what each analysis in this design supports and what it does not.

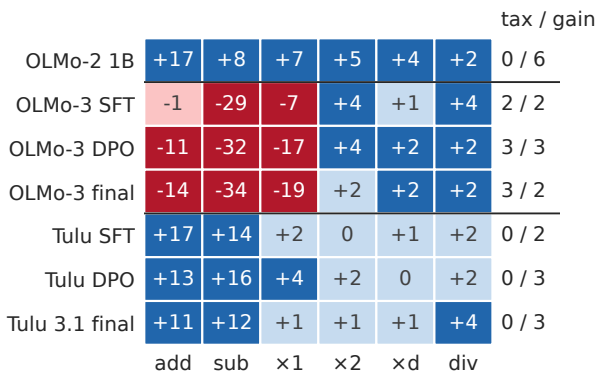

Figure 2: Fresh-grid fixed-route checkpoint effects for the seven documented contrasts across six tasks, in rounded percentage points. Blue and red cells show positive and negative effects; saturated tiles are Holm-significant, light tiles are not; the right column reports each row's Holm-significant tax / gain count. Rows are grouped by base checkpoint: OLMo-2, OLMo-3, and Llama-3.1/Tulu.

4 Results

4.1 Post-training does not impose a uniform arithmetic tax

On the fresh grids, the seven documented post-training contrasts come from three bases. Holm-significant tax / gain counts are 0 / 6 for OLMo-2 1B, 8 / 7 for OLMo-3 7B, and 0 / 8 for Llama-3.1/Tulu (Figure 2). Every significant fixed-route tax occurs in OLMo-3: subtraction and single-digit multiplication, plus addition after DPO and final release. All persist under the bundled comparison. OLMo-3's fixed native-route contrast is unobserved. Eight of the 24 reversal cells are documented contrasts: Tulu SFT 2, DPO 3, final 3, OLMo-2 none.

OLMo-3 final's fresh-grid fixed-route point estimates are small and positive on $d \times 2$, $d \times d$, and division ($+1.5$–$+2.4$ pp; $18.55 \to 20.05\%$, $2.65 \to 4.50\%$, and $5.30 \to 7.65\%$). Only $d \times d$ and division are Holm-significant gains, both at low absolute accuracy (Appendix Table 10). Across all 96 cells, 19 Holm-significant fixed-route effects have $|\Delta| < 5$ pp (5 deficits, 14 gains; Appendix D.1).

The OLMo-3 task pattern persists on the digit widths observed in both grids; these cumulative contrasts use stage-specific grids, not paired stage increments (Appendix D).

4.2 Checkpoint effects depend on the measurement route

Checkpoint–route interactions are Holm-significant in 12 of 14 cells above the 1% floor and 45 of 54 overall (median $+8.00$ pp; Figure 3). In 16 cells the routes give opposite Holm-significant directions, three above the floor. OLMo-3 and Tulu are excluded for tokenizer incompatibility; the fourth condition was registered after the other three were observed (Appendix D).

Post-outcome bootstrap-t and Wald/score re-analyses agree with all 54 decisions. Pair-cluster and leave-one-pair-out analyses retain significance and a positive median (Appendix D).

4.3 Bundled and fixed-route comparisons can support opposite conclusions

Under the registered decoding protocol, the bundled and fixed-completion comparisons reach opposite Holm-significant conclusions, with significant route support, in 24 of 96 fresh-grid replication cells (Figure 4). All 24 remain reversals after a single Holm correction over all 288 tests. Among these cells, the median absolute differences are 11.3 pp for the fixed-completion comparison and 18.0 pp for the bundled comparison. In a post-hoc analysis with Holm correction over all 345 observed digit-width units per family, each of these 24 cells also contains at least one width-level reversal. Appendices D and D.1 report the full replication and sensitivity analyses.

A pre-specified same-item audit compared constrained decoding with free generation (first-integer scoring; 500 items at one width per cell). SmolLM3 and Tulu SFT reverse under both contracts; Qwen3.5 4B and Gemma 4 E2B under neither. This establishes persistence in the two tested slices, not contract independence or a replication rate for the 24 cells (Appendix F.6).

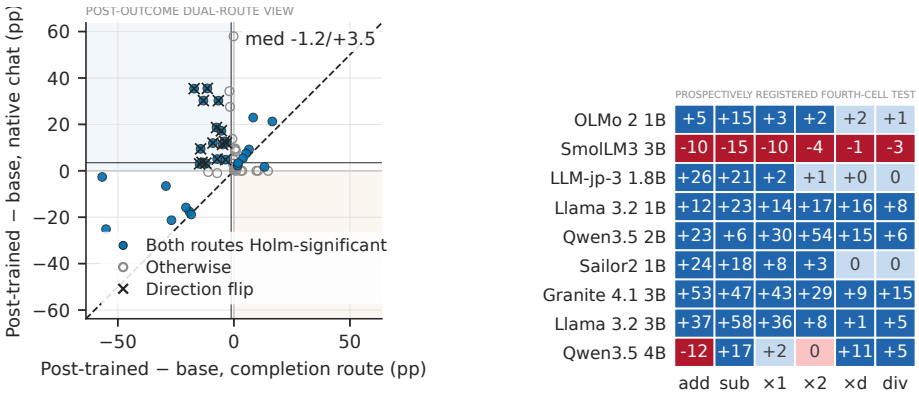

Figure 3: Factorial breadth evidence: 12 of 14 cells above 1% base chat-template accuracy have significant interactions (eligible cells: Qwen3.5 2B, 6; 4B, 6; Llama 3.2 3B, 2). Overall, 45 of 54 are significant, with 40 base-floor cells in the total. OLMo-3 and Tulu are excluded for tokenizer incompatibility. *Left:* completion versus chat-template checkpoint contrasts (pp); diagonal: no interaction; crosses: 16 opposite Holm-significant directions, 13 at the base chat floor. *Right:* interaction θ (pp); saturated tiles are Holm-significant; blue positive, red negative. The interaction was registered before fourth-condition outcomes; the dual-route display is post-outcome (Appendix Table 13).

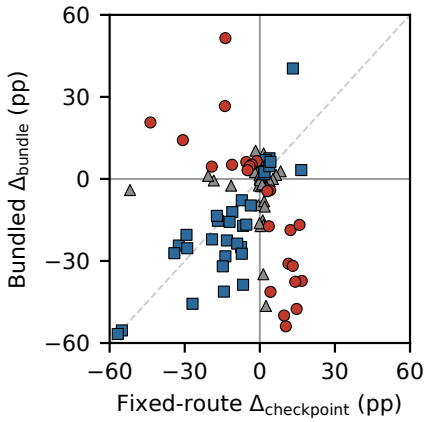

Figure 4: Fixed-route checkpoint effect against the bundled effect, 96 fresh-grid cells (pp; dashed line: equality). Red circles: the 24 cells whose Holm-corrected families support opposite conclusions; blue squares: same sign, both significant (33); grey triangles: at least one not significant (39). Direction changes between two fixed routes are the separate question of Section 4.2.

4.4 Prompt-form sensitivity of subtraction deficits

For OLMo-3 we replace `{a}-{b}=` by `What is {a} minus {b}?\nAnswer:`, after `Output ONLY a number.\n`; this changes the operator wording, sentence, and answer cue. Under the question form with free-generation first-integer scoring, no deficit is detected on the three-digit slice (99.2% vs. 100.0%, n.s.; Table 2A). The 14-digit loss remains (52.4% vs. 1.4%, -51.0 pp; Table 2A). Of 500 final outputs, 496 contain a parser-recognised integer (7 correct, 489 wrong; group-alignment exclusions: Appendix F.5); crediting all 4 unparsed outputs gives at most 2.2%. The loss appears at SFT (3.0%) and persists at DPO (1.2%) and final (1.4%; Figure 5C): 253 of 259 lost answers first fail at SFT and remain wrong (Appendix F.5). It persists under last-integer and target-present readouts. Wrong non-negative numerals usually fail from the leading or second group, across about four groups, with only $\approx 12\%$ showing a ± 1 borrow signature (Appendix F.5).

For Llama-3.1/Tulu, the question form helps the base more. Fixed-route gains of $+5.8/+18.4$ pp become -1.0 pp at 9 digits (n.s.) and a significant -8.4 pp at 15 digits. Bundled question-form estimates are also negative ($-34.4/-36.6$ pp, descriptive). The route deficit keeps its registered sign and persists under last-integer and target-present readouts (-27.4 to -30.0 pp; Table 2B; Appendix F.5).

Table 2: Prompt-form audit: free generation (greedy, no mask or stop set, 128-token budget), first-integer scoring on 500 paired subtraction items per slice. Differences (pp) have Newcombe 95% intervals. (A) Post-trained minus base; OLMo-3 14-digit rows use batch-size-matched base, final, and stage outputs; the original batched-base values are in the supplementary archive. (B) Native minus shared at the Tulu final checkpoint. †: Holm-significant within the corresponding family (Appendix Table 8).

(A) Pair	Slice ($n = 500$)	Form	Base (%)	Post-trained (%)	Δ (pp) [95% CI]
OLMo-3	3-digit	Expression	80.6	63.8	-16.8 [$-21.7, -11.8$]†
OLMo-3	3-digit	Question	99.2	100.0	$+0.8$ [$-0.1, +2.0$]
OLMo-3	14-digit	Expression	27.4	0.2	-27.2 [$-31.3, -23.4$]†
OLMo-3	14-digit	Question	52.4	1.4	-51.0 [$-55.4, -46.4$]†
OLMo-3 SFT	14-digit	Question	52.4	3.0	-49.4 [$-53.9, -44.6$]†
OLMo-3 DPO	14-digit	Question	52.4	1.2	-51.2 [$-55.6, -46.5$]†
Llama/Tulu	9-digit	Expression	84.4	90.2	$+5.8$ [$+2.6, +9.1$]†
Llama/Tulu	9-digit	Question	94.8	93.8	-1.0 [$-2.8, +0.7$]
Llama/Tulu	15-digit	Expression	17.2	35.6	$+18.4$ [$+13.8, +23.0$]†
Llama/Tulu	15-digit	Question	40.4	32.0	-8.4 [$-12.5, -4.2$]†
(B) Route contrast (native − shared) at the Tulu final checkpoint					
Checkpoint	Slice ($n = 500$)	Form	Shared (%)	Native (%)	Δ (pp) [95% CI]
Tulu final	9-digit	Question	93.8	60.4	-33.4 [$-37.8, -28.9$]†
Tulu final	15-digit	Question	32.0	3.8	-28.2 [$-32.4, -24.1$]†

4.5 Secondary diagnostic: base-content transfer on OLMo-3

Expression-form constrained measurement places 89.0% of subtraction losses at SFT. A secondary base-state dictionary transplant raises accuracy by 27.5–50.0 points to 35.0–72.5% on 40 never-used lost items per task shared across stages; strict cross-item controls reach at most 2.5%. Pointwise intervals favor dictionary over random transfer for subtraction at every stage and multiplication at DPO/final. Exact paired tests with Holm over six cells support the advantage for subtraction at every stage and multiplication at none. Between-task differences remain exploratory, and random-subspace overlap limits specificity (Appendix G). No tested operation restores a full 14-digit subtraction answer (0/10 at every stage), although the dictionary transplant corrects the leading group in 7/10; question-form outputs were not intervened on (Appendix G). Stage measurement uses bfloat16 rotary-embedding buffers, transplantation float32 with within-condition controls (Figure 5). The transfer recovers some selected lost answers; it does not identify why they were lost.

5 Limitations

The persistent OLMo-3 14-digit loss was measured under shared completion with the question form; its native-route question-form, reasoning-budget, and scratchpad conditions were not evaluated. Base checkpoints use zero-shot completion; the single few-shot arm was uninformative, and digit widths are calibrated on the base only. Full main-battery generations are digest-only in the supplement, which ships an item-level score matrix and ten raw rows per route for each reversal cell.

The primary registered-instrument estimates are conditional on the registered serialization, numeric grammar, greedy decoding, and strict last-line readout; the free-generation and prompt-form audits use the first-integer readout without the mask. The matched comparison arms were registered before their outcomes; their readout sensitivity is cell-specific (Appendix F.1–F.2). With both mask and stop set removed for free generation, the raw strict readout is at floor and may retain an end-of-sequence marker in the scored string; the scoring and stop convention do not transfer unchanged. The first-integer comparison aligns the answer position (Appendix F.5); the representative arm covers four cells at one digit length each, two slices show no fresh-grid reversal under either contract, and SmolLM3's replication depends on the readout. The stage measurement's whole-model cast changed rotary-embedding buffer precision relative to the other programs. For the audited SFT/DPO slice, each condition reproduces its outputs, with correctness agreement on 499/500 items per checkpoint; conclusions are stated per instrument (Appendix H).

Recomputing on the digit widths observed in both grids leaves the 24 reversal cells and all interaction signs unchanged (44 of 54 significant; Appendix D). The panel is a set of checkpoints we could

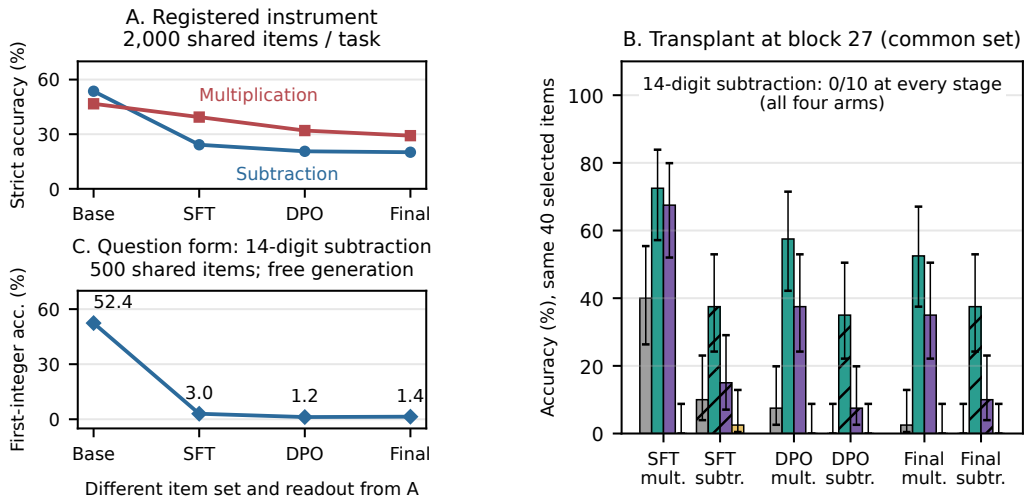

Figure 5: OLMo-3 stage measurement and secondary transplant diagnostic. *A:* expression-form constrained strict accuracy on 2,000 shared items per task. *B:* block-27 accuracy on 40 never-used base-correct/final-wrong items per task with a leading-group competitor, shared across stages: grey, no intervention; teal, same-item base dictionary transplant; purple, rank-matched random-subspace transplant; gold, strict cross-item control at the same component norm. Solid/hatched bars: multiplication/subtraction; Wilson 95% intervals. All operations score 0/10 on 14-digit subtraction at every stage. A uses bfloat16 rotary-embedding buffers; B uses float32 and its own controls. *C:* question-form free generation, float32 buffers, first-integer scoring on 500 shared 14-digit subtraction items at batch size 1. C differs from A in items and readout; B is a selected population (Appendix H; Tables 17 and 26).

observe, not a census; route contrasts compare rendered interfaces at one checkpoint and are associational. The factorial covers tokenizer-compatible pairs only; $A(B, R_n)$ is a diagnostic interface, not a deployment route. Execution precedence and the rendering defect are disclosed in Appendix H; dual-route, pair-level, and difficulty-slice analyses are post-outcome. The OLMo-3 transplant uses one lineage, route, subspace, and dose on base-correct/final-wrong items with a leading-group competitor. Its matched controls show recovery of some selected lost answers, but an advantage over random-subspace transfer for multiplication is not established. The intervention does not identify the deficit's natural source, unchanged downstream processing, or a unique answer-carrying subspace (Appendix G).

6 Conclusion

Comparisons between released checkpoints can indicate either a gain or a deficit, depending on the evaluation conditions. Under our registered protocol, bundled and fixed-completion comparisons lead to opposite Holm-significant conclusions in 24 of 96 replication cells. The contribution is the observed frequency and stability of such reversals within this panel under the registered contract, not their algebraic possibility. The factorial adds Holm-significant checkpoint–route interactions above the base-chat floor in the tokenizer-compatible pairs, which exclude OLMo-3 and Tulu. Some deficits are no longer detected after changing the prompt, whereas OLMo-3's 14-digit subtraction deficit persists under question-form free generation on the shared-completion route, with most lost answers first failing at SFT; a secondary transplant recovers some selected lost answers but none of ten 14-digit expression-form items. These results concern conditional answer accuracy, not route-independent ability. Evaluations should report checkpoint and route comparisons separately, alongside the prompt, decoding, stopping, and answer-extraction rules.

REPRODUCIBILITY STATEMENT

The anonymous supplementary archive is completely enumerated and SHA-256-verified by the shipped `MANIFEST`. It contains source, frozen registrations and designs, aggregated study artifacts, figures, compact per-item causal-extension rows, and the compact score matrix for the main battery; full main-battery generations and full activation readouts are digest-only. The causal-extension boundary is explicit: its compact per-item rows are shipped, while full activations are represented by digests only. For the prompt-form audit (including stage measurements and matched route comparisons) and common-set results, compact item-level records (item id, rendered-prompt digest, raw output, scoring, execution conditions) ship in the supplement and are indexed by the crosswalk in `REPRODUCE.md`. The accompanying `REPRODUCE.md` states these boundaries and gives exact commands; analysis-only table and figure regeneration runs on a laptop CPU in minutes, and the measured wall-clock of the clean-room reproduction of the supplement profile is recorded there. Retained inference on one NVIDIA RTX A4000 consumed about 436 GPU-hours, including loading and disclosed restarts; the largest components were original measurement, replication, and ablation/factorial runs. The exact per-programme breakdown is in `REPRODUCE.md`. The matched-audit runner now records per-load runtime fingerprints (parameter dtypes and buffer dtypes and hashes) for future executions. Fitted lens weights are digest-anchored and are not required to reproduce the shipped analyses.

AI-USE STATEMENT

AI systems assisted implementation, analysis tooling, and drafting under human direction. They were additionally used for methodological critique of the evaluation design and of the statistical presentation, for literature discovery, and for evaluation-integrity auditing—including the discovery of the route-rendering defect disclosed in Appendix H—as well as repair planning and result-interpretation support, all under human direction. All AI-suggested analyses, audits, text, and references were verified by the authors, who defined the scientific questions, measurement contract, data boundaries, interpretation, and final text, and who take full responsibility for the content of this paper.

REFERENCES

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 2022.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12, 2024.

Haotong Yang, Yi Hu, Shijia Kang, Zhouchen Lin, and Muhan Zhang. Number Cookbook: Number understanding of language models and how to improve it. In *International Conference on Learning Representations*, 2025.

Rahmad Mahendra, Damiano Spina, Lawrence Cavedon, and Karin Verspoor. Evaluating numeracy of language models as a natural language inference task. In *Findings of NAACL*, pages 8351–8376, 2025.

Shulin Huang, Linyi Yang, Yan Song, Shuang Chen, Leyang Cui, Ziyu Wan, Qingcheng Zeng, Ying Wen, Kun Shao, Weinan Zhang, Jun Wang, and Yue Zhang. ThinkBench: Dynamic out-of-distribution evaluation for robust LLM reasoning. arXiv:2502.16268, 2025.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale. In *International Conference on Learning Representations*, 2026.

Lintang Sutawika, Hailey Schoelkopf, Leo Gao, Baber Abbasi, Stella Biderman, Jonathan Tow, et al. EleutherAI/lm-evaluation-harness: v0.4.3. Zenodo, July 2024. `https://doi.org/10.5281/zenodo.12608602`.

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

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

Robert G. Newcombe. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17(22):2635–2650, 1998.

Hongzhi Luan, Changxin Tian, Zhaoxin Huan, Xiaolu Zhang, Kunlong Chen, Zhiqiang Zhang, and Jun Zhou. BOSE: A systematic evaluation method optimized for base models. arXiv:2503.00812, 2025.

Erfan Baghaei Potraghloo, Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. One token away from collapse: The fragility of instruction-tuned helpfulness. arXiv:2604.13006, 2026.

Selen Erkan, Bastian Boll, Kristian Kersting, Björn Deiseroth, and Letitia Parcalabescu. Soft-prompt tuning for fair and efficient LLM benchmark evaluation. arXiv:2606.12117, 2026.

Prateek Munjal, Clement Christophe, Ronnie Rajan, and Praveenkumar Kanithi. Do instruction-tuned models always perform better than base models? Evidence from math and domain-shifted benchmarks. arXiv:2601.13244, 2026.

Mario Sanz-Guerrero, Manuel Mager, and Katharina von der Wense. Large language models are overconfident in their own responses. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 31406–31418, 2026.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. Flaw or artifact? Rethinking prompt sensitivity in evaluating LLMs. In *Proceedings of EMNLP*, pages 19889–19899, 2025.

Nicholas Sadjoli, Tim Siefken, Atin Ghosh, Yifan Mai, and Daniel Dahlmeier. Optimization before Evaluation: Evaluation with Unoptimized Prompts Can be Misleading. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 619–638, 2025. `https://aclanthology.org/2025.acl-industry.44/`.

Lukas Thede, Stefan Winzeck, Zeynep Akata, and Jonathan Richard Schwarz. CapTrack: Multifaceted evaluation of forgetting in LLM post-training. arXiv:2603.06610, 2026.

Shi Ying Chang, Chiok Yew Ho, Yichen Li, and Yintong Huo. Lost in the flow with code talkers: Unveiling the instruction-tuning tax of large language models in code tasks. arXiv:2606.08676, 2026.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? An interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Andy Zou, Long Phan, Sarah Chen, et al. Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405, 2023.

Wes Gurnee, Nicholas Sofroniew, et al. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026.

Irina Bigoulaeva, Jonas Rohweder, Subhabrata Dutta, and Iryna Gurevych. Patches of nonlinearity: Instruction vectors in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Hongzhe Du, Weikai Li, Min Cai, Karim Saraipour, Zimin Zhang, Himabindu Lakkaraju, Yizhou Sun, and Shichang Zhang. How post-training reshapes LLMs: A mechanistic view on knowledge, truthfulness, refusal, and confidence. In *Conference on Language Modeling*, 2025. arXiv:2504.02904.

Gangwei Jiang, Caigao Jiang, Zhaoyi Li, Siqiao Xue, Jun Zhou, Linqi Song, Defu Lian, and Ying Wei. Unlocking the power of function vectors for characterizing and mitigating catastrophic forgetting in continual instruction tuning. In *International Conference on Learning Representations*, 2025.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations*, 2024.

Dan Shi, Zhuowen Han, Simon Ostermann, Renren Jin, Josef van Genabith, and Deyi Xiong. Why does reinforcement learning generalize? A feature-level mechanistic study of post-training in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026.

Yihan Wang, Andrew Bai, Nanyun Peng, and Cho-Jui Hsieh. On the loss of context awareness in general instruction fine-tuning. In *Advances in Neural Information Processing Systems*, 2025.

Honglin Zhang, Qianyue Hao, Fengli Xu, and Yong Li. Reinforcement learning fine-tuning enhances activation intensity and diversity in the internal circuitry of LLMs. In *International Conference on Learning Representations*, 2026.

Yifan Zhou. Same targets, different computation: How post-training divides work across model layers. arXiv:2605.07284v2, August 2026.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of EMNLP*, 2023.

Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. Grammar-aligned decoding. In *Advances in Neural Information Processing Systems*, 2024.

Samy Jelassi, Stéphane d'Ascoli, Carles Domingo-Enrich, Yuhuai Wu, Yuanzhi Li, and François Charton. Length generalization in arithmetic transformers. arXiv:2306.15400, 2023.

Hanseul Cho, Jaeyoung Cha, Pranjal Awasthi, Srinadh Bhojanapalli, Anupam Gupta, and Chulhee Yun. Position coupling: Improving length generalization of arithmetic transformers using task structure. In *Advances in Neural Information Processing Systems*, 2024.

Hanseul Cho, Jaeyoung Cha, Srinadh Bhojanapalli, and Chulhee Yun. Arithmetic transformers can length-generalize in both operand length and count. In *International Conference on Learning Representations*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, et al. LIMA: Less Is More for Alignment. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.11206.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. In *International Conference on Learning Representations*, 2024. arXiv:2312.01552.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-Tuning Enhances Existing Mechanisms: A Case Study on Entity Tracking. In *International Conference on Learning Representations*, 2024. arXiv:2402.14811.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis. In *Proceedings of EMNLP*, pages 7035–7052, 2023. https://aclanthology.org/2023.emnlp-main.435/.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic Without Algorithms: Language Models Solve Math with a Bag of Heuristics. In *International Conference on Learning Representations*, 2025. arXiv:2410.21272.

SmolLM3 Team. SmolLM3: Smol, multilingual, long-context reasoner. Hugging Face blog, https://huggingface.co/blog/smollm3, 2025.

Team OLMo. 2 OLMo 2 Furious. arXiv:2501.00656, 2025a.

Team OLMo. OLMo 3. Technical report, Allen Institute for AI, 2025b.

Allen Institute for AI. Llama-3.1-Tulu-3.1-8B model card. Hugging Face, https://huggingface.co/allenai/Llama-3.1-Tulu-3.1-8B, 2025a.

Allen Institute for AI. OLMo-2-0425-1B-Instruct model card. Hugging Face, https://huggingface.co/allenai/OLMo-2-0425-1B-Instruct, 2025b.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, et al. Tülu 3: Pushing frontiers in open language model post-training. arXiv:2411.15124, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. arXiv:2407.21783, 2024.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, et al. Sailor2: Sailing in South-East Asia with inclusive multilingual LLMs. arXiv:2502.12982, 2025.

Hirokazu Kiyomaru and Takashi Kodama. LLM-jp-3-1.8b-instruct model card. Hugging Face, https://huggingface.co/llm-jp/llm-jp-3-1.8b-instruct, 2024.

Qwen Team. Qwen3.5-2B-Base, Qwen3.5-2B, Qwen3.5-4B-Base, and Qwen3.5-4B model cards. Hugging Face model cards, https://huggingface.co/Qwen/Qwen3.5-2B-Base, https://huggingface.co/Qwen/Qwen3.5-2B, https://huggingface.co/Qwen/Qwen3.5-4B-Base, https://huggingface.co/Qwen/Qwen3.5-4B, 2026.

Meta AI. Llama 3.2 1B, 1B-Instruct, 3B, and 3B-Instruct model cards. Hugging Face model cards, https://huggingface.co/meta-llama/Llama-3.2-1B, https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct, https://huggingface.co/meta-llama/Llama-3.2-3B, https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct, 2024.

Gemma Team. Gemma 4 Technical Report. arXiv:2607.02770, 2026.

Google. Gemma 4 E2B and E2B-it model cards (google/gemma-4-E2B, google/gemma-4-E2B-it). Hugging Face model cards, https://huggingface.co/google/gemma-4-E2B, https://huggingface.co/google/gemma-4-E2B-it, 2026.

IBM Granite. Granite 4.1 3B base and final model cards (ibm-granite/granite-4.1-3b-base, ibm-granite/granite-4.1-3b). Hugging Face model cards, https://huggingface.co/ibm-granite/granite-4.1-3b-base, https://huggingface.co/ibm-granite/granite-4.1-3b, 2026.

APPENDIX ROADMAP

Table 3 locates the models, measurement contracts, results, and limitations. The supplementary crosswalk in `REPRODUCE.md` gives full paths and regeneration commands; its source archive retains the superseded tables and exploratory detail.

Table 3: Appendix roadmap by reader question.

Reader question	Appendix section(s)
Models, training deltas, and selection	A Panel lineage; B.2 Census selection and disposition
Task generation and calibration	B.3 Frozen item-generation rules; B.4 Registered calibration window-selection rule
Full effects	C Full path-decomposition estimands
Absolute accuracies and difficulty	D Replication, per-digit heterogeneity, readout and difficulty disclosures
Factorial and robustness	D Factorial completion and common-support sensitivity; D.1 Reversal heterogeneity
Failure modes	E Descriptive failure-mode decomposition of route effects
OLMo-3 stage localization and transplant	G Stage-resolved measurement and base-content transplant on OLMo-3
Matched readout controls	F Matched readout controls; F.1 Unconstrained decoding; F.2 No-comma serialization; F.4 Free-generation termination audit; F.5 Prompt-form audit ; F.6 Representative-cell free-generation arm
Protocol deviations and reproducibility	H Protocol deviations, corrective analyses, and reproducibility limitations

A PANEL LINEAGE

Tables 4 and 5 record, for every ordinary contrast, the pinned base and post-trained checkpoints (Hugging Face repository id and short revision SHA from the frozen design), the source of the registered route template, the public documentation of the training delta, and the registered openness tier. Public recipe descriptions and verified access to data and intermediate stage checkpoints are distinct dimensions of this evidence.

Table 4: Pinned checkpoint lineage for the 16 ordinary contrasts. Revisions are short SHA prefixes of the frozen design's pinned commits. The registered native route renders each post-trained checkpoint's official chat template under the per-family registered system configuration at the pinned revision; bases and shared-completion cells use the shared visible prompt with no template. Documented training deltas and adjudicated evidence for the same contrasts are in Table 5.

Contrast	Base (revision)	Post-trained (revision)
OLMo 2 1B	`allenai/OLMo-2-0425-1B` (a1847df)	`...-1B-Instruct` (48d788e)
LLM-jp-3 1.8B	`llm-jp/llm-jp-3-1.8b` (60acda3)	`...-instruct` (413b792)
Llama 3.2 1B	`meta-llama/Llama-3.2-1B` (4e20de3)	`...-1B-Instruct` (9213176)
Qwen3.5 2B	`Qwen/Qwen3.5-2B-Base` (b1485b2)	`Qwen/Qwen3.5-2B` (15852e8)
Sailor2 1B	`sail/Sailor2-1B` (75c2856)	`...-1B-Chat` (51b48ec)
Gemma 4 E2B	`google/gemma-4-E2B` (63db66a)	`...-E2B-it` (3e22461)
Granite 4.1 3B	`ibm-granite/granite-4.1-3b-base` (dacb9cb)	`...-4.1-3b` (c065040)
Llama 3.2 3B	`meta-llama/Llama-3.2-3B` (13afe51)	`...-3B-Instruct` (0cb88a4)
Qwen3.5 4B	`Qwen/Qwen3.5-4B-Base` (1001bb4)	`Qwen/Qwen3.5-4B` (851bf6e)
SmolLM3 3B	`HuggingFaceTB/SmolLM3-3B-checkpoints` (0485ec1)	`.../SmolLM3-3B` (a07cc9a)
OLMo 3 7B final	`allenai/Olmo-3-1025-7B` (a81bae4)	`...-7B-Instruct` (096bb54)
Tulu 3.1 final	`meta-llama/Llama-3.1-8B` (d04e592)	`allenai/Llama-3.1-Tulu-3.1-8B` (46239c2)
OLMo 3 7B SFT	`allenai/Olmo-3-1025-7B` (a81bae4)	`...-7B-Instruct-SFT` (e1452fc)
OLMo 3 7B DPO	`allenai/Olmo-3-1025-7B` (a81bae4)	`...-7B-Instruct-DPO` (5c88782)
Tulu 3 SFT	`meta-llama/Llama-3.1-8B` (d04e592)	`allenai/Llama-3.1-Tulu-3-8B-SFT` (f2a0b46)
Tulu 3 DPO	`meta-llama/Llama-3.1-8B` (d04e592)	`allenai/Llama-3.1-Tulu-3-8B-DPO` (a7beb67)

Table 5: Public documentation and registered openness tiers for the contrasts in Table 4. T1: fully itemized recipe, public post-training data and public intermediate stage checkpoints. T2: a technical report and/or code exists, but intermediate stage checkpoints were not verified public and/or post-training data were not public. T3: card-level description only. The documented subset comprises the seven T1 base-to-post-training contrasts under the adjudication frozen on 2026-08-11 before any outcome; tiers and subset membership were not re-adjudicated post hoc. SmolLM3 remains T1 but is outside this subset because its reference is a mid-training checkpoint, not a pretrained base. T2 rows may have public recipe descriptions without verified intermediate checkpoints. Citations identify public sources; availability statements report the adjudication evidence.

Contrast	Public documentation of the compared delta	Registered tier; pair date
OLMo 2 1B	SFT on Tülu 3 data, then DPO and RLVR from the OLMo-2-0425-1B base (Team OLMo, 2025a; Allen Institute for AI, 2025b)	T1; 2025-04-29
LLM-jp-3 1.8B	Card describes SFT and names the full public instruction-tuning dataset list and pretraining corpus; no intermediate stage checkpoints or dedicated LLM-jp-3 technical report linked on the endpoint card (Kiyomaru and Kodama, 2024)	T2; 2024-09-23
Llama 3.2 1B	Card describes multi-round SFT, rejection sampling and DPO; no public post-training data or intermediate checkpoints, and no Llama-3.2-specific technical report (Meta AI, 2024)	T3; 2024-09-18
Qwen3.5 2B	Card identifies the base and post-trained endpoint and links a blog; no model-specific technical report on the card, public post-training data or intermediate checkpoints at adjudication (Qwen Team, 2026)	T3; 2026-02-28
Sailor2 1B	Two-stage SFT and LR-DPO recipe with hyperparameters, SFT/DPO datasets and code in the technical report; intermediate stage checkpoints not verified at adjudication (Dou et al., 2025)	T2; 2024-12-03
Gemma 4 E2B	Technical report covers the family; E2B post-training data and intermediate checkpoints not public (Gemma Team, 2026; Google, 2026)	T2; 2026-03-02
Granite 4.1 3B	Cards describe SFT and reinforcement-learning alignment using open instruction data and non-public internal synthetic data; base card links the Granite 4.x report, endpoint report unposted at adjudication; no intermediate checkpoints (IBM Granite, 2026)	T2; 2026-04-06
Llama 3.2 3B	Card describes multi-round SFT, rejection sampling and DPO; no public post-training data or intermediate checkpoints, and no Llama-3.2-specific technical report (Meta AI, 2024)	T3; 2024-09-18
Qwen3.5 4B	Card identifies the base and post-trained endpoint and links a blog; no model-specific technical report on the card, public post-training data or intermediate checkpoints at adjudication (Qwen Team, 2026)	T3; 2026-02-27
SmolLM3 3B	Mid-training reference to final release: SFT followed by Anchored Preference Optimization (APO); the reference is not a pretrained base, so this T1 pair is outside the documented subset (SmolLM3 Team, 2025)	T1; 2025-07-20
OLMo 3 7B final	SFT, DPO and RLVR from the OLMo-3 base (Team OLMo, 2025b)	T1; 2025-11-19
Tulu 3.1 final	Tulu 3 SFT and DPO, followed by the Tulu 3.1 RLVR stage (GRPO, from the Tulu 3 DPO checkpoint); the final-stage recipe differs from Tulu 3 (PPO) (Lambert et al., 2024; Allen Institute for AI, 2025a; Grattafiori et al., 2024)	T1; 2025-02-07
OLMo 3 7B SFT	SFT stage of Instruct recipe (Team OLMo, 2025b)	T1; stage row
OLMo 3 7B DPO	SFT followed by DPO in the Instruct recipe (Team OLMo, 2025b)	T1; stage row
Tulu 3 SFT	SFT stage of Tulu 3 recipe (Lambert et al., 2024)	T1; stage row
Tulu 3 DPO	SFT followed by DPO in the Tulu 3 recipe (Lambert et al., 2024)	T1; stage row

B PANEL AND PROTOCOL DETAILS

Table 6 lists the panel composition and each row's analysis role. The 16 ordinary contrasts contain 12 Layer-A rows and four Layer-B stage rows. Layer A includes 10 quota-selected rows and two size-pair mandate rows. The six tasks use paired grids with 2,000 scored items per contrast-task cell, arranged as four equal-sized digit strata. Calibration is base-only and precedes confirmatory scoring. Qualification items, calibration items, and confirmatory items are disjoint. The frozen item-generation rules—operand ranges, non-degeneracy and uniqueness constraints, division's inverse construction, and the qualification/calibration/confirmatory role-pool separation—are recorded with generator and design source hashes in the supplementary item-generation specification.

Table 6: Panel composition and analysis role.

Panel component	Analysis role	Contrasts
Layer A quota-selected	60-test bundled family	10
Layer A size-pair mandate	descriptive size family	2
Layer B stage rows	24-test cumulative family	4
Total ordinary panel	three 96-cell path-decomposition families	16

Table 7: Census disposition under the deterministic selection rule. Left: the 16 selected contrasts with size band, adjudicated openness tier, and selection role. Right: the 49 non-selected candidates counted by gate reason. Gate labels: pair documentation (base identification, official organization release, post-training-only delta), packaging (loadable under the frozen text-only loader contract), pinning/license (commit-pinnable revisions with research-usable licenses), release window (later checkpoint inside 2024-08-10 to 2026-08-10), reasoning-branch qualification (Think-branch instrument qualification present).

Selected contrast	Band	Tier	Selection role
OLMo 2 1B	small	T1	representative quota
LLM-jp-3 1.8B	small	T2	representative quota
Sailor2 1B	small	T2	representative quota
Qwen3.5 2B	small	T3	representative quota
Llama 3.2 1B	small	T3	size-pair mandate
SmolLM3 3B (mid→final)	middle	T1	representative quota
Gemma 4 E2B	middle	T2	representative quota
Granite 4.1 3B	middle	T2	representative quota
Qwen3.5 4B	middle	T3	representative quota
Llama 3.2 3B	middle	T3	size-pair mandate
OLMo 3 7B final	large	T1	representative quota
Tulu 3.1 final	large	T1	representative quota
OLMo 3 7B SFT / DPO	large	T1	Layer-B stage rows
Tulu 3 SFT / DPO	large	T1	Layer-B stage rows

Non-selected disposition	Rows
Excluded: release window	20
Excluded: pair documentation	11
Excluded: packaging	1
Uncertain: pair documentation	9
Uncertain: pinning/license	5
Uncertain: reasoning-branch qual.	1
Eligible, not admitted (quota/budget)	2
Total non-selected	49

B.1 Route and scoring views

The base route is direct completion. The shared route for the post-trained checkpoint uses the same visible arithmetic prompt without a model-specific chat wrapper. The registered native chat-template route (the registered native route in the main text) applies the post-trained checkpoint's official chat template under a per-family registered system configuration: an empty system turn where the template accepts one, and the registered family fallback otherwise — no system turn for the Llama-family templates, and a minimal fixed system instruction with thinking disabled for SmolLM3. The exact rendering for each post-trained checkpoint is inventoried in Appendix H. The three route records are preserved separately. Each row retains rendered input, raw output, stop or cap state, strict parse, tolerant parse, conditional correctness, and failure category. Empty output, malformed output, refusals, cap termination, lost reasoning boundaries, and valid wrong numerals are not collapsed.

For every paired difference we report a 95% Newcombe paired-score interval (Newcombe, 1998) (method 10: Wilson score limits combined through a continuity-corrected within-pair correlation estimate; defined even when one discordant direction is zero). Generation-cap terminations are zero on all ordinary-route rows; headline denominators are the full 2,000 items per cell (capped rows counted incorrect), and pairwise cap exclusion arises only in the Think-route views (the cap-exclusion audit is retained in the supplement). Recomputing all three families under tolerant scoring leaves every Holm decision, effect sign, and reversal unchanged (Appendix D): the results are unchanged between the two retained parsers under the constrained instrument; the retained failure analysis nevertheless shows that many route losses are syntactically valid but shorter-than-target numerals (Appendix E).

B.2 Census gates, deterministic selection, and disposition

The frozen 65-candidate census fills quotas of 4 small ($[0, 2.5\text{B}]$), 4 middle ($(2.5\text{B}, 5.5\text{B}]$), and 2 large ($(5.5\text{B}, 8.5\text{B}]$) rows, reserving eligible Llama 3.2 1B→3B and Qwen3.5 2B→4B size-pair mandates before filling and adding Layer-B stage rows after top-ups. Within bands, filled small to large, eligible candidates rank by openness (T1 > T2 > T3 > undetermined), snapshot clarity (exact revision > named version > ambiguous > undetermined), developer absent from selected rows, newest pair date, ascending planned GPU-hours, then lexicographic post-trained checkpoint ID. Table 7 gives selected rows and exclusion reasons; the full 65-row gate trace and 8.0/8.5/9.0B size-bound sensitivity remain in the supplement. Environment receipts identify an NVIDIA RTX A4000, PyTorch 2.11.0+cu128, Python 3.12.3, and bfloat16; confirmatory summaries bind design, item-grid, and row-file hashes and requested CUDA, without per-cell hardware records.

B.3 FROZEN ITEM-GENERATION RULES

The frozen v3 registration fixes seed 20260813, six tasks, a maximum digit width of 20, and 500 items per sampled digit point, giving 2,000 scored items per contrast-task cell over the four calibrated digit points. Generation is a deterministic local pseudorandom stream; role pools are generated in the fixed order qualification → calibration → confirmatory with one shared seen-key set, so the three pools are disjoint by construction. The generator rules, each bound to the executable source and design hashes in the supplementary item-generation specification:

- *Operand ranges.* Addition and subtraction draw each operand uniformly from 0–9 at one digit and $[10^{d-1}, 10^d - 1]$ at $d \geq 2$ digits (no leading zeros); multiplication and division use the non-degenerate range excluding 0 and 1 (2–9 at one digit).
- *Subtraction.* Operand pairs are resampled until unequal, then ordered $a > b$, so the target is positive; there is no carry- or borrow-conditioned sampling (carry/borrow counts are recorded as metadata only).
- *Multiplication.* The right operand has a fixed width of one digit, two digits, or the left operand's width ($d \times 1$, $d \times 2$, $d \times d$); both operands are sampled independently.
- *Division.* The sampled pair is (q, b) at the requested width; the visible dividend is constructed as $a = q \times b$ and the target is q, so every item is exactly divisible, the divisor is never 0 or 1, and $q = b$ is allowed and recorded.
- *Duplicates.* Items are rejected until unique under the key (d, a, b)—for division (d, q, b), because the dividend is derived—with the shared seen-key set extending the rule across the three role pools.

B.4 REGISTERED CALIBRATION WINDOW-SELECTION RULE

Calibration measures base completion only; it fixes difficulty and variance without retuning the qualified instrument. The supplementary calibration specification gives the full specification and source-bound branches; the goodness-of-fit amendment is disclosed in Appendix H.

Registered calibration and four-point selection

1. **Search and stop.** Visit digit widths in ascending order from the first $d \in \{1, \ldots, 20\}$ with capacity at least $100 + 560$ (100 calibration; 30 qualification, 30 holdout, 500 confirmatory reserved). Stop after two consecutive accuracies below 0.25 or at $d = 20$. Omit generation-cap rows from the fitted binomial denominator.
2. **Crossing.** Fit $\mathrm{sigmoid}(a - sd)$, $s \geq 0$, by binomial maximum likelihood. On non-convergence, $s = 0$, or residual-deviance goodness-of-fit $p < 0.05$ against $\chi^2_{n_{\mathrm{points}}-2}$ (when degrees of freedom are positive), use weighted non-increasing isotonic PAVA and its exact/interpolated 0.5 crossing instead.
3. **Seal and uncertainty.** A crossing inside/below/above the measured grid is bracketed/left-censored/right-censored; an undefined crossing is ambiguous. Resample items within digit points ($B = 2{,}000$, seed 20260810), refit the full cascade, and use crossing replicates for the 95% percentile D50 interval. If the point estimate crosses but no replicate does, reseal ambiguous. For censored cells, flag boundary uncertainty if the fixed boundary accuracy's unconditional bootstrap interval straddles 0.5.
4. **Four points.** For a bracketed crossing, take the last pre-transition point at accuracy ≥ 0.75, the two crossing points, and the first post-transition point at ≤ 0.25. If these are not four distinct admissible points, retain valid roles and fill by the nearest admissible combination to $D50 + [-1.5, -0.5, 0.5, 1.5]$; if that fails, use the non-four-role fallback. Left/right censoring selects the four easiest/hardest admissible points and reports $D50 <$ floor/$>$ cap. Ambiguous cells or failed bracketed placements use four points nearest the available sealed-grid center when D50 is absent; never invent or extrapolate D50.
5. **Capacity.** At each selected point require capacity minus the verified global qualification–calibration exclusion union to be at least 500. Require four unique points (not necessarily adjacent); otherwise fail closed.

For the registered Llama 3.2 and Qwen3.5 size pairs, member-level placement is followed by a common task × digit grid using the available Base D50 center (or the sealed-grid center when both members are censored), projected to four globally admissible digits; the common grid takes precedence over a provisional member window. The special $d = 1$ floor cell is a separately labeled descriptive, calibration-only cell that never enters the primary calibration pool or D50 grid. A placement that cannot produce four unique points fails closed.

Table 8: Inferential family hierarchy. Each decomposition family is separately Holm-corrected over 96 cells per tier: post-hoc discovery on the original grids and pre-registered confirmation on fresh grids. Audit families retain their own slice/readout scopes; stage-arm confirmation recomputes the two stage contrasts with the single-stream base. Common-set dictionary-minus-none tests and the dictionary-minus-random exact-test sensitivity are post-hoc; registered control comparisons retain pointwise intervals (Appendix G). Dates and information states appear in Appendix H; prompt-arm registrations and records are indexed in `REPRODUCE.md`.

Claim family	Estimand	Cells	Holm scope	Status
Fixed-route checkpoint	$A(E, R_s) - A(B, R_s)$	96×2	96 cells per tier	discovery + registered confirmation
Route at fixed E	$A(E, R_n) - A(E, R_s)$	96×2	96 cells per tier	discovery + registered confirmation
Bundled post-trained vs base	$A(E, R_n) - A(B, R_s)$	96×2	96 cells per tier	discovery + registered confirmation
Quota-selected bundle	$A(E, R_n) - A(B, R_s)$	60	60 quota-selected cells	pre-specified primary
Cumulative stage (registered)	$A(E, R_n) - A(B, R_s)$	24	24 stage cells	pre-specified
Cumulative stage (shared)	$A(E, R_s) - A(B, R_s)$	24	24 shared stage cells	post-hoc secondary
Within-series size pair	larger − smaller E	12	12 size-pair cells	pre-specified
Think branch	Think: post-trained vs base	6	separate Think family	pre-specified, separate
Completion: route at base	$A(B, R_n) - A(B, R_s)$	54	60 registered slots	registered completion
Completion: interaction	route×ckpt θ	54	60 registered slots	registered; floor-bound
Completion: interaction, B=200k	same θ	54	54 + 60-slot mirror	post-hoc secondary
Fresh completion: interaction	same θ, fresh	54	54 cells	pre-registered
Fresh completion: route at base	$A(B, R_n) - A(B, R_s)$	54	54 cells	registered secondary
Matched control: decoding	constrained − unconstrained	72	72 cells	registered matched
Matched control: serialization	grouped − no-comma	54	54 cells	registered matched
Prompt-form audit slices	post-trained − base, two pairs × two forms	4×4	4 per slice/readout; two slices, two readouts	registered audit/addendum
Free-generation audit	post-trained − base, three pairs × two tasks	6 per view	6 separately per readout	registered audit
Question-form stage arm	SFT/DPO − base	2	2 primary contrasts	registered; expression secondary
Batch-size-matched confirmation	SFT/DPO − single-stream base	2	same 2 primary contrasts	registered recomputation
Native-route confirmation	Tulu native − Tulu shared / base shared	2×2	2 per slice; first integer	registered; other views descriptive
Representative-cell free-generation audit	fixed-route checkpoint and bundled contrasts, four cells	4×2	2 per cell (family size not fixed at registration); joint 8-test Holm as post-hoc sensitivity	registered audit
Common-set transplant	dictionary − none; dictionary − random, two tasks × three stages each	6×2	6 cells separately per comparison	dictionary − none: post-hoc; dictionary − random, six cells: post-hoc exact-test sensitivity

Note. Six Qwen3.5 4B native-route cells appearing in the route, bundled, quota-selected, and size-pair summaries are protocol-corrected replacement rows executed after a post-outcome route-integrity audit (Appendix H).

C Full path-decomposition estimands

The supplementary crosswalk in `REPRODUCE.md` indexes this section's estimand, classification, per-contrast, lineage, threshold, and direction records with full paths and regeneration commands.

The estimand records contain one row for each of the 16 contrasts, six tasks, and three estimands of the ordered three-condition decomposition: effect, Newcombe paired-score 95% interval, exact conditional McNemar p, Holm-adjusted p, discordant counts, route labels, and layer. The classification records apply the main text's deterministic rule to all 96 cells. The three 96-row discovery families are independently corrected post-hoc reanalyses.

Table 9 aggregates the Holm-significant fixed-route deficit/gain counts, mean per-task effects, and conclusion-reversal counts, with the seven documented post-training rows grouped by their three base checkpoints. Unrounded means satisfy $\bar{\Delta}_{\text{bundle}} = \bar{\Delta}_{\text{ckpt}} + \bar{\Delta}_{\text{route}}$; the records also retain medians. Summed per documented lineage, discovery gives OLMo-2 1B: 0 tax / 6 gain; OLMo-3 7B: 8 tax / 0 gain; Llama-3.1 8B: 0 tax / 12 gain Holm-significant cells. The pre-registered fresh-grid checkpoint family gives 0 tax / 6 gain, 8 tax / 7 gain, and 0 tax / 8 gain, respectively: every significant documented tax again occurs in OLMo-3. The discovery bundled contrast reproduces the checkpoint sign in 43 of 96 cells and differs in 53 (52 strictly opposite, one zero-delta). Full quota-selected bundled-family values are retained in the supplement.

Table 9: Per-contrast aggregates of the three estimands on the original confirmatory grids (discovery tier; the fresh-grid effects are in Figure 2). Deficit/Gain are Holm-significant fixed-route task counts (of six); the Δ columns are mean per-task effects in pp, rounded to two decimals (the unrounded means satisfy $\bar{\Delta}_{\text{bundle}} = \bar{\Delta}_{\text{ckpt}} + \bar{\Delta}_{\text{route}}$ exactly); Rev. is the conclusion-reversal task count. Rows list the seven documented post-training contrasts first, grouped by base checkpoint, then the SmolLM3 reference-to-final row, then the tier T2/T3 rows (outside the documented subset) in panel order.

Contrast	Base checkpoint	Deficit	Gain	Mean Δ_{ckpt}	Mean Δ_{route}	Mean Δ_{bundle}	Rev.
Documented post-training delta (seven rows, three bases)							
OLMo 2 1B	OLMo-2 1B	0	6	$+6.69$	-5.52	$+1.17$	1
OLMo 3 7B SFT	OLMo-3 7B	2	0	-5.60	-0.94	-6.54	0
OLMo 3 7B DPO	OLMo-3 7B	3	0	-9.64	$+4.72$	-4.92	0
OLMo 3 7B final	OLMo-3 7B	3	0	-11.62	$+2.65$	-8.97	0
Tulu 3 SFT	Llama-3.1 8B	0	3	$+5.82$	-38.44	-32.63	3
Tulu 3 DPO	Llama-3.1 8B	0	5	$+6.53$	-30.65	-24.12	5
Tulu 3.1 final	Llama-3.1 8B	0	4	$+5.51$	-27.40	-21.89	4
Reference-to-final							
SmolLM3 3B	SmolLM3 mid-training	0	6	$+8.11$	-45.68	-37.57	6
Tier T2/T3 (not in the documented subset)							
LLM-jp-3 1.8B	LLM-jp-3 1.8B	2	2	-1.22	-6.97	-8.19	1
Llama 3.2 1B	Llama 3.2 1B	4	0	-3.98	-1.58	-5.56	2
Qwen3.5 2B	Qwen3.5 2B Base	6	0	-29.52	-2.37	-31.88	0
Sailor2 1B	Sailor2 1B	3	0	-4.55	-8.73	-13.28	0
Gemma 4 E2B	Gemma 4 E2B	6	0	-31.91	$+44.22$	$+12.31$	5
Granite 4.1 3B	Granite 4.1 3B	4	0	-7.27	$+5.82$	-1.45	1
Llama 3.2 3B	Llama 3.2 3B	1	0	-0.99	-4.58	-5.58	0
Qwen3.5 4B	Qwen3.5 4B Base	5	1	-10.77	$+28.25$	$+17.48$	4

The supplementary replication records provide per-cell discovery and fresh-grid effects with paired 95% intervals for the 42 documented cells (seven contrasts from three bases); Figure 2 shows the main-text tiles.

C.1 Discovery-tier contrast details

The fixed-route discovery counts remain 39 deficits and 27 gains under a single global 288-test Holm correction. The documented-lineage scope comprises seven contrasts from three bases (Figure 2); the lineage sensitivity records distinguish it from the broader openness filter.

D Pre-registered fresh-grid replication

Original-grid factorial and qualification. Ten tokenizer-compatible pairs were registered, excluding six OLMo-3/Tulu rows; nine passed six-task qualification (30 rows per task; 1,620 total) and raw-output review. Gemma 4 E2B failed token-ID equality (27 versus 26 IDs; second BOS at position 1, base ID 2 versus post-trained ID 105), excluding six cells assigned $p = 1$ in the 60-slot Holm families; the 54-cell mirror and tolerant scoring agree. All 54 route-at-base contrasts are significant (48 negative; six Qwen3.5 4B positive). The registered $B = 2{,}000$ bootstrap lacked Holm resolution and rejected none; the $B = 200{,}000$ post-outcome reanalysis, pointwise intervals, and original-grid counts remain supplementary sensitivity records, preserving the registered decisions. Full original-grid and fresh-grid estimates are indexed in the supplementary crosswalk.

Registered design. Frozen outcome-blind on 2026-08-18 alongside factorial completion, before either outcome, replication covers 16 contrasts $\times$ six tasks on 96 fresh 2,000-item grids under sealed v3 windows and generator (seed family 20260818). Base completion, post-trained shared, and native routes give 288 cells and 576,000 rows. The 192,000 unique keys have zero overlap with qualification, calibration, confirmatory, or ablation manifests, or with other replication grids. Finite two-digit pools could not supply 500 keys to every sequentially allocated grid: subtraction was nearly exhausted, and later subtraction, two-digit/$d \times d$ multiplication, and division grids reached zero. The registered capacity rule redistributes shortfalls among remaining points to retain 2,000 items, records allocations, and fails closed on overlap. Original grids have four points; 39 of 96 fresh cells have three observed digit points, all lacking digit 2, as do 18 of 54 factorial cells. Across all 96 fresh grids, 78 have no digit-2 row after capacity-aware redistribution; 39 of these still have four observed points. The three 96-cell exact conditional McNemar families, Holm within family, use the discovery estimands and registered reversal/route-support rules.

Per-digit heterogeneity. Per-digit effects can change sign or magnitude despite a stable pooled effect: among the 66 Holm-significant fixed-route discovery cells, 51 span more than $10\,\mathrm{pp}$ and 13 contain both positive and negative sampled points. OLMo-3 final addition changes from $-52.8\,\mathrm{pp}$ at ten digits to $+9.6\,\mathrm{pp}$ at eleven; Granite 4.1 3B addition changes from $-3.8\,\mathrm{pp}$ at two digits to -20.6–$-23.4\,\mathrm{pp}$ at the sampled 13–17-digit points. These descriptive slices use pointwise Newcombe intervals without an additional multiplicity claim. The supplementary crosswalk indexes the full per-digit effects and variation, fixed-route four-state counts ($n = 2{,}000$ per cell; zero cap exclusions), and cumulative-stage records. Original stage grids share no item IDs; paired stage localization instead uses the re-measured OLMo-3 items (Appendix G).

For OLMo-3 addition at 10 and 11 operand digits, base/final shared-route accuracies are respectively $73.2/20.4\%$ and $47.6/57.2\%$ in discovery, and $67.6/23.0\%$ and $44.8/56.2\%$ in fresh replication; final-checkpoint errors in the fresh slices include 148 short-by-one and 161 wrong-leading-group outputs at 10 digits, versus 3 short-by-one, 97 wrong-leading-group, and 43 longer-than-target outputs at 11. For target numerals, width 11 marks a leading comma-group boundary ($1;3;3;3 \to 2;3;3;3$), with a single-digit leading chunk at 10 digits and a two-digit chunk at 11 under the pinned tokenizer; sums can be one digit wider than their operands, and this records the boundary without establishing a cause.

Descriptive common-grid profiles cover ten contrasts, 120 curve cells and 1,100 point-cells, all estimable, without an inferential family; full per-digit records are indexed in `REPRODUCE.md`.

Readout and difficulty. Strict last-line and tolerant numeric scoring give identical effect signs and Holm decisions in all three 96-cell decomposition families, with zero effect discrepancy. D50 retains censored and non-estimable cells, size-pair analyses are secondary, and the non-estimable Think branch is disclosed in Appendix F.3.

Difficulty-support sensitivity. Recomputations in the supplementary digit-support records use equal digit weights and the grids' common support, with stratified paired-Wald tests replacing exact McNemar and paired-bootstrap tests; both weighting and test change. All 24 of 96 reversal cells persist under both weightings. Discovery–confirmation Holm agreement is 251, 247, and 266 of 288 family-cell rows under registered, equal-weight, and common-support weighting. Interaction significance drops from 45 to 44 of 54 (Qwen3.5 2B division), with all signs unchanged.

Common-support recount. For each contrast–task cell, we restrict both grids to their intersection of observed digit points and give those points equal weight. The stratified paired-Wald recount applies Holm separately over all 96 cells in each family and grid before extracting the documented contrasts (Table 11). OLMo-3 retains taxes on subtraction and single-digit multiplication at all three stages and on addition after DPO and final; gains occur only on $d \times 2$, $d \times d$, and division. Gain counts depend on difficulty support, explaining the discrepancy with counts on the registered grids; the common-support recount also uses the test described above. The residual OLMo-3 discovery–fresh gain differences on common support occur within those multiplication tasks, without changing the operation-level tax/gain partition.

Family results and tier agreement. Table 12 compares canonical single-BOS discovery and confirmation families. Fresh-grid classifications are 32 amplification, 14 partial-cancellation, 24 reversal, 24 opposite-sign-unresolved, and 2 zero-sign cells (discovery per-family: 28/15/32/20/1; global: 28/15/31/21/1). Reversals number 24 under both per-family and global 288-test Holm scopes (discovery: 32 and 31); all 24 have significant route support, and 23 overlap discovery reversals. Effect signs agree in 271 of 288 cells (94.1%), Holm decisions in 251 (87.2%). Checkpoint/route/bundle correlations are Pearson $r = 0.98/0.96/0.96$, Spearman $\rho = 0.94/0.97/0.97$; median absolute discrepancies are $1.53/1.65/1.68\,\mathrm{pp}$. Cell estimates, classification, tier comparison, and tolerant mirrors remain in the supplementary replication records.

Registered factorial completion on the fresh grids. Registered 2026-08-27 and run 2026-08-28, this completion adds $A(B, R_n)$ to the observed three-condition replication grids: nine compatible pairs, 54 runnable of 60 slots, with Gemma's token-ID exclusion carried forward. The instrument and 2,000 paired items per cell are unchanged. Interaction inference was fixed before fourth-condition outcomes or their interaction statistics, using $B = 200{,}000$ paired bootstrap replicates (the earlier

Table 10: Fresh-grid absolute accuracies (%) for the seven documented post-training contrasts, with 2,000 paired items per task cell. The four columns distinguish checkpoint and route; $A(B, R_n)$ is not observed in this three-condition replication (—), and the separate compatible-pair factorial is reported in Table 13.

Pair	Task	$A(B, R_s)$	$A(E, R_s)$	$A(B, R_n)$	$A(E, R_n)$
OLMo-2 1B	add	18.00	34.55	—	21.25
	sub	20.20	28.50	—	23.00
	$d \times 1$	7.75	14.25	—	9.20
	$d \times 2$	7.50	12.75	—	7.50
	$d \times d$	6.20	10.10	—	5.40
	div	3.30	4.95	—	2.30
OLMo-3 SFT	add	51.80	51.25	—	42.20
	sub	53.60	24.60	—	28.25
	$d \times 1$	45.40	38.70	—	28.25
	$d \times 2$	18.80	22.60	—	23.45
	$d \times d$	2.90	4.25	—	4.75
	div	4.85	9.10	—	11.10
OLMo-3 DPO	add	51.10	40.10	—	38.95
	sub	54.35	22.00	—	30.00
	$d \times 1$	48.05	31.30	—	32.75
	$d \times 2$	18.50	22.35	—	26.05
	$d \times d$	3.40	5.15	—	5.10
	div	6.20	8.65	—	11.30
OLMo-3 final	add	51.20	37.45	—	22.90
	sub	54.15	20.00	—	27.00
	$d \times 1$	47.85	28.75	—	25.75
	$d \times 2$	18.55	20.05	—	27.90
	$d \times d$	2.65	4.50	—	5.15
	div	5.30	7.65	—	12.20
Tulu SFT	add	46.55	63.30	—	9.25
	sub	49.10	63.25	—	11.50
	$d \times 1$	48.10	50.55	—	1.65
	$d \times 2$	17.55	17.85	—	0.30
	$d \times d$	2.15	2.65	—	0.00
	div	10.30	12.25	—	0.20
Tulu DPO	add	50.15	63.30	—	18.35
	sub	47.95	63.80	—	31.15
	$d \times 1$	47.25	51.50	—	5.95
	$d \times 2$	17.60	19.10	—	0.75
	$d \times d$	2.40	2.80	—	0.20
	div	10.30	12.00	—	2.15
Tulu final	add	48.75	60.20	—	17.75
	sub	49.35	61.60	—	30.65
	$d \times 1$	47.65	49.05	—	12.80
	$d \times 2$	16.90	18.05	—	1.75
	$d \times d$	3.00	3.55	—	0.35
	div	9.70	13.80	—	5.55

$B = 2{,}000$ rule was not reused), Holm over 54 cells, and a secondary route-at-base exact McNemar family.

For strict-correct indicator a_i, $\theta_i = [a_i(E, R_n) - a_i(E, R_s)] - [a_i(B, R_n) - a_i(B, R_s)] \in \{-2, \ldots, 2\}$ and $\hat{\theta} = \bar{\theta}_i$. Each replicate draws 2,000 item indices with replacement, jointly for all four outcomes (NumPy `default_rng`, seed $[20260827, \text{cell index}]$). The registered two-sided percentile-bootstrap tail probability is

$$\min\left(1, 2\min\left(\frac{1 + \#\{\hat{\theta}^* \le 0\}}{B+1}, \frac{1 + \#\{\hat{\theta}^* \ge 0\}}{B+1}\right)\right).$$

Table 11: Fixed-completion checkpoint tax/gain recount on common discovery–fresh digit support. Entries are tax / gain / not-directional counts after 96-cell Holm correction. The upper panel counts six tasks per contrast; the lower counts the three cumulative OLMo-3 contrasts per task. Not-directional does not imply equivalence.

Contrast or task	Discovery	Fresh
OLMo-2 1B	0 / 6 / 0	0 / 6 / 0
OLMo-3 SFT	2 / 2 / 2	2 / 3 / 1
OLMo-3 DPO	3 / 2 / 1	3 / 3 / 0
OLMo-3 final	3 / 2 / 1	3 / 2 / 1
Tulu SFT	0 / 2 / 4	0 / 2 / 4
Tulu DPO	0 / 4 / 2	0 / 3 / 3
Tulu final	0 / 3 / 3	0 / 3 / 3
OLMo-3 task totals		
Addition	2 / 0 / 1	2 / 0 / 1
Subtraction	3 / 0 / 0	3 / 0 / 0
$d \times 1$	3 / 0 / 0	3 / 0 / 0
$d \times 2$	0 / 1 / 2	0 / 2 / 1
$d \times d$	0 / 2 / 1	0 / 3 / 0
Division	0 / 3 / 0	0 / 3 / 0

Table 12: Pre-registered replication families versus their discovery counterparts. Sig. is the Holm-significant count in the named 96-cell family with its negative/positive split; sign and Holm agreement are cell-level agreement rates between the tiers; medians are family median effects in pp.

	Sig. (−/+)		Tier agreement		Median (pp)	
Family	Confirmation	Discovery	Sign	Holm	Confirmation	Discovery
Fixed-route checkpoint	67 (39/28)	66 (39/27)	90.6%	84.4%	$+0.175$	-0.125
Route at fixed E	79 (50/29)	89 (52/37)	94.8%	87.5%	-2.30	-2.50
Bundled post-trained vs base	79 (53/26)	87 (57/30)	96.9%	89.6%	-2.625	-10.050

Table 13: Fresh-grid factorial completion: per-pair six-task mean absolute accuracies (%), median checkpoint contrasts on each route (pp), and task-level floor flags. F/N means base chat-template accuracy $\leq 1\%$/$> 1\%$, in task order add/sub/$d \times 1$/$d \times 2$/$d \times d$/div; Sig. counts Holm-significant interactions and Flip counts opposite Holm-significant checkpoint directions. Dual-route summaries are post-outcome; the interaction family was registered before the fourth-condition outcomes. Full 54-cell values and original-grid medians are indexed in the supplementary crosswalk.

	Mean accuracy (%)				Median Δ (pp)				
Pair	B, R_s	E, R_s	B, R_n	E, R_n	R_s	R_n	F/N flags	Sig.	Flip
OLMo 2 1B	10.49	17.52	0.01	11.44	$+5.88$	$+8.35$	FFFFFF	4	0
SmolLM3 3B	29.20	36.26	0.03	0.02	$+6.65$	$+0.00$	FFFFFF	6	0
LLM-jp-3 1.8B	15.35	13.33	0.03	6.30	$+0.55$	$+2.45$	FFFFFF	3	2
Llama 3.2 1B	17.29	13.68	0.00	11.52	-4.48	$+12.00$	FFFFFF	6	4
Qwen3.5 2B	58.48	27.25	33.75	24.86	-28.10	-4.60	NNNNNN	6	1
Sailor2 1B	17.79	12.27	0.00	3.51	-2.00	$+3.03$	FFFFFF	4	3
Granite 4.1 3B	27.77	20.09	0.00	25.06	-7.33	$+28.90$	FFFFFF	6	4
Llama 3.2 3B	32.09	30.81	1.23	24.10	-0.15	$+19.33$	NFNFFF	6	1
Qwen3.5 4B	37.28	25.63	62.20	54.32	-16.08	-8.15	NNNNNN	4	1

Zero ties enter both tails; there is no null centering. The attainable floor $2/200{,}001 \approx 1.0 \times 10^{-5} < 0.05/54$ was checked before registration; 95% intervals use the $2.5/97.5$ percentiles.

Interaction is significant in 45 of 54 cells (38 positive, 7 negative; median $+8.00$ pp); every non-significant $|\theta| \leq 1.65$ pp, with identical tolerant decisions. Route-at-base is significant in 53 of 54 (median -9.18 pp; 48 negative, six Qwen3.5 4B positive estimates). Against the canonical original-grid $B = 200{,}000$ reanalysis, signs agree in 51 of 54, Holm decisions in 50, Pearson $r = 0.939$, and median absolute discrepancy is 2.08 pp.

Post-outcome robustness resamples nine checkpoint-pair clusters ($B = 20{,}000$, fixed seed): median/mean interaction 95% intervals are $[+1.38, +15.33]/[+5.06, +20.14]$ pp. Leaving out any pair retains 39–42 of 48 significant cells and a positive median. Equal digit weighting preserves every significant cell's sign (largest shift 13.8 pp); 84.8% of 164 significant-cell slices agree with the pooled sign (78.8% of all 198 slices), all points agree in 30 of 45 significant cells, and at least three agree in

24 of 29 four-point significant cells. The other 16 significant cells lack two-digit points in subtraction, two-digit/$d \times d$ multiplication, or division. Slices are descriptive (median $n = 500$, range 11–667).

Table 13 shows the four mean accuracies and floor flags; the full 54-cell values ship. For OLMo-2 subtraction the four accuracies are 20.2/28.5/0.05/23.0%; for Qwen3.5 2B subtraction, 69.9/43.0/45.5/24.2%. Of 40 cells with base native accuracy $\leq 1\%$, 33 have significant interactions; of 14 non-floor cells, 12 do. Of 16 opposite Holm-significant checkpoint-direction flips, 13 are floor and 3 non-floor: Llama-3.2-3B addition (base native 6%), Qwen3.5 2B $d \times d$ (3%), and Qwen3.5 4B subtraction (84%).

Interaction re-analyses. The prospectively registered held-out fourth-condition confirmation was registered after the other three conditions were observed, with the fourth-condition outcome unobserved. Its registered bootstrap decision was Holm-significant in 45 of 54 cells. Two independent calibrated post-outcome mirrors were also Holm-significant in 45 of 54 cells each—a null-centered studentized bootstrap-t mirror and a Wald/score mirror—and each agreed with the registered decisions in 54 of 54 cells; these are sensitivity analyses of retained rows.

D.1 Within-cell heterogeneity and cluster robustness of the reversals

The primary digit-width analysis applies Holm separately to the checkpoint, route, and bundled families over all 345 observed units of the 96 fresh-grid cells (1,035 tests), before selecting pooled reversal cells, yielding 56 width-level reversal units with opposite corrected checkpoint/bundled directions and significant route support. Of these, 50 lie in the 24 pooled reversal cells: all 24 contain at least one, and 4 qualify at every observed width (Table 14); the other 6 lie in 6 of the 72 non-reversal cells, none qualifying at every observed width. This all-unit pre-selection view is a post-outcome analysis of digit-width heterogeneity.

Within the 24 pooled reversal cells, the bundled sign is consistent at all observed widths in 16 cells, the checkpoint sign in 18, and their signs are opposite at every width in 13; in 10 of those cells, both contrasts also have nominal $p < 0.05$ at every width. In a post-selection within-cell view, correcting each family only over the 87 units in those 24 selected cells (261 tests) yields 52 reversal units across all 24 cells, with 6 qualifying at every observed width; this conditional description does not estimate reversal frequency over all panel widths.

Absolute checkpoint effects have minimum/median/maximum $1.25/11.275/43.70\,\mathrm{pp}$, bundled effects $3.20/18.00/53.90\,\mathrm{pp}$, and route effects $7.45/36.625/65.25\,\mathrm{pp}$; 15 cells have both checkpoint and bundled magnitudes at least $5\,\mathrm{pp}$, and 12 at least $10\,\mathrm{pp}$. Recomputing each family's Holm correction after omitting each of the 16 pairs leaves 19–24 reversals among the remaining 90 cells (median 23.5); the minima omit Gemma or SmolLM3. Across 12 distinct base lineages, reversal tallies are SmolLM3 5, Qwen3.5 2B 0, Qwen3.5 4B 3, OLMo-2 1B 0, OLMo-3 7B 0, Gemma 4 E2B 5, Granite 4.1 3B 1, LLM-jp-3 1.8B 0, Llama-3.1/Tulu 8, Llama 3.2 1B 2, Llama 3.2 3B 0, and Sailor2 1B 0. The three OLMo-3 stage pairs share one base lineage, as do the three Tulu pairs (SFT/DPO/final: 2/3/3 reversals), so pair omission does not imply independent lineage replication.

The raw checkpoint and bundled point estimates share a sign in 46 of the 96 fresh-grid cells; the classification counts behind that total, and the overlap of 23 of the 32 discovery reversals, are reported in Appendix D. Across all 96 fresh-grid cells, 19 Holm-significant fixed-route checkpoint effects have $|\Delta| < 5\,\mathrm{pp}$ (5 deficits and 14 gains); the per-cell effect sizes behind this count are in the supplementary effect-size records rather than in a printed table.

The introduction's Gemma example also has a separate 4-digit, 500-item audit slice with a constrained bundled contrast of $-31.00\,\mathrm{pp}$ (Appendix F.6), an instance of within-cell heterogeneity relative to the positive pooled contrast.

Table 14: The 24 fresh-grid pooled reversal cells: pooled checkpoint, bundled, and route effects (pp), observed digit widths, and width-level consistency. B/C: bundled/checkpoint sign consistent across every observed width. H: opposite Holm-significant checkpoint/bundled directions and significant route support at every observed width, using the pre-selection correction over all 345 units per family (1,035 tests); four cells qualify.

Pair	Task	Δ_{ckpt}	Δ_{bundle}	Δ_{route}	Widths	B	C	H
Gemma 4 E2B	sub	-30.75	$+14.20$	$+44.95$	10,11,12	No	Yes	No
	$d \times 1$	-43.70	$+20.65$	$+64.35$	4,7,8,9	No	Yes	No
	$d \times 2$	-13.95	$+26.65$	$+40.60$	2,3,4,5	Yes	Yes	No
	$d \times d$	-1.25	$+6.50$	$+7.75$	3,4,5	No	No	No
	div	-5.40	$+6.20$	$+11.60$	3,4,5	Yes	Yes	No
Granite 4.1 3B	div	-3.20	$+5.25$	$+8.45$	3,4,5	No	Yes	No
Tulu final	add	$+11.45$	-31.00	-42.45	12,13,14,18	Yes	Yes	No
	sub	$+12.25$	-18.70	-30.95	9,12,13,15	Yes	Yes	No
	div	$+4.10$	-4.15	-8.25	3,4,5	Yes	No	No
Tulu DPO	add	$+13.15$	-31.80	-44.95	12,13,14,18	Yes	Yes	No
	sub	$+15.85$	-16.80	-32.65	9,12,13,15	Yes	Yes	Yes
	$d \times 1$	$+4.25$	-41.30	-45.55	4,5,6,11	Yes	No	No
Tulu SFT	add	$+16.75$	-37.30	-54.05	12,13,14,18	Yes	Yes	No
	sub	$+14.15$	-37.60	-51.75	9,12,13,15	Yes	Yes	Yes
Llama 3.2 1B	$d \times 2$	-4.05	$+4.65$	$+8.70$	2,3,4,5	Yes	No	No
	$d \times d$	-4.90	$+3.20$	$+8.10$	2,3,4,5	No	No	No
Qwen3.5 4B	sub	-13.75	$+51.50$	$+65.25$	3,4,5	Yes	Yes	No
	$d \times 1$	-19.20	$+4.50$	$+23.70$	9,10,11,12	No	Yes	No
	$d \times d$	-11.10	$+5.20$	$+16.30$	3,4,5	No	Yes	No
SmolLM3 3B	add	$+10.40$	-53.90	-64.30	5,7,8,10	Yes	Yes	No
	sub	$+14.70$	-47.55	-62.25	5,8,9,12	Yes	Yes	Yes
	$d \times 1$	$+9.70$	-49.95	-59.65	4,5,6,9	Yes	Yes	Yes
	$d \times 2$	$+3.60$	-17.30	-20.90	3,4,5	Yes	Yes	No
	div	$+3.00$	-4.45	-7.45	3,4,5	No	No	No

E DESCRIPTIVE FAILURE-MODE DECOMPOSITION OF ROUTE EFFECTS

This exploratory, outcome-informed analysis of retained rows has no inferential family; input/output hashes and full decompositions are in the supplementary failure-mode records.

Layer 1: paired transitions under the constrained instrument. The losing emission on each discordant route pair is a shorter-than-target numeral, a full-length wrong number (at least target-length digits), or cap/loop. Shorter-than-target describes emitted length despite permission to continue; it does not establish interrupted correct computation. Such numerals account for about half of route losses and dominate the largest contrasts (Table 15); gains mostly replace full-length wrong numbers, with zero cap/loop throughout.

Layer 2: unmasked emissions under unconstrained decoding. On the unconstrained-decoding rows (four contrasts, six tasks, fresh different grids, disclosed; 4,800 native-route rows per contrast), 9,572 of 19,200 native-route emissions (49.9%) contain no numeric token, including the open-reasoning-marker rows: SmolLM3 4,800 of 4,800 and the Llama-3.1/Tulu row 4,768 of 4,800, against 3 of 4,800 (Gemma 4 E2B) and 1 of 4,800 (Qwen3.5 2B). The no-numeric outcome—

Table 15: Descriptive route transitions: item counts and shorter-than-target / full-length-wrong shares. Post-trained: 96 discovery cells; base: 54 factorial cells. Top five selects contrasts by absolute median route effect; cap/loop is zero throughout.

Side	All losses	Top-five losses	All gains
Post-trained, items	29,009	17,438	18,543
Short / full wrong	43.3 / 56.7%	58.2 / 41.8%	23.1 / 76.9%
Base, items	24,609	17,054	3,685
Short / full wrong	55.9 / 44.1%	61.5 / 38.5%	8.2 / 91.8%

format-compliance failure, and for the marker rows reasoning-boundary loss—is thus concentrated in specific contrasts rather than a uniform native-route property; constrained decoding masks it by forcing numeric emission.

Identification limits. The assistant-role prior is not separately identifiable from template wrapping in the retained data—the factorial-completion post-trained-template probe cells bound only the combined wrapper effect at fixed weights; Layer 2 covers four contrasts on a fresh, different item grid; and no view carries a familywise claim.

F MATCHED READOUT CONTROLS

The original decoding and serialization ablation arms were exploratory and outcome-informed, whereas the matched-control completion on identical items was registered before its outcomes. Neither enters the 96-cell primary confirmatory families: the paired decoding and serialization comparisons use separate Holm families, with item-key equality, paired accuracy differences, exact conditional McNemar tests, and Newcombe intervals. Tolerant mirrors are retained alongside the strict results.

F.1 UNCONSTRAINED DECODING

In the unconstrained-decoding arm, the retained evidence is the identical-item paired analysis below, with raw per-cell failure taxonomies retained and the paired unconstrained-decoding estimates indexed in the supplementary crosswalk.

The registered paired family on identical items finds that the constrained instrument beats unconstrained decoding by a median $+1.69\,\mathrm{pp}$ (38 of 72 cells Holm-significant: 35 positive, 3 negative; range -9.0 to $+38.5\,\mathrm{pp}$). The paired effect is route-dependent: base completion median $+6.63\,\mathrm{pp}$ (13 of 24 significant, all positive), post-trained shared $+2.38\,\mathrm{pp}$ (13 of 24, all positive), and post-trained native median $0.00\,\mathrm{pp}$ with heterogeneous pairs—the Tulu pair's significant native cells recover $+2.0$ to $+26.6\,\mathrm{pp}$ under the grammar, while the SmolLM3 native cells stay at or below 1 of 800 under both decoders. The three significant negative cells are all Gemma 4 E2B native (up to $-9.0\,\mathrm{pp}$), where free-running decoding beats the grammar as the preferred emission format diverges from it.

F.2 NO-COMMA SERIALIZATION

The no-comma serialization arm uses identical item keys; its registered paired estimates are indexed in the supplementary crosswalk. A registered tokenizer-only token-count table (384 rows; no inference) records how each pinned tokenizer segments grouped and ungrouped operands; this table is indexed in the same crosswalk.

The registered paired family on identical item keys finds no systematic serialization direction: comma grouping minus no-comma has median $+0.00\,\mathrm{pp}$, but 25 of 54 cells are Holm-significant with mixed signs (15 favor grouping, 10 favor no-comma; range -24.0 to $+44.25\,\mathrm{pp}$), concentrated on the base-completion route (median $+2.75\,\mathrm{pp}$, 8 of 18 significant, all positive) while the post-trained shared route has median $0.00\,\mathrm{pp}$ and the post-trained native route median $-0.19\,\mathrm{pp}$. Serialization choice is therefore materially cell-specific—consistent with tokenizer-dependent segmentation.

Readout sensitivity is cell-specific rather than captured by an overall median. In the identical-item comparisons, decoding preserves both the decision and sign in 21 of 24 cells, changes the decision without changing the sign in 2, and flips the sign in 1; serialization preserves both in 15 of 18, changes the decision only in 2, and flips the sign in 1. The effects can differ by tens of percentage points: for SmolLM3 addition, the registered unconstrained arm gives $-6.25\,\mathrm{pp}$ versus $+12.75\,\mathrm{pp}$ for its grouped-constrained mirror, while for Qwen3.5-2B subtraction, the registered no-comma arm gives $+1.75\,\mathrm{pp}$ versus $-22.00\,\mathrm{pp}$ for its comma-grouped mirror. The readout-stable conclusions are that most cell-level decisions and signs persist and that readout effects remain cell-specific; individual magnitudes, significance decisions, and selected directions are not universally stable. The registered "unconstrained" arm removes the vocabulary mask but keeps the registered stop set, while its grouped-

constrained condition is the matched mirror; it is not free-form or chain-of-thought generation and does not generalize to free-response arithmetic.

F.3 Reasoning-budget arm and truncation

The reasoning-budget arm is non-estimable and truncated: the natural-termination qualification subset is selection-conditioned ($n = 7$), while retained budget points had zero natural terminations and were dominated by cap-forced readouts. Remaining cells were not executed after universal cap-hitting; this branch supports no primary arithmetic claim.

F.4 Free-generation termination audit (registered)

The earlier unconstrained arm removes the vocabulary mask but keeps the stop set. Registered on 2026-09-09 before its data, this audit removes both under the unchanged shared-completion prompt and greedy decoding, terminating only at the model's EOS or the token budget. It covers subtraction and single-digit multiplication for OLMo-3 base/post-trained, Llama-3.1/Tulu base/final, and SmolLM3 mid-training/final. After an initial 624-row launch exposed the runtime cost, scope was reduced before outcome analysis to the first 1,000 confirmatory items per cell and 128 new tokens; the abandoned rows are excluded. The paired comparisons use Newcombe intervals, exact McNemar, and Holm over six contrasts separately for each readout; registered-instrument scores are recomputed on the same items. Raw outputs, termination states, numeral presence, strict last-line, first-integer, and last-integer scores are retained.

Outcome. The registered primary criterion—maintaining every strict contrast's sign and Holm decision—fails: 8,818 of 12,000 rows reach the budget, strict accuracy is at most 0.9%, and no strict contrast is significant (Table 21). Numeral-first continuations drift into further equations or prose, so the last line no longer scores the answer; numeral-only outputs can fail because the decoded EOS marker remains in the scored string. Stripping that marker does not align the endpoint: strict accuracy then measures stopping after an answer and would reverse the OLMo-3 subtraction contrast to $+23.2\,\mathrm{pp}$. OLMo-3 post-trained also emits no numeral in 470 subtraction and 651 multiplication rows, typically ending after a question mark. The first-integer readout aligns the answer position and maintains all six signs and Holm decisions, including Tulu's null multiplication contrast; five effect estimates are within $2\,\mathrm{pp}$ of the registered value. OLMo-3 multiplication is the exception ($-12.6\,\mathrm{pp}$ registered versus $-36.6\,\mathrm{pp}$ free): its first-integer accuracy falls from 42.7% to 17.8% when the mask and stop set are removed together. The last-integer view instead reverses Tulu subtraction to $-22.0\,\mathrm{pp}$ and yields a $+29.0\,\mathrm{pp}$ multiplication gain, reflecting later continuations rather than the answer.

Per-model EOS and budget counts, including OLMo-3 single-stream replacements, are indexed in the supplementary crosswalk; the supplementary crosswalk also locates the archived table with all short-output, numeral, and accuracy columns. SmolLM3's system-text repetitions occur after the answer, not as pre-answer reasoning. Single-stream OLMo-3 checks moved the post-trained counts by at most 3 rows and changed no Holm decision; base counts moved by up to 11 rows. The records retain both executions.

Nonresponse versus wrong numerals. Linking the OLMo-3 post-trained outputs to the registered scores (Table 16), numeral-emitting rows agree in 96.0% of subtraction cases (509 of 530) and 96.6% of multiplication cases (337 of 349). No-numeral rows are registered-correct in 58 of 470 subtraction cases (12%) and 239 of 651 multiplication cases (37%); the latter supply 56% of the 427 registered-correct multiplication rows. The audit does not separate refusal from absent computation or the mask from termination handling. Its failed strict criterion concerns transfer of the scoring convention; the first-integer comparison concerns checkpoint differences at the aligned answer position.

F.5 Prompt-form audit (registered)

Registered after the free-generation audit and before any of its own data existed, this audit fixed the first-500 slice in its original design and the last-500 (14-/15-digit) slice in a registered addendum, also before its data. The registration records ship in the supplementary archive. This audit asks whether fixed-route subtraction contrasts depend on changing from the expression-completion form

Table 16: OLMo-3 post-trained checkpoint, same 1,000 items per task: free-generation outcome (rows) against the registered strict outcome under the constrained instrument (columns).

Task	Free generation	Registered correct	Registered wrong
subtraction	no numeral (470)	58	412
subtraction	numeral, correct (340)	332	8
subtraction	numeral, wrong (190)	13	177
mult.	no numeral (651)	239	412
mult.	numeral, correct (178)	177	1
mult.	numeral, wrong (171)	11	160

Table 17: Question-form free generation on 500 OLMo-3 14-digit subtraction items per checkpoint; base uses single-stream execution. First-wrong counts run from the leading through fifth group; wrong-group counts run from one through five and six-or-more groups. Both distributions and the ± 1 count use wrong non-negative numerals only; excluded rows include negative predictions.

Checkpoint	Exact	Wrong	Excluded	First wrong (1–5)	Wrong groups (1–6+)	± 1
base	262	238	0	14/46/73/88/17	179/47/10/1/1/0	181
SFT	15	473	12	171/262/32/7/1	19/17/62/273/99/3	51
DPO	6	475	19	209/234/28/3/1	8/19/79/266/100/3	59
final	7	481	12	206/244/28/3/0	7/20/66/291/93/4	60

to a question form. Two pairs (OLMo-3 7B base/post-trained; Llama-3.1 8B/Tulu 3.1) were re-measured on two slices of their confirmatory subtraction grids in registered item order—the first 500 items (three-digit operands for OLMo-3, nine-digit for Llama/Tulu) and the last 500 (14- and 15-digit operands)—under the expression form `Output ONLY a number.\n{a}-{b}=` and the question form `Output ONLY a number.\nWhat is {a} minus {b}?\nAnswer:`. Each form was read out with the registered constrained decoder and under free generation (greedy, no mask, no stop set, 128-token budget, first-integer view). Expression-form constrained rows reproduced the confirmatory raw outputs byte for byte on checked items, and question-form raw outputs were read before contrasts were interpreted. Table 2 reports the representative free-generation contrasts; Table 19 retains the constrained comparison and its question-form failures; original batched-base values are in the supplementary archive. The representative 14-digit OLMo-3 values use the confirmation arm's single-stream base and the addendum's single-stream final restatement; the SFT/DPO outputs are reused from the stage arm, not re-generated (Appendix F.5).

Outcome. The question-form constrained decoder is mismatched: every checkpoint prefers a standalone space after `Answer:`, which the grammar masks; those continuations are not interpreted as checkpoint effects. Under question-form free generation (Table 2), no deficit is detected on the three-digit OLMo-3 slice, while the 14-digit loss persists: of 500 post-trained outputs, 496 contain a parser-recognised integer (7 correct, 489 wrong); counting all 4 unparsed rows as correct gives at most 2.2%, versus the single-stream base 52.4%. Tulu's gains become negative because its base benefits more from the question form. Thus fixed-route effects depend on prompt form; the near-ceiling three-digit result shows no observed loss, not capability equivalence.

Digit-group localisation. On the 500 shared 14-digit subtraction items, we split non-negative integer predictions and targets into canonical base-1,000 groups and align them from the left; target lengths range from 11 to 14 digits. Table 17 localises question-form free-generation errors. Among 500 final question-form outputs, 496 contain a parser-recognised integer (7 exact and 489 wrong); the group-alignment view additionally excludes the 8 parsed negative predictions and the 4 rows without a first integer, leaving 481 wrong and 12 excluded. At SFT, DPO, and final, errors usually start in the leading or second group and have a modal count of four wrong groups. The base more often fails later and usually in one group. A borrow signature means at least one aligned wrong group differs from its target by exactly ± 1; it is a descriptive signature, not a diagnosed causal chain.

Under expression-form constrained decoding in the prompt-form audit, exact/wrong counts are base 146/354, SFT 6/494, DPO 1/499, and final 1/499, with no excluded numerals. Post-trained errors start in the leading group in 456/494, 478/499, and 480/499 cases (92–96%); most affect four or five groups. The base starts there in 136/354 cases and has one wrong group in 191/354. The original registered stage rows, with distinct buffer precision (Appendix H), show the same pattern: exact/wrong counts are 142/358, 7/493, 2/498, and 1/499, and leading-first errors number 138, 458, 479, and 481, respectively. The concentration of post-trained leading-group errors is consistent with

Table 18: First-loss stage on the same 500 14-digit subtraction items, restricted to each instrument's base-correct/final-wrong items (n). Free-generation rows use the single-stream base; the registered-instrument row uses the original stage-measurement records.

Form	Readout	n	SFT	DPO	Final
Question	Free generation, first integer	259	253	6	0
Expression	Free generation, first integer	136	132	4	0
Expression	Registered constrained, strict	141	139	1	1

With the single-stream base, all 253 question-form and 132 expression-form SFT losses remain wrong at DPO and final.

the transplant population's leading-group-competitor selection. Across the post-trained question and expression views, the ± 1 signature occurs in only about 11–18% of wrong numerals (60/481, approximately 12%, for question-form final), so a single propagated borrow error does not describe the bulk of failures.

Batched decoding. Llama-3.1/Tulu single-stream checks diverged only after the answer, but OLMo-3 post-trained flipped its first token in 2 of 16 rows; its single-stream re-generations changed 7–23% of texts, aggregates by at most 3 rows, and no Holm decision. In the separate 1,000-item termination audit the post-trained bound is 3 and the base bound 11; in the later 500-item prompt confirmation the base bound is 2. Original values and execution histories are in the supplementary archive.

Stage arm. Registered before its data, this arm evaluates pinned OLMo-3 SFT/DPO revisions on the same last 500 subtraction items (14 digits), both forms and readouts, with batch size 1. The confirmation arm supplies the single-stream base, and the addendum supplies the single-stream final; SFT/DPO outputs are reused. Question-form first-integer counts (unparsed in parentheses) are base 262 (0), SFT 15 (6), DPO 6 (10), final 7 (4); expression-form counts are 137 (0), 11 (255), 1 (472), 1 (451). Unparsed includes malformed numerals and nonresponse. Constrained expression-form SFT/DPO counts are 6/1; question-form counts are 0/0 and are instrument mismatches. Raw wrong/malformed outputs and mismatches were reviewed before interpretation.

Table 18 compares first-loss stages within each instrument's base-correct/final-wrong subset of the same 500 items. With the single-stream base under question-form free generation, all 253 items first lost at SFT remain wrong at DPO and final, and all 6 first lost at DPO remain wrong at final. The dominant first-loss stage is SFT under both free-generation forms and the registered instrument; the 259-item question-form and 141-item registered-instrument lost sets are different populations.

Question-form readout audit. Using the shipped compact records indexed in `REPRODUCE.md`, we combine the single-stream base re-execution of the confirmation arm, the existing single-stream SFT/DPO outputs of the stage arm, and the single-stream final outputs of the addendum; counts on 500 shared 14-digit items per checkpoint are, in order: first-integer correct / last-integer correct / target string present / rows restating the first operand (the comma-grouped a appears in the raw output) / first integer equals an operand / no first integer. Final gives $7/7/7/0/0/4$; SFT $15/15/17/1/1/6$; DPO $6/6/7/0/0/10$; and the re-executed base $262/31/262/437/0/0$. Base continuations repeat equations after answering, explaining low last-integer correctness and the registered choice of the first integer for free generation. The final checkpoint's wrong numerals are not operand pick-ups, and neither alternative readout removes the 14-digit deficit.

Execution conditions and the failed byte-level reproduction check are reported in Appendix H.

Confirmation arm. Registered before its data, the confirmation arm of the prompt-form audit re-generated the OLMo-3 base checkpoint on the same last 500 subtraction items (14 digits) under expression-form constrained decoding and free generation with both forms, and the Tulu final checkpoint on the same first 500 (9-digit) and last 500 (15-digit) items under question-form free generation through the registered native route (its chat template with the registered system configuration).

All new cells ran single-stream (bfloat16 parameters, SDPA, NVIDIA RTX A4000, PyTorch 2.11.0, Transformers 5.12.1); free generation was greedy with model EOS or a 128-token budget. Both prompt-form programs omit the stage-measurement driver's post-load cast (Appendix H). Against the single-stream base, question-form SFT/DPO contrasts are -49.4 pp $[-53.9, -44.6]$ and -51.2 pp $[-55.6, -46.5]$, both significant in the same two-test Holm family; secondary expression contrasts are

Table 19: Prompt-form audit under constrained decoding: paired checkpoint contrasts on the same first/last 500 subtraction items (pp, paired Newcombe 95% intervals; †: Holm-significant in the original four-contrast slice family). Expression-form rows are valid comparisons; question-form rows are instrument mismatches and their daggers are suppressed without recomputing the families. Free-generation comparisons are in Table 2. Accuracies in %.

Slice, readout	Pair	Form	Base	Post-trained	Contrast
(A) Checkpoint contrasts					
first 500, constrained	OLMo-3	expression	87.8	75.2	-12.6 [-17.0, -8.2]†
first 500, constrained	Llama/Tulu	expression	84.4	89.8	$+5.4$ [$+2.2$, $+8.7$]†
last 500, constrained	OLMo-3	expression	29.2	0.2	-29.0 [-33.1, -25.1]†
last 500, constrained	Llama/Tulu	expression	17.2	35.4	$+18.2$ [$+13.7$, $+22.6$]†
(B) Instrument-mismatch diagnostic (not a checkpoint effect)					
first 500, constrained	OLMo-3	question	64.2	15.6	-48.6 [-53.3, -43.4]
first 500, constrained	Llama/Tulu	question	66.0	59.6	-6.4 [-11.0, -1.8]
last 500, constrained	OLMo-3	question	20.0	0.0	-20.0 [-23.7, -16.6]
last 500, constrained	Llama/Tulu	question	26.4	20.6	-5.8 [-9.8, -1.8]

Table 20: Tulu final native-minus-shared question-form free-generation sensitivity on the same 500 subtraction items per slice. Native counts are correct last integers or target-string matches; contrasts are pp with paired Newcombe 95% intervals, without adjusted tests. Primary first-integer rows are in Table 2B; the registered constrained references are -27.8 [-32.2, -23.3] and -35.2 [-39.5, -31.0] pp.

	9 digits (first 500)		15 digits (last 500)	
Readout	Native	Route contrast	Native	Route contrast
Last integer	324/500	-29.0 [-33.4, -24.6]	23/500	-29.6 [-33.9, -25.3]
Target string present	333/500	-27.4 [-31.7, -23.1]	23/500	-30.0 [-34.3, -25.7]

Table 21: Free-generation termination audit: paired checkpoint contrasts (post-trained minus base, pp, Newcombe 95% intervals) on the same 1,000 items per cell under the registered instrument and under free generation with three readouts. †: Holm-significant within its six-contrast family. Strict floor: registered strict last-line accuracy on the complete continuation under free generation (base / post-trained, %).

Contrast	Task	Registered (strict)	Free: first integer	Free: last integer	Strict floor
OLMo-3 7B	subtr.	-34.2 [-37.5, -30.7]†	-35.7 [-39.2, -32.1]†	-13.3 [-16.9, -9.6]†	0.0 / 0.1
OLMo-3 7B	mult.	-12.6 [-16.1, -9.0]†	-36.6 [-40.1, -32.9]†	$+0.0$ [-3.1, $+3.1$]	0.1 / 0.0
Llama-3.1/Tulu 3.1	subtr.	$+11.2$ [$+8.2$, $+14.2$]†	$+11.2$ [$+8.1$, $+14.3$]†	-22.0 [-26.1, -17.8]†	0.9 / 0.2
Llama-3.1/Tulu 3.1	mult.	$+0.3$ [-2.3, $+2.9$]	$+0.3$ [-2.2, $+2.8$]	$+29.0$ [$+25.5$, $+32.4$]†	0.0 / 0.0
SmolLM3 3B	subtr.	$+13.2$ [$+10.6$, $+15.8$]†	$+14.7$ [$+11.9$, $+17.5$]†	$+6.2$ [$+3.0$, $+9.4$]†	0.3 / 0.1
SmolLM3 3B	mult.	$+4.3$ [$+1.6$, $+7.0$]†	$+5.5$ [$+2.6$, $+8.4$]†	$+19.7$ [$+16.6$, $+22.8$]†	0.1 / 0.1

-25.2 pp [-29.3, -21.2] and -27.2 pp [-31.3, -23.4]. Base expression-form constrained accuracy is 146/500 (29.2%), matching the confirmatory count, with byte-identical outputs on all three checked rows. Table 2 and Figure 5C use the single-stream final-minus-base comparisons; Table 18 retains the first-loss counts. All four last-500 free-generation Holm decisions and the dominant first-loss stage are unchanged.

For Tulu final, the primary native-minus-shared first-integer contrasts appear in Table 2B; both are significant within their registered two-contrast slice families (native Tulu versus shared Tulu and versus shared base). They keep the constrained reference's sign, differing by 5.6 and 7.0 pp, without establishing equivalence. Table 20 retains the descriptive last-integer and target-present sensitivities: the route deficit persists under all three readouts.

Native-route outputs restate the operands in prose on 22 nine-digit and 5 fifteen-digit rows (e.g., "The result of … minus … is …"), so the first-integer readout scores the operand as wrong even when the final number is right. Eight nine-digit rows begin with an underscore before the numeral; 19 fifteen-digit rows are empty after removing the EOS marker. Every row in both slices ends by EOS; none reaches the token budget. The supplementary raw review retains verbatim examples of these behaviours.

Table 22: Representative-cell contract comparison on the same 500-item slices: Qwen 2-digit subtraction, Gemma multiplication with a 4-digit multiplicand, SmolLM3 5-digit addition, and Tulu SFT 12-digit addition. C: constrained strict scoring, counts $a/c/d$; F: free first-integer scoring, counts $a/b/c/d$ (correct out of 500). Effects are pp with paired Newcombe 95% intervals and exact McNemar p_H, Holm-corrected over the two contrasts within each cell and contract; † marks significance at 0.05. Rev. indicates opposite Holm-significant checkpoint/bundled directions on the slice.

Cell	Contract	Correct counts	Δ_{ckpt} $(c-a)$ [95% CI]; p_H	Δ_{bundle} $(d-a)$ [95% CI]; p_H	Rev.
Qwen3.5 4B	C	426/432/500	$+1.20\,[-2.56, +4.97]$ 0.60	$+14.80\,[+11.86, +18.18]$† 2.1×10^{-22}	No
	F	497/500/15/500	$-96.40\,[-97.64, -94.19]$† 3.2×10^{-145}	$+0.60\,[-0.26, +1.75]$ 0.25	No
Gemma 4 E2B	C	430/0/275	$-86.00\,[-88.77, -82.60]$† 1.4×10^{-129}	$-31.00\,[-35.85, -25.94]$† 9.0×10^{-30}	No
	F	429/442/0/284	$-85.80\,[-88.59, -82.38]$† 2.9×10^{-129}	$-29.00\,[-33.77, -24.05]$† 1.1×10^{-27}	No
SmolLM3	C	380/423/0	$+8.60\,[+4.96, +12.28]$† 5.7×10^{-6}	$-76.00\,[-79.54, -71.99]$† 1.6×10^{-114}	Yes
	F	368/408/434/272	$+13.20\,[+9.40, +17.04]$† 2.3×10^{-11}	$-19.20\,[-24.40, -13.84]$† 1.5×10^{-11}	Yes
Tulu 3 SFT	C	366/399/102	$+6.60\,[+3.23, +9.99]$† 1.8×10^{-4}	$-52.80\,[-57.45, -47.63]$† 3.7×10^{-63}	Yes
	F	369/6/394/90	$+5.00\,[+1.52, +8.49]$† 0.0066	$-55.80\,[-60.35, -50.69]$† 9.0×10^{-69}	Yes

F.6 Representative-cell free-generation arm (registered)

Selection and routes. Registered before selection and before its data, the representative-cell free-generation arm ranks the 24 fresh-grid reversal cells by decreasing $|\Delta_{\text{bundle}} - \Delta_{\text{checkpoint}}|$ using the registered values, takes the first cell per contrast, skips contrasts already represented, and excludes the Tulu subtraction cells covered by the confirmation arm. Each selected cell's first 500 confirmatory items have one digit length: Qwen3.5 4B subtraction uses 2-digit operands; Gemma 4 E2B single-digit multiplication uses a 4-digit multiplicand; SmolLM3 mid-to-final addition uses 5-digit operands; and Llama 3.1 8B to Tulu 3 SFT addition uses 12-digit operands. These 500-item slices are evaluated through four routes: (a) base/shared expression form, zero-shot; (b) base/3-shot completion; (c) post-trained/shared expression form; and (d) post-trained/registered native route, with the expression as user text. Three exemplars are drawn by a seeded rule from the task's calibration pool, disjoint from evaluated items; each exemplar's exact target occupies the line immediately following its expression. For SmolLM3, "base" denotes the mid-training reference.

Readouts and decision rule. The primary readout is the first integer after the prompt; last integer and target-string presence anywhere in the continuation are sensitivities, with presence not necessarily an asserted answer. The predeclared comparison asks whether $c - a$ and $d - a$ retain their full-grid signs and Holm decisions, and whether the few-shot base changes the bundled sign ($d - b$). The registration specifies Holm decisions without a family size; we use two exact McNemar tests per cell/readout at $\alpha = 0.05$, with Newcombe method-10 95% intervals. Route ($d - c$) and few-shot ($d - b$, $b - a$) contrasts are unadjusted. Relative to the full grid, SmolLM3 and Tulu SFT retain both signs and decisions, Qwen retains signs but loses bundled significance, and Gemma retains bundled significance with the opposite sign; the same-item comparison is in Table 22.

Same-item constrained recount. On the same 500 items, SmolLM3 and Tulu SFT reverse under both contracts; Qwen and Gemma under neither (Table 22). Gemma's effects agree closely across contracts. Qwen uses the repaired native record with the closed reasoning block (Appendix H); the superseded record is excluded. Its native route remains 500/500 correct while post-trained shared correctness falls from 432/500 to 15/500 in capped reasoning prose and base shared rises from 426/500 to 497/500. Thus its checkpoint contrast changes from $+1.20$ pp (n.s.) to -96.40 pp and its bundled contrast from $+14.80$ pp to $+0.60$ pp (n.s.).

Sensitivity readouts. SmolLM3 retains the reversal under target presence ($+14.00/-13.40$ pp, checkpoint/bundle) but not last-integer scoring ($+7.00/+49.40$ pp); Tulu retains it under all three readouts. Qwen and Gemma retain no full-grid reversal under any free readout; Gemma base shared

is 0/500 last-integer-correct. Joint eight-test Holm correction changes no decision under any readout (smallest first-integer margin: Tulu checkpoint $p_H = 0.0131$). With the few-shot base, Qwen remains nonsignificant ($+0.00$ versus $+0.60$ pp), Gemma and SmolLM3 remain negative, and Tulu's $d - b$ is $+16.80$ pp $[+13.38, +20.47]$ versus -55.80 pp zero-shot, but its few-shot instrument failure makes this contrast uninformative; full route/few-shot contrasts and sensitivity intervals are in the supplementary archive.

Raw-output review. Outputs were read before paired scoring was interpreted. Qwen's post-trained shared route emits a thinking marker and reasoning prose on 499/500 rows, and all 500 reach the 128-token cap: the first integer is often an intermediate value (15/500 correct), although 405/500 contain the target string. This is a readout/budget outcome of the shared expression form on a reasoning-tuned checkpoint; this first-integer decline alone does not establish a contract-independent loss of arithmetic ability, and the presence of the target string in 405/500 continuations does not establish that it is preserved. Gemma's post-trained shared route scores 0/500 under every readout; 26 rows contain no parser-recognised integer (18 repeated numeric/comma degeneration, 8 other), and the remaining continuations omit the target. This shared-route degeneration explains only Gemma's checkpoint contrast; its bundled contrast compares post-trained native with base shared and does not involve post-trained shared outputs. For Llama 3.1 8B base with 3-shot completion, 489/500 rows emit only <|end_of_text|> immediately after the rendered prompt, with no new number or newline continuation; 11 rows reach the cap. All 500 inputs match the specified exemplar rendering, have identical length (78 tokens), and record batch size 16; immediate EOS occurs at every batch position 0–15; the recorded counts do not associate immediate termination with a unique batch position or a prompt-length difference. The exemplar targets were verified exact; immediate termination is the model's behaviour under this rendering, making the few-shot contrast uninformative, and it was not rerun with a different rendering, which would have been outcome-informed tuning. Tulu's native route immediately emits an end marker on 126/500 rows (no integer); the remaining rows contain numeric continuations, and 104/500 contain the target string. Verbatim examples in the supplementary review include Qwen's "Thinking Process:", Gemma's repeated "8,8,8,8", and the complete output <|end_of_text|> for both immediate-termination subsets.

Interpretation boundary. (a) Same-slice reversals persist under the 128-token free-generation first-integer contract in two representative cells; this does not establish contract independence in general, since SmolLM3 loses the reversal under last-integer scoring. (b) Gemma's negative bundled sign on the slice concerns base shared versus post-trained native, and its difference from the positive full-grid sign cannot be attributed to post-trained shared-route degeneration. (c) Selection by the largest $|\Delta_{\text{bundle}} - \Delta_{\text{checkpoint}}|$, one cell per contrast family, is not a random sample of the 24 reversal cells and does not estimate their replication rate. (d) SmolLM3 retains signs and Holm decisions across constrained and free first-integer scoring, but its bundled effect changes from -76.00 pp to -19.20 pp as post-trained native correctness rises from 0/500 to 272/500; sign and decision retention is distinct from effect-size stability.

Termination and provenance. Qwen base shared and post-trained native end at EOS on every item; both Gemma base routes always hit the cap. SmolLM3 base/post-trained shared hit the cap on 499/500 and 477/500 items; Llama/Tulu base shared on 493/500, while post-trained shared ends at EOS on 381/500. All routes use free greedy generation, cap 128, and batch size 16. Compact item records with prompt digests, raw outputs, readouts, termination and execution conditions ship through the `REPRODUCE.md` crosswalk; launch and resume histories are archived.

G Stage-resolved measurement and base-content transplant on OLMo-3

Stage-resolved measurement. The four OLMo-3 7B checkpoints (base, SFT, DPO, final release) were measured freshly on one hardware stack on the 2,000 subtraction and 2,000 single-digit-multiplication items of the confirmatory grids under the shared-completion route, with the registered constrained numeric grammar and strict scoring. Unlike the non-overlapping stage grids of Appendix D, all four checkpoints answer the same items here, so the stage at which each item is first lost is observable. Table 23 gives the per-stage accuracies and the first-loss stage of the base-correct/final-wrong items: 630 of 708 subtraction losses (89.0%) first occur at SFT. For base and final, accuracy

Table 23: OLMo-3 shared-route accuracy by post-training stage (2,000 common items per task; strict scoring), the first stage at which base-correct/final-wrong items become incorrect, and the number of base-wrong/final-correct items.

	Accuracy (%)				First loss (items)			
Task	Base	SFT	DPO	Final	SFT	DPO	Final	Base-wrong/final-correct
Subtraction	53.6	24.2	20.6	20.1	630	62	16	38
Multiplication	46.7	39.4	32.0	29.2	336	116	49	151

Table 24: Digit-wise stage measurement and common-set dictionary transplant under the registered instrument. Lost: base-correct/final-wrong items among 500 stage-measurement items per width. Triples give SFT/DPO/final counts. Common n: items among the 40-item transplant set per task; the last column counts correct answers after transplant.

Task	Digits	Lost	First loss	Common n	Transplant correct
Subtraction	3	98	51 / 35 / 12	3	3 / 3 / 3
	12	285	257 / 25 / 3	14	7 / 7 / 8
	13	184	183 / 1 / 0	13	5 / 4 / 4
	14	141	139 / 1 / 1	10	0 / 0 / 0
Multiplication	6	120	82 / 25 / 13	10	8 / 7 / 5
	7	116	67 / 30 / 19	7	6 / 5 / 5
	8	168	108 / 46 / 14	18	14 / 10 / 10
	14	97	79 / 15 / 3	5	1 / 1 / 1

differs from the retained confirmatory rows by at most 0.2 percentage points; exact-answer agreement with the retained rows is 96.1% and 86.3% (subtraction, base and final) and 94.1% and 90.9% (multiplication), so final-subtraction agreement falls below the registered non-gating 90% threshold. Correctness disagreements are small and approximately balanced between the two runs (fresh-only 4–13 versus retained-only 7–15 items per cell); their cause was not experimentally isolated.

Correctness trajectories. The full four-checkpoint correctness trajectories (base, SFT, DPO, final; 1 = correct) are dominated by monotone loss: subtraction 1000 (627), 1100 (62), 1110 (16) against 1010 (3), 1011 (7), 1101 (5), 1001 (4); multiplication 1000 (334), 1100 (116), 1110 (49) against 1011 (15), 1101 (8), 1001 (6). Items that base answers wrongly and the final release correctly number 38 in subtraction (0111: 32) and 151 in multiplication (0111: 126; 0011: 17).

Digit-wise breakdown. The stage measurement contains 500 items per digit width per task. Table 24 separates stage first losses and the common transplant set by width. The common set contains 10 of 40 subtraction items and 5 of 40 multiplication items at 14 digits. On those 14-digit subtraction items, no intervention, the dictionary transplant, the strict cross-item control, and the random-subspace control all yield zero correct answers at every stage. Subtraction restoration therefore comes entirely from the 3-, 12-, and 13-digit items. This limits the link between the prompt-form and transplant experiments: agreement on the dominant first-loss stage does not establish restoration of the persistent 14-digit subtraction deficit by the tested block-27 transplant. The digit-wise recount is descriptive and uses existing items, without a new selection or test family.

Leading-group correctness after transplant. Table 25 separates leading-group correctness from full-answer correctness on the same common set. For 14-digit subtraction, after the block-27 dictionary transplant seven of ten leading groups are correct at every stage—six newly corrected at SFT and seven at DPO and final—while no full answer is correct. The intervention restores the leading group on these seven items while leaving errors in later groups; it does not restore the persistent full-answer deficit. The other widths and controls show how often local restoration extends to the complete answer.

Instrument and qualification. The intervention reads the shared-completion residual stream at the equals position. We use unembedding-based logit-lens directions through final norm at that block output; public Jacobian-lens fits (Gurnee et al., 2026) failed all-layer qualification, and none was fitted here. Each checkpoint had 120 retained base-correct items (60 selection, 60 validation), disjoint across checkpoints and splits. The criterion ranks the correct leading group in the numeric vocabulary's top three at the selected block. Initial qualification pooled shared and native routes and

Table 25: Leading-group correct counts on the common transplant set, by task, digit width, and arm. Triples are SFT/DPO/final; n is shared across stages and arms. The last column gives dictionary-transplant full-answer correct counts. All 96 digit–stage–task–arm rows have available outputs. Multiplication is $d \times 1$.

Task	Digits	n	None	Dictionary	Strict cross-item	Random	Dictionary full
Subtraction	3	3	2/0/0	3/3/3	0/0/0	2/1/2	3/3/3
	12	14	3/1/0	12/11/11	0/0/0	6/5/4	7/7/8
	13	13	0/0/0	12/11/10	1/0/0	1/1/1	5/4/4
	14	10	1/0/0	7/7/7	0/0/0	3/2/2	0/0/0
Multiplication	6	10	5/1/0	8/7/7	0/1/1	6/6/6	8/7/5
	7	7	3/1/0	7/7/7	0/0/0	5/4/4	6/5/5
	8	18	8/1/1	17/16/15	0/0/0	18/13/11	14/10/10
	14	5	0/1/1	4/4/4	0/0/0	4/3/3	1/1/1

all items, failing at every checkpoint. Before intervention outcomes, the rule was restricted to shared completion and each checkpoint's strictly correct items, requiring at least 20 per split and validation hit rate ≥ 0.8; existing readouts were revalidated without generation. Selected block/hits: base 32, 60/60; SFT 31, 35/35; DPO 32, 26/29; final 32, 21/21. Only blocks 31–32 are qualified semantic readouts; intervention blocks 25–28 came from exploratory trajectories. Among wrong answers, block-30 top-three hits are 52%/41%/34% at SFT/DPO/final (descriptive). The planned multi-step $(a + b) \times c + d$ family failed one-shot (base/final 0 of 40) and was disqualified.

Operations. Let h_E be the receiver (post-trained) residual state at block ℓ and the equals position, and h_B the same item's full base-checkpoint state at the same block and position. Let D be the answer-independent dictionary subspace spanned by the unit logit-lens directions of the 1,000 three-digit group values and the ten digits, and P_D its orthogonal projector; the projector has rank 1,000 at every checkpoint and block because the digit directions coincide with the single-digit group values. All projections use the receiver checkpoint's own projector. The dictionary transplant is

$$h_E^{(D)} = h_E - P_D h_E + \|P_D h_E\| \frac{P_D h_B}{\|P_D h_B\|}.$$

Its control uses a matched other item's base state in the same equation, with the total perturbation rescaled to the treatment perturbation norm, so the control is not an exact subspace replacement. In base-checkpoint trials the receiver is h_B, so the transplant is the identity, verified on all 300 base rows to a maximum deviation of 6×10^{-8}. Perturbations are applied through a forward hook on the block output and answers are regenerated with the registered constrained decoder; every row records the answers before and after treatment and control. For the leading-group target, items whose final-release emitted answer keeps the correct leading group have no competitor and are excluded from the exploration and held-out pools (32 of 708 subtraction and 171 of 501 base-correct/final-wrong multiplication items); rows whose arm is unreachable are recorded as unavailable and excluded from effect estimates.

Exploration and first held-out replication. Before outcomes, the inventory was reduced to leading-group target, dose 1.0, three blocks per task (25–27 multiplication, 26–28 subtraction), and 50 exploration items per task. Block 27 was selected exploratorily, not by semantic-readout qualification: six candidates per task from 180 operation–control comparisons; base multiplication wrong-group push was ineligible (12 available arms). Initial held-out draws excluded only each candidate's exploration items, leaving 14 items used in other operations' exploration and 98 of 421 used in readout qualification. The never-used re-tally preserves all eleven intended directions with pointwise intervals excluding zero; full early outcomes, intervals, and corrected exclusions remain in the supplementary archive.

Exploratory oracle, removal, wrong-group push, amplification, counterfactual, half-dose, and carry/borrow results are retained in the supplementary archive. The unqualified carry/borrow readout supplied no confirmed candidate and establishes neither presence nor absence of a carry effect.

Common-set confirmation with a strict cross-item control. Registered before its data, this confirmation keeps block 27, the dictionary, dose 1.0, and transplant operation fixed. One seeded set of 40 items per task is drawn from base-correct/final-wrong items with a leading-group competitor,

Table 26: Common-set block-27 transplant and controls, dose 1.0: the same 40 never-used items per task at every checkpoint. Entries are correct counts (repaired/broken relative to None). Dictionary transplants the same item's base dictionary component; Strict cross-item uses a different item's base component; Random uses a seeded rank-1,000 subspace. Differences are accuracy proportions with pointwise Newcombe paired 95% intervals (method 10), not Holm-adjusted intervals. Dictionary-minus-None is post-hoc; the control comparisons were registered before their data. Rescaled-control and remaining pairwise intervals are in the supplementary records. Post-hoc exact paired tests with Holm correction over six dictionary-minus-random cells are significant for subtraction at every stage and multiplication at none; dictionary-minus-none is significant in all six.

Checkpoint	Task	None	Dictionary	Strict cross-item	Random	Dictionary−None [CI]	Dictionary−Random [CI]
SFT	subtr.	4	15 (11/0)	1 (1/4)	6 (2/0)	$+0.275$ [$+0.124$, $+0.420$]	$+0.225$ [$+0.084$, $+0.362$]
SFT	mult.	16	29 (14/1)	0 (0/16)	27 (11/0)	$+0.325$ [$+0.147$, $+0.471$]	$+0.050$ [-0.059, $+0.159$]
DPO	subtr.	0	14 (14/0)	0 (0/0)	3 (3/0)	$+0.350$ [$+0.194$, $+0.505$]	$+0.275$ [$+0.124$, $+0.423$]
DPO	mult.	3	23 (21/1)	0 (0/3)	15 (13/1)	$+0.500$ [$+0.303$, $+0.648$]	$+0.200$ [$+0.032$, $+0.349$]
final	subtr.	0	15 (15/0)	0 (0/0)	4 (4/0)	$+0.375$ [$+0.216$, $+0.530$]	$+0.275$ [$+0.124$, $+0.420$]
final	mult.	1	21 (21/1)	0 (0/1)	14 (13/0)	$+0.500$ [$+0.304$, $+0.650$]	$+0.175$ [$+0.013$, $+0.321$]

excluding every item used anywhere in the program: 197 exploration, 480 qualification, 431 held-out replication, and 22 ineligible-candidate items, leaving 51 multiplication and 295 subtraction items. The same items run at SFT, DPO, and final. The rescaled control's total-perturbation matching did not strictly replace the dictionary component, motivating the strict control: a seeded within-task derangement supplies another item's base state in the transplant equation, without rescaling. Treatment and strict control therefore replace the same component at the same norm. Table 26 reports correct and repaired/broken counts. The rescaled arm yields 1–6 correct answers (1–3 repaired); its full counts and intervals remain in the common-set records. All twelve dictionary-minus-rescaled/strict pointwise 95% intervals exclude zero. The no-intervention arm re-executes each item: one final-multiplication item selected as final-wrong is correct on re-execution. Dictionary-minus-none intervals are a post-hoc sensitivity view; exact McNemar tests give $p \leq 0.00098$ in all six cells (largest Holm-adjusted $p = 0.0020$). A second execution on Apple M4 Max (MPS) agrees on correctness in 948 of 960 rows and emitted answers in 851 of 960; every cell retains its direction and interpretation.

Exact-test sensitivity. Post-hoc exact paired McNemar tests with Holm correction over the six dictionary-minus-random cells are significant for subtraction at SFT, DPO, and final ($p_H = 0.0156/0.0059/0.0059$), but for multiplication at none (exact $p = 0.625/0.0386/0.0654$; $p_H = 0.625/0.116/0.131$; dictionary versus random counts $29/27$, $23/15$, $21/14$). Dictionary-minus-none remains Holm-significant in all six cells (all $p_H \leq 0.002$), corrected as a separate six-cell family. These exploratory comparisons characterize the selected intervention population and do not establish a stage-specific natural mechanism.

Rank-matched random-subspace control. The strict control equalises the replaced component and its norm but not the distance of the post-intervention state from the receiver's original state. A further control, registered before its data, uses a seeded random orthonormal subspace R of the same rank (1,000) as D, with orthogonal projector P_R, on the same common set, checkpoints, block, and dose (Table 26):

$$h_E^{(R)} = h_E - P_R h_E + \|P_R h_E\| \frac{P_R h_B}{\|P_R h_B\|}.$$

The control projects the full base state; R is not orthogonal to D ($P_D P_R \neq 0$ in general), so it does not isolate dictionary-independent transfer. Geometry is closely matched: median perturbation norms are 18.1 versus 18.2 (multiplication) and 22.0 versus 21.8 (subtraction), post-intervention cosines 0.904 versus 0.905 and 0.861 versus 0.859, and dictionary-component norms 20.5 versus 20.6 and 19.0 versus 19.3. In subtraction, random transfer repairs 2–4 of 40 answers versus 11–15 for the dictionary; random-minus-none differences are $+0.05$–$+0.10$ with intervals including zero (upper limits up to $+0.23$). Dictionary-minus-random intervals exclude zero at every stage, with 31 checkpoint–item cells repaired only by the dictionary and none only by the random control. The registered rule calls this intermediate, not strict specificity: a small broad-transfer benefit remains possible. In multiplication, random transfer repairs 11–13 of 40 ($+0.275$–$+0.325$ versus none, intervals excluding zero); 35 cells are repaired by both, 21 only by the dictionary, and 2 only by random transfer. The pointwise paired intervals favor the dictionary at DPO and final but include

Table 27: Dictionary-minus-random accuracy effects and their subtraction-minus-multiplication difference δ (pp; within-task paired, between-task independent item bootstrap, pointwise 95% intervals). Each task has 40 common items at each stage; multiplication is $d \times 1$.

Stage	Subtraction	Multiplication	Task difference δ
SFT	$+22.5$ [$+10.0$, $+35.0$]	$+5.0$ [-5.0, $+15.0$]	$+17.5$ [$+2.5$, $+35.0$]
DPO	$+27.5$ [$+15.0$, $+42.5$]	$+20.0$ [$+5.0$, $+35.0$]	$+7.5$ [-12.5, $+27.5$]
final	$+27.5$ [$+15.0$, $+42.5$]	$+17.5$ [$+2.5$, $+32.5$]	$+10.0$ [-10.0, $+30.0$]

zero at SFT; none of the multiplication comparisons is significant under the six-cell exact-test Holm sensitivity (Table 26).

Task difference of subspace specificity. We directly compare dictionary-minus-random effects between tasks, $\delta = (A_D - A_R)_{\text{sub}} - (A_D - A_R)_{\text{mult}}$. The post-hoc bootstrap uses 10,000 draws (seed 20260912): within each task the 40 paired dictionary/random outcomes are resampled together, the two tasks are resampled independently (their item sets are distinct), and δ is the difference of the two independent resamples; the same item sets are reused across stages without pooling. Table 27 gives pointwise percentile 95% intervals. The between-task differences are $+17.5$ pp [$+2.5, +35.0$] at SFT, $+7.5$ pp [$-12.5, +27.5$] at DPO, and $+10.0$ pp [$-10.0, +30.0$] at final. Only the SFT pointwise interval excludes zero; these comparisons remain exploratory, without a Holm-adjusted between-task test. Differing within-task significance alone does not establish a task difference or a stage-specific natural mechanism.

Scope. These interventions test sufficiency within one lineage, route, instrument, and registered subspace; replication and common-set confirmation cover block 27 only. They do not identify the deficit's natural cause, unchanged downstream processing, or a unique answer-carrying subspace. The exploratory amplification null does not exclude a weaker or differently oriented post-trained representation. The supplement retains the exploration log, compact per-row outputs, digest-bound states, exclusions, donors, registrations, norms, cosines, trajectories, re-tallies, and sensitivity analyses.

H Protocol deviations, corrective analyses, and reproducibility limitations

Table 28 lists the protocol deviations, their fixes, and the remaining limits. Protocol/grids froze on 2026-08-13; prospective registrations covered factorial/replication (2026-08-18), the fresh fourth condition before its outcome (2026-08-27), and matched controls (2026-08-29). Failure-mode and high-resolution interaction analyses (2026-08-24–25) were post-outcome. BOS duplication arose from retokenizing an already wrapped string with special tokens enabled; an external audit and pinned-tokenizer input-ID hashes established it. Canonical comparisons used exact McNemar with within-family Holm; all main arithmetic families use replacement cells. The five stop-differing pairs are Llama 3.2 1B/3B, Qwen3.5 2B/4B, and Sailor2, which replace base EOS with chat stops; six OLMo-3/Tulu rows add chat tokens. The controlled buffer replay used NVIDIA RTX A4000, PyTorch 2.11.0, Transformers 5.12.1, SDPA, and batch size 1; numerical conditions for stage measurements and transplants remain distinct. Full amendments, historical effects, replay output counts, repetition/restart checks, and program-level receipts are indexed in `REPRODUCE.md`.

Table 28: Protocol deviations, fixes, and remaining limits. Unaffected cells and frozen item grids were verified unchanged; original values, run counts, and restart histories remain in the supplementary archive.

Issue	Scope	Fix	Effect on conclusions	Remaining limitation
D50 ambiguity (2026-08-13)	Exploratory difficulty summaries only	Post-calibration estimator cascade: monotone logistic, isotonic fallback, $\alpha = 0.05$	Accuracy estimands and grids unaffected; sealed states identical at $\alpha \in \{0.001, 0.01, 0.05\}$	Censored and non-estimable states retained
Calibration overlap (2026-08-14)	Qualification overlap: 34 primary points (30 of 72 cells, 65 rows), 13 floor cells	Fresh-item reruns (100 rows per primary point); zero-overlap tree audit	Qualification pools unchanged; confirmatory grids verified disjoint	Original calibration rows remain historical
Missing native cells (2026-08-16)	Generator assigned shared completion to every post-trained cell	Add native cells on frozen items and revisions; retain earlier shared rows	No model, item, prompt, route policy, decoder, or parser selected from shared outcomes	Original decomposition post-hoc; fresh-grid counterpart pre-registered
Qwen thinking (2026-08-18)	Six Qwen3.5 4B native cells and summaries; open reasoning marker passed qualification (0/60 easiest-addition correct)	Rerun with `enable_thinking=false`; require closed delimiters	Earlier size-pair conclusion reversed; route extreme invalidated; repaired cells used throughout	Structural validity alone had missed a readout failure
Duplicated BOS	Six checkpoints: Llama-3.2 1B/3B, Tulu SFT/DPO/final, llm-jp; affected native cells and lens program	Retokenize with `add_special_tokens=False`; gate BOS and equality to tokenized chat template; requalify and rerun all affected cells	Canonical families replace affected values; shared family unchanged, 24/96 reversals, 45/54 interactions; base-native floor persists (0.0–1.2% pair means)	Pairwise identity missed identical duplication; defective-BOS results remain instrument-conditional
Route configuration	Native prefixes and system placement	`add_generation_prompt=true`; preserve/omit empty system, or no-system/minimal fallback; thinking off for both Qwen checkpoints and SmolLM3	Qwen prefixes contain a closed empty reasoning block; unaffected BOS renderings byte-identical	Route includes template, system, and termination; prefix/revision/preamble hashes ship
Stop-contract differences	16 pairs: numeric vocabularies agree in 16, full contracts 11, maps 10, special-token inventories 9	Production-faithful audit supersedes all-automata-identical claim; union/intersection re-scoring for five differing pairs	No strict-score changes: union 120,000 rows; intersection 106,605 evaluable rows	13,395 need regeneration; scoring does not change generation-time candidates; package contrast, not weight-only
Buffer precision	Whole-model stage cast changes only RoPE `inv_freq`/ `original_inv_freq` buffers float32→bfloat16; parameters already bfloat16	Controlled cast/no-cast replay; identical revisions, files, inputs, masks, grammar, stop set, 17-token cap	Each condition reproduces its historical SFT/DPO outputs on 500/500 items; strict/first-integer correctness agree on 499/500; numerically distinct instruments	Original stage library versions unrecorded; first-loss and transplant conditions retained; transplant not rerun
Runtime receipts	Battery, ablations, matched controls, factorial, fresh grids: shared loader; stage: whole-model cast; lens/transplants and prompt/free audits: no cast	Code paths imply float32 buffers except stage bfloat16; current audits record per-load parameter/buffer fingerprints	Replay isolates buffer precision despite matched parameter dtypes; no evidence of an added shared-runner cast	Historical battery receipts bind no code commit: buffer condition remains "undetermined from receipts"; no per-cell hardware receipts