Everything in Moderation: Marginal Per-Domain Coverage Optima and Gap-Persistent Fixed-Budget Supervised Reweighting in Multi-Domain Mid-Training

Abstract

Mid-training's per-domain composition is usually set by availability, not design. We test whether downstream alignment repairs that on Qwen3-8B-Base, with a mid-training-only 4B arm that tests the findings' scale-dependence, across five rule-disjoint KOR-Bench domains: 30 allocations on the simplex, five seeds each. The internal mid-training token share leaves a between-domain gap structure alignment shifts but does not close. First, every domain's coverage marginal has an interior optimum: the moderate band (10–40%) is best for all five—group means, though the boundaries are not independent of the fitted peaks—with a family-wise quadratic-interiority permutation test at $P \approx 0.010$, the decline above it resolving in only one of the five. Coverage is collinear with repetition, and on the tested domain a share $\times$ pool-fraction experiment quadruples the epoch count while accuracy moves at most 1 pp, so repetition does not carry the optimum there. Second, the gaps survive: compensatory SFT bridges $0/240$ pairs at 5 pp, and a budget-feasible permutation null still puts it below the null ($P = 0.005$)—fewer gaps closed than chance—as does a uniform control. Third, zero coverage collapses a domain below its no-mid-training baseline (Counterfactual 86.0% $\rightarrow$ 45.6%), though not all (Cipher gains 8.8 pp); SFT removes the collapse, and two ladders attribute it to the domain's absence (8B; the 4B arm does not reproduce it). On the six withheld allocations, the mid-training ranking survives alignment ($r = +0.975$) and the fitted curves add no measurable increment ($r = +0.967$; both at $n = 6$), and a mixture baseline lands within 1 pp of θ^*.

1 Introduction

The multi-stage training paradigm (pre-training $\rightarrow$ mid-training $\rightarrow$ Supervised Fine-Tuning (SFT) $\rightarrow$ Reinforcement Learning (RL)) assumes each stage refines the last, but early design choices may create constraints later stages cannot undo. We test one: **per-domain data coverage at mid-training**, asking whether, in a fixed total-token mixture and a fixed downstream recipe, SFT or RL reduces coverage-associated differences. The alignment arm is a policy family, not a single recipe: the non-closure is a measured property of budget placement, not an artifact of one weak formula (Appendix H.3, H.2). One budget and one model family are the scope; whether alignment in general can repair coverage gaps is not settled here.

Prior work leaves the coverage-allocation dimension untested: studies of pretraining exposure and post-training behavior describe stage-level effects, and data-mixture methods optimize domain proportions against a loss objective, returning either a single allocation chosen once or, in the dynamic variants, a schedule—rather than allocating a fixed mid-training budget. None isolates *per-domain mid-training coverage* under a held-fixed downstream recipe (Appendix A).

We address this gap with a controlled study on Qwen3-8B-Base using five rule-disjoint domains from KOR-Bench (Ma et al., 2025) (§3.1). We construct 24 configurations spanning the five-domain simplex, so each per-domain curve is a mixture-level marginal rather than a dose (§5.2). We evaluate on KOR-Bench and three external benchmarks (CounterBench (Chen et al., 2026), ProofWriter (Tafjord et al., 2021), ZebraLogic (Lin et al., 2025)); three findings emerge (Figure 1; Sections 4.1–4.4): mid-training coverage leaves a between-domain gap structure in a fixed-budget multi-stage pipeline that the later stages we tested shift accuracy around but do not close.

Finding 1 [held-out: six withheld allocations close $0/60$ pairs]: The gaps survive the reweightings that keep the average gain; fully closing them forfeits that gain (§4.2, §3.4). Reweighting SFT data toward coverage-deficient domains raises accuracy in 116 of 120 cells yet bridges $0/240$ pairwise gaps at 5 pp and only $30/240$ at 10%; an equal-budget uniform control behaves almost identically. A permutation null holding the observed gain magnitudes fixed

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.

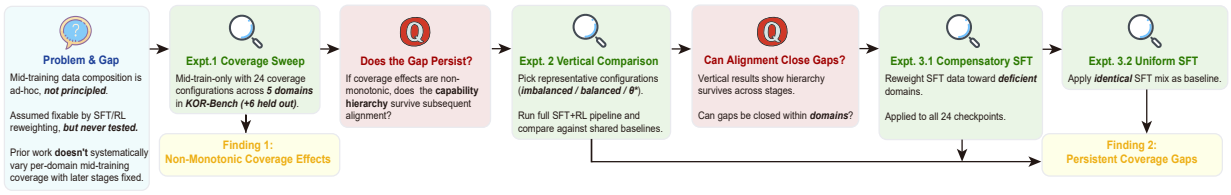

Figure 1: Experimental flow. The downstream recipe (mid-training→SFT→RL) is held fixed while the five-domain coverage mixture varies: 24 sweep configurations (plus six held-out allocations) establish per-domain interior optima (Finding 3, §4.1); three allocations carry the vertical comparison through the full pipeline (§4.2); compensatory vs. uniform SFT on all 24 checkpoints tests gap closure (Finding 1, §3.4).

bridges 13.8 ± 3.3 pairs at 5 pp ($P < 0.001$; Appendix H.2). Sharpening within the family closes gaps only by forfeiting that gain ($0/60 \rightarrow 12/60$ pairs while the mean gain falls $+4.34 \rightarrow +2.26$ pp; Appendix H.3). The short GSPO stage leaves that ordering intact (§5.1).

Finding 2 [single scale: 8B only, not reproduced at 4B]: Zero coverage selectively collapses a domain below its no-mid-training baseline, and the tested recipe removes the collapse (§4.2). Zero coverage collapses three of the four domains it is applied to, including the high-prior ones (Counterfactual 86.0%→45.6%; Operation 60.8%→33.2%; Logic 49.2%→22.1%), but not the lowest-prior of the four (Cipher 13.2%→22.0%, +8.8 pp), so the effect is not a uniform property of zero coverage. The tested recipe removes it (Table 8): one SFT pass puts Operation $+5.2 \pm 0.3$ pp above baseline, and Counterfactual within seed noise of zero (-0.8 ± 1.2 pp), positive after RL ($+1.2 \pm 3.4$ pp, not statistically resolved). The FineWeb-Edu-only control (§5.3) degrades all five domains, less on the two zeroed ones (Counterfactual -34.8 vs. -40.4 pp; Operation -15.2 vs. -27.6 pp), so the selectivity is not generic drift. Two ladder controls attribute the collapse to the domain's absence rather than to the others' dominance or to what replaces the dropped domain (Appendix J).

Finding 3 [curve shapes held out; peak locations and bands in-sample; the decline above the band resolves in one of five]: Every domain's own coverage marginal rises to a moderate band (§4.1). The moderate band (10–40%) is best for all five domains in-sample (Table 11), with a permutation test for quadratic interiority giving $P \approx 0.010$ (Appendix G.1); the band boundaries were chosen after the fitted peaks were read and the "all five" result is not robust to moving them, so the low-versus-moderate contrast rather than the exact band is the better-supported half. Fitted split-Gaussian curves at 8B place the peaks between 9.9% and 35.1%. Six withheld allocations reproduce the *curve shapes* out-of-sample, closely for four domains, only approximately for Operation, while fitting on the 24 sweep allocations alone predicts post-RL accuracy at $r = +0.967$ ($p = 0.002$; Appendix G.4). The fitted peak *locations* are what held-out data do not validate: a mid-training-only 4B replication reproduces the band ordering while the peaks shift with scale (Appendix C). The claim is per-domain marginal, not joint: the compositional surface's stationary point reads as a saddle in every domain—a provisional classification (§5.2, Appendix G.2).

2 Methodology

We fix the planned subsequent training stages (SFT, RL) across configurations so mid-training coverage is the intended intervention; our focus is not the recipe itself but how the mid-training data distribution shapes later behavior. The reasoning benchmarks are described in §3.1.

2.1 Training Stages

Mid-training performs standard continued causal language modeling on $\mathcal{D}_{\text{mid}}$ from base parameters θ_0, yielding checkpoint θ_{mid}. SFT optimizes conditional next-token likelihood over problem-response pairs (Wang et al., 2023); the θ_{mid} vs. θ_0 start points and the row pairing that makes each Table 1 comparison like-for-like are enumerated in Appendix H. After SFT we apply Group-based Sequence-level Policy Optimization (GSPO) (Zheng et al., 2025), a sequence-level group-based RL with a binary verifier reward; full equations are in Appendix D.1.

2.2 Data Construction

The training data are *newly synthesised instances* derived from the benchmark rule definitions, not reused benchmark items: split isolation (by instance ID for KOR-Bench; by depth for ProofWriter) and format/information deduplication guarantee that no training instance coincides with an evaluation item (Appendix F). The external component was generated, verified, and filtered through the same pipeline and contains no original benchmark items. The pipeline's seven numbered components (Figure 6, Appendix D.9) are symbolic for the $> 90\%$ of items the deterministic solvers cover; the remaining $< 10\%$ take the optional model branch, in which Claude Fable 5 supplies the answer and trace (Appendix B.2).

Mid-training controls token exposure rather than sample counts, at a fixed total budget (≈1.5B tokens, two epochs—about 0.19 tokens per parameter, well below the budgets we cite, so this is a short adaptation rather than full continued pretraining; Table 5) mixing an internal five-domain component with a fixed external component. Roughly a third of the external tokens (34.8%) come from the ProofWriter rule family, so we treat ProofWriter as same-family exposure rather than a zero-exposure transfer test (§4.4). The FineWeb-Edu-only baseline (Penedo et al., 2024) (§4.1) uses the same total budget with FineWeb-Edu as the entire corpus, replacing both the internal five-domain component and the fixed external component; SFT traces come from the same two sources as the mid-training traces—the deterministic solvers, or Claude Fable 5 on the $< 10\%$ of items the solvers do not cover (token-share summary: Appendix E).

2.3 Coverage Quantification

Domain coverage $\mathrm{Cov}(d)$ is the fraction of internal mid-training tokens allocated to domain d; the external component is fixed (Table 5, Appendix E), so the five shares sum to 100% and a coverage change is a pure reallocation. Each domain draws uniformly over its own 25 rule types, and each rule's pool is filed into six difficulty tiers (D0–D5) on a monotone problem-scale ladder (Appendix B.2); the sweep varies exposure, not the within-domain rule-type composition. Limits are treated where they bite: closed-composition coupling (§5.2, Appendix G.2) and unmatched unique-example diversity, repetition, data quality, and trace length (§3.3, Appendix K). We track two deltas: $\Delta_{\mathrm{mid}}(d) = \mathrm{Acc}_{\mathrm{midtrain+sft+rl}}(d) - \mathrm{Acc}_{\mathrm{sft+rl}}(d)$ and $\Delta_{\mathrm{rl}}(d) = \mathrm{Acc}_{\mathrm{midtrain+sft+rl}}(d) - \mathrm{Acc}_{\mathrm{midtrain+sft}}(d)$; evaluation details are in Appendix B.

3 Experimental Setup

3.1 Reasoning Benchmarks

Primary benchmark: KOR-Bench. We mainly use KOR-Bench (Ma et al., 2025), a structured benchmark for knowledge-orthogonal reasoning spanning five domains—ciphers, custom mathematical operations, formal logic, constraint puzzles, and counterfactual reasoning—each comprising 25 rule types with verifiable ground-truth answers, and our generated corpus files each rule's instances into six difficulty tiers (Appendix B.2).

The domains are rule-disjoint but not independent in the statistical or neural sense (Appendix A), so per-domain coverage is a controlled mixture coordinate rather than an independent causal dose (§5.2). KOR-Bench is our in-distribution primary metric and the three external benchmarks are consistency probes within one symbolic-rule setting (Appendices F, I).

External benchmarks. Three external benchmarks are evaluated on the same five-seed checkpoints under their native protocols: *ProofWriter* (Tafjord et al., 2021) (deductive reasoning; D5 exposed, D6–D9 held out by depth), *ZebraLogic* (Lin et al., 2025) (MC and Grid), and *CounterBench* (Chen et al., 2026) (zero direct KOR-Bench coverage). ProofWriter-family exposure is held fixed across configurations (34.8% of external tokens), while template- and rule-family-level overlap remains uncontrolled (Appendix I). Each benchmark's ranking association with the KOR-Bench average, with its exposure term, is reported over the nine fixed configurations (§4.4).

3.2 Model and Hyperparameters

All trainable configurations start from Qwen3-8B-Base without instruction tuning (Yang et al., 2025). The scale was chosen for measured post-training headroom (Appendix B.3); the 4B arm tests whether the pattern is scale- or

headroom-specific (Appendix C). Mid-training and SFT both use full fine-tuning under a 1:5 budget ratio that keeps mid-training dominant (Appendix H); all hyperparameters are in Table 5 (Appendices B, B.1). RL uses GSPO (Zheng et al., 2025) at a fixed 200-step budget (hyperparameters in Table 5; objective in Appendix D.1; algorithm choice in Appendix D.2). Its reward is a *binary outcome verifier*—1 if the final answer matches, 0 otherwise—so no dense per-token signal exists and only whole-answer outcomes can be reinforced (Appendix D.7). This leg is therefore a secondary observation rather than a headline claim; a pass@k probe separates its two candidate readings (§4.2; Appendix D.7).

Each coverage-sweep configuration is a mean $\pm$ 1 SD over five seeds on the held-out KOR-Bench test split, which carries 200 items per domain (Table 16); full-pipeline rows (Table 1) and external benchmarks are likewise five-seed means with cross-seed SDs, and the 95% bootstrap CIs are evaluation-sample estimates (Appendices B, D.4). Seeds are matched across pipeline rows: the same five seed indices at every stage. Run accounting (630 training runs; 12 further interior allocations budgeted separately) is in Appendix B.1.

3.3 Mid-Training Data Coverage Study

We train 24 controlled domain-coverage configurations (Table 16, Figure 2) spanning the five-domain simplex, measuring evaluation accuracy before any SFT/RL intervention. Six further allocations are withheld from the curve fit and carried through the complete mid-training→compensatory SFT→RL pipeline at five seeds each (Table 16, RL Evaluation rows), forming the out-of-sample arm (§4.1, Appendix H, G.4); the pool bookkeeping for the band comparison and the interiority permutation test—which allocations enter the pool, and that both tests are in-sample—is in Appendix G.1.

In Figure 2, coverage percentages reflect *internal* allocation only (the fixed external component contributes shared formal-deduction exposure, which is why Logic retains non-zero accuracy at 0% internal coverage; Appendix G). The sweep varies each domain's share at a constant token quota, and the difficulty composition is fixed by the tier construction, while answer length and trace length remain unbalanced across configurations (Appendix G).

3.4 Pairwise Gap-Closure Analysis

Compensatory SFT is applied to all 24 checkpoints under a fixed budget, $r_d \propto \max(0,\, 60\% - M_d)$ (Appendix H); we summarize gap closure over the $24 \times \binom{5}{2} = 240$ domain pairs before/after compensation (Appendix H.2). Because the configurations are purposively selected design points, the counts are not population estimates over random allocations, so we test them against a within-configuration permutation null holding the observed gain magnitudes fixed (2,000 draws; Appendix H.2).

3.5 Full Training Pipeline Comparison

Table 1 compares three shared no-mid-training baselines against the three mid-training stages. The row pairing that makes each comparison like-for-like is given in the caption of Table 1 and enumerated in Appendix H. The compensatory and uniform passes of Appendix H are not on the same scale as the SFT rows of Table 1, and the difference is the supervised token budget rather than the data: Table 1's SFT is the full alignment pass, trained toward the best attainable score, whereas those two diagnostic passes share one smaller fixed budget and differ only in how it is allocated across domains. The two families are therefore not compared numerically. Qwen3-8B-Instruct (Yang et al., 2025) is a non-comparable external reference only (Appendix D.6).

4 Results

4.1 The Impact of Mid-Training Data Coverage

Figure 2 plots each domain's accuracy against its mid-training proportion across the 24 normal configurations (§3.3); FineWeb-Edu and θ^* are excluded from the fit.

The fits only *locate* the optima; the moderation claim rests on the group means below.

The pattern is anticipated rather than surprising: an interior marginal under a closed budget is the generic prediction of the capacity-competition accounts we cite, so what we add is its measurement in post-alignment per-domain accuracy rather than the existence of the optimum (Appendix A).

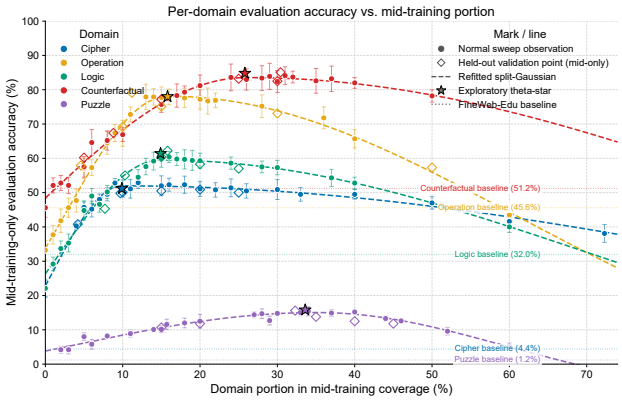

Figure 2: Per-domain mid-training-only accuracy vs. the domain's own mid-training portion. Filled points: 23 of the 24 normal configurations (Table 16; 23 shown, all 24 fitted). Hollow diamonds: six held-out allocations, withheld from the split-Gaussian fit and carried through the complete pipeline; they track the fitted curves at seed-noise level except Operation's low-coverage points, ≈ 6 pp above the fitted tail (Appendix G.4). Vertical bars: ± 1 SD across five seeds. Dashed lines: split-Gaussian fits; stars: exploratory θ^*; dotted lines: FineWeb-Edu baselines.

Group means by coverage band. Binning each domain's coverage into low ($< 10\%$), moderate (10–40%) and high ($> 40\%$), the moderate band is best for all five domains (Table 11). The band boundaries were chosen after the fitted peaks were read, and the "all five" result is boundary-sensitive (at *high*$>30\%$ three of five domains remain); the low-versus-moderate contrast survives every boundary we tested and is the half the moderation claim rests on. A whole-row permutation test finds all five domains concave with an interior vertex in only 0.98% of $20{,}000$ draws ($P \approx 0.010$; both tests are in-sample with respect to the trained pool, and the high band is thin, $n = 1$–5; Appendix G.1). Testing the moderate-versus-high contrast directly resolves the decline for one domain of five (Cipher, $+8.8$ pp, $[+0.5, +17.0]$) and for no other—two high bands are single allocations, the other two intervals span zero (Table 12).

Coverage is not repetition, and the sweep could not have shown that. At a fixed token budget a domain's share and its epoch count are one degree of freedom, not two, so no reparameterisation of this design separates them (Appendix J). A follow-up quarters a domain's pool at fixed processed tokens, moving Counterfactual by $+0.1$ pp (moderate share) and -1.0 pp (high share) against a 5.6–6.7 pp share contrast; the manipulation has the power to detect a repetition-driven optimum and does not (Table 28; Appendix J.2). On the domain tested the interior optimum is a coverage effect; the sweep's own repetition channel is bounded by this test, not closed.

Out-of-sample validation on held-out allocations. The six held-out allocations, withheld from the fit and carried through the full mid-training+SFT+RL pipeline, confirm the curves out-of-sample for the *curve shape* (peak locations remain unvalidated). Fitting on the 24 sweep configurations alone also predicts their *post-RL* accuracy ($r = +0.967$, $p = 0.002$; Appendix G.4), so a coverage choice made at mid-training remains visible after alignment. The fitted curves are not what carries that signal: the *realized* mid-training-only ranking predicts post-RL accuracy at $r = +0.975$ on the same six allocations, so at $n = 6$ the fits add no measurable increment (Appendix G.4), and the claim is the survival of the ranking, not a virtue of the curves. The relative-gain objective that selects θ^* ranks them only weakly ($r = +0.441$; Appendix G.4).

Fitted marginal optima. The fitted peaks of the split-Gaussian curves (Appendix G.3, Table 14) are Cipher 9.9% (seed-level bootstrap 95% interval $[7.4, 12.4]$), Operation 15.9% $[14.1, 17.5]$, Logic 15.0% $[12.7, 16.5]$, Counterfactual 26.0% $[22.2, 30.7]$ and Puzzle 35.1% $[31.9, 38.5]$. The five peaks sum to $\approx 102\%$ of the budget, which we do not read as a design rule (§5.2; Appendix G.1). From these marginals we derive an illustrative allocation θ^* (Eq. 4)—a descriptive probe read off the fitted curves rather than a validated allocation rule—whose *fitted* objective is $\approx +9.3\%$ relative gain over balanced; the re-trained θ^* gives $+4.36$ pp over SFT+RL (Table 1)—a model-selected candidate whose ordering is not established at this seed count (§4.2). The reported θ^* is the rounded fitted-peak allocation rather than strictly the argmax of Eq. 4, differing by at most 1.5 pp on any coordinate (Appendix G.3). A RegMix-style

predictor fit on this sweep and carried through the identical protocol lands within 1 pp of θ^*, and all three allocations—predictor, θ^*, balanced—are within 1 pp of one another (Appendix J.5). The 95%-of-peak intervals (Appendix G.3, Table 13) are function-level thresholds for this sweep, not confidence intervals.

4.2 Vertical Comparison: Mid-Training Contribution Across Stages

The selected full-pipeline comparisons are exploratory.

We carry three configurations through the identical full SFT+RL pipeline, restricted for compute reasons: **Expt. 1** (imbalanced), **Expt. 2** (balanced, 20% per domain), and **Expt. 3** (exploratory θ^*; §4.1); their mixture coordinates are the block headers of Table 1. They span a coverage contrast but are not representative of the sweep (Figure 3).

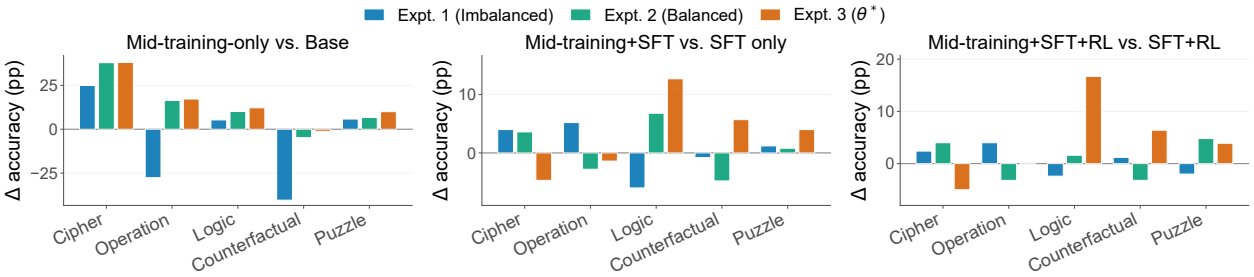

Figure 3: Pipeline-stage breakdown: per-domain accuracy gain (Δ, pp) across three stages (data from Table 1). **Left:** Mid-training-only vs. Base ($\Delta = \text{Acc}_{\text{mid}} - \text{Acc}_{\text{base}}$). **Middle:** Mid-training+SFT vs. SFT only. **Right:** Mid-training+SFT+RL vs. SFT+RL. Expt. 1 (imbalanced), Expt. 2 (balanced), Expt. 3 (θ^*). Note that the y-axis range differs across the three panels.

Config	Overall	Cipher	Operation	Logic	Counterfactual	Puzzle
No-mid-training baselines (shared across experiments)						
Base	43.0 ± 1.0	13.2 ± 1.9	60.8 ± 2.0	49.2 ± 1.7	86.0 ± 2.6	5.8 ± 0.7
SFT only	64.08 ± 1.8	66.0 ± 2.0	88.8 ± 2.7	59.2 ± 2.3	86.4 ± 3.0	20.0 ± 2.0
SFT+RL	65.28 ± 1.9	68.0 ± 2.3	88.8 ± 3.0	59.2 ± 2.2	87.6 ± 3.0	22.8 ± 1.5
Expt. 1 — Imbalanced: Cipher 72.3%, Operation 0%, Logic 12.0%, Counterfactual 0%, Puzzle 15.7%						
Mid-training only	36.60 ± 1.8	38.1 ± 2.6	33.2 ± 2.8	54.5 ± 2.2	45.6 ± 2.8	11.6 ± 1.5
Mid-training+SFT	64.80 ± 2.5	70.0 ± 2.2	94.0 ± 2.7	53.2 ± 1.9	85.6 ± 2.8	21.2 ± 1.4
Mid-training+SFT+RL	65.92 ± 1.5	70.4 ± 2.1	92.8 ± 3.0	56.8 ± 2.0	88.8 ± 3.5	20.8 ± 2.0
Expt. 2 — Balanced: 20% per domain (4/5 domains within fitted 95%-of-peak intervals)						
Mid-training only	56.26 ± 1.5	51.1 ± 2.2	77.2 ± 2.9	59.3 ± 2.2	81.2 ± 3.1	12.5 ± 2.0
Mid-training+SFT	64.80 ± 1.7	69.6 ± 2.1	86.0 ± 2.9	66.0 ± 2.1	81.6 ± 2.9	20.8 ± 1.9
Mid-training+SFT+RL	66.08 ± 2.3	72.0 ± 2.3	85.6 ± 3.2	60.8 ± 2.1	84.4 ± 3.3	27.6 ± 1.5
*Expt. 3 — Exploratory θ^*: Cipher 9.9%, Operation 15.8%, Logic 14.9%, Counterfactual 25.8%, Puzzle 33.6% (**all five** within fitted 95%-of-peak intervals). Re-trained full pipeline.*						
Mid-training only	58.24 ± 1.5	51.2 ± 2.1	78.0 ± 2.8	61.4 ± 2.1	84.8 ± 3.1	15.8 ± 1.9
Mid-training+SFT	67.34 ± 2.1	61.3 ± 2.2	87.4 ± 2.9	71.9 ± 2.3	92.1 ± 3.5	24.0 ± 1.9
Mid-training+SFT+RL	69.64 ± 2.6	63.0 ± 2.3	88.6 ± 2.5	75.9 ± 2.2	94.0 ± 3.3	26.7 ± 1.3

Table 1: KOR-Bench zero-shot accuracy (%). Base, SFT only, and SFT+RL are shared no-mid-training baselines. All rows of this table use five seeds and one fixed SFT+RL recipe on identical rule-type-balanced SFT data, so corresponding rows differ only in the inserted mid-training mixture; the separate compensatory and uniform passes of Appendix H use a smaller fixed supervised budget and are not on this scale (Appendix H, Table 5). The full 24-configuration sweep is in Appendix Table 16. Values are five-seed means $\pm$ 1 SD; post-RL differences among the overall scores are descriptive (§4.2), and the three blocks are not representative of the remaining trained allocations. The θ^* block is fit-selected, not independently evaluated (Appendix G.4). The Qwen3-8B-Instruct reference is in Appendix D.6.

Mid-training-only gains are largest at low Base prior in the two extremes, though gains are not monotonically ordered by Base prior across the three experiments (Figure 3, left; per-experiment per-domain gains in Table 1). θ^*

gives the largest mid-training contribution: $+4.36$ pp over the shared SFT+RL baseline, against $+0.80$ pp (balanced) and $+0.64$ pp (imbalanced)—$5.4\times$ the balanced contribution, and the only one above either row's cross-seed SD (± 2.6 and ± 1.9 pp). The five per-domain movements behind that $+4.36$ pp figure, which partially cancel, are in Appendix K.

Under the Welch procedure, Δ_{mid} (§2.3) is the strongest of the four overall comparisons: 69.64 ± 2.6 vs. 65.28 ± 1.9 on the θ^* row ($+4.36$ pp, nominally $p \approx 0.019$ before selection or multiplicity correction). Seeds are matched across pipeline rows (§3.2), so these unpaired statistics are conservative: the *aggregate* ordering does not survive a four-way control, while the per-domain movements are better supported (Appendix K). The ordering is already visible before RL (mid+SFT: 67.34 vs. 64.80 and 64.80; SFT-only 64.08).

At mid-training-only the θ^* (9.9%) and balanced (20%) allocations are indistinguishable on Cipher (51.2 ± 2.1 vs. 51.1 ± 2.2), as the fitted plateau predicts (Table 13); the 8.3 pp separation appears only after the identical SFT pass and persists after RL. Since the two rows differ in all five mixture coordinates, we read this as the mixture contrast producing a Cipher effect the mid-training-only marginal did not show, rather than as a relocation of Cipher's own optimum (the simplex confound of §5.2).

RL neither creates nor disturbs the ordering: GSPO adds a positive increment in all ten configurations we ran (Expts. 1–3 and the shared baseline, $+1.12$ to $+2.30$ pp; the six held-out allocations, mean $+1.19$ pp), and over the six held-out allocations it preserves the mid-training ranking ($r = +0.975$). We make no significance claim on the increment: it is modest. Its two candidate readings are not, however, inseparable. A pass@k probe from the same post-SFT policy reaches near-ceiling on the collapsed domains at $k = 64$ (Operation $41.0\% \rightarrow 98.2\%$; Appendix D.7), so the correct chains are already inside the policy's support and the shortfall is an optimization-budget effect rather than a competence limit on the verifier. The Qwen3-8B-Instruct reference is a scale reference, not like-for-like: our θ^* row (69.64) sits above every instruct reference we evaluated at or below 30B-A3B (Appendix D.6).

How much of the mid-training-only collapse survives alignment? The two largest collapses are *mid-training-only* and are reversed in sign by a single SFT pass in Expt. 1 (Table 8). The remaining per-domain movements are small: a residual deficit persists only on Logic and Puzzle (-2.4 and -2.0 pp), and Counterfactual turns positive only by the RL row, within its SD (Table 8). The interference effect (Finding 2) is thus checkpoint-level. After RL no domain retains more than $+4.0$ pp of an effect whose largest magnitude was 40.4 pp (Table 8).

4.3 Horizontal Comparison: Can Alignment Close Mid-Training Gaps?

We test whether reweighting SFT data toward under-covered domains can correct the imbalance installed at mid-training, comparing a compensatory pass with a uniform-SFT control at a fixed token budget.

Compensatory SFT. The fixed compensation policy of §3.4 raises absolute scores in 116 of 120 configuration–domain cells (mean $+4.32$ pp). Closure over the 240 pairs is similar: 30/240 under a 10% ratio threshold and 0/240 under a 5 pp threshold (against the 41 pairs closable in range, 0/41) (Appendix H.2, Figure 4). A permutation null that reallocates those same gains at random bridges far more: 13.8 ± 3.3 pairs at 5 pp and 77.9 ± 8.5 at 10%, against 0 and 30 for the policy ($P < 0.001$, one-sided; Appendix H.2). That margin depends on which reallocations the null admits: a budget-feasible restriction still leaves the observed count below the null, but at $P = 0.005$ and $P = 0.043$ rather than $P < 0.001$, and the claim does not hold under the strictest per-domain capacity restriction ($P = 0.65$), which is near-degenerate on this sweep (Appendix H.2, Table 17). The limit is placement within this policy family, not budget size: on the fitted gain curves, 148 of the 240 pairs (62%) can in principle be narrowed by at least 5 pp by *some* budget-feasible allocation of the same SFT tokens—a figure that requires evaluating the gain fits at $r = 100\%$, $3.3\times$ beyond the largest remedial ratio (30%) at which any curve was fitted, and that over-predicts the one extreme placement we measured—winner-take-all closes $12/60 = 20\%$—which forfeits roughly half the mean gain (Appendix H.3). The six held-out allocations (§4.1) bridge $0/60$ pairs at 5 pp.

Uniform SFT control. A uniform SFT pass (identical per-domain mix; three epochs) covered all 24 checkpoints (Eval. (Uniform) rows in Table 16; Appendix H). The two passes are indistinguishable on the aggregates we compare: mean gain $+4.20$ vs. $+4.32$ pp and pairwise closure alike. Per domain, compensation's only material advantage is Counterfactual ($+6.37$ vs. $+4.49$ pp), with the other four domains agreeing within 0.63 pp and the $120/120$ vs. $116/120$ cell-count difference reflecting the compensation rule's zero allocations rather than a quality difference (Appendix H, Table 19). Whether the SFT budget is spent uniformly or by coverage is therefore not what blocks closure: the deficits survive both alike.

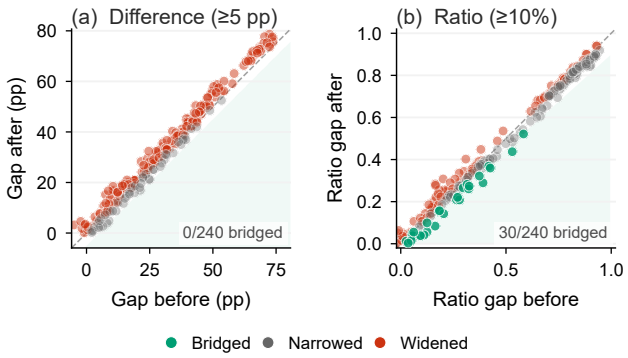

Figure 4: Pairwise gap closure after compensatory SFT across the 24 coverage-sweep checkpoints. **(a)** Difference metric (pp), threshold ≥ 5 pp closure. **(b)** Ratio metric, threshold $\geq 10\%$ relative closure. Points below the diagonal indicate gap reduction: $0/240$ pairs bridged under the difference metric, $30/240$ under the ratio metric. Because the 240 pairs share the 24 configurations, we compare both counts to a permutation null that reallocates the observed compensatory gains with the arithmetic budget cap held fixed: the observed bridged-pair counts lie *below* the null at both metrics ($P < 0.001$, one-sided; Appendix H.2).

Sharpening the policy does close gaps, at a measured price. The γ=1 rule belongs to the family $r_d(\gamma) \propto [\max(0, T - M_d)]^\gamma$ ($T = 60\%$), which we ran over $\gamma \in \{1, 2, 4, \infty\}$ on the six configurations with the most closable-but-unclosed pairs (Appendix H.3). Closure rises monotonically from $0/60$ at γ=1 to $12/60$ at winner-take-all, but average accuracy falls from $+4.34$ to $+2.26$ pp (Table 18). That trade-off is expected of a worst-case-oriented reweighting (Sagawa et al., 2020); Appendix H.3 measures it.

4.4 External Consistency Checks: Held-Out-Depth, Same-Family Exposure, and a Zero-Exposure Column

The gap pattern in §4.2 may reflect shared task structure, so we add three further benchmarks—ProofWriter (Tafjord et al., 2021), ZebraLogic (Lin et al., 2025), and CounterBench (Chen et al., 2026). Their rank associations with the KOR-Bench average are $\rho = +0.34$ (ProofWriter), $+0.53$ (ZebraLogic Grid), $+0.67$ (CounterBench) and -0.12 (ZebraLogic MC). These probe consistency within one symbolic-rule setting, not generalization to a held-out reasoning domain, and each column carries a different exposure or coverage caveat; per-depth, per-house, per-type and exposure results, including CounterBench's per-stratum deltas (Appendix Figure 9), are in Appendix I.

5 Discussion

The two findings use different reference frames—Finding 1 compares a configuration's own domains, Finding 2 compares each domain with its no-mid-training baseline—so a domain can move by tens of points under SFT while the between-domain spread Finding 1 tracks does not. With that fixed: coverage-induced gaps are robust to the *placement* of the alignment budget: fixed-budget compensatory and equal-budget uniform SFT passes leave them essentially intact. The supervised *dose* is never varied across arms (Appendix J). A permutation null shows the gains close fewer gaps than a random reallocation of the same magnitudes—below chance rather than merely neutral, though the margin narrows under a budget-feasible null—and the sharpening sweep shows that closing them as far as the family can costs roughly half the mean gain, while partial closure is cheaper (Appendix H.2, H.3). This does not license the general claim that alignment cannot repair coverage choices (Appendix D.7, K). Checkpoint-level negative transfer is selective—in Expt. 1 the zeroed domains collapse while the other three *gain*—whereas drift is global: the FineWeb-Edu-only control falls below the untouched Base model on all five domains (§5.3).

What would disconfirm the headline claim. The findings are not equally exposed: Finding 2 is the best-controlled rather than the most fragile, but also the least replicated—the mid-training-only 4B arm does not reproduce it, and a

second scale whose base prior on the probed domains is not near ceiling would disconfirm it. What would disconfirm Finding 1, and what Finding 3 does and does not rest on, are in Appendix K, together with Table 33.

5.1 Why the Tested Alignment Passes Did Not Close the Gaps

The non-closure is measured, not explained. Accounts from budget competition to optimization mismatch remain compatible with it, but the data bound the effect rather than account for it: we measure neither representations, gradient conflict, nor parameter overlap, so a capability-envelope explanation stays a hypothesis.

5.2 The Simplex Confound

Coverage is zero-sum: the five shares sum to 100%, and the unvarying external corpus contributes shared exposure, so each per-domain curve is a *mixture-level marginal*, not a dose-response. A compositional reanalysis returns a *saddle in all five domains* rather than a joint optimum—provisional, since the negative curvature is shallow and quadratic misspecification is untested (Appendix G.2, B.4). Finding 3's optima are therefore per-domain marginals.

5.3 Two Regimes of Degradation: Selective Coverage Collapse and Global Drift from the FineWeb-Edu Control

The FineWeb-Edu-only control (Table 16) degrades all five domains, but falls short of the zero-coverage collapse on both domains Expt. 1 zeroes (-34.8 vs. -40.4 pp on Counterfactual; -15.2 vs. -27.6 pp on Operation; Table 8), so generic drift does not account for the selective component—though the two arms differ in corpus composition as well as coverage, making this a comparison, not a bound. It is also the weakest model on three of the four external columns (Table 21).

6 Conclusion

Three findings hold here. First, coverage-induced domain gaps survive the tested post-training procedures: neither the compensatory policy nor its equal-budget control closes them ($0/240$ bridged pairs at 5 pp, against the 41 pairs closable within the fitted range; $0/60$ on the six held-out allocations; Appendix H.2), and a permutation null shows the gains close fewer gaps than a random reallocation of the same magnitudes—below the null under every restriction except the strictest per-domain capacity one, which is near-degenerate on this sweep (Table 17). Second, zero mid-training coverage collapses mid-training-only accuracy for three of the four domains it is applied to, including the high-prior ones, but not the lowest-prior Cipher ($+8.8$ pp); two ladders attribute the collapse to the domain's absence rather than to the others' dominance or to what replaces it (Appendix J), and the tested recipe reverses it on the two domains Expt. 1 zeroes (Table 8) at the cost of a -6.0 ± 0.8 pp deficit on Logic, whose coverage was 12%; the mid-training-only 4B arm does not reproduce the collapse (Appendix K). Third, every domain has an interior coverage optimum: the moderate band is best for all five under the reported boundaries (Appendix G.1), the fitted peaks lie between 9.9% and 35.1%, and a share$\times$pool-fraction experiment rules out repetition (Table 28). These optima are per-domain marginals, not a jointly achievable mixture (§5.2, Appendix G.2). Coverage should thus be an audited design decision at this budget and scale, decidable before any alignment run: on the six withheld allocations the mid-training ranking survives SFT and RL ($r = +0.975$), and the fitted curves do not improve on that ranking, though θ^* itself is unvalidated out-of-sample (Appendix G.4). Gap closure is reachable in principle under some budget-feasible allocation of the same tokens, but not at $\gamma{=}1$, and sharpening within the family buys it only by forfeiting mean gain (Appendix H.3).

6.1 Limitations

Four limits bound these results, each stated at the point of its result. *Scale*: one model family, with a mid-training-only 4B replication. *Benchmark scope*: KOR-Bench is in-distribution; the external benchmarks are consistency checks, not transfer tests. *Design*: coverage is a closed composition, so each curve is a mixture-level marginal and no joint optimum is identified (§5.2). *Inference*: the fitted peaks, the 95%-of-peak intervals and θ^* are read off the sweep that produced them and are not held-out validated, and the alignment arm is one policy family at one supervised budget. Every limitation is collected with its evidence tier in Table 33 and Appendix K.

References

M. Abdin et al. Phi-4 technical report. arXiv:2412.08905, 2024.

Anthropic. Claude Fable 5 model card. Anthropic, 2025.

J. Aitchison. The statistical analysis of compositional data. Chapman and Hall, London, 1986.

Z. Azerbayev et al. Llemma: An open language model for mathematics. In *ICLR*, 2024.

Z. Cai et al. InternLM2 technical report. arXiv:2403.17297, 2024.

J. Chen et al. Unlock the correlation between SFT and RL in training code LLMs. arXiv:2406.10305, 2024.

Z. Chen et al. Self-play fine-tuning converts weak language models to strong language models. In *ICLR*, 2024.

J. Chen et al. Step-wise adaptive integration of SFT and RL for task-specific LLMs. arXiv:2505.13026, 2025.

S. Chen, X. Ouyang, M. A. L. Pearce, T. Hartvigsen, and J. R. Schwarz. ADMIRE-BayesOpt: Accelerated data mixture re-weighting for language models with Bayesian optimization. In *TMLR*, 2025.

Y. Chen, V. K. Singh, J. Ma, and R. Tang. CounterBench: Evaluating and improving counterfactual reasoning in large language models. In *AAAI*, 2026.

T. Chu, Y. Zhai, J. Yang, S. Tong, S. Xie, D. Schuurmans, Q. V. Le, S. Levine, and Y. Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. arXiv:2501.17161, 2025.

P. Clark et al. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv:1803.05457, 2018.

K. Cobbe et al. Training verifiers to solve math word problems. arXiv:2110.14168, 2021.

R. Dai and S. Zheng. Explaining data mixing scaling laws. In *ICML*, 2026.

T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024.

Y. Deng et al. Supervised RL: From expert trajectories to step-wise reasoning. arXiv:2510.25992, 2025.

A. Dubey et al. The Llama 3 herd of models. arXiv:2407.21783, 2024.

S. Fan, M. Pagliardini, and M. Jaggi. DoGE: Domain reweighting with generalization estimation. In *ICML*, 2024.

R. M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.

L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The Pile: An 800GB dataset of diverse text for language modeling. arXiv:2101.00027, 2021.

D. Guo et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via RL. arXiv:2501.12948, 2025.

S. Gururangan, A. Marasović, S. Swayamditta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, 2020.

S. Han et al. FOLIO: Natural language reasoning with first-order logic. In *EMNLP*, 2024.

A. Havrilla et al. Teaching large language models to reason with RL. arXiv:2403.04642, 2024.

D. Hendrycks et al. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS*, 2021.

Y. Huang et al. LoRA-PAR: A flexible dual-system LoRA partitioning approach. In *EMNLP*, 2025.

J. Huang et al. ReMiT: RL-guided mid-training for iterative LLM evolution. arXiv:2602.03075, 2026.

A. Ibrahim, B. Thérien, K. Gupta, M. L. Richter, Q. Anthony, T. Lesort, E. Belilovsky, and I. Rish. Simple and scalable strategies to continually pre-train large language models. arXiv:2403.08763, 2024.

Y. Ishibashi et al. Mining hidden thoughts from texts: Evaluating continual pretraining with synthetic data for LLM reasoning. arXiv:2505.10182, 2025.

F. Kang, Y. Sun, B. Wen, S. Chen, D. Song, R. Mahmood, and R. Jia. AutoScale: Scale-aware data mixing for pre-training LLMs. In *COLM*, 2025.

Z. Ke et al. A survey of frontiers in LLM reasoning. In *TMLR*, 2025.

N. Kim and T. Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *EMNLP*, 2020.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

J. Li, X. Gu, and J. Zhang. Capacity-aware mixture law enables efficient LLM data optimization. arXiv:2603.08022, 2026.

B. Y. Lin, R. Le Bras, K. Richardson, A. Sabharwal, R. Poovendran, P. Clark, and Y. Choi. ZebraLogic: On the scaling limits of LLMs for logical reasoning. In *ICML*, 2025.

J. Liu, L. Cui, H. Liu, D. Huang, Y. Wang, and Y. Zhang. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *IJCAI*, 2020.

Q. Liu et al. RegMix: Data mixture as regression for language model pre-training. In *ICLR*, 2025.
Z. Lu et al. MathCoder2: Better math reasoning from continued pretraining on model-translated mathematical code. arXiv:2410.08196, 2024.
Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv:2308.08747, 2023.
H. Luo et al. WizardMath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. arXiv:2308.09583, 2023.
K. Ma, X. Du, Y. Wang, H. Zhang, Z. Wen, X. Qu, J. Yang, J. Liu, M. Liu, X. Yue, W. Huang, and G. Zhang. KOR-Bench: Benchmarking language models on knowledge-orthogonal reasoning tasks. In *ICLR*, 2025.
K. Matsutani. RL squeezes, SFT expands: A comparative study of reasoning LLMs. arXiv:2509.21128, 2025.
N. Muennighoff et al. Scaling data-constrained language models. In *NeurIPS*, 2023.
L. Ouyang et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.
G. Penedo, H. Kydlíček, L. B. allal, A. Lozhkov, M. Mitchell, C. Raffel, L. V. Werra, and T. Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. arXiv:2406.17557, 2024.
Y. Qin et al. daVinci-LLM: Towards the science of pretraining. arXiv:2603.27164, 2026.
R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.
S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC*, 2020.
Q. Ren et al. Rethinking generalization in reasoning SFT: A conditional analysis on optimization, data, and model capability. arXiv:2604.06628, 2026.
L. Ruis et al. Procedural knowledge in pretraining drives reasoning in large language models. arXiv:2411.12580, 2024.
S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.
S. Seto, P. Ablin, A. Filippova, J. Ye, L. Béthune, A. Katharopoulos, and D. Grangier. Optimal splitting of language models from mixtures to specialized domains. In *ICLR Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2026.
Z. Shao et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300, 2024.
Z. Shen et al. SlimPajama-DC: Understanding data combinations for LLM training. arXiv:2309.10818, 2023.
M. Shukor, L. Béthune, D. Busbridge, D. Grangier, E. Fini, A. El-Nouby, and P. Ablin. Scaling laws for optimal data mixtures. In *NeurIPS*, 2025.
K. Sinha, S. Sodhani, J. Dong, J. Pineau, and W. L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *EMNLP-IJCNLP*, 2019.
L. Soldaini et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *ACL*, 2024.
M. Suzgun et al. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of ACL*, 2023.
O. Tafjord, B. D. Mishra, and P. Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *ACL-IJCNLP*, 2021.
Team KIMI. Kimi k1.5: Scaling reinforcement learning with LLMs. arXiv:2501.12599, 2025.
S. Toshniwal et al. OpenMathInstruct-1: A 1.8 million math instruction tuning dataset. In *NeurIPS*, 2024.
C. Tu et al. A survey on LLM mid-training. arXiv:2510.23081, 2025.
Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In *ACL*, 2023.
Z. Wang et al. RLSR: Reinforcement learning with supervised reward outperforms SFT in instruction following. arXiv:2510.14200, 2025.
J. Wang, C. Tian, K. Chen, Z. Liu, J. Mao, W. X. Zhao, Z. Zhang, and J. Zhou. MergeMix: Optimizing mid-training data mixtures via learnable model merging. arXiv:2601.17858, 2026.
J. Wei et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.
Xiaomi LLM-Core. MiMo: Unlocking the reasoning potential of language model. arXiv:2505.07608, 2025.
S. M. Xie, H. Pham, X. Dong, N. Du, H. Liu, Y. Lu, P. Liang, Q. V. Le, T. Ma, and A. W. Yu. DoReMi: Optimizing data mixtures speeds up language model pretraining. In *NeurIPS*, 2023.

S. M. Xie, S. Santurkar, T. Ma, and P. Liang. Data selection for language models via importance resampling. In *NeurIPS*, 2023.
S. Yadlowsky, L. Doshi, and N. Tripuraneni. Pretraining data mixtures enable narrow model selection capabilities in transformer models. arXiv:2311.00871, 2023.
C. Yang et al. The fine line: Navigating LLM pretraining with down-streaming capability analysis. arXiv:2404.01204, 2024.
A. Yang et al. Qwen3 technical report. arXiv:2505.09388, 2025.
J. Ye et al. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *ICLR*, 2025.
T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020.
L. Yu et al. MetaMath: Bootstrap your own mathematical questions for large language models. arXiv:2309.12284, 2023.
C. Zhang, G. Neubig, and X. Yue. On the interplay of pre-training, mid-training, and RL on reasoning language models. arXiv:2512.07783, 2025.
R. Zhao, A. Meterez, S. Kakade, C. Pehlevan, S. Jelassi, and E. Malach. Echo chamber: RL post-training amplifies behaviors learned in pretraining. In *COLM*, 2025.
K. Zhao, Z. Miao, A. Aizawa, and Y. Tsuruoka. RegMix-D: Dynamic data mixing via proxy training trajectories. arXiv:2606.18663, 2026.
C. Zheng et al. Group sequence policy optimization. arXiv:2507.18071, 2025.
C. Zhou et al. LIMA: Less is more for alignment. In *NeurIPS*, 2023.
R. Zhou et al. RuleArena: A benchmark for rule-guided reasoning with LLMs in real-world scenarios. In *ACL*, 2025.

A Related Work

A.1 Multi-Stage Training for LLM Reasoning

Modern reasoning models are typically trained through a sequence of continued pretraining, SFT, and RL. The foundational Reinforcement Learning from Human Feedback (RLHF) framework (Ouyang et al., 2022) established the standard three-stage recipe of pretraining, supervised fine-tuning on demonstrations, and RLHF, which most subsequent systems adopt. Guo et al. (2025) show that RL can elicit strong reasoning behavior when applied to a capable base model, while their full DeepSeek-R1 pipeline still relies on cold-start SFT and additional supervised data. Shao et al. (2024) demonstrate that domain-specific continued pretraining can substantially improve mathematical reasoning before later post-training. Kimi k1.5 further scales reinforcement learning in a multi-stage setting (Team KIMI, 2025). These works establish the effectiveness of stage-wise training, but they usually report aggregate benchmark scores and do not isolate how mid-training data imbalance shapes domain-level outcomes.

Broader model reports also reinforce the importance of training stages and data coverages. Llama 3 emphasizes the role of large-scale pretraining and post-training recipes (Dubey et al., 2024). InternLM2 (Cai et al., 2024) explicitly incorporates a progressive training strategy with carefully curated mid-training data for reasoning, making domain allocation a first-class engineering concern. Phi-4 highlights the importance of high-quality synthetic and curated data for reasoning-oriented models (Abdin et al., 2024). Our work differs from these system reports by treating mid-training data composition itself as the object of study.

A.2 The Role of Mid-Training

Mid-training is increasingly recognized as a stage with its own objectives and failure modes. Tu et al. (2025) survey recent work and frame mid-training as a bridge between broad pretraining and downstream alignment. Zhang et al. (2025) provide controlled evidence that earlier training exposure determines how much later RL can generalize. Their finding motivates our central question, but their setting is deliberately controlled and synthetic. We instead examine a naturally imbalanced multi-domain corpus.

Several studies suggest that continued pretraining can inject useful reasoning structure. Ruis et al. (2024) argue that procedural knowledge in pretraining data drives reasoning performance. Ishibashi et al. (2025) study continual pretraining with synthetic data for reasoning. Yu et al. (2023) show that bootstrapped mathematical data can improve

math reasoning. In a related direction, Lu et al. (2024) use model-translated mathematical code for continued pretraining. The clearest single-domain exemplar of this approach is Llemma (Azerbayev et al., 2024), which continues pretraining Code Llama on the 55B-token Proof-Pile-2 corpus and achieves substantial gains on mathematical benchmarks; this work establishes that domain-specific continued pretraining on curated reasoning corpora is a reliable path to capability specialization. Our work extends this line to a multi-domain setting where allocation across domains is the primary planned coordinate.

The over-coverage half of the dose-response effect may be related to interference in multi-task optimisation (French, 1999; Kirkpatrick et al., 2017; Yu et al., 2020). In sequential continual learning, training intensively on a new task causes gradient updates to overwrite the parameter regions encoding previously learned tasks—a phenomenon termed catastrophic forgetting (French, 1999; Luo et al., 2023a). Scaling and data-curation strategies for continual pre-training have been studied as mitigation levers (Ibrahim et al., 2024), but they address sequential adaptation rather than allocation among simultaneous task domains. Our setting differs structurally: all domains are trained simultaneously rather than sequentially. However, imbalanced coverage may produce an analogous gradient-dominance effect: the high-coverage domain contributes the majority of the gradient signal, pushing shared representations toward its objectives and reducing the signal from low-coverage domains. We do not directly measure gradient conflict, parameter overlap, or Elastic Weight Consolidation (EWC)-style mitigation, so this explanation should be read as a hypothesis rather than a demonstrated mechanism.

The KOR-Bench domains have *semantic rule-disjointness*: each domain defines a self-contained symbolic system whose rules do not rely on concepts or operations from other domains. This reduces one source of semantic-transfer confounding, but it does not make domains independent in the statistical or neural sense. Domains still share language, tokenization, output formats, model parameters, and potentially general reasoning procedures. We therefore treat the proportions as a controlled mixture coordinate, not as an independent causal dose. The present design does not separately identify coverage from unique-example diversity, repetition, sequence length, or template- and rule-family-level overlap. The external component is identical in composition and token count across the sweep configurations, so it enters as a shared level offset rather than a per-configuration difference (§4.4, §5.2); the FineWeb-Edu-only arm is the one arm in which corpus composition changes, and it is reported as a drift control rather than a coverage contrast (§5.3).

The quality and format of mid-training data matter as much as the quantity. Wei et al. (2022) demonstrate that chain-of-thought reasoning traces—step-by-step explanations provided as training signal—substantially improve multi-step reasoning in large models; our pipeline embeds analogous structured traces in every generated instance. Luo et al. (2023b) show that instruction evolution (Evol-Instruct) applied to mathematical seed problems produces progressively harder variants that boost fine-tuned model performance, demonstrating that data quality amplification can substitute for raw volume. Toshniwal et al. (2024) scale this approach with OpenMathInstruct-1, a 1.8M-problem corpus synthesized by Mixtral-8x7B targeting the MATH benchmark (Hendrycks et al., 2021), which has become a canonical measure of mathematical reasoning capability. These studies collectively show that structured, verified, trace-annotated problems are the fundamental currency of mid-training effectiveness—a principle our multi-domain study extends to five heterogeneous reasoning types.

Data composition is another important thread. Yang et al. (2024) analyze how pretraining choices affect downstream capabilities. Qin et al. (2026) argue that pretraining establishes capability ceilings that later stages may not easily exceed. Xiaomi's MiMo report also emphasizes unlocking reasoning potential through carefully designed training data (Xiaomi LLM-Core, 2025). Our study adds a domain-level view: the question is not only whether mid-training helps overall, but which domains benefit and which domains are left behind.

A.3 Data Mixture and Coverage Optimization

The question of how to allocate data across domains is well-studied in the pretraining literature. Gururangan et al. (2020) establish that domain-adaptive continued pretraining on even modest domain-specific data substantially improves downstream task performance, and that the benefit scales with the mismatch between the general-web pretraining distribution and the target domain. Xie et al. (2023a) propose DoReMi, a Group Distributionally Robust Optimization (DRO)-based method that adapts domain weights during proxy training to equalize per-domain excess loss and then trains the target model with those weights held fixed; their framework directly addresses the uniform-vs.-non-uniform weighting question that our sweep operationalizes empirically. Fan et al. (2024) extend this line with DoGE, which reweights domains by their estimated contribution to a generalization objective via a small proxy model; DoGE and our θ^* both produce a single recommended allocation, but from opposite directions—DoGE optimizes a

gradient-based generalization estimate without training at the target allocation, whereas θ^* is read off measured per-domain accuracy curves. A controlled comparison on the same five-domain corpus would be informative and is not attempted here.

At the corpus level, large-scale composition ablations in the Pile (Gao et al., 2021) and Dolma (Soldaini et al., 2024) demonstrate that data-source mix is among the strongest predictors of benchmark performance. Complementary work addresses *which* data to include rather than how to weight each domain. Xie et al. (2023b) propose Data Selection via Importance Resampling (DSIR), selecting pretraining documents whose n-gram distributions match a target domain, effectively enabling soft domain filtering without hard corpus boundaries. Shen et al. (2023) construct SlimPajama-DC by rebalancing and deduplicating the RedPajama corpus, showing that domain rebalancing at the corpus level yields consistent downstream improvements. Muennighoff et al. (2023) examine the complementary question of data repetition: when a domain corpus is small, training for multiple epochs can partially substitute for additional data, though with diminishing returns. Note the direction in which this bound binds: it bites at the *high* end of our sweep, where a domain's token allocation is largest relative to its own pool and the pool is most likely to be exhausted, whereas a low share draws few tokens and does not approach it. That is the opposite of the intuition the low-coverage regime invites, and it is one reason a decline above the moderate band is not evidence for a capacity limit on its own. More fine-grained theoretical treatments have followed: Ye et al. (2025) establish quantitative scaling laws that predict language modeling loss as a function of domain mixture proportions, and Liu et al. (2025) propose RegMix, which trains 512 small proxy models across diverse domain combinations and fits a regression predictor to identify the optimal large-scale mixture—a principled complement to our empirical sweep. Both return a single recommended allocation rather than a response curve, and neither is dynamic: DoReMi fits group-DRO weights on a proxy model and then trains with those weights fixed, and RegMix fits a regression predictor over hundreds of small proxy runs and trains once on the mixture that predictor ranks highest. Our contribution is a domain-resolved diagnostic of how finite-budget mixture configurations are associated with downstream accuracy in this controlled setting, not a general dose-response law. The distance from our object is one of target rather than of optimization machinery: these methods select against a proxy model's loss, whereas the quantity we report is per-domain downstream accuracy measured after the alignment stages. More recent mixture-optimization methods such as MergeMix (Wang et al., 2026) propose using model merging weights as a low-cost proxy for domain performance—a different optimization strategy. A controlled comparison of these methods—and of the more recent predictors discussed next—in the present five-domain setup is future work: the current sweep does not test their recommended allocations on this corpus, nor does it establish that the moderate band carries beyond the trained pool (the band comparison and the interiority test are in-sample with respect to that pool; Appendix G.1).

Several recent methods pursue the same allocation question with cheaper or more adaptive machinery. Shukor et al. (2025) fit a scaling law that predicts loss jointly in model size, token budget, and the domain-weight vector, then minimize it over the simplex to obtain budget-optimal weights; Kang et al. (2025) fit a parametric loss model at small token budgets and extrapolate the resulting compute-optimal compositions to larger scales; Li et al. (2026) introduce an explicit capacity budget, deriving a mixture law in which model capacity couples nonlinearly with the mixture; and Dai and Zheng (2026) give a theoretical account of two mechanisms—capacity competition, in which finite model capacity couples domain losses globally, and noise reduction, which shifts optimal weights toward harder domains. Two of these extend the search rather than the law: Zhao et al. (2026) generalize RegMix from a single optimal allocation to a time-varying schedule by exploiting the full loss trajectories that proxy runs produce, and Chen et al. (2025b) recast mixture selection as black-box hyperparameter optimization and accelerate it with multi-fidelity Bayesian optimization. What separates these methods from the present study is the object being optimized rather than the optimization method: they search for or predict a mixture against an aggregate objective—a language-modeling loss, or in ADMIRE-BayesOpt's instruction-tuning setting an aggregate validation score over target benchmarks—and return a single recommended allocation. Our sweep instead spans low, moderate, and high coverage for every domain so that each domain's response curve is identifiable, and its product is the shape of that curve rather than a recommended mixture. The two families differ in what they leave available, and we state the difference rather than a blanket claim. A *search* procedure concentrates its evaluations near promising mixtures, so per-domain response curves—and the 95%-of-peak intervals read off them, which are descriptive function-level thresholds for this sweep rather than confidence intervals—are neither reported nor identified by its design. A *fitted mixture law* of the Shukor et al. or AutoScale form is a function of the whole weight vector, so its per-domain marginals are in principle available; what it does not supply is our object, because it is fit to proxy-model loss rather than to post-alignment per-domain accuracy, and at pretraining rather than mid-training scale.

Two further strands bound how far the present result should be read. Seto et al. (2026) study how to allocate

compute between a shared pretraining phase and per-domain specialization; because our design keeps one shared model throughout, specialization is a design alternative we do not test, and it is one reason the gap-closure result should not be read as a statement about specialized models. Yadlowsky et al. (2023) supply an earlier controlled precedent for the premise that mixture composition can induce selective behavior, reporting that transformers trained on particular pretraining mixtures acquire narrow, in-context model-selection capabilities rather than broadly transferable ones.

A.4 SFT and RL: Stage Combination and Refinement

SFT and RL are often treated as mechanisms for alignment and refinement rather than as sources of entirely new capabilities. Havrilla et al. (2024) observe that RL can struggle to explore far beyond solutions already reachable by the supervised model. Chen et al. (2024b) show that self-play fine-tuning can improve weak models but also converges without continually expanding the data distribution. These findings are consistent with the idea that later stages inherit constraints from earlier stages.

Recent work further studies how SFT and RL should be combined. Wang et al. (2025) compare reinforcement learning and supervised training in instruction following. Chen et al. (2025a) propose step-wise adaptive integration of SFT and RL. Deng et al. (2025) connect expert trajectories with step-wise reinforcement learning. Ren et al. (2026) revisit generalization in reasoning SFT, while Matsutani (2025) contrasts the effects of SFT and RL on reasoning models. Zhao et al. (2025) show that RL post-training amplifies behaviours the pretraining distribution already contained, and Chu et al. (2025) that SFT primarily teaches format while RL generalizes from capabilities the model already has; both describe stage-level effects rather than a per-domain allocation. Two foundational contributions bound the effectiveness of post-pretraining stages. LIMA (Zhou et al., 2023) demonstrates that as few as 1,000 carefully curated demonstrations can match much larger SFT corpora in instruction-following quality, implying that the capability headroom pre-established by earlier training stages is the primary bottleneck rather than fine-tuning volume. Direct Preference Optimization (DPO) (Rafailov et al., 2023) reformulates preference alignment as contrastive supervised learning, removing the reward model and substantially lowering the engineering complexity of the RL stage; the wider adoption of DPO has made controlled earlier-stage ablations such as ours easier to reproduce. Our experiments address both directions of this question. First, later stages do not reliably compensate for domain imbalance introduced at mid-training: even when SFT data is augmented toward coverage-deficient domains, difference-based pairwise gaps fail to close beyond the 5 pp repair threshold (0/240 pairs under a 5 pp difference metric), and ratio-based gaps show only limited average closure (30/240 under a 10% ratio). An equal-budget uniform-SFT control behaves almost identically (0/240 and 32/240 under the same metrics; Appendix H), so the non-closure is not an artifact of the compensation formula. Per-domain fits of the gain against the remedial allocation ($\Delta \approx a \log(1 + b\, r) + c$) rise within each domain yet saturate at domain-specific ceilings, and pooling across domains explains little of the cell-level variance ($R^2 \approx 0.31$; Figure 7, Table 20): redirecting SFT data does not reset the mid-training starting point. Second, the fitted-curve allocation θ^* carries the largest mid-training contribution of the three configurations trained end to end: mid-training adds $+4.36$ pp over no-mid-training SFT+RL in Expt. 3, against $+0.8$ pp under balanced coverage and $+0.6$ pp under imbalanced coverage. Band membership does not account for this gap—the balanced 20% design point is four of five domains in-band, one fewer than θ^*'s five of five (Table 1; §4.1)—so we read the gap between them as the mixture contrast redistributing accuracy across domains (§4.2), not as a property of the fitted intervals, which are function-level thresholds for the present sweep rather than validated allocation rules.

A.5 Reasoning Evaluation and Domain-Level Analysis

Reasoning ability is often reported as a single aggregate number, but aggregate scores can obscure where gains and losses occur. Two widely used benchmarks illustrate this masking problem: GSM8K (Cobbe et al., 2021) reduces grade-school arithmetic reasoning to a single pass@1 accuracy number, while BIG-Bench Hard (Suzgun et al., 2023) aggregates 23 diverse tasks—spanning inductive, deductive, spatial, and algorithmic reasoning—into a single score that can conceal divergent per-domain trends of the kind our domain-level analysis reveals. Ke et al. (2025) survey reasoning benchmarks broadly and emphasize the diversity of tasks used to evaluate LLM reasoning. Prior work on code and reasoning also decomposes broad competence into smaller behavioral units, for example in analyses of code-model functions (Chen et al., 2024a) and parameter-efficient reasoning modules (Huang et al., 2025). Huang et al. (2026) further use RL signals to guide mid-training data selection. We share the motivation of looking beyond a single global score, but our analysis is organized around domain-level mid-training coverage and downstream domain-level behavior.

A.6 Knowledge-Orthogonal and Structured Reasoning Benchmarks

Evaluating LLMs on structured reasoning tasks that require applying explicit rules rather than retrieving memorized world knowledge has gained increasing attention. Han et al. (2024) introduce FOLIO, a first-order logic reasoning dataset in which models must derive entailment from formal logical statements without relying on surface-level heuristics; its difficulty for frontier models motivates the mid-training coverage study we conduct on the Logic domain. Liu et al. (2020) present LogiQA, a multiple-choice logical reasoning benchmark drawn from Chinese civil-service examinations covering categorical, conditional, and disjunctive reasoning types; its structural diversity is representative of the rule-following challenge we study. Sinha et al. (2019) construct CLUTRR to test systematic generalization in relational rule-following by withholding chain-length compositions from the training distribution—a compositional generalization challenge closely related to the formal-deduction structure of our Logic and Puzzle domains. More recently, Zhou et al. (2025) propose RuleArena, evaluating models on complex real-world rule systems (airline policies, tax codes, NBA regulations), requiring the same kind of rule-adherence that KOR-Bench (Ma et al., 2025) formalizes as knowledge-orthogonal reasoning. COGS (Kim and Linzen, 2020) probes compositional generalization in semantic parsing by constructing test cases whose structural complexity systematically exceeds the training distribution; the compositional gap it reveals—where models fail on combinations of primitives they handle individually—parallels KOR-Bench's rule-composition design and motivates the emphasis on generalization over memorization. The AI2 Reasoning Challenge (ARC) (Clark et al., 2018) partitions elementary science questions into a retrieval-solvable Easy set and an inference-requiring Challenge set, providing an early demonstration that even modest reasoning demands stratify models in ways aggregate accuracy cannot capture.

Our work differs from these benchmark contributions in its focus on the *training side*: rather than proposing new evaluation criteria, we ask how the proportion of domain-specific data during mid-training determines whether the corresponding structured reasoning capabilities are reliably acquired. The knowledge-orthogonal property of KOR-Bench—which makes the domains semantically rule-disjoint, so that training gains in one domain need not carry to another—makes it a particularly controlled instrument for studying how per-domain coverage is associated with domain-level accuracy.

B Training and Evaluation Details

B.1 Compute

All training used full fine-tuning on Qwen3-8B-Base with DeepSpeed ZeRO-3 and bf16; evaluation used vLLM on 8 GPUs. Data generation is a separate, smaller compute item: the deterministic solvers run on CPU, and the model branch (Appendix B.2) is invoked only on the $< 10\%$ of pool items the solvers do not cover. At a mid-training corpus of ≈750M unique tokens that is ≈75M generated tokens, or ≈5–10 GPU-hours at the same assumed throughput, i.e. still negligible against training. We report the study's scale as run counts and token budgets, which are exactly determined by the design, rather than as GPU-hours, which we did not instrument per run. An order-of-magnitude estimate is nevertheless recoverable from those budgets: the primary study trains roughly 2.4×10^{11} tokens at mid-training (160×1.5B), 1.3×10^{11} over the supervised passes (420×0.3B) and a smaller step-bounded RL remainder, about 3.7×10^{11} tokens in total. At an assumed sustained throughput of 2–4×10^3 tokens/s per GPU for full fine-tuning of an 8B model under ZeRO-3 on 8×H20 (144 GB), this corresponds to roughly 2.6×10^4–5.1×10^4 GPU-hours. This is an estimate from reported budgets under a stated throughput assumption, not a measurement, and it scales inversely with that assumption.

Mid-training accounts for 160 runs at a fixed ≈1.5B-token budget over two epochs each: 24 sweep configurations, six held-out allocations, the θ^* allocation and the FineWeb-Edu baseline, at five seeds apiece. The supervised stage adds 155 compensatory passes and 155 matched uniform-control passes (both over the 24 sweep and six held-out checkpoints plus θ^*, five seeds each), 20 standard-SFT passes for the three full-pipeline configurations and the shared SFT-only baseline, and a further 90 passes for the three additional sharpening branches of Appendix H.3 ($\gamma \in \{2, 4, \infty\}$) over six configurations. The RL stage adds 50 GSPO runs at a fixed 200-step budget: the three full-pipeline configurations, the six held-out allocations, and the shared SFT+RL baseline, five seeds each. This totals 630 training runs (160 mid-training, 420 supervised, 50 RL) for the primary 8B study, of which mid-training dominates the cost because it is the only stage operating on a multi-billion-token corpus. Two validation sub-experiments are budgeted separately and on a smaller scale: Experiment A (Appendix B.4) adds 12 interior allocations × 3 seeds = 36 mid-training-only runs, and the 4B replication (Appendix C) adds 8 allocations × 3 seeds = 24 mid-training-only

runs at the smaller model.

The design deliberately spends most of that budget on breadth at the mid-training stage rather than depth at the RL stage, which is why the RL leg is short (200 steps) and its contribution is scoped accordingly. A study prioritising the RL question differently would invert this allocation. For a group planning a replication, the cheapest informative subset is the coverage sweep plus the compensatory and uniform SFT passes—roughly the first two thirds of the runs above—since Findings 1 and 3 are both identified without any RL.

B.2 Difficulty Tiers and the Model Branch

Six-tier difficulty ladder. Each rule's pool is generated across six difficulty tiers D0–D5 on a monotone problem-scale ladder. The tier label indexes the generation-time scale parameter, from the smallest at D0 to the largest at D5, so the tiers are ordered by construction rather than estimated after the fact. The tiers are held to a fixed composition within each domain, so a coverage change reallocates tokens across domains without changing the difficulty mix of any domain's sampled data. The character ranges shown in Figure 6 (1–3, 2–8, 8–12) are illustrative of the cipher domain's scale parameter; the ladder itself is defined per domain on that domain's own scale (grid size for Puzzle, derivation depth for Logic, operand count for Operation, world/rule count for Counterfactual).

The model branch. The pipeline's component 5 has two branches. The deterministic solvers (5a) cover $> 90\%$ of generated items and supply both the answer and the reasoning trace. The remaining $< 10\%$—items whose rule the solvers do not implement—take the model branch (5b), in which Claude Fable 5 (Anthropic, 2025) supplies the answer and the trace. Three properties of this branch matter for interpreting the results. First, *it does not replace symbolic generation*: it is invoked only where a solver is unavailable, so the bulk of every corpus is solver-generated and the branch is a coverage completion rather than the primary source of supervision. Second, *it is answer-only, not difficulty-only*: unlike a difficulty labeller, the branch emits the ground-truth answer and the chain-of-thought trace for the items it handles, which is why the corpus is described as solver-generated with a model-completed minority rather than as purely symbolic. Third, *its outputs are verified*: every branch output passes the same component-6 verification (extract, compare against the solver-independent reference where one exists, deduplicate, format check) as a solver output, and is discarded rather than admitted if verification fails.

Reproducibility of the branch. The branch is a closed model, so the corpus depends on a specific model version and decoding configuration. We report the model identifier, the decoding parameters and the branch invocation seed alongside the released corpus, and we note the limit plainly: a reconstruction using a different model or decoding setting will not reproduce the $< 10\%$ of items that take this branch, whereas the $> 90\%$ solver-generated majority is exactly reproducible from the rule definitions.

Scope of the claims. Two claims in the main text are unaffected by the branch and two are scoped by it. Unaffected: the decontamination guarantees (Appendix F), because the branch sees only freshly synthesised pool items and never an evaluation instance; and the coverage definition (§2.3), which is a token-share quantity independent of how any item was solved. Scoped: statements that the corpus is *symbolic* are now statements about the $> 90\%$ majority, and the residual $< 10\%$ is model-generated and therefore inherits that model's failure modes on the rules the solvers do not cover. A reader should treat the branch as a bounded, disclosed, verified minority of the corpus rather than as an unexamined source of supervision.

B.3 Model Scale Selection

We chose Qwen3-8B-Base after measuring, rather than assuming, where a mid-training intervention has room to show itself. Table 6 evaluates the Qwen3 family under our KOR-Bench protocol, and the relevant quantity is the headroom a base checkpoint still has before post-training: the base-to-instruct gap is $+21.2$ pp at 4B ($41.88 \rightarrow 63.08$) and $+20.8$ pp at 8B ($43.00 \rightarrow 63.80$), but only $+10.4$ pp at 30B-A3B ($56.48 \rightarrow 66.88$). At the top of the range the base models have already closed most of that distance on their own—14B-Base reaches 60.56 and 32B-Base 60.68, against an instruct band of roughly 63–67—so any mid-training reallocation competes for a shrinking remainder and its per-domain effects are compressed toward the ceiling. At the bottom, Qwen3-1.5B-Base scores 22.84 overall with Cipher at 1.0 and Puzzle at 4.2, i.e. at the floor on two of the five domains, where allocation differences cannot register

above seed noise. The 4B and 8B checkpoints are the two with the largest measured post-training headroom, carrying roughly twice that of 30B-A3B; 8B is the larger of the two and the one we could afford to run at full sweep scale (24 configurations $\times$ five seeds, plus six held-out allocations and the downstream passes). This is a statement about where the effect is measurable, not a claim that 8B is representative: whether the interior optima and the non-closure result reproduce at other scales is untested here.

What a 4B replication would have to show. Because 4B carries essentially the same post-training headroom as 8B ($+21.2$ vs. $+20.8$ pp), it is the sharpest available test of whether our findings are scale-specific or headroom-specific: a failure to reproduce at 4B could not be explained away as a ceiling or floor artifact. A minimal replication needs three components. *(i) Coverage sweep.* Train 8–12 allocations at 4B that span the low ($< 10\%$), moderate (10–40%) and high ($> 40\%$) band for every domain—a subset of the allocations in Table 16 suffices—at three seeds, evaluated mid-training-only. This supports both the band comparison of Appendix G.1, which involves no curve fitting, and the quadratic interiority permutation test on which our moderation claim actually rests. *(ii) Alignment arm.* Apply the compensatory and uniform SFT passes to those same 4B checkpoints and recompute the pairwise closure counts together with the permutation null of Appendix H.2; this separates a property of the recipe from a property of the 8B model. *(iii) Peak comparison.* On allocations common to both scales, compare fitted peak locations at 4B and 8B. The three possible outcomes are diagnostic: peaks reproducing within seed noise would indicate the optimum is a property of the data mixture rather than the model; peaks shifting systematically with scale would make the fitted intervals scale-dependent and require the allocation guidance to be re-derived per scale; and an absence of interior optima at 4B, given its matched headroom, would localise the effect to 8B and falsify the generality we do not currently claim. A full-pipeline (SFT+RL) 4B arm is not required for any of these three tests, and would be needed only to check the θ^* ordering at a second scale.

B.4 Experiment A: An Interior-Simplex Filler Sub-Sweep

Our compositional reanalysis (Appendix G.2) finds a saddle, not an interior maximum, because the 30 trained allocations cluster near the simplex's edges and mid-ridge, so the curvature of the interior cannot be resolved. To determine whether a jointly optimal mixture exists—which would let the per-domain moderation result of Finding 3 be promoted to a joint claim, or would establish that it cannot be—we therefore trained 12 new allocations confined to the interior (every domain share in $[12, 34]\%$, each at least 5 pp in Euclidean (L_2) distance over the five shares from every existing allocation), drawn by Dirichlet($\alpha = 8$) sampling and screened for feasibility. Table 2 reports them (mid-training-only, three seeds). **The result is consistent with the coda analysis and does not overturn it:** adding these 12 interior points to the 30-allocation pool and refitting the compositional surface leaves every domain's stationary point reading as a saddle ($R^2 = 0.78$–0.96 over the 42-pool), though the negative curvature is shallow and quadratic misspecification is untested (Appendix G.2)—the interior data neither establish a joint optimum nor exclude a shallow flat region. The per-domain moderation result is correspondingly confirmed and delimited: over the full 42-allocation pool the moderate band remains the highest for all five domains (low/moderate/high means, e.g. Cipher $43.8/51.3/42.2$, Operation $51.9/75.8/50.4$), and the 12 interior points also supply an additional batch of out-of-sample curve-validation points. Because no joint optimum is identified, the moderation finding stays a per-domain marginal statement and is not promoted to a single recommended mixture.

C Study II: A Cross-Scale Replication at 4B

The scale rationale of Appendix B.3 shows that the post-training headroom—the quantity that determines whether a mid-training reallocation can produce measurable per-domain differences—is largest at 4B ($+21.2$ pp) and 8B ($+20.8$ pp) and roughly half that at 30B-A3B ($+10.4$ pp), while 1.5B sits on the floor (Cipher 1.0, Puzzle 4.2) and 14B/32B near the ceiling (60.6 vs. an instruct band of roughly 63–67). Only the 4B and 8B checkpoints therefore lie in a regime where the effect can be observed. Because 4B carries almost the same headroom as 8B, a failure to reproduce at 4B cannot be written off as a ceiling or floor artifact, which makes it the sharpest available test of whether our findings are scale-specific or headroom-specific.

We therefore ran a mid-training-only replication at Qwen3-4B-Base over the eight allocations of Table 3, chosen so that *every domain* has low ($< 10\%$), moderate (10–40%) and high ($> 40\%$) coverage samples across the set; the allocations reuse the corresponding 8B design points where they exist so the two scales can be compared on

Table 2: Experiment A: 12 interior allocations (all components in 12–34%, ≥ 5 pp from every existing allocation), each trained mid-training-only over three seeds. Each row block reports the **Portion** (internal mid-training token share, %), the **Evaluation** (mid-training-only accuracy, mean $\pm$ SD across the three seeds), and the **Remedial SFT** allocation from Eq. 5. Adding these to the 30-allocation pool and refitting the compositional surface leaves a saddle in every domain, so no joint interior optimum is established (§ 5.2).

Domains	Cipher	Operation	Logic	Counterfactual	Puzzle
Config 1: Portion	12.9%	24.4%	29.0%	14.1%	19.5%
Evaluation	$52.9 \pm 1.6\%$	$76.4 \pm 3.0\%$	$57.2 \pm 1.8\%$	$74.2 \pm 2.4\%$	$12.0 \pm 1.2\%$
Remedial SFT	23.5%	17.8%	15.5%	22.9%	20.2%
Config 2: Portion	14.5%	18.3%	12.9%	21.2%	33.0%
Evaluation	$51.4 \pm 1.3\%$	$76.9 \pm 2.3\%$	$59.1 \pm 1.5\%$	$82.4 \pm 1.8\%$	$15.1 \pm 3.2\%$
Remedial SFT	22.7%	20.8%	23.5%	19.4%	13.5%
Config 3: Portion	12.5%	15.9%	27.5%	21.6%	22.5%
Evaluation	$51.5 \pm 2.1\%$	$77.2 \pm 2.9\%$	$57.6 \pm 0.6\%$	$82.6 \pm 3.0\%$	$13.6 \pm 1.3\%$
Remedial SFT	23.8%	22.0%	16.3%	19.2%	18.7%
Config 4: Portion	17.5%	22.0%	23.2%	12.0%	25.3%
Evaluation	$53.6 \pm 1.2\%$	$78.1 \pm 2.5\%$	$59.4 \pm 0.8\%$	$74.0 \pm 2.3\%$	$13.7 \pm 3.3\%$
Remedial SFT	21.3%	19.0%	18.4%	24.0%	17.3%
Config 5: Portion	30.4%	14.6%	12.9%	17.4%	24.6%
Evaluation	$50.7 \pm 0.7\%$	$77.4 \pm 0.7\%$	$58.6 \pm 2.8\%$	$78.0 \pm 3.4\%$	$14.4 \pm 1.8\%$
Remedial SFT	14.8%	22.7%	23.5%	21.3%	17.7%
Config 6: Portion	25.7%	24.5%	15.6%	16.6%	17.6%
Evaluation	$50.9 \pm 2.5\%$	$77.8 \pm 2.5\%$	$58.5 \pm 2.1\%$	$77.3 \pm 2.8\%$	$11.1 \pm 2.9\%$
Remedial SFT	17.2%	17.8%	22.2%	21.7%	21.2%
Config 7: Portion	24.9%	13.8%	18.6%	25.8%	16.9%
Evaluation	$50.4 \pm 1.1\%$	$77.9 \pm 3.1\%$	$59.5 \pm 1.1\%$	$81.9 \pm 2.8\%$	$11.5 \pm 2.9\%$
Remedial SFT	17.5%	23.1%	20.7%	17.1%	21.5%
Config 8: Portion	18.9%	24.7%	18.4%	15.9%	22.1%
Evaluation	$52.5 \pm 2.6\%$	$76.2 \pm 0.6\%$	$59.7 \pm 3.2\%$	$76.9 \pm 1.0\%$	$12.9 \pm 1.3\%$
Remedial SFT	20.5%	17.7%	20.8%	22.1%	18.9%
Config 9: Portion	24.3%	16.6%	16.4%	20.8%	21.9%
Evaluation	$51.5 \pm 2.2\%$	$77.2 \pm 2.6\%$	$60.4 \pm 2.5\%$	$80.9 \pm 1.2\%$	$12.8 \pm 1.5\%$
Remedial SFT	17.9%	21.7%	21.8%	19.6%	19.1%
Config 10: Portion	33.4%	17.4%	17.4%	12.7%	19.1%
Evaluation	$50.8 \pm 1.0\%$	$78.1 \pm 0.6\%$	$58.6 \pm 1.8\%$	$73.3 \pm 2.4\%$	$11.1 \pm 3.1\%$
Remedial SFT	13.3%	21.3%	21.3%	23.7%	20.5%
Config 11: Portion	13.2%	17.7%	23.8%	25.3%	20.1%
Evaluation	$51.8 \pm 3.4\%$	$78.9 \pm 1.0\%$	$58.8 \pm 2.8\%$	$83.5 \pm 3.4\%$	$13.0 \pm 3.2\%$
Remedial SFT	23.4%	21.2%	18.1%	17.4%	20.0%
Config 12: Portion	14.8%	19.2%	29.9%	15.4%	20.7%
Evaluation	$52.2 \pm 2.4\%$	$77.4 \pm 1.4\%$	$56.9 \pm 2.5\%$	$75.7 \pm 2.9\%$	$13.4 \pm 3.2\%$
Remedial SFT	22.6%	20.4%	15.1%	22.3%	19.6%

matched points. **The moderation pattern reproduces at 4B:** the moderate band again yields the highest mean accuracy in all five domains (e.g. Cipher $36.3/47.9/38.0$, Operation $49.3/90.8/61.1$), though the effect is far weaker for Counterfactual, whose three bands ($91.0/91.3/90.0$) span just 1.3 pp because its base prior is already near ceiling; the three-parameter quadratic is concave with an interior vertex in all five domains. The peak locations, however, shift with scale, and not in one direction: the fitted peaks at 4B (Cipher 8.8, Operation 17.2, Logic 18.8, Counterfactual 8.5, Puzzle 31.6; the Counterfactual estimate is dominated by the near-flat response and its base prior of 90.0, and Puzzle by a single high-coverage point) differ from the 8B peaks by -1.1, $+1.3$, and $+3.8$ pp for Cipher, Operation, and Logic, -3.5 pp for Puzzle, and -17.5 pp for Counterfactual (deltas against the rounded peaks of Table 13). This is the scale-dependence outcome flagged in Appendix B.3: the interior optima are present at both scales, but their locations are not scale-invariant, so allocation guidance derived at one scale should not be transferred directly to another. We did not run the SFT or RL leg at 4B, so the alignment pass (compensatory vs. uniform closure, and the permutation null of Appendix H.2) and the θ^* ordering were not re-tested at the second scale.

The band comparison and the quadratic interiority test (read off the mid-training-only evaluations) and the peak-location comparison are therefore the two of the three tests in Appendix B.3 that the current 4B data support; the alignment-pass test and the θ^* ordering remain to be run at the second scale.

Table 3: Study II: eight cross-scale allocations for the Qwen3-4B-Base replication, each trained mid-training-only. Each row block reports the **Portion** (internal mid-training token share, %), the **Bands** (each domain's low/moderate/high coverage), the **Evaluation** (mid-training-only accuracy, mean $\pm$ SD), and the **Remedial SFT** allocation from Eq. 5. The 4B run was mid-training-only, so no post-SFT columns are reported; the moderation pattern reproduces at 4B (moderate band best in all five domains, concave quadratic with an interior vertex), but the fitted peak locations shift with scale.

Domains	Cipher	Operation	Logic	Counterfactual	Puzzle
Config 1: Portion	72.3%	0%	12%	0%	15.7%
Bands (L/M/H)	H	L	M	L	M
Evaluation	$38.0 \pm 0.8\%$	$44.4 \pm 1.8\%$	$45.2 \pm 3.0\%$	$90.5 \pm 1.0\%$	$7.4 \pm 2.5\%$
Remedial SFT	0%	28.3%	22.6%	28.3%	20.9%
Config 2: Portion	20%	20%	20%	20%	20%
Bands (L/M/H)	M	M	M	M	M
Evaluation	$45.1 \pm 0.8\%$	$96.9 \pm 1.3\%$	$49.9 \pm 1.5\%$	$91.0 \pm 0.8\%$	$9.1 \pm 0.7\%$
Remedial SFT	20%	20%	20%	20%	20%
Config 3: Portion	0%	10%	40%	10%	40%
Bands (L/M/H)	L	M	M	M	M
Evaluation	$25.2 \pm 1.8\%$	$75.3 \pm 3.1\%$	$42.2 \pm 0.7\%$	$91.9 \pm 3.0\%$	$10.7 \pm 1.9\%$
Remedial SFT	30%	25%	10%	25%	10%
Config 4: Portion	5%	15%	5%	15%	60%
Bands (L/M/H)	L	M	L	M	H
Evaluation	$43.8 \pm 0.6\%$	$96.3 \pm 1.7\%$	$37.1 \pm 3.1\%$	$91.7 \pm 1.4\%$	$2.3 \pm 2.0\%$
Remedial SFT	27.5%	22.5%	27.5%	22.5%	0%
Config 5: Portion	33%	5%	6%	50%	6%
Bands (L/M/H)	M	L	L	H	L
Evaluation	$48.7 \pm 1.1\%$	$55.9 \pm 1.6\%$	$35.9 \pm 0.7\%$	$90.0 \pm 2.4\%$	$2.3 \pm 3.0\%$
Remedial SFT	13.5%	27.5%	27%	5%	27%
Config 6: Portion	4%	3%	30%	29%	34%
Bands (L/M/H)	L	L	M	M	M
Evaluation	$39.8 \pm 0.8\%$	$47.6 \pm 0.9\%$	$49.4 \pm 2.2\%$	$90.6 \pm 0.8\%$	$12.6 \pm 3.0\%$
Remedial SFT	28%	28.5%	15%	15.5%	13%
Config 7: Portion	25%	45%	10%	10%	10%
Bands (L/M/H)	M	H	M	M	M
Evaluation	$48.3 \pm 2.1\%$	$61.1 \pm 2.5\%$	$45.7 \pm 1.9\%$	$91.5 \pm 1.7\%$	$4.6 \pm 3.1\%$
Remedial SFT	17.5%	7.5%	25%	25%	25%
Config 8: Portion	20%	25%	45%	5%	5%
Bands (L/M/H)	M	M	H	L	L
Evaluation	$49.3 \pm 0.9\%$	$94.9 \pm 0.9\%$	$40.5 \pm 0.7\%$	$91.4 \pm 1.8\%$	$3.8 \pm 1.4\%$
Remedial SFT	20%	17.5%	7.5%	27.5%	27.5%

D RL, Evaluation, and Artifact Details

The subsections below cover the RL objective and its algorithm choice, the training hyperparameters and seed protocol, the evaluation and answer-extraction protocol, the external reference models, the RL budget, artifact availability, and the study-roadmap figures.

D.1 GSPO Objective

Group-based Sequence-level Policy Optimization (GSPO) (Zheng et al., 2025) is a group-relative policy gradient method. For each prompt q, G responses $\{o_1, \ldots, o_G\}$ are sampled from the current policy π_θ. Each response

receives a binary verifier reward $R_i \in \{0, 1\}$ (1 if correct, 0 otherwise). The advantage for response o_i is computed via group-level normalisation:

$$A_i = \frac{R_i - \text{mean}(\{R_1, \ldots, R_G\})}{\text{std}(\{R_1, \ldots, R_G\})}, \tag{1}$$

and the policy is updated by maximising the clipped surrogate objective:

$$\mathcal{J}_{\text{GSPO}}(\theta) = \mathbb{E}_{q,\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\min(\rho_i A_i,\ \text{clip}(\rho_i, 1-\varepsilon, 1+\varepsilon)A_i)\right], \tag{2}$$

where $\rho_i = \pi_\theta(o_i \mid q)/\pi_{\text{old}}(o_i \mid q)$ is the *sequence-level* probability ratio. Under the standard autoregressive factorization this is the product of per-token ratios, $\rho_i = \prod_t \pi_\theta(o_{i,t} \mid q, o_{i,<t})/\pi_{\text{old}}(o_{i,t} \mid q, o_{i,<t})$, equivalently the sum of per-token log-ratios in log space. $\varepsilon = 0.2$ is the clip ratio, and θ_{old} denotes the policy parameters before the update. When all rewards within a group are equal, $\text{std}(\{R_1, \ldots, R_G\}) = 0$ and the normalized advantage in Eq. 1 is undefined. Because a group with all-equal binary rewards carries no preference signal, its treatment affects only that group's gradient contribution and not the relative ordering of updates across other groups; the exact handling in our runs is implementation-specific and is acknowledged here as an implementation detail rather than a design choice. We use group size $G = 8$. Training proceeds for 200 steps with a peak learning rate of 1×10^{-6} under the cosine schedule with 5% warmup (Table 5). RL rollouts are sampled with the same 16,384-token sequence cutoff as evaluation, so rollout truncation does not differentially bias verifier rewards for long reasoning traces (Appendix D.7).

D.2 Choice of RL Algorithm

The RL leg is a secondary observation in this study, but its algorithm choice is not arbitrary, and we record the reasoning here. Our setting imposes three constraints: we wanted an *on-policy* method, since the RL question is whether exploration can surface correct-but-low-probability chains rather than whether a fixed dataset can be reweighted—the latter being what the SFT arm already tests; the RL stage has to run against a *binary outcome verifier* rather than a learned reward model; and it must operate on *very long* chain-of-thought traces—rollouts are sampled at the same 16,384-token cutoff as evaluation (Appendix D.1). Table 4 summarises how the four standard options sit against those constraints.

Table 4: RL algorithm options against the constraints of this study. The table records design reasoning, not a measured comparison: we ran GSPO only, and no algorithm ablation is reported here.

Algorithm	Position in this setting	Reason
GRPO	Default for a dense model	Samples several long CoTs per prompt and scores them by group-relative advantage on final-answer correctness; needs no critic, so it is lighter than PPO in memory and engineering. Reinforces traces that reach the answer and suppresses wrong enumeration.
GSPO	**Chosen here**	A stability-oriented variant of GRPO that applies one importance ratio per *sequence* rather than one per token (Zheng et al., 2025). Better matched to long CoTs.
DPO	Cold start / format alignment only	Offline: it imitates a fixed preference set and cannot explore new search paths during training. Suitable for teaching a model to enumerate, backtrack, verify, then answer—but it cannot deepen search, so it cannot test the online-repair question we ask.
PPO	Fallback, not preferred	Highest online-exploration ceiling, but requires a critic and a reward model. On 16,384-token traces the value estimate is poorly conditioned and token-level credit assignment is noisy, at high cost. Worth it only with a strong process reward model (PRM), which we do not have.

Qwen3-8B-Base is dense, so GRPO would be the default choice for the model class, and we did not run an algorithm comparison that could settle the question empirically. Two features of this specific setting move the choice to GSPO. First, the traces are long: at a 16,384-token rollout cutoff, GRPO's surrogate clips an importance ratio at every token position, so the clipping decision is made thousands of times along a single response, whereas GSPO applies one ratio per response and clips once for the whole trace. Second, we observed the failure mode that motivates a sequence-level ratio at all: extending the schedule beyond 200 steps produced cross-seed instability (reward divergence without a consistent directional trend) rather than a stable higher-performing policy (Appendix D.7). We read that observation as a reason to hold the budget fixed, not as evidence that GSPO outperforms GRPO—the instability was itself observed

under GSPO, so it cannot arbitrate between the two. We also note that the stability argument most often made for GSPO in the literature (MoE routing collapse) does not apply here: Qwen3-8B-Base is dense, so the trace-length argument is the only one of the two that bears on this setting. Keeping the alignment stage on our own pipeline's configuration makes it a like-for-like continuation rather than a foreign intervention, which matters when the claim is about what *this* downstream recipe can and cannot repair.

DPO is ruled out on a more basic ground: it is offline preference imitation, training only on chains already present in its preference data, so it adds no exploration channel and cannot speak to whether on-policy search would reach chains the data does not contain. It also consumes preference pairs, which a binary correctness verifier supplies only by construction. PPO is ruled out on cost and conditioning: it needs a critic whose value estimates are unreliable over 16,384-token traces, and we have a binary outcome verifier rather than the process reward model that would justify that cost.

Two scope limits follow, and we state them rather than let the rationale imply more than it shows. First, **we did not ablate the RL algorithm**: GRPO, DPO and PPO were not run, so the entries above are design reasoning, not measured comparisons, and nothing here establishes that another algorithm would have produced the same $+1.1$–$+2.3$ pp increment or the same non-closure. Second, GRPO and GSPO share the group-relative advantage estimator and differ mainly in the level at which the importance ratio is computed, so the RL leg should be read as evidence about a *group-relative* family at a fixed 200-step budget rather than about GSPO specifically. Because the RL gains are small and the leg is scoped accordingly (§5.1), this ambiguity does not touch the paper's headline claims, which rest on the mid-training sweep and the SFT arm.

D.3 Training Hyperparameters

Table 5 lists the complete hyperparameter set for all three stages. The compensatory-SFT pass and its uniform-SFT control use the SFT column's three-epoch protocol at a smaller fixed total token budget, matched across the two diagnostic passes and below the SFT stage's $\sim$0.3B, which is why the two families are not on the same scale (Appendix H); the RL stage uses a fixed 200-step GSPO schedule (Appendix D.7).

Parameter	Mid-Training	SFT	RL (GSPO)
Base model	Qwen3-8B-Base (base)	Mid-training ckpt.	Mid-training+SFT ckpt.
Stage type	pt (cont. pretraining)	sft	rl (GSPO)
Fine-tuning type	Full	Full	Full
Learning rate	1e-5	5e-5	1e-6
Per-device batch size	4	4	1
Gradient accumulation steps	4	4	16
Effective batch size	16	16	16
Sequence length (cutoff)	12,000	12,000	16,384
Training volume	$\sim$1.5B tokens (2 ep.)	$\sim$0.3B tokens (36,537 inst., 3 ep.)	200 steps
Epochs	2.0	3.0	— (step-based)
Optimizer	AdamW (ZeRO-3)	AdamW (ZeRO-3)	AdamW (ZeRO-3)
LR schedule	warmup_stable_decay	cosine	cosine
Warmup ratio	0.05	0.05	0.05
Precision	bf16	bf16	bf16
Gradient checkpointing	enabled	enabled	enabled
Flash attention	FA2	FA2	FA2
Packing	enabled	enabled	disabled
Validation split	1%	5%	reward-based
Chat template	None (completion)	qwen (chat)	qwen (chat)
RL-specific hyperparameters			
Group size G	—	—	8
Clip ratio ε	—	—	0.2
Reward type	—	—	binary verifier

Table 5: Training hyperparameters for all three stages. Mid-training and SFT use LLaMA-Factory with full fine-tuning, DeepSpeed ZeRO-3 (Rajbhandari et al., 2020) and FlashAttention-2 (Dao et al., 2024). RL uses GSPO (Zheng et al., 2025) with binary verifier rewards at a fixed 200-step budget. The compensatory-SFT pass and its uniform-SFT control are separate supervised runs that use the same 3-epoch protocol and total token budget as each other (Appendix H). FA2 = FlashAttention-2.

D.4 Seed Selection and Model Selection Protocol

For the mid-training coverage sweep (Table 16), each configuration is trained from five random seeds and evaluated on a held-out KOR-Bench test split that was never used during training or model selection; we report the mean $\pm$ 1 SD across the five seeds, so the SDs capture cross-seed variance for the mid-training-only, compensatory-SFT, and uniform-SFT evaluations. **The same five seed indices are used at every stage**: a mid-training configuration, its compensatory and uniform SFT successors, its RL successor, and the shared no-mid-training baselines all draw seed $s \in \{1, \ldots, 5\}$ from the same index set, so every pipeline row is seed-paired with the baseline it is compared against and the per-stage deltas of Table 8 are per-seed differences. The full-pipeline rows in Table 1 (Mid-training+SFT, Mid-training+SFT+RL) are likewise five-seed means, and Table 1 reports the corresponding cross-seed SDs in every row and cell; the external-benchmark measurements are five-seed means, and bootstrap 95% CIs over per-question resamples, where reported, capture evaluation-sample variance and do not replace the cross-seed uncertainty. All values in this manuscript are five-seed means unless stated otherwise.

D.5 Answer Extraction, Decoding, and Evaluation Protocol

We extract the final answer via regex `[[...]]` (last match). For Operation, we additionally apply SymPy-based equivalence matching. For Logic, whitespace and punctuation are stripped before comparison. All evaluations use greedy decoding (temperature = 0.0, max_tokens = 16,384) with vLLM on 8 GPUs. Overall accuracy is the macro-average of domain-level accuracies $\mathrm{Acc}(d) = \frac{1}{N_d} \sum_{i=1}^{N_d} \mathbb{1}[\hat{y}_i = y_i^*]$. For difficulty-stratified external benchmarks, per-stratum accuracy is $\mathrm{Acc}(d, \ell) = \frac{1}{N_{d,\ell}} \sum_{i:\ell_i=\ell} \mathbb{1}[\hat{y}_i = y_i^*]$, where ℓ is proof depth, house count, or reasoning type. Confidence intervals are computed using bootstrap resampling for KOR-Bench domain results and asymptotic intervals for large external strata. External benchmarks are evaluated using their native protocols.

D.6 External Reference Models (Qwen3 Family)

Table 6 evaluates the Qwen3 family (Yang et al., 2025) under the same KOR-Bench protocol, for two purposes. None of these models is trained in our pipeline and their instruction-tuning and RL recipes differ from the recipe studied here, so no value in the table is comparable to our pipeline rows—in particular the Qwen3-8B-Instruct score of 63.80% should not be read as evidence about any of our allocations. The table's role is instead (i) to place our trained checkpoints against publicly available reference points, and (ii) to supply the base-to-instruct headroom measurements that motivate the choice of scale in Appendix B.3.

Table 6: KOR-Bench zero-shot accuracy (%) of the Qwen3 model family (1.5B/4B/8B/14B/30B-A3B/32B, base and instruct), evaluated with the same protocol as Table 1. None of these models is trained in our pipeline; their instruction-tuning and RL recipes differ from the recipe studied here, so no value is comparable to our pipeline rows. The base-to-instruct pairs supply the post-training headroom used in the model-scale rationale (Appendix B.3). Two Logic/Counterfactual cells for Qwen3-30B-A3B-Instruct were not measured and are marked "—"; that row's Overall is therefore not the macro-average of the five domain columns.

Model	Overall	Cipher	Oper.	Logic	Counterf.	Puzzle
Qwen3-1.5B-Base	22.8 ± 1.2	1.0 ± 1.7	7.0 ± 1.7	20.2 ± 1.8	81.8 ± 2.5	4.2 ± 0.5
Qwen3-4B-Base	41.9 ± 0.9	18.2 ± 1.0	60.0 ± 2.6	39.2 ± 1.6	90.0 ± 2.5	2.0 ± 1.0
Qwen3-4B-Instruct	63.1 ± 1.7	59.8 ± 1.1	93.4 ± 1.8	49.0 ± 1.5	88.2 ± 2.2	25.0 ± 0.7
Qwen3-8B-Base	43.0 ± 1.0	13.2 ± 1.9	60.8 ± 2.0	49.2 ± 1.7	86.0 ± 2.6	5.8 ± 0.7
Qwen3-8B-Instruct	63.8 ± 0.9	66.1 ± 1.5	74.5 ± 2.4	57.0 ± 1.2	93.3 ± 2.4	28.1 ± 0.7
Qwen3-14B-Base	60.6 ± 1.1	50.4 ± 1.6	88.8 ± 2.5	56.0 ± 1.4	93.2 ± 2.2	14.4 ± 0.6
Qwen3-30B-A3B-Base	56.5 ± 1.4	43.2 ± 1.1	85.6 ± 1.8	45.2 ± 1.9	96.4 ± 2.3	12.0 ± 0.8
Qwen3-30B-A3B-Instruct	66.9 ± 1.8	65.4 ± 1.4	91.6 ± 2.5	—	—	28.8 ± 0.5
Qwen3-32B-Base	60.7 ± 1.4	51.8 ± 1.6	87.8 ± 1.9	57.0 ± 1.6	90.6 ± 2.2	16.2 ± 0.8

D.7 RL Budget: Interpretive Limitations and Training Instability

The small RL gains observed across all three experiments ($+1.1$–$+2.3$ pp from GSPO; the largest, $+2.30$ pp, occurs under the exploratory θ^* allocation) raise the question of whether a larger RL budget would alter the compensatory conclusion. Two interpretations are consistent with the data. **Interpretation 1 — Competence-limited RL.** The 200-step GSPO stage may be genuinely limited by the model's current competence: verifier-based RL reinforces existing chains rather than discovering new symbolic procedures, so domains that mid-training left weak receive little RL gain regardless of budget. **Interpretation 2 — Insufficient RL budget.** 200 steps may simply be insufficient to manifest RL's corrective potential; a longer schedule could in principle reshape the domain gaps by giving the model more opportunities to explore and for the verifier to surface correct but low-probability chains in weak domains.

Breaking the indistinguishability with a pass@k probe. The two interpretations are not in fact inseparable, and a sampling probe separates them. We drew k completions from the same post-SFT policy at $T > 0$ and measured pass@k on the collapsed domains (Table 7). On the Expt. 1 checkpoint, where mid-training left Operation at 39.5% and Counterfactual at 56.5%, pass@64 reaches 98.2% and 99.4% respectively, against pass@1 of 41.0% and 58.2%. A pass@64 near ceiling means the correct chain is already in the policy's support: the model can produce the right answer, and what the 200-step budget fails to do is concentrate probability mass on it rather than discover it. That rules out Interpretation 1 in its strong form—the verifier is not competence-limited, because there is nothing left for it to find—and leaves Interpretation 2, an optimization-budget account, as the surviving reading. The same probe on the balanced arm shows the effect is not specific to the collapsed domains: Puzzle rises from 12.4% to 70.5%, the widest gap of any domain in either arm.

Table 7: Pass@k from the post-SFT policy, five seeds. Pass@1 reproduces the greedy mid-training+SFT accuracy within noise, so the sampling temperature does not itself move the operating point. The collapsed domains of Expt. 1 reach near-ceiling pass@64, which places the correct chain inside the policy's support.

Arm	Domain	mid+SFT	k=1	k=2	k=4	k=8	k=16	k=64
balanced	Cipher	55.0	55.2	72.0	82.9	93.5	95.3	99.2
	Operation	83.0	82.8	92.7	97.7	98.7	99.9	100.0
	Logic	64.0	62.7	79.8	90.8	96.8	98.2	99.5
	Counterfactual	88.0	88.5	95.9	99.0	99.6	99.8	100.0
	Puzzle	14.5	12.4	23.3	32.7	45.4	56.2	70.5
Expt. 1	Cipher	39.0	40.5	53.4	71.4	84.3	90.9	97.9
	Operation	39.5	41.0	55.7	71.5	82.0	89.9	98.2
	Logic	62.5	65.5	77.4	89.7	95.3	96.6	99.7
	Counterfactual	56.5	58.2	72.6	84.7	93.8	96.8	99.4
	Puzzle	13.0	11.8	22.0	32.4	42.1	51.2	61.7

We treat the RL leg as a secondary observation rather than a headline claim for two reasons, neither of which is the indistinguishability above. First, the reward is a binary correctness verifier, so no dense per-token reward signal can be extracted from the chain-of-thought; RL can only reinforce whole-answer outcomes. Second, extending the GSPO budget is not a cleanly tunable knob in practice: longer schedules were unstable across seeds (reward divergence without a consistent directional trend) rather than simply maturing into a stable higher-performing policy. Two distinct RL settings appear in this paper and we separate them explicitly. (i) The three full-pipeline configurations of Table 1 use the standard pipeline (mid-training $\rightarrow$ rule-type-balanced SFT $\rightarrow$ RL). (ii) The six held-out allocations are carried through mid-training $\rightarrow$ *compensatory* SFT $\rightarrow$ RL, and it is these runs that produce the **RL Evaluation** rows of Table 16 and the out-of-sample results reported in §4.1 (mean $+1.19$ pp over the compensatory pass; $0/60$ pairs closed at the 5 pp threshold and $7/60$ at the 10% ratio after the complete pipeline). Early exploratory attempts to add RL after compensatory SFT *on the 24 sweep configurations* showed cross-seed instability and are excluded from the reported sweep results, under the same criterion used elsewhere (reward divergence or no consistent directional trend within the 200-step budget); that exclusion applies to those sweep runs only and not to the six held-out allocations, whose RL runs are reported in full. Whether a compensatory-SFT-plus-RL schedule would close gaps across the whole 24-configuration sweep therefore remains open, while on the six held-out allocations it demonstrably does not. We scope the RL stage to the tested 200-step schedule and do not read its small gains as evidence about post-training repair in general. Figure 5 illustrates the pattern on the shared no-mid-training baseline run: the per-step group-mean

verifier reward fluctuates between $\approx 62\%$ and 67% across the 200 steps without a consistent upward trend, and the evaluation gain over the whole stage is small ($+1.20$ pp, SFT-only 64.08% to SFT+RL 65.28%).

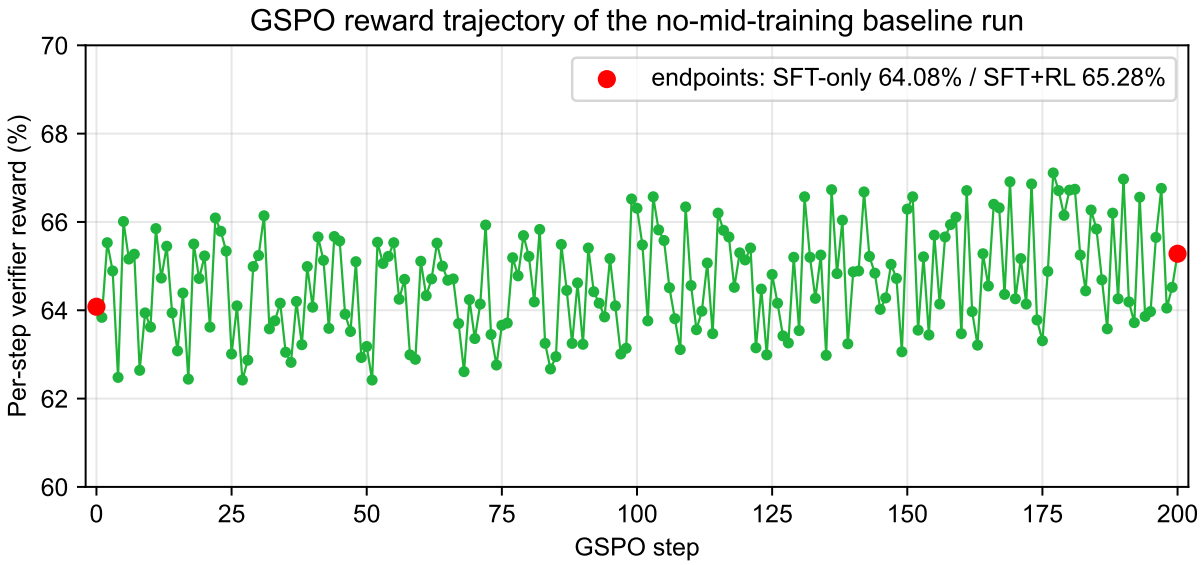

Figure 5: Per-step group-mean verifier reward of the GSPO stage for the shared no-mid-training baseline run (SFT$\rightarrow$RL, 200-step schedule). Red endpoints mark the SFT-only evaluation accuracy at step 0 (64.08%) and the final SFT+RL accuracy at step 200 (65.28%). The reward wanders in a $\approx$62–67% band without a consistent upward trend, matching the small evaluation gain of the RL stage ($+1.20$ pp).

D.8 Code and Data Availability

To support replication, we will release the training configurations, the data-generation pipeline (rule library, synthesizer, deterministic solvers, the model branch with its model identifier and decoding settings, verification, and preprocessing), the evaluation harness (prompt templates, answer-extraction rules, and the six-tier difficulty ladder), the analysis scripts that produce every reported table and figure, and per-seed results for all reported tables, together with the RL reward trajectories of all runs, including those excluded under the criteria of Appendix D.7. Because the model branch is closed, we also release the item-level record of which pool items took branch 5b and the branch's raw output for each, so that a reconstruction can substitute its own completion model on exactly the same item set and hold the $> 90\%$ solver-generated majority fixed.

D.9 Pipeline and Study-Roadmap Figures

The synthetic data-generation pipeline (Figure 6) and the compensatory-SFT gain fits (Figure 7) are reproduced here (§2).

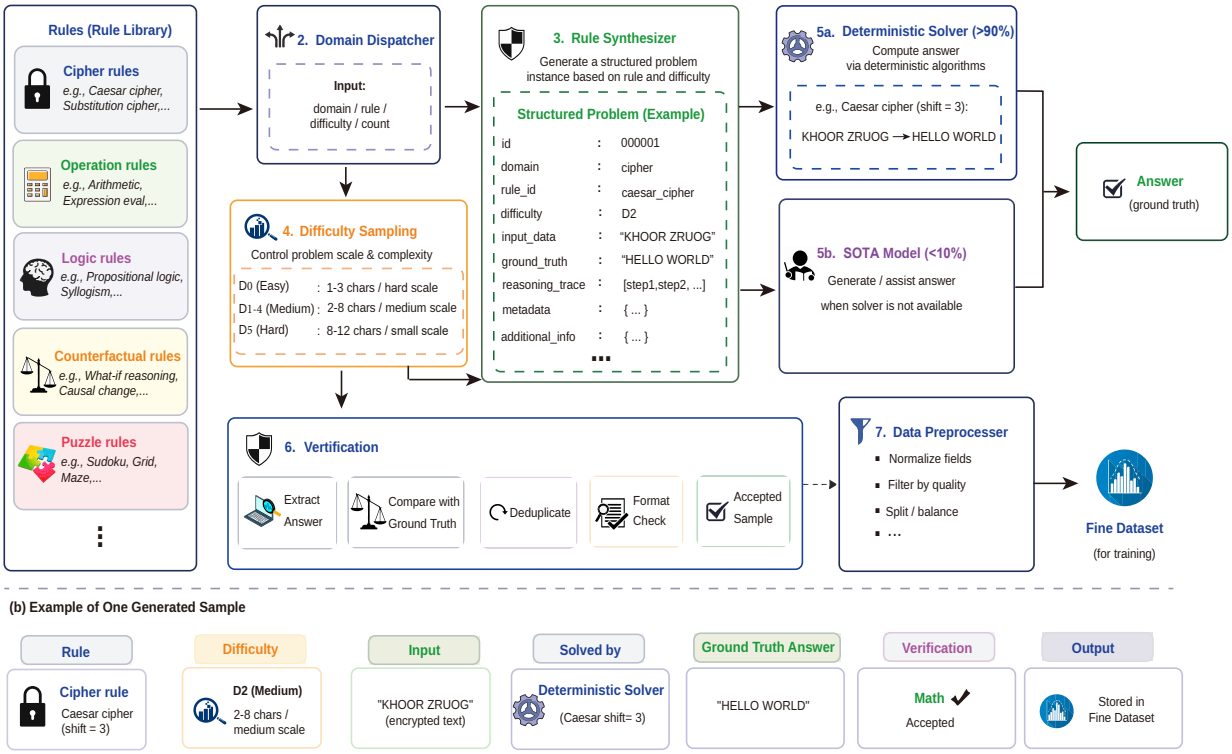

Figure 6: Synthetic data generation pipeline. The rule library supplies problem templates; the synthesizer generates fresh instances across the six difficulty tiers (Appendix B.2); answers and reasoning traces come from the deterministic solvers ($> 90\%$ of items) or, where no solver is available, from Claude Fable 5 ($< 10\%$); all outputs pass verification, split isolation, and deduplication.

Table 8: Per-domain accuracy change (Δ, pp) vs. the same-stage no-mid-training baseline at each pipeline stage, for Expt. 1 (imbalanced). Mid-only vs. Base; Mid+SFT vs. SFT only; Mid+SFT+RL vs. SFT+RL. All entries in pp. The mid-training-only collapse (Operation -27.6, Counterfactual -40.4) is reversed in sign by SFT/RL—strongly for Operation, and within seed noise for Counterfactual; a smaller residual deficit persists on Logic and Puzzle. The same SFT pass that reverses the two collapses also installs a larger deficit on Logic (-6.0 ± 0.8 pp at the mid+SFT stage), which is resolved at 7.5 times its cross-seed SD. The Puzzle SDs in the first two columns are bounded below by $|\mathrm{SD}(X) - \mathrm{SD}(Y)|$ for seed-matched rows (Table 1: $|1.5 - 0.7| = 0.8$ and $|1.4 - 2.0| = 0.6$) and are reported at that bound. Rows are five-seed means of per-seed differences against the matched-seed baseline.

Domain	Mid-only vs. Base	Mid+SFT vs. SFT only	Mid+SFT+RL vs. SFT+RL
Cipher	$+24.9 \pm 1.8$	$+4.0 \pm 1.0$	$+2.4 \pm 4.4$
Operation	-27.6 ± 1.1	$+5.2 \pm 0.3$	$+4.0 \pm 2.8$
Logic	$+5.3 \pm 0.9$	-6.0 ± 0.8	-2.4 ± 2.5
Counterfactual	-40.4 ± 2.4	-0.8 ± 1.2	$+1.2 \pm 3.4$
Puzzle	$+5.8 \pm 0.8$	$+1.2 \pm 0.6$	-2.0 ± 3.0
Overall	-6.40	$+0.72$	$+0.64$

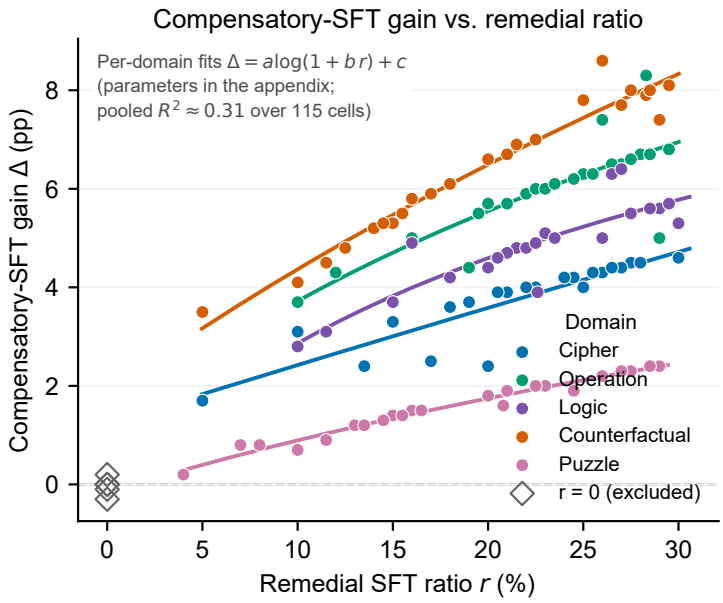

Figure 7: Compensatory-SFT gain Δ = Eval. (Remedial)−Evaluation vs. the domain's remedial SFT ratio r, from Table 16 (115 cells with $r > 0$, coloured by domain; the five $r = 0$ cells are open diamonds and excluded from the fits). Solid curves: per-domain fits $\Delta \approx a \log(1 + b\,r) + c$ (Table 20). The gain rises with r and saturates at a domain-specific ceiling; pooling across domains leaves most cell-level variance unexplained (pooled $R^2 \approx 0.31$). Descriptive, not confirmatory.

E Data Accounting

Coverage as measured in this study is the *token share* of each domain within the internal five-domain component (Table 16, Portion rows). Because the downstream recipe and the total token budget are held fixed, the coverage percentages fully determine the per-domain token allocation of that component; Table 9 records the corresponding allocation for the two configurations used in the full-pipeline comparison (Expt. 1 and the balanced allocation). The composition and provenance of the fixed external corpus component are described in §2.2; the internal/external token split and the per-component budgets are fixed across all configurations.

Table 9: Per-configuration data accounting for the mid-training coverage sweep. Coverage percentages are the measured internal-component token shares (Table 16); the internal/external token ratio and the per-component token budgets are fixed across all 24 configurations.

Quantity	Cipher	Operation	Logic	Counterfactual	Puzzles
Internal token share (e.g. balanced config)	20%	20%	20%	20%	20%
Internal token share (e.g. Expt. 1)	72.3%	0%	12.0%	0%	15.7%

F Decontamination and Split Isolation

1. **ProofWriter-family depth split.** The external mid-training corpus contains only regenerated ProofWriter-family items of depth at most 5; evaluation uses depths 5 to 9. Depth 5 is trained *and* evaluated—it is explicitly labeled as in distribution, and carries 18,498 of the 20,030 ProofWriter evaluation items (92%)—so the depth split isolates D6–D9 only, not D5.

2. **KOR-Bench instance split.** SFT and evaluation data use different random seeds. No identical instance appears in both; evaluation instances are held out by instance ID.

3. **External benchmark isolation.** No CounterBench-family items and no ProofWriter-family items of depths 6–9 appear in any training corpus. ZebraLogic-family items contribute 201 mid-training samples (0.3% of the total mid-training corpus, not of the internal five-domain component); these are distinct instances from the evaluation set but same-family exposure, so ZebraLogic is not a zero-exposure benchmark (Appendix I).

4. **Base model contamination.** We do not control for Qwen3-8B-Base's original pretraining exposure to public benchmarks, and this applies to every checkpoint reported here. The isolation claims in items 1–3—including the family-level "zero-exposure" label applied to CounterBench in §4.4—assert the absence of benchmark-family items from *our* training corpora only; they do not certify that the base model never encountered such items during pretraining.

5. **Model-branch isolation.** Answers and traces come from the deterministic solvers for the $> 90\%$ of items they cover and from Claude Fable 5 for the remainder. The model branch never sees an evaluation instance: it is invoked only on freshly synthesised pool items whose instance IDs lie outside every held-out evaluation split (Appendix F item 2), and no generated text is reused across the training/evaluation boundary.

G Mid-Training Coverage Sweep

Mid-training-only gains over Base are largest at low Base prior in the two extremes, though not monotonically: Puzzle (Base 5.8) gains only $+5.8$/$+6.7$/$+10.0$ pp across Expts. 1–3, Cipher (Base 13.2) gains $+24.9$/$+37.9$/$+38.0$ pp, and Operation (Base 60.8) exceeds Logic (Base 49.2) in two of the three.

Table 16 reports the full per-configuration results of the 24 coverage sweep variants (§4.1) together with the FineWeb-Edu baseline and the exploratory θ^* allocation. Each row block gives: Portion (domain allocation), Evaluation (mid-training-only accuracy, five-seed mean $\pm$ 1 SD), Remedial SFT (compensatory allocation from Eq. 5), Eval. (Remedial) (accuracy after the compensatory SFT pass), and Eval. (Uniform) (accuracy after the matched uniform-SFT control; Appendix H). All 24 configurations include both SFT-pass results. The external corpus derives 34.8% of its tokens from the ProofWriter rule family, contributing shared formal-deduction exposure that explains why Logic's curve sits above zero at 0% internal coverage; Puzzle has no zero-coverage cell (its minimum internal share in the sweep is 2%, where it scores near its floor).

Per-domain fitted-region interpretation. Cipher, Operation, and Logic have broad 95%-of-peak intervals (upper bounds near 30–36%), consistent with their large fitted σ_R values; Cipher's interval [7.4%, 36.0%] spans the wide plateau observed in Figure 2 rather than a sharp peak. Counterfactual's fitted 95%-of-peak interval [17.9%, 47.7%] is consistent with a broad fitted peak near 26.0%. Puzzle's interval [27.2%, 41.2%] requires relatively high coverage; its fitted peak accuracy (15.0%) is low, and whether a larger model could raise it remains open.

Simplex design validity. Each domain's x-coordinate in Figure 2 is its own internal proportion while the other four vary jointly. The curves describe exploratory mixture-level associations rather than isolated causal functions. A compositional or simplex-aware joint response-surface analysis (in the sense of compositional data analysis (Aitchison, 1986)) would be needed for stronger allocation claims; we report a preliminary such analysis in Appendix G.2, which finds a saddle rather than an interior maximum even after the 12 interior allocations of Appendix B.4 are added to the pool. A factorial design that independently varies each domain's coverage remains future work. Any monotonicity or concavity tests below are likewise descriptive diagnostics, and should not be read as population-level causal tests.

Concavity test (descriptive). We fit a quadratic $y_d = \beta_0 + \beta_1\theta + \beta_2\theta^2$ per domain ($n = 24$) as an exploratory shape summary. The negative $\hat{\beta}_2$ values in Table 10 indicate a concave trend within this sweep, but they are reported strictly as descriptive shape descriptors and are not tests of curvature: accuracy is a bounded proportion, observations share training configurations, and the five domain fits are not independent. A binomial/beta-binomial or clustered compositional analysis with multiplicity control would be required before making inferential claims about curvature; such an analysis is left for future work.

Table 10: Per-domain quadratic regression concavity check, descriptive only. OLS fit $y = \beta_0 + \beta_1\theta + \beta_2\theta^2$ on all 24 normal coverage configurations per domain. $\hat{\beta}_2$ is reported as a shape descriptor only and is *not* used as a curvature test; we therefore omit its standard error, which would invite a t-test we do not make. The calibrated test of interiority is the permutation test of Appendix G.1 (see §G).

Domain	$\hat{\beta}_2$	R^2_{quad}
Cipher	-0.013	0.494
Operation	-0.048	0.858
Logic	-0.033	0.816
Counterfactual	-0.034	0.979
Puzzle	-0.012	0.956

G.1 Model-Free Check: Coverage Bands

The split-Gaussian summaries below are five-parameter fits, so a reader may reasonably ask how much of the inverted-U shape is a property of the model rather than of the data. We therefore repeat the moderation claim without any curve fitting. Pooling all 30 trained allocations (the 24 sweep configurations plus the six held-out allocations), we bin each domain's own coverage into *low* ($< 10\%$), *moderate* (10–40%), and *high* ($> 40\%$) and compare the mean mid-training-only accuracy in each band (Table 11). Under these boundaries the moderate band is the best of the three for every one of the five domains, and accuracy is lower both below and above it. The boundaries are not independent of the fit—they were chosen after the fitted peaks were read, and they bracket the peak range 9.9–35.1%—so we report the boundary sensitivity explicitly rather than claiming the band result is fit-free: moving the upper boundary from 40% to 35% leaves Counterfactual's moderate band behind its high band by 0.1 pp, and moving it to 30% flips both Counterfactual and Puzzle, leaving three of five domains. The *low*-versus-moderate contrast is the half that survives every boundary we tested, and it is the half the moderation claim rests on. The high band is also thinly populated for several domains ($n = 1$–5); the 12 interior allocations of Appendix B.4, all in the moderate band, do not add high-coverage points but thicken the moderate band (to $n = 29$–32). Within these limits, the interior optimum is visible in the raw group means and does not depend on the parametric fit, and it survives the addition of the 12 interior allocations (moderate best for 5/5 domains over the 42-allocation pool). One bookkeeping point deserves stating plainly: the six held-out allocations serve two roles. They are withheld from the split-Gaussian curve fitting, which is what makes the out-of-sample test of §4.1 valid, but they are included in the 30-allocation pool used for the band comparison above, for the quadratic-interiority permutation test (which we also run on the 42-allocation pool, once Appendix B.4 is added), and for the compositional surface of §G.2, because those analyses are not fitted to the curves and benefit from the extra design points. No statistic is therefore both fitted on and validated against the same allocations, but the band and interiority results are in-sample with respect to the pool (30 or, once Appendix B.4 is included, 42) and should be read as such. A permutation test sharpens this into a calibrated statement. For each domain we fit a three-parameter quadratic OLS to the 42-allocation pool (30 base plus the 12 interior allocations of Appendix B.4) and ask whether it is concave ($\hat{\beta}_2 < 0$) with a vertex strictly inside the tested coverage range; all five domains satisfy both conditions. Under a null that permutes whole accuracy rows across allocations—destroying the coverage–accuracy link while leaving the simplex design and the cross-domain correlation structure exactly intact—an average of 1.78 of five domains do so, and all five do so in 0.98% of 20,000 draws ($P \approx 0.010$). We use the quadratic here only to test concavity and interiority, not to locate the optima: a symmetric parabola pulls its vertex toward the centre of the coverage range, so its vertex positions are not comparable to the split-Gaussian peaks reported in Table 13. Two further permutation results bear on how much the split-Gaussian fit itself can be trusted, and we report both, including the one that cuts against us. First, the fitted curves capture real structure: under the same whole-row permutation the mean per-domain R^2 of the split-Gaussian is 0.174 ± 0.055, against 0.982 for the observed data (no permutation draw out of 250 reached the observed value, $P < 0.004$). Second, and less favourably, the *count* of interior peaks is not a usable statistic for a five-parameter split-Gaussian: a permuted dataset still yields interior peaks in all five domains 64% of the time, because a flexible asymmetric curve can place a maximum inside the range even on noise. This is precisely why the interiority test above is run on the three-parameter quadratic, which cannot manufacture an interior optimum so easily, and why we do not treat "all five peaks are interior" as evidence on its own. Finally, the five fitted peaks sum to $\approx 102\%$, i.e. they are close to jointly realisable within the 100% budget. A referee might read a sum near the budget as the fingerprint of a simplex artifact; the permutation null does not support that reading, since permuted data give a peak sum of $121.5 \pm 33.3\%$ that does not concentrate near 100%, and only 2.8% of draws land as close to the budget

as the observed sum. We nonetheless treat the near-feasibility as a numerical coincidence worth reporting rather than as a design rule, because the sum of five marginal optima has no guaranteed relationship to the joint optimum under the constraint (§5.2).

Table 11: Mean mid-training-only accuracy (%) by coverage band, pooling all 30 trained allocations (24 sweep + 6 held-out). Each cell gives the mean, its SD across the allocations in that band, and the allocation count n; a blank SD marks a single-allocation cell, and the *Low, excl.* 0% column reports means without SDs because it is a recomputation of the preceding column rather than an independent band. The moderate band is the best of the three for all five domains under these boundaries; the boundaries themselves were chosen after the fitted peaks were read, and the result is not robust to moving them—at *low*<10, *high*>30 only three of five domains have the moderate band best, and at *high*>35 Counterfactual's moderate-band margin is 0.1 pp. No curve fitting is involved in the group means. The third column recomputes the low band with the four zero-coverage cells (Cipher, Operation, Logic and Counterfactual at 0%) removed. The low-versus-moderate gap then narrows from 7.2 to 4.5 pp (Cipher), 22.8 to 20.9 (Operation), 18.6 to 16.4 (Logic) and 23.7 to 22.2 (Counterfactual), and is unchanged for Puzzle, which has no zero-coverage cell.

Domain	Low ($<10\%$)	Low, excl. 0%	Moderate (10–40%)	High ($>40\%$)
Cipher	43.8 (9.1, n=9)	46.5 (n=8)	**51.0** (0.9, n=18)	42.2 (4.5, n=3)
Operation	51.9 (11.3, n=11)	53.7 (n=10)	**74.7** (4.1, n=17)	50.4 (9.8, n=2)
Logic	39.6 (9.9, n=9)	41.8 (n=8)	**58.1** (2.4, n=20)	40.0 (, n=1)
Counterfactual	57.5 (7.4, n=9)	59.0 (n=8)	**81.2** (4.2, n=20)	78.2 (, n=1)
Puzzle	6.1 (2.0, n=5)	6.1 (n=5)	**13.1** (2.1, n=20)	10.5 (3.2, n=5)

Is the decline above the band established, or only a plateau? The band comparison establishes the low-versus-moderate contrast. It does not, on its own, establish the decline above the band, and a saturating monotone response would produce a similar raw pattern, so we test the upper half directly rather than let the parametric fit carry it. Two checks, reported in Table 12.

First, the moderate-versus-high contrast per domain, as a two-group comparison using the across-allocation SDs of each band. It is individually resolved for *one* of the five domains: Cipher, $+8.8$ pp $[+0.5, +17.0]$. For Operation the point estimate is large ($+24.3$ pp) but rests on two high-band allocations and its interval spans zero; for Puzzle the contrast is $+2.6$ pp $[-1.3, +6.5]$; and for **Logic and Counterfactual the high band is a single allocation each**, so no interval is defined at all. The family-wise interiority test of the previous paragraph is a conjunction over five domains and remains the only calibrated statement about the decline; per domain it is not resolved for three of them.

Second, a shape comparison within one quadratic family and at equal parameter count: the best *monotone* member (vertex pinned at or beyond the right edge of the observed range) against the best *unimodal* member (vertex free), both scored by leave-one-allocation-out MAE. Allowing the curve to turn down improves held-out prediction in all five domains, by $+0.3$ to $+6.9$ pp MAE over the 30-allocation pool (the largest gains are Counterfactual $+6.9$ and Logic $+3.1$). We report both checks because they are not redundant and they point the same way only in part: the parametric comparison favours a declining shape everywhere, while the model-free contrast individually resolves the decline in one domain out of five. Nothing here revives the decline for Logic and Counterfactual, whose high bands cannot be tested with this design; what it supports is the family-wise statement, not five separate per-domain ones.

G.2 Compositional (Aitchison) Reanalysis

Because coverage is a closed composition, an analysis that treats the five shares as free variables is not strictly appropriate, and we report here the compositional analysis the simplex discussion (§5.2) calls for. We map each allocation to four isometric log-ratio (ilr) coordinates after multiplicative replacement of zero shares, and regress each domain's mid-training-only accuracy on a response surface in ilr space over the 42-allocation pool (the 30 sweep and held-out allocations plus the 12 interior allocations of Appendix B.4).

The result is informative in both directions. A full quadratic ilr surface fits well ($R^2 = 0.78$–0.96 per domain over the 42-allocation pool), which confirms that allocation carries real signal when modelled compositionally rather than marginally. But its stationary point is a saddle in every domain, not an interior maximum. The limitation on that classification is *shallowness* rather than estimator noise, and the two are worth separating. The saddle survives every one of the 210 leave-one-allocation-out refits and every one of $2{,}000$ draws in which each accuracy is perturbed by its own cross-seed SD; what is not resolved is the *magnitude* of the negative curvature, since the fitted Hessian eigenvalues

Table 12: Moderate-versus-high band contrast (left), and the monotone-versus-unimodal shape comparison (right), over the 30-allocation pool. *Left:* mean mid-training-only accuracy per band with the allocation count, the difference (moderate − high), and a Welch 95% interval from the across-allocation SDs; a positive difference means the high band is worse, and a dash marks the two domains whose high band is a single allocation, where no interval is defined. *Right:* leave-one-allocation-out MAE for the best monotone and best unimodal member of one quadratic family at equal parameter count (vertex pinned at or beyond the right edge versus vertex free); a negative gain means the unimodal fit predicts held-out allocations better. The decline is individually resolved for one domain, untestable for two, and not resolved for the remaining two.

	Band contrast (pp)				LOO MAE (pp)	
Domain	Moderate	High	Diff.	95% CI	Monotone	Unimodal
Cipher	51.0 (n=18)	42.2 (n=3)	+8.8	[+0.5, +17.0]	4.26	**3.97**
Operation	74.7 (n=17)	50.4 (n=2)	+24.3	[−5.7, +54.2]	12.87	**12.21**
Logic	58.1 (n=20)	40.0 (n=1)	—	—	12.74	**9.61**
Counterfactual	81.2 (n=20)	78.2 (n=1)	—	—	15.26	**8.33**
Puzzle	13.1 (n=20)	10.5 (n=5)	+2.6	[−1.3, +6.5]	3.37	**2.94**

span $|\lambda| \approx 4 \times 10^{-3}$ to 1.6×10^{-1} accuracy units per ilr^2 (largest-to-smallest ratio 2.7–9.0 across domains), i.e. the surface is close to flat along its decreasing directions. The eigenvalue spectrum is invariant to the choice of sequential binary partition, because ilr coordinates built from different partitions are related by orthogonal transformations, so the one unpublished choice in this analysis does not affect the classification. A reduced surface without cross terms (9 parameters) is more stable and predominantly concave: 15 of the 20 ilr curvature coefficients are negative (Logic 4 of 4; Cipher, Operation and Counterfactual 3 of 4; Puzzle 2 of 4), so the surface curves downward in most directions but not all.

Two limits keep the conclusion provisional, and neither is estimator variance. A quadratic response is a misspecification we do not test, so the fitted Hessian need not be the Hessian of the true surface; and a shallow negative direction means a small joint improvement over the per-domain marginals cannot be excluded. The saddle classification itself, however, is stable under the resampling we can do.

We therefore do not claim a joint interior optimum over the simplex. The moderation result of Appendix G.1 is a statement about per-domain marginals—each domain's own coverage band—and it stands on the band comparison, which involves no curve fitting, and the permutation test, neither of which requires a joint surface. Establishing a joint compositional optimum would require more allocations than the 42 trained here. The 12 interior allocations of Appendix B.4 were added precisely to test this and did not produce a joint maximum—every domain's stationary point remained a saddle—so we flag additional simplex coverage, not a different estimator, as the natural next design rather than a gap the present data can close.

G.3 Fitting Procedure

Peak locations carry a seed-level bootstrap interval: we resample the five per-seed evaluations with replacement, refit the split Gaussian on each resample, and report the 2.5–97.5 percentile range of $\hat{\mu}_d$ over 2,000 draws. These intervals capture seed and fit variability only; they are not held-out validation of the peak locations.

Each domain's dose–response curve is summarised by the following *asymmetric linear-space Gaussian* (split Gaussian):

$$f(\theta) = a \exp\left(-\tfrac{1}{2}\left(\tfrac{\theta-\mu}{\sigma(\theta)}\right)^2\right) + c,$$
$$\sigma(\theta) = \begin{cases} \sigma_L, & \theta \le \mu, \\ \sigma_R, & \theta > \mu, \end{cases} \tag{3}$$

where the two widths σ_L and σ_R are free parameters with no ordering imposed, with five free parameters (a, μ, σ_L, σ_R, c) per domain. In the fitted curves, $\sigma_R > \sigma_L$ (steeper left rise than right decline) for four domains, while Puzzle's fitted curve has $\sigma_L > \sigma_R$. Parameters are estimated by weighted nonlinear least squares (`scipy.optimize.curve_fit`). Configurations in which a domain's internal proportion is $\theta_d \le 5\%$ or $\theta_d \ge 55\%$ are assigned half-weight ($w = 0.5$);

all others receive $w = 1.0$. This heuristic down-weighting reduces the leverage of extreme, near-degenerate points, but the exact thresholds are not claimed to be optimal.

Peak locations μ and peak accuracies per domain are listed in Table 13. The 95%-of-peak moderate region for each domain is the set $\{\theta_d \mid f_d(\theta_d) \geq 0.95 \cdot f_d(\mu)\}$, solved numerically from the fitted split-Gaussian curve (Eq. 3); the resulting intervals are listed in Table 13. The peaks are stable to refitting, but they remain function-level thresholds for this sweep rather than confidence intervals, and are not validated allocation rules.

Domain	Peak accuracy (%)	Peak location μ (%)	95%-of-peak interval
Cipher	52.0	9.9	[7.4, 36.0]
Operation	78.1	15.9	[11.9, 29.4]
Logic	59.5	15.0	[11.1, 31.4]
Counterfactual	83.5	26.0	[17.9, 47.7]
Puzzle	15.0	35.1	[27.2, 41.2]

Table 13: Fitted 95%-of-peak intervals from split-Gaussian summaries. Each interval is the θ_d range satisfying $f_d(\theta_d) \geq 0.95 \cdot f_d(\mu)$; the threshold is inclusive ($\geq$) and bounds are displayed rounded to one decimal place. These are descriptive function-level thresholds from the Qwen3-8B/KOR-Bench sweep under the simplex constraint (see Appendix G); they are not confidence intervals and not validated allocation rules. For Cipher, the fitted curve is dominated by a broad plateau (≈51–53% over ≈8–24% coverage); the fitted peak at ≈9.9% is not a distinct mode and should not be read as a sharp optimum.

Sensitivity to the down-weighting threshold. Because the $w = 0.5$ threshold at $\theta_d \leq 5\%$ or $\theta_d \geq 55\%$ is a heuristic, we refit the split-Gaussian across four alternative thresholds ($\leq 0/ \geq 100$, $\leq 3/ \geq 57$, $\leq 5/ \geq 55$, $\leq 8/ \geq 52$). The fitted peaks are essentially unchanged: Cipher $\mu \in [9.6, 9.9]$, Operation $[15.9, 15.9]$, Logic $[14.8, 15.3]$, Counterfactual $[26.0, 26.2]$, Puzzle $[34.9, 35.4]$—a spread of at most 0.5 pp. The reported peaks are therefore not an artifact of this particular down-weighting choice; note that this sensitivity check was itself run inside the down-weighted framework, since the extreme points it re-weights are exactly the points that define each curve's right arm. The $\theta_d \leq 5\%$ and $\theta_d \geq 55\%$ thresholds are used throughout.

The illustrative allocation θ^* reported in §4.1 is obtained by solving:

$$\theta^* = \arg\max_{\theta} \sum_d \frac{f_d(\theta_d) - b_d}{b_d} \quad \text{s.t.} \quad \sum_d \theta_d = 100\%,\ \theta_d \geq 0, \tag{4}$$

where b_d is the FineWeb-Edu-only baseline accuracy, solved with Sequential Least Squares Programming (SLSQP; 2,000 random initialisations). Because this objective is defined on fitted marginal curves, the resulting allocation should be interpreted as a descriptive probe rather than a globally optimal mixture.

Reconciliation with Eq. 4. The θ^* reported in §4.1 sits on the fitted peaks, not strictly at the argmax of Eq. 4. The two differ because the objective is normalised by b_d: at the constrained optimum each domain is pulled off its own peak by $\delta_d \propto b_d / |f_d''(\mu_d)|$. Across the five domains b_d spans a factor of 43 (1.2 for Puzzle to 51.2 for Counterfactual) while the fitted 95%-of-peak widths span only a factor of 2.1, so the pull is largest on Counterfactual and smallest on Puzzle. Re-solving Eq. 4 on a reconstruction of the fitted curves—the fitted σ_L, σ_R are not published, so we recovered them from the tabulated 95%-of-peak endpoints, which are given to one decimal place and therefore pin each width only to about ± 0.6 pp—gives a peak offset of ≈ 1.2 pp on Counterfactual and ≈ 0.2 pp on Puzzle. Perturbing the widths across that rounding uncertainty leaves the argmax stable to within 0.1 pp on every coordinate, so the ordering is not a reconstruction artifact: Counterfactual's optimum sits 1.0–1.2 pp below its own peak while Puzzle's sits essentially on its peak (34.9–35.0%). The reported θ^* reverses this, offsetting Counterfactual by only 0.2 pp and Puzzle by 1.5 pp. The reported allocation is therefore the rounded fitted-peak allocation rather than the argmax; the two differ by at most 1.5 pp on any coordinate. We report it this way rather than publishing the exact argmax because the offset is smaller than the resolution the fit supports, and we flag the discrepancy here so that a reader re-solving Eq. 4 is not misled.

Goodness of fit. The split-Gaussian substantially outperforms a flat baseline on every domain. Weighted Root Mean Square Error (RMSE, in pp), split-Gaussian vs. weighted constant: Cipher 0.84 vs. 5.32; Operation 1.15 vs. 12.51; Logic 1.44 vs. 9.10; Counterfactual 1.18 vs. 10.60; Puzzle 0.63 vs. 3.28.

Model form comparison. Table 14 compares the split-Gaussian with a quadratic OLS alternative ($y = \beta_0 + \beta_1\theta + \beta_2\theta^2$, three parameters). The quadratic recovers concave shape with fewer assumptions and is the form we use for the calibrated interiority test of Appendix G.1, precisely because its three parameters cannot manufacture an interior optimum from noise; the $\hat{\beta}_2$ values of Table 10 are shape descriptors accompanying that test, not a curvature test in themselves. For describing the curves, the split-Gaussian is preferred on evidence rather than flexibility: it attains both the lower weighted in-sample RMSE and the lower leave-one-out MAE in all five domains, and the LOO-MAE comparison does not reward its extra parameters. It additionally captures the left–right asymmetry the quadratic cannot represent. Both forms agree that the curves are non-monotonic within this sweep, with the strongest fit for Counterfactual.

Table 14: Model fit comparison for the per-domain coverage–accuracy curves ($n = 24$). All three forms are scored on the same footing: RMSE is the weighted in-sample RMSE under the identical weighting scheme (Appendix G.3), and LOO-MAE is leave-one-out mean absolute error, which is the out-of-sample criterion and does not reward extra parameters. The split-Gaussian attains the lower LOO-MAE in all five domains, so its advantage is not an artifact of its larger parameter count. RMSE and LOO-MAE in pp.

	Flat baseline		Quadratic OLS		Split-Gaussian	
Domain	RMSE	LOO-MAE	RMSE	LOO-MAE	RMSE	LOO-MAE
Cipher	5.32	4.28	3.95	4.22	0.84	1.10
Operation	12.51	12.34	5.20	7.67	1.15	1.01
Logic	9.10	8.65	4.28	6.36	1.44	1.16
Counterfactual	10.60	10.49	1.72	1.95	1.18	1.09
Puzzle	3.28	3.16	0.71	0.66	0.63	0.55

Leave-one-out cross-validation. We apply LOO-CV over the 24 normal non-baseline configurations. For each fold we refit the split Gaussian on the other 23 configurations using the same deterministic multi-start procedure (four σ seed pairs, two intercept seeds, and five μ seeds drawn from the sorted-top-3 targets plus a uniform grid) and record the prediction error on the held-out configuration.

Table 15 reports LOO-MAE, the range of fitted μ across folds ($\Delta\mu$), and the fraction of held-out points within $\pm 2\times\mathrm{RMSE}_{\mathrm{full}}$ (in-band), all over the 24 normal non-baseline configurations. LOO-MAE is small and tight (0.55–1.16 pp), and peak location is stable across folds for every domain ($\Delta\mu \leq 1.78$ pp). These diagnostics assess interpolation for the finite sweep only; they do not validate peak locations, moderate bands, or the jointly selected θ^* on new mixtures.

Table 15: Leave-one-out cross-validation results for the split-Gaussian dose-response fits. LOO-MAE: mean absolute error on held-out; μ_{full}: peak from full-data fit; $\Delta\mu$ LOO: range of μ across LOO folds; In-band: fraction within $\pm 2\times\mathrm{RMSE}_{\mathrm{full}}$.

Domain	LOO-MAE (pp)	μ_{full} (%)	$\Delta\mu$ LOO (pp)	In-band
Cipher	1.10	9.9	1.04	21/24 (88%)
Operation	1.01	15.9	0.94	21/24 (88%)
Logic	1.16	15.0	1.78	20/24 (83%)
Counterfactual	1.09	26.0	0.58	22/24 (92%)
Puzzle	0.55	35.1	1.11	22/24 (92%)

One sweep configuration (Cipher 26%, Operation 16%, Logic 17%, Counterfactual 30%, Puzzle 11%), added after an initial set of eight configurations to broaden simplex coverage, realizes 82.4% Counterfactual accuracy at 30% Counterfactual coverage, within the cluster of Counterfactual results the sweep yields across 24–32% coverage (84.1%, 83.9%, 83.7% at 31%, 29% and 32% coverage; Table 16, all within seed noise of the fitted peak of 83.5%). Because these configurations cluster around the fitted level rather than singling out this design point, the fitted Counterfactual

peak (≈26%; Table 13) is supported by a group of points rather than by a single post-hoc observation. The pattern remains directionally consistent with the relatively higher Counterfactual allocation in the illustrative θ^* (a separate, analytically derived allocation trained as an independent checkpoint outside the 24 sweep configurations; Table 16), but a single configuration does not validate the fitted Counterfactual peak. To confirm that this design point does not drive the fitted curves, we refit all five domains with it removed. The peaks move by at most 0.20 pp (Cipher $9.95 \rightarrow 10.00$, Operation $15.90 \rightarrow 15.95$, Logic unchanged at 14.95, Counterfactual $26.05 \rightarrow 26.25$, Puzzle unchanged at 35.10) and their sum changes from 101.95% to 102.25%, so every peak location and every 95%-of-peak interval reported here is materially unchanged by excluding the post-hoc configuration.

Table 16: Mid-training coverage sweep results with the compensatory-SFT pass and its uniform-SFT control. Each row block reports: **Portion** (domain coverage proportions), **Evaluation** (mid-training-only accuracy), **Remedial SFT** (compensatory allocation from Eq. 5), **Eval. (Remedial)** (accuracy after the compensatory pass), and **Eval. (Uniform)** (accuracy after the matched uniform control, whose data mix is identical across configurations; Appendix H); the six held-out blocks additionally carry an **RL Evaluation** row (accuracy after the full mid-training+SFT+RL pipeline, Appendix H). Values are reported as mean ± 1 SD across the five seeds. The **Exploratory** θ^* block (Expt. 3 in Table 1) reports the analytically derived θ^* allocation (9.9/15.8/14.9/25.8/33.6), its target Remedial SFT allocation from Eq. 5, and the results of the dedicated re-run (five seeds); it does not coincide with any of the 24 sweep configurations. The **FineWeb-Edu** baseline block precedes six **held-out allocations** that are withheld from the split-Gaussian curve fit—and from that fit only: they remain in the 30-allocation pool behind the band comparison and the interiority permutation test, which are in-sample with respect to that pool (Appendix G.1)—and are used as out-of-sample validation (mid-training-only, SFT-stage and RL-stage results; §4.1, Appendix H).

Domains	Cipher	Operation	Logic	Counterfactual	Puzzle
Portion	72.3%	0.0%	12.0%	0.0%	15.7%
Evaluation	$38.1 \pm 2.6\%$	$33.2 \pm 2.8\%$	$54.5 \pm 2.2\%$	$45.6 \pm 2.8\%$	$11.6 \pm 1.5\%$
Remedial SFT	$0.0\%^{\dagger}$	28.3%	22.6%	28.3%	20.8%
Eval. (Remedial)	$38.3 \pm 0.5\%$	$41.5 \pm 0.4\%$	$58.4 \pm 0.8\%$	$53.5 \pm 0.8\%$	$13.2 \pm 0.3\%$
Eval. (Uniform)	$41.5 \pm 0.7\%$	$41.6 \pm 0.3\%$	$59.3 \pm 0.5\%$	$50.8 \pm 0.2\%$	$13.8 \pm 0.7\%$
Portion	26.0%	16.0%	17.0%	30.0%	11.0%
Evaluation	$50.5 \pm 2.1\%$	$77.9 \pm 2.3\%$	$59.8 \pm 1.9\%$	$82.4 \pm 3.2\%$	$8.9 \pm 1.0\%$
Remedial SFT	17.0%	22.0%	21.5%	15.0%	24.5%
Eval. (Remedial)	$53.0 \pm 0.6\%$	$83.8 \pm 0.4\%$	$64.6 \pm 1.0\%$	$87.7 \pm 0.8\%$	$10.8 \pm 0.3\%$
Eval. (Uniform)	$54.2 \pm 0.7\%$	$83.4 \pm 0.3\%$	$64.4 \pm 1.0\%$	$86.5 \pm 0.8\%$	$11.2 \pm 0.9\%$
Portion	30.0%	18.0%	16.0%	18.0%	18.0%
Evaluation	$50.9 \pm 2.9\%$	$77.6 \pm 2.8\%$	$60.4 \pm 1.6\%$	$79.3 \pm 1.8\%$	$12.0 \pm 1.4\%$
Remedial SFT	15.0%	21.0%	22.0%	21.0%	21.0%
Eval. (Remedial)	$54.2 \pm 0.6\%$	$83.3 \pm 0.3\%$	$65.2 \pm 0.6\%$	$86.0 \pm 1.0\%$	$13.9 \pm 0.5\%$
Eval. (Uniform)	$54.6 \pm 0.4\%$	$83.0 \pm 0.3\%$	$65.1 \pm 1.1\%$	$83.5 \pm 0.8\%$	$14.1 \pm 0.6\%$
Portion	20.0%	20.0%	20.0%	20.0%	20.0%
Evaluation	$51.1 \pm 2.2\%$	$77.2 \pm 2.9\%$	$59.3 \pm 2.2\%$	$81.2 \pm 3.1\%$	$12.5 \pm 2.0\%$
Remedial SFT	20.0%	20.0%	20.0%	20.0%	20.0%
Eval. (Remedial)	$53.5 \pm 0.9\%$	$82.9 \pm 0.3\%$	$63.7 \pm 0.5\%$	$87.8 \pm 0.9\%$	$14.3 \pm 0.2\%$
Eval. (Uniform)	$54.9 \pm 0.7\%$	$82.6 \pm 0.5\%$	$63.9 \pm 0.7\%$	$85.4 \pm 0.6\%$	$14.6 \pm 0.8\%$
Portion	0.0%	10.0%	40.0%	10.0%	40.0%
Evaluation	$22.0 \pm 2.3\%$	$67.3 \pm 3.8\%$	$52.8 \pm 1.7\%$	$66.9 \pm 2.0\%$	$15.2 \pm 0.8\%$
Remedial SFT	30.0%	25.0%	10.0%	25.0%	10.0%
Eval. (Remedial)	$26.6 \pm 0.7\%$	$73.6 \pm 0.3\%$	$55.6 \pm 1.0\%$	$74.7 \pm 0.5\%$	$15.9 \pm 0.4\%$
Eval. (Uniform)	$26.8 \pm 0.8\%$	$72.9 \pm 0.2\%$	$57.2 \pm 0.4\%$	$71.3 \pm 0.8\%$	$17.1 \pm 1.0\%$
Portion	10.0%	8.0%	37.0%	8.0%	37.0%
Evaluation	$51.2 \pm 2.5\%$	$64.5 \pm 3.3\%$	$54.3 \pm 2.5\%$	$65.2 \pm 2.8\%$	$14.9 \pm 1.0\%$
Remedial SFT	25.0%	26.0%	11.5%	26.0%	11.5%
Eval. (Remedial)	$55.2 \pm 0.8\%$	$71.9 \pm 0.5\%$	$57.4 \pm 0.5\%$	$73.8 \pm 0.6\%$	$15.8 \pm 0.1\%$
Eval. (Uniform)	$55.2 \pm 0.8\%$	$70.2 \pm 0.8\%$	$58.7 \pm 0.4\%$	$69.7 \pm 0.9\%$	$16.8 \pm 0.7\%$
Portion	5.0%	15.0%	5.0%	15.0%	60.0%
Evaluation	$45.7 \pm 1.6\%$	$77.7 \pm 3.1\%$	$44.8 \pm 2.8\%$	$75.7 \pm 2.3\%$	$5.3 \pm 0.9\%$
Remedial SFT	27.5%	22.5%	27.5%	22.5%	0.0%
Eval. (Remedial)	$50.2 \pm 0.5\%$	$83.7 \pm 0.4\%$	$50.3 \pm 0.7\%$	$82.7 \pm 0.4\%$	$5.0 \pm 0.5\%$
Eval. (Uniform)	$49.9 \pm 0.5\%$	$83.2 \pm 0.6\%$	$49.8 \pm 0.8\%$	$80.0 \pm 0.7\%$	$7.0 \pm 0.8\%$
Portion	50.0%	40.0%	0.0%	5.0%	5.0%
Evaluation	$47.0 \pm 1.8\%$	$65.7 \pm 2.7\%$	$22.1 \pm 2.7\%$	$57.5 \pm 3.0\%$	$8.0 \pm 1.1\%$
Remedial SFT	5.0%	10.0%	30.0%	27.5%	27.5%
Eval. (Remedial)	$48.7 \pm 0.7\%$	$69.4 \pm 0.5\%$	$27.4 \pm 0.8\%$	$65.5 \pm 0.4\%$	$10.3 \pm 0.1\%$

Continued on next page

Table 16 (continued)

Domains	Cipher	Operation	Logic	Counterfactual	Puzzle
Eval. (Uniform)	$50.5 \pm 0.5\%$	$70.9 \pm 0.6\%$	$27.7 \pm 1.0\%$	$62.6 \pm 0.9\%$	$10.5 \pm 0.9\%$
Portion	40.0%	4.0%	8.0%	40.0%	8.0%
Evaluation	$49.5 \pm 1.4\%$	$47.7 \pm 3.4\%$	$50.2 \pm 1.9\%$	$82.0 \pm 1.6\%$	$8.2 \pm 0.7\%$
Remedial SFT	10.0%	28.0%	26.0%	10.0%	26.0%
Eval. (Remedial)	$52.6 \pm 0.8\%$	$54.4 \pm 1.0\%$	$55.2 \pm 0.6\%$	$86.1 \pm 1.1\%$	$10.4 \pm 0.1\%$
Eval. (Uniform)	$53.1 \pm 0.4\%$	$55.6 \pm 0.4\%$	$55.1 \pm 0.5\%$	$86.0 \pm 0.9\%$	$10.6 \pm 0.9\%$
Portion	8.0%	36.0%	2.0%	2.0%	52.0%
Evaluation	$49.5 \pm 1.9\%$	$71.8 \pm 3.2\%$	$33.7 \pm 2.2\%$	$52.8 \pm 2.2\%$	$9.6 \pm 1.1\%$
Remedial SFT	26.0%	12.0%	29.0%	29.0%	4.0%
Eval. (Remedial)	$53.8 \pm 0.7\%$	$76.1 \pm 0.4\%$	$39.3 \pm 0.5\%$	$60.2 \pm 0.8\%$	$9.8 \pm 0.4\%$
Eval. (Uniform)	$53.6 \pm 0.6\%$	$77.0 \pm 0.7\%$	$38.9 \pm 0.9\%$	$60.6 \pm 0.4\%$	$11.4 \pm 0.6\%$
Portion	9.0%	14.0%	14.0%	17.0%	46.0%
Evaluation	$52.8 \pm 1.3\%$	$77.9 \pm 2.4\%$	$59.2 \pm 2.6\%$	$78.3 \pm 3.4\%$	$12.6 \pm 0.7\%$
Remedial SFT	25.5%	23.0%	23.0%	21.5%	7.0%
Eval. (Remedial)	$57.1 \pm 0.9\%$	$83.9 \pm 0.9\%$	$64.3 \pm 0.5\%$	$85.2 \pm 1.1\%$	$13.4 \pm 0.4\%$
Eval. (Uniform)	$56.8 \pm 0.4\%$	$83.4 \pm 0.6\%$	$63.9 \pm 0.7\%$	$82.5 \pm 0.4\%$	$14.4 \pm 0.6\%$
Portion	11.0%	2.0%	15.0%	28.0%	44.0%
Evaluation	$51.1 \pm 2.7\%$	$41.8 \pm 3.4\%$	$59.9 \pm 2.5\%$	$83.4 \pm 3.6\%$	$13.3 \pm 0.8\%$
Remedial SFT	24.5%	29.0%	22.5%	16.0%	8.0%
Eval. (Remedial)	$55.3 \pm 0.9\%$	$46.8 \pm 0.8\%$	$64.8 \pm 1.2\%$	$89.2 \pm 0.9\%$	$14.1 \pm 0.4\%$
Eval. (Uniform)	$55.1 \pm 0.5\%$	$47.8 \pm 0.3\%$	$64.6 \pm 0.6\%$	$87.5 \pm 0.3\%$	$15.1 \pm 0.7\%$
Portion	4.0%	3.0%	30.0%	29.0%	34.0%
Evaluation	$40.3 \pm 1.1\%$	$45.6 \pm 2.8\%$	$57.3 \pm 2.1\%$	$83.9 \pm 3.8\%$	$15.4 \pm 1.3\%$
Remedial SFT	28.0%	28.5%	15.0%	15.5%	13.0%
Eval. (Remedial)	$44.8 \pm 0.6\%$	$52.3 \pm 0.7\%$	$61.0 \pm 1.0\%$	$89.4 \pm 0.6\%$	$16.6 \pm 0.4\%$
Eval. (Uniform)	$44.6 \pm 0.3\%$	$51.5 \pm 0.1\%$	$61.8 \pm 0.3\%$	$88.0 \pm 0.5\%$	$17.3 \pm 0.5\%$
Portion	15.0%	6.0%	19.0%	32.0%	28.0%
Evaluation	$52.0 \pm 2.9\%$	$57.3 \pm 2.3\%$	$59.4 \pm 2.9\%$	$83.7 \pm 1.7\%$	$14.7 \pm 1.4\%$
Remedial SFT	22.5%	27.0%	20.5%	14.0%	16.0%
Eval. (Remedial)	$56.0 \pm 0.5\%$	$63.8 \pm 0.7\%$	$64.0 \pm 1.2\%$	$88.9 \pm 1.1\%$	$16.2 \pm 0.2\%$
Eval. (Uniform)	$55.9 \pm 0.8\%$	$63.1 \pm 0.8\%$	$64.0 \pm 0.5\%$	$87.7 \pm 0.6\%$	$16.7 \pm 0.9\%$
Portion	60.0%	22.0%	13.0%	3.0%	2.0%
Evaluation	$41.6 \pm 1.4\%$	$76.9 \pm 2.1\%$	$57.6 \pm 2.4\%$	$52.1 \pm 2.4\%$	$4.2 \pm 1.0\%$
Remedial SFT	0.0%	19.0%	23.5%	28.5%	29.0%
Eval. (Remedial)	$41.6 \pm 0.8\%$	$81.3 \pm 0.8\%$	$62.6 \pm 0.8\%$	$60.1 \pm 1.0\%$	$6.6 \pm 0.1\%$
Eval. (Uniform)	$45.0 \pm 0.3\%$	$82.3 \pm 0.6\%$	$62.3 \pm 0.8\%$	$56.8 \pm 0.6\%$	$6.9 \pm 0.7\%$
Portion	12.0%	60.0%	24.0%	1.0%	3.0%
Evaluation	$52.9 \pm 1.7\%$	$43.5 \pm 3.2\%$	$58.6 \pm 1.8\%$	$52.1 \pm 2.2\%$	$4.2 \pm 1.2\%$
Remedial SFT	24.0%	0.0%	18.0%	29.5%	28.5%
Eval. (Remedial)	$57.1 \pm 0.6\%$	$43.5 \pm 0.7\%$	$62.8 \pm 1.2\%$	$60.2 \pm 0.7\%$	$6.6 \pm 0.2\%$
Eval. (Uniform)	$56.9 \pm 0.5\%$	$48.5 \pm 0.4\%$	$63.1 \pm 0.6\%$	$57.1 \pm 0.5\%$	$6.8 \pm 0.8\%$
Portion	6.0%	28.0%	28.0%	24.0%	14.0%
Evaluation	$45.2 \pm 2.3\%$	$75.2 \pm 3.3\%$	$57.5 \pm 1.7\%$	$83.6 \pm 3.9\%$	$10.1 \pm 0.6\%$
Remedial SFT	27.0%	16.0%	16.0%	18.0%	23.0%
Eval. (Remedial)	$49.6 \pm 0.5\%$	$80.2 \pm 0.5\%$	$62.4 \pm 1.1\%$	$89.7 \pm 0.5\%$	$12.1 \pm 0.5\%$
Eval. (Uniform)	$49.4 \pm 0.4\%$	$80.5 \pm 0.4\%$	$62.0 \pm 0.9\%$	$87.7 \pm 1.0\%$	$12.3 \pm 0.5\%$
Portion	7.0%	11.0%	18.0%	37.0%	27.0%
Evaluation	$48.0 \pm 1.6\%$	$72.8 \pm 2.2\%$	$59.7 \pm 2.8\%$	$83.2 \pm 3.7\%$	$14.4 \pm 0.9\%$
Remedial SFT	26.5%	24.5%	21.0%	11.5%	16.5%
Eval. (Remedial)	$52.4 \pm 0.8\%$	$79.0 \pm 0.5\%$	$64.4 \pm 0.9\%$	$87.7 \pm 0.6\%$	$15.9 \pm 0.5\%$
Eval. (Uniform)	$52.1 \pm 0.6\%$	$78.4 \pm 0.5\%$	$64.3 \pm 0.7\%$	$87.2 \pm 0.7\%$	$16.4 \pm 1.0\%$
Portion	18.0%	1.0%	60.0%	6.0%	15.0%
Evaluation	$52.3 \pm 2.5\%$	$37.7 \pm 2.7\%$	$40.0 \pm 1.6\%$	$64.6 \pm 3.8\%$	$10.2 \pm 1.0\%$
Remedial SFT	21.0%	29.5%	0.0%	27.0%	22.5%
Eval. (Remedial)	$56.2 \pm 0.8\%$	$44.5 \pm 0.5\%$	$39.9 \pm 0.9\%$	$72.3 \pm 0.5\%$	$12.2 \pm 0.2\%$
Eval. (Uniform)	$56.1 \pm 0.7\%$	$43.9 \pm 0.8\%$	$44.2 \pm 0.8\%$	$70.2 \pm 0.3\%$	$12.4 \pm 0.9\%$
Portion	24.0%	13.0%	7.0%	26.0%	30.0%
Evaluation	$51.4 \pm 2.6\%$	$77.9 \pm 3.7\%$	$46.6 \pm 2.3\%$	$83.0 \pm 3.1\%$	$14.8 \pm 0.6\%$

Continued on next page

Table 16 (continued)

Domains	Cipher	Operation	Logic	Counterfactual	Puzzle
Remedial SFT	18.0%	23.5%	26.5%	17.0%	15.0%
Eval. (Remedial)	$55.0 \pm 0.8\%$	$84.0 \pm 0.6\%$	$52.9 \pm 0.9\%$	$88.9 \pm 0.7\%$	$16.2 \pm 0.2\%$
Eval. (Uniform)	$55.1 \pm 0.6\%$	$83.4 \pm 0.7\%$	$51.5 \pm 0.6\%$	$87.1 \pm 0.3\%$	$16.8 \pm 0.8\%$
Portion	19.0%	7.0%	19.0%	24.0%	31.0%
Evaluation	$51.3 \pm 1.2\%$	$60.2 \pm 3.6\%$	$59.7 \pm 2.4\%$	$82.5 \pm 1.7\%$	$15.1 \pm 0.8\%$
Remedial SFT	20.5%	26.5%	20.5%	18.0%	14.5%
Eval. (Remedial)	$55.2 \pm 0.9\%$	$66.7 \pm 1.0\%$	$64.3 \pm 0.7\%$	$88.6 \pm 0.5\%$	$16.4 \pm 0.3\%$
Eval. (Uniform)	$55.1 \pm 0.6\%$	$65.9 \pm 0.7\%$	$64.3 \pm 1.0\%$	$86.6 \pm 0.5\%$	$17.1 \pm 0.6\%$
Portion	22.0%	9.0%	1.0%	35.0%	33.0%
Evaluation	$50.8 \pm 2.1\%$	$67.4 \pm 2.5\%$	$29.2 \pm 2.0\%$	$82.6 \pm 3.6\%$	$15.0 \pm 0.9\%$
Remedial SFT	19.0%	25.5%	29.5%	12.5%	13.5%
Eval. (Remedial)	$54.5 \pm 0.6\%$	$73.7 \pm 0.6\%$	$34.9 \pm 0.8\%$	$87.4 \pm 0.9\%$	$16.2 \pm 0.5\%$
Eval. (Uniform)	$54.6 \pm 0.7\%$	$73.0 \pm 0.2\%$	$34.6 \pm 0.3\%$	$86.6 \pm 0.5\%$	$16.9 \pm 0.8\%$
Portion	16.0%	21.0%	3.0%	31.0%	29.0%
Evaluation	$52.3 \pm 1.9\%$	$76.7 \pm 3.8\%$	$35.3 \pm 2.6\%$	$84.1 \pm 2.6\%$	$12.7 \pm 1.3\%$
Remedial SFT	22.0%	19.5%	28.5%	14.5%	15.5%
Eval. (Remedial)	$56.3 \pm 0.5\%$	$82.2 \pm 0.3\%$	$40.9 \pm 1.0\%$	$89.4 \pm 1.1\%$	$14.1 \pm 0.1\%$
Eval. (Uniform)	$56.2 \pm 0.6\%$	$82.1 \pm 0.2\%$	$40.4 \pm 0.5\%$	$88.2 \pm 0.7\%$	$14.7 \pm 1.0\%$
Portion	33.0%	5.0%	6.0%	50.0%	6.0%
Evaluation	$49.5 \pm 2.0\%$	$57.1 \pm 3.0\%$	$49.0 \pm 2.2\%$	$78.2 \pm 1.8\%$	$5.8 \pm 1.4\%$
Remedial SFT	13.5%	27.5%	27.0%	5.0%	27.0%
Eval. (Remedial)	$51.9 \pm 0.5\%$	$63.7 \pm 0.6\%$	$55.4 \pm 0.6\%$	$81.7 \pm 0.5\%$	$8.1 \pm 0.4\%$
Eval. (Uniform)	$53.1 \pm 0.5\%$	$62.9 \pm 0.5\%$	$54.0 \pm 1.0\%$	$82.1 \pm 0.8\%$	$8.3 \pm 0.5\%$
Exploratory θ (Expt. 3): Cipher 9.9%, Oper. 15.8%, Logic 14.9%, Counterf. 25.8%, Puzzle 33.6%; dedicated re-run, five seeds.*					
Portion	9.9%	15.8%	14.9%	25.8%	33.6%
Evaluation	$51.2 \pm 2.1\%$	$78.0 \pm 2.8\%$	$61.4 \pm 2.1\%$	$84.8 \pm 3.1\%$	$15.8 \pm 1.9\%$
Remedial SFT	25.1%	22.1%	22.6%	17.1%	13.2%
Eval. (Remedial)	$55.7 \pm 0.9\%$	$84.2 \pm 0.9\%$	$66.2 \pm 1.1\%$	$90.3 \pm 0.9\%$	$17.2 \pm 0.2\%$
Eval. (Uniform)	$55.2 \pm 0.3\%$	$83.6 \pm 0.9\%$	$66.1 \pm 0.9\%$	$88.9 \pm 0.4\%$	$17.7 \pm 0.6\%$
FineWeb-Edu only (100%); no reasoning-domain data.					
Evaluation	$4.4 \pm 2.2\%$	$45.6 \pm 2.9\%$	$32.0 \pm 2.8\%$	$51.2 \pm 2.3\%$	$1.2 \pm 1.2\%$
6 held-out allocations (withheld from the curve fit; used for out-of-sample validation).					
Portion	10.0%	10.0%	15.0%	30.0%	35.0%
Evaluation	$50.1 \pm 1.7\%$	$69.2 \pm 2.4\%$	$60.4 \pm 1.2\%$	$82.5 \pm 1.6\%$	$13.8 \pm 1.7\%$
Remedial SFT	25.0%	25.0%	22.5%	15.0%	12.5%
Eval. (Remedial)	$55.3 \pm 1.1\%$	$74.8 \pm 2.7\%$	$64.9 \pm 1.9\%$	$87.7 \pm 0.9\%$	$14.9 \pm 1.2\%$
Eval. (Uniform)	$53.1 \pm 1.3\%$	$75.2 \pm 1.7\%$	$65.6 \pm 1.2\%$	$85.9 \pm 1.8\%$	$16.1 \pm 3.4\%$
RL Evaluation	$55.9 \pm 2.1\%$	$75.9 \pm 3.2\%$	$66.1 \pm 2.8\%$	$88.8 \pm 3.9\%$	$15.2 \pm 4.3\%$
Portion	9.7%	50.0%	10.3%	15.0%	15.0%
Evaluation	$49.8 \pm 2.0\%$	$57.3 \pm 1.7\%$	$55.0 \pm 2.3\%$	$77.2 \pm 1.9\%$	$10.7 \pm 0.9\%$
Remedial SFT	25.2%	5.0%	24.8%	22.5%	22.5%
Eval. (Remedial)	$54.0 \pm 1.5\%$	$59.1 \pm 1.8\%$	$60.7 \pm 1.4\%$	$84.7 \pm 2.1\%$	$12.6 \pm 1.3\%$
Eval. (Uniform)	$53.9 \pm 1.2\%$	$61.8 \pm 1.6\%$	$59.9 \pm 1.3\%$	$81.9 \pm 1.5\%$	$13.0 \pm 1.0\%$
RL Evaluation	$55.2 \pm 2.2\%$	$61.0 \pm 2.9\%$	$62.8 \pm 4.2\%$	$85.2 \pm 3.1\%$	$13.3 \pm 2.5\%$
Portion	15.0%	11.2%	20.0%	8.8%	45.0%
Evaluation	$50.5 \pm 1.3\%$	$79.1 \pm 2.7\%$	$58.3 \pm 1.9\%$	$67.4 \pm 1.5\%$	$11.8 \pm 1.2\%$
Remedial SFT	22.5%	24.4%	20.0%	25.6%	7.5%
Eval. (Remedial)	$54.4 \pm 1.2\%$	$85.0 \pm 4.2\%$	$62.9 \pm 1.6\%$	$75.0 \pm 2.8\%$	$12.2 \pm 1.4\%$
Eval. (Uniform)	$54.5 \pm 1.9\%$	$84.9 \pm 1.7\%$	$63.2 \pm 1.7\%$	$72.8 \pm 1.9\%$	$13.7 \pm 2.5\%$
RL Evaluation	$55.2 \pm 3.1\%$	$86.1 \pm 4.2\%$	$65.7 \pm 4.5\%$	$75.6 \pm 1.8\%$	$13.2 \pm 3.5\%$
Portion	4.2%	15.0%	15.8%	25.0%	40.0%
Evaluation	$40.8 \pm 4.0\%$	$75.1 \pm 2.5\%$	$62.2 \pm 2.8\%$	$83.2 \pm 2.1\%$	$12.5 \pm 1.4\%$
Remedial SFT	27.9%	22.5%	22.1%	17.5%	10.0%
Eval. (Remedial)	$45.3 \pm 1.7\%$	$81.0 \pm 1.0\%$	$67.1 \pm 3.5\%$	$89.2 \pm 1.2\%$	$12.9 \pm 1.4\%$
Eval. (Uniform)	$45.0 \pm 1.6\%$	$80.8 \pm 3.2\%$	$67.0 \pm 1.0\%$	$87.5 \pm 2.9\%$	$14.5 \pm 2.4\%$
RL Evaluation	$47.0 \pm 2.3\%$	$81.8 \pm 2.6\%$	$68.2 \pm 3.1\%$	$90.1 \pm 1.7\%$	$14.9 \pm 3.1\%$
Portion	20.0%	4.6%	25.0%	30.4%	20.0%
Evaluation	$50.9 \pm 1.2\%$	$57.9 \pm 1.3\%$	$57.0 \pm 2.5\%$	$85.1 \pm 1.8\%$	$11.8 \pm 0.3\%$

Continued on next page

Table 16 (continued)

Domains	Cipher	Operation	Logic	Counterfactual	Puzzle
Remedial SFT	20.0%	27.7%	17.5%	14.8%	20.0%
Eval. (Remedial)	$54.5 \pm 2.3\%$	$64.4 \pm 2.0\%$	$61.2 \pm 1.7\%$	$90.5 \pm 1.1\%$	$13.5 \pm 2.1\%$
Eval. (Uniform)	$54.8 \pm 1.2\%$	$63.9 \pm 1.3\%$	$61.5 \pm 1.0\%$	$89.2 \pm 0.7\%$	$14.0 \pm 3.0\%$
RL Evaluation	$55.2 \pm 2.7\%$	$65.1 \pm 2.3\%$	$62.6 \pm 2.8\%$	$90.9 \pm 2.0\%$	$14.1 \pm 2.4\%$
Portion	25.0%	30.0%	7.7%	5.0%	32.3%
Evaluation	$50.0 \pm 1.4\%$	$73.1 \pm 1.9\%$	$45.3 \pm 2.2\%$	$60.2 \pm 2.5\%$	$15.6 \pm 0.7\%$
Remedial SFT	17.5%	15.0%	26.2%	27.5%	13.8%
Eval. (Remedial)	$56.3 \pm 2.2\%$	$78.0 \pm 3.6\%$	$50.9 \pm 1.7\%$	$68.1 \pm 3.4\%$	$16.8 \pm 2.3\%$
Eval. (Uniform)	$56.9 \pm 1.5\%$	$77.3 \pm 1.3\%$	$50.5 \pm 1.4\%$	$65.9 \pm 1.3\%$	$17.2 \pm 1.0\%$
RL Evaluation	$57.0 \pm 4.1\%$	$79.3 \pm 2.9\%$	$53.7 \pm 2.9\%$	$71.0 \pm 2.9\%$	$17.4 \pm 3.7\%$

† Domains whose mid-training coverage is at or above the 60% threshold receive $r = 0$: Configs 1 (Cipher 72.3%), 7 (Puzzle 60%), 15 (Cipher 60%), 16 (Operation 60%), and 19 (Logic 60%). Config 1 Cipher's zero-allocation cell gains only $+0.2$ pp (§H). Values are mean $\pm$ 1 SD across the five seeds.

G.4 Out-of-Sample Predictive Validity of the Fitted Curves

The six held-out allocations (Table 16, held-out blocks) are trained at allocations withheld from the curve fitting and carried through the complete mid-training+SFT+RL pipeline. The split-Gaussian fits generalize to them: mean absolute held-out residuals are 1.1 pp for Counterfactual (all six points within ± 2 fit-$RMSE$), 1.0 pp for Puzzle, 1.3 pp for Cipher, and 1.8 pp for Logic (five of six), while Operation is the weakest match (3.0 pp; two low-coverage points realize $\approx +6$ pp above the curve), so Operation's left tail rises somewhat faster than the split-Gaussian assumes. Applying the same fixed compensatory budget to the six held-out checkpoints gives gains in all 30 configuration–domain cells (mean $+4.5$ pp) yet closes $0/60$ pairs at the 5 pp threshold and only 5 of 59 rank-consistent pairs at the 10% ratio (uniform control: $0/60$ and $8/60$), and compensation's only material advantage over uniform remains Counterfactual ($+6.6$ vs. $+4.6$ pp). The RL stage adds $+1.19$ pp on average—in line with the $+1.1$ to $+2.3$ pp seen in the main experiments—and leaves the picture unchanged: after the complete pipeline the held-out allocations still close $0/60$ pairs at the 5 pp threshold and $7/60$ at the 10% ratio.

The held-out allocations also provide the only direct test of the premise underlying θ^*: that curves fitted at the mid-training stage carry information about *final* accuracy. Fitting on the 24 sweep configurations alone and predicting the six withheld allocations, the predicted overall accuracy tracks the realized mid-training-only value almost exactly (Pearson $r = +0.954$, Spearman $\rho = +1.000$, mean absolute error 0.64 pp) and—the point that matters for allocation choice—it also predicts the realized *post-RL* overall accuracy after the complete mid-training+SFT+RL pipeline ($r = +0.967$, $p = 0.002$; $\rho = +0.886$). The mid-training ranking is largely preserved through post-training ($r = +0.975$ between mid-only and post-RL overall), so a coverage choice made before SFT is still visible after RL. Reading those two numbers together matters for what the fit contributes: at $+0.975$ the *realized* mid-training-only ranking predicts post-RL accuracy slightly better than the fitted curves do ($+0.967$), so at $n = 6$ the curve-fitting step adds no measurable predictive value over simply evaluating the six mid-training checkpoints and ranking them, and the two correlations are not distinguishable at this sample size. What the held-out arm establishes is therefore that the mid-training ranking survives the downstream stages, not that the parametric curves are what makes it predictable. The predictive relationship is not carried by any single held-out point: dropping each of the six in turn moves the mid-training-only correlation only within $[+0.941, +0.973]$ and the post-RL correlation within $[+0.954, +0.987]$. At $n = 6$ the resampling intervals are wide and we do not read them as tight bounds: a percentile bootstrap over the six allocations spans most of the admissible range for both correlations, and we report the leave-one-out ranges above as the more informative stability check.

For completeness, θ^* reaches 62.72 overall on the compensatory-SFT scale while the six held-out allocations span 54.02–59.52, but this is *not* an out-of-sample validation of θ^*: θ^* was chosen as the argmax of the fitted objective and is evaluated here under a different downstream recipe (rule-type-balanced SFT in Table 1 versus the compensatory pass applied to the held-out checkpoints), so the comparison records the internal consistency of an argmax across two SFT recipes rather than evidence that a fit-selected allocation generalises. One component does *not* survive this test: the relative-gain objective of Eq. 4, which is what actually selects θ^*, ranks the six held-out allocations only weakly ($r = +0.441$, $p = 0.38$; $\rho = +0.257$), and unlike the predicted-accuracy correlations it is fragile: leaving out one allocation moves it anywhere in $[+0.058, +0.750]$ and its bootstrap interval $[-0.875, +0.976]$ spans zero. The baseline normalisation b_d appears to be what degrades it, so the fitted curves are the trustworthy object here while the

b_d-normalised objective layered on top of them is not; an allocation rule should optimise predicted accuracy directly rather than the relative-gain objective. What remains outside this test is θ^*'s own full-pipeline row on the matched protocol: its $+4.36$ pp figure in Table 1 follows the standard rule-type-balanced SFT, whereas the held-out allocations were carried through the compensatory pass, so the two are not run on identical downstream recipes.

H Compensatory SFT: Domain-Augmented Alignment Does Not Close Mid-Training Gaps

H.1 Experimental Design

To test whether *compensatory* SFT can close inter-domain gaps, we construct a domain-weighted SFT dataset:

$$r_d = \frac{\max(0,\ 60 - M_d)}{\sum_{j=1}^{5} \max(0,\ 60 - M_j)} \times 100\%, \tag{5}$$

where M_d is the mid-training token percentage for domain d (read directly from each coverage configuration) and r_d is the target SFT proportion for domain d.

One feature of Eq. 5 deserves stating plainly, because it bounds how the compensatory arm should be read. Its threshold $T = 60\%$ lies above *every* fitted peak in this sweep (the peaks span 9.9–35.1%), so the rule treats any share below 60% as a shortfall and directs remedial weight toward domains that the fitted curves already place at or beyond their own optimum. The rule is therefore a coverage-shortfall rule calibrated to a 60% reference share rather than one that targets each domain's fitted optimum: it is not an instantiation of θ^*, and the compensatory arm should be read as a generic deficit-targeted reweighting rather than as the policy the moderation result would recommend. Domains whose mid-training coverage is at or above 60% receive no compensatory SFT allocation ($r_d = 0$); all other domains receive SFT budget in proportion to their shortfall from the 60% threshold, with the allocation vector normalised to sum to 100%. The threshold 60% was chosen because any single domain exceeding 60% of the five-domain token budget already constitutes extreme over-allocation; every domain below this threshold is treated as under-served and eligible for remediation. Applying Eq. 5 to all 24 configurations, five configurations have a domain at or above the 60% threshold (Configs 1, 7, 15, 16, and 19; see Table 16); in all other configurations, all five domains fall below 60% and therefore receive positive compensatory SFT. SFT starts from θ_{mid} in the Mid-training+SFT+RL configuration and from θ_0 in the SFT+RL baseline, so the two arms differ only by the inserted mid-training checkpoint.

The row pairing behind Table 1 is worth stating because it is what makes each comparison like-for-like: the SFT data is identical across the eight SFT-bearing rows (the two shared baselines plus the Mid-training+SFT and Mid-training+SFT+RL rows of each of the three experiments), so Mid-training+SFT vs. SFT only isolates the effect of inserting a mid-training checkpoint, and after RL the analogue is Mid-training+SFT+RL vs. SFT+RL (Table 8).

Total SFT token count is held equal to the uniform baseline so that only the cross-domain distribution changes, not the overall training volume. Two scope limits on the resulting non-closure follow from this design. First, the supervised stage is deliberately the smaller of the two: the compensatory and uniform passes train on ≈0.3B tokens against mid-training's ≈1.5B, a ratio of roughly 1:5 (Table 5). That ratio is chosen so that mid-training remains the dominant stage of the paradigm under study—a compensation pass large enough to rival mid-training would relocate the allocation decision into the alignment stage rather than test whether alignment can repair the mid-training one—while remaining large enough that the cross-domain distribution of the SFT data is consequential: the compensatory arm is a ≈0.3B-token intervention on a mixture set by mid-training, not a token-level no-op, and it moves per-domain accuracy by up to $+6.4$ pp (Table 19). The tested quantity is the cross-domain *placement* of that fixed supervised budget, not the supervised dose, which is never varied across arms; whether a substantially larger supervised budget would close the gaps is therefore untested here. Second, all counts above are computed between a mid-training checkpoint and its post-SFT successor; Appendix H.2 reports the same metric on the no-mid-training path, which supplies that missing contrast class. The uniform-SFT control uses the identical rule-type-balanced data mix for every configuration, with the same total token budget and the same 3-epoch schedule as the compensatory pass, so the two passes differ only in the cross-domain allocation of the SFT data; both are separate supervised runs from the SFT rows of Table 1 (Appendix B.1) and are compared only against each other.

H.2 Capability Gap Metric: Pairwise Closure Framework

We evaluate whether compensatory SFT closes inter-domain gaps by comparing every unordered pair of domains within each of the 24 coverage-sweep checkpoints, giving $24 \times \binom{5}{2} = 240$ pairs. The pair members are ordered by their post-compensatory accuracies (see below), so the gap is always reported as the stronger-minus-weaker domain after compensation. Because pairs share both configurations and domains, all pairwise closure summaries are descriptive and require configuration-clustered inference for population-level claims. For each pair (i, j) with mid-training accuracies a_i^{mid}, a_j^{mid} and post-compensatory-SFT accuracies a_i^{comp}, a_j^{comp}, sorted such that $a_i^{\text{comp}} \leq a_j^{\text{comp}}$, we compute two gap measures:

$$G_{ij}^{\text{diff}}(a) = a_j - a_i, \tag{6}$$

$$G_{ij}^{\text{ratio}}(a) = 1 - \frac{a_i}{a_j}. \tag{7}$$

The closure rate for gap measure G is

$$C_{ij} = \frac{G_{ij}(a^{\text{mid}}) - G_{ij}(a^{\text{comp}})}{G_{ij}(a^{\text{mid}})}. \tag{8}$$

A pair is counted as *bridged* (meaningfully repaired) if the gap narrows by at least a metric-specific margin: an absolute reduction of at least 5 pp in the difference gap, or a relative closure rate C_{ij} of at least 10% in the ratio gap. A stricter 18% ratio threshold is examined only as a sensitivity check. These thresholds are calibrated to the scale of the intervention and of the seed noise in this sweep rather than to a statistical null (§H), and we interpret the resulting bridged-pair counts descriptively. Equal absolute gains naturally produce larger relative gains for weaker domains. A pair whose members reverse their relative order between the mid-training and post-compensatory checkpoints deserves a word, because it is the one case the two metrics handle differently. Such a pair has a negative pre-SFT ratio gap, $G_{ij}^{\text{ratio}}(a^{\text{mid}}) < 0$, so dividing by it would change the sign of C_{ij} and could register a spurious closure. The ratio denominator is therefore floored at 10^{-3}, which turns a reversed pairś closure rate into a large negative number: *a rank-crossed pair can never be counted as bridged*. Seven of the 240 pairs reverse in this way (seven under the uniform control; one of the 60 held-out pairs); all seven stay in the denominator, so the ratio counts are taken over the full 240 pairs, with 0 of the 7 reversed pairs bridged. No pre-SFT gap is exactly zero—the smallest is 0.40 pp and 11 of 480 pairs fall below 1 pp—so the floor never binds on a pair that keeps its order. Under the difference metric (5 pp) the reversal is immaterial for the same reason, since a reversed pair cannot reach the 5 pp threshold either: $0/240$ pairs are bridged. Under the ratio metric (10%), $30/240$ pairs are bridged. These counts are descriptive because the 240 pairs share all training configurations and are not independent Bernoulli trials; the thresholds were selected for exploratory sensitivity analysis rather than preregistered confirmatory testing. We therefore do not interpret the associated nominal binomial p-values as population-level evidence. The results indicate that this fixed compensatory policy produces absolute gains and limited, heterogeneous gap changes, but they do not establish that stronger alignment interventions would fail. The same pattern holds on the six held-out allocations withheld from the fit ($6 \times \binom{5}{2} = 60$ pairs): $0/60$ bridged at 5 pp and 5 of 59 rank-consistent pairs at 10% (uniform $0/60$ and $8/60$, no pair reversed), so the non-closure generalizes beyond the 24-configuration sample.

Contrast class: the same metric on the no-mid-training path. Every count above is computed between a mid-training checkpoint and its post-SFT successor, so on its own it cannot distinguish "this supervised stage closes no gaps" from "mid-training-installed gaps are the ones that survive". The shared no-mid-training baselines of Table 1 supply the missing contrast: they use the same rule-type-balanced SFT *mix* as the uniform control—though at the larger full-alignment supervised budget of Table 1 rather than the sweep's smaller fixed budget—applied to Base rather than to a mid-training checkpoint (20 standard-SFT passes, Appendix B.1). Applying the identical metric to that path, 4 of 10 pairs are bridged from Base to SFT only and 5 of 10 from Base to SFT+RL, against $0/240$ for the compensatory pass and $0/240$ for its equal-budget uniform control on the mid-training path. The no-mid-training gaps being narrowed are far the larger ones—Base spreads 72.8 pp between Cipher and Counterfactual, and SFT lifts the two floor domains ($13.2 \rightarrow 66.0$ and $5.8 \rightarrow 20.0$)—so we read this as a direction rather than a like-for-like test: the same supervised stage that compresses the pretraining-installed spread leaves the coverage-installed spread where it found it, which is what the gap-persistent reading of §4.2 predicts. Two limits apply. The contrast is a single configuration (ten pairs) against a pool of 24 (240 pairs), so it carries no test statistic and we report it descriptively.

And it is the *uniform* pass that runs on both paths; a compensatory-policy arm on the no-mid-training path—which would isolate the policy from the stage—would require training new passes and is left to future work.

Threshold sensitivity. Figure 8 sweeps both thresholds over the 240 pairs: the bridged count falls from 57/240 at the most permissive ratio threshold (5%) and 39/240 at the most permissive difference threshold (0.5 pp) to $0/240$ from 4.5 pp upward. To calibrate these counts we use a permutation null rather than a "half of all pairs" reference line, which is not a null expectation: under a true no-effect null the expected number of *bridged* pairs is near zero, not half. The null permutes, within each configuration, which domain receives which compensatory gain, holding the observed gain magnitudes—and hence the arithmetic budget cap—exactly fixed while destroying the alignment between a domain's gain and its position within each pair (2,000 reallocations). It therefore preserves the marginal gain distribution but not the joint structure of the gain vector across domains. With only 24 configurations, a null that preserved that joint structure as well would have to be estimated from the same 24 gain vectors it is meant to test against, so we report the within-configuration null and scope the resulting P-values to it. A random reallocation of the very same gains would bridge 13.8 ± 3.3 pairs under the 5 pp difference metric and 77.9 ± 8.5 under the 10% ratio metric; the compensatory policy bridges 0 and 30 respectively ($P < 0.001$ in both cases, one-sided). The observed non-closure is therefore not a mechanical consequence of the capped budget: the same gains, reallocated at random, would have closed substantially more gaps than the compensatory policy does.

How much does the choice of null matter? The within-configuration null holds the arithmetic budget cap fixed but still admits reallocations no plausible policy would produce—it can hand a large observed gain to an already-saturated domain, for instance. We therefore repeat the test under four restrictions that narrow only the set of admissible placements, and report all five in Table 17. Three reallocate the observed gain magnitudes without changing them, restricting which domain may receive which gain; the fourth instead redraws the remedial allocation itself, so that every draw is a placement the policy family could actually have produced.

- **Capacity-aware (headroom).** Domain d may receive gain g only if g does not exceed its headroom $100 - a_d^{\mathrm{mid}}$. This restriction turns out to be *vacuous* on this sweep: the smallest headroom anywhere is 15.9 pp (Counterfactual) against a largest observed compensatory gain of 8.6 pp, so all 120 permutations stay admissible in every configuration and the null is numerically identical to the unrestricted one.
- **Need-monotone.** The placement must at least weakly follow the compensatory logic—a domain with a larger remedial need $\max(0, 60 - M_d)$ never receives a strictly smaller gain. The observed count stays well below this null.
- **Budget-feasible.** Rather than permuting gains, we redraw the remedial allocation r on the simplex subject to $r_d \leq 30\%$, the largest share the policy ever allocated, and map each draw through the paper's own fitted per-domain gain curves $\Delta_d(r) = a_d \log(1 + b_d r) + c_d$ (Table 20). Every draw is then a placement this policy family could actually have produced, so the null no longer assumes gains are freely transferable across domains. The observed count is *still* below the null, but by a much smaller margin than under the unrestricted null.
- **Per-domain demonstrated capacity.** Domain d may receive only a gain magnitude it was itself observed to achieve in some configuration. This is the strictest reading of "capacity-aware" and it is nearly degenerate here—14 of the 24 configurations admit the identity permutation alone, and the full admissible set is 1–4 of 120 per configuration—so for most configurations it removes reallocation altogether. Under it the observed count is in line with the null.

Two consequences follow, and they cut in opposite directions. First, the capacity objection does not bite in its natural form: because no observed compensatory gain comes close to any domain's headroom, the headroom constraint removes no permutation at all, and a placement null restricted to "plausible" headroom-respecting reallocations is the same null. Second, and against our own headline, the "worse than chance" reading is weaker than the unrestricted null suggests. It survives the need-monotone and budget-feasible restrictions—including the version in which every draw is a placement the policy family could genuinely have produced—but at $P = 0.005$ (difference metric) and $P = 0.043$ (ratio metric) rather than $P < 0.001$, and the ratio-metric margin is marginal. Under the strictest per-domain capacity restriction the claim does not hold, though that restriction admits almost no reallocation on this sweep and so is closer to a no-op than to a competing hypothesis. We therefore scope the claim to reallocations that hold the gain magnitudes

Table 17: Pairwise gap closure under the reported null and four restricted nulls. Every null uses $\geq 2{,}000$ draws. Nulls B, C and D reallocate the observed gain magnitudes without changing them, restricting which domain may receive which gain and hence holding the arithmetic budget cap fixed; only the budget-feasible null E redraws the remedial allocation and maps it through the fitted gain curves of Table 20. P is one-sided and asks how often the null closes as few or fewer pairs than the observed policy. The observed policy closes $0/240$ pairs at the difference metric and $30/240$ at the ratio metric, and row A is an independent redraw of the null reported in the text, reproducing it to within Monte Carlo error (13.8 ± 3.3 and 77.9 ± 8.5 there; 13.7 ± 3.2 and 77.9 ± 8.2 here).

Null	Restriction	Bridged at 5 pp		Bridged at 10%	
		mean $\pm$ SD	P	mean $\pm$ SD	P
A (as reported)	none	13.7 ± 3.2	< 0.001	77.9 ± 8.2	< 0.001
B	headroom $g \leq 100 - a^{\text{mid}}$	13.7 ± 3.2	< 0.001	77.8 ± 8.4	< 0.001
D	need-monotone placement	8.5 ± 0.5	< 0.001	81.5 ± 2.7	< 0.001
E	budget-feasible ($r \leq 30\%$)	6.1 ± 2.7	0.005	40.1 ± 5.8	0.043
C	per-domain demonstrated capacity	0.7 ± 1.0	0.65	31.5 ± 1.5	0.26

fixed or are drawn from the policy family's own fitted curves; we do not claim that the tested placement is worse than *any* plausible alternative.

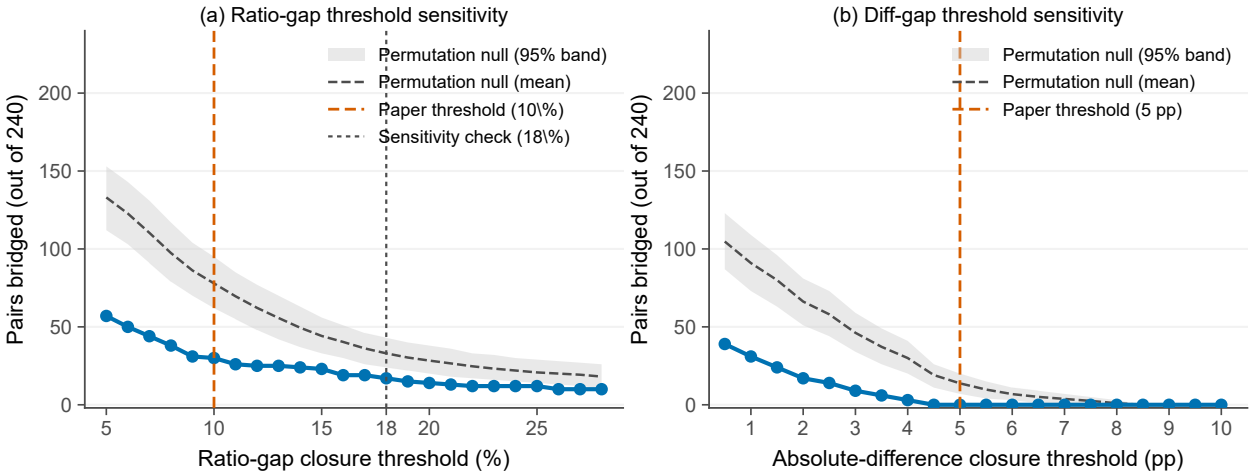

Figure 8: Threshold sensitivity for pairwise gap-closure over the 240 domain pairs. Left: bridged pairs vs. ratio threshold (5%–28%). Right: same for difference threshold (0.5–10 pp). The dashed line and shaded band are the mean and 95% interval of a permutation null that reallocates the observed compensatory gains across domains within each configuration (2,000 draws), holding the gain magnitudes and the arithmetic budget cap fixed. The observed counts lie *below* this null at both metrics ($P < 0.001$), so the non-closure is not a mechanical consequence of the capped budget.

H.3 A Stronger Reweighting Arm: Sharpening the Compensatory Policy

Finding 1 is stated for a policy family whose within-configuration gain differential is arithmetically bounded at ≈ 8 pp, so a reader may reasonably ask whether the non-closure is an artifact of a policy too weak to succeed. Two analyses in this appendix already bear on that question. First, the feasible set is large: evaluated on the fitted gain curves, 148 of the 240 domain pairs (62%) could in principle be narrowed by at least 5 pp by *some* budget-feasible allocation of the same SFT tokens—for each pair the largest attainable narrowing orients the pair by mid-training accuracy, puts the entire remedial budget on the weaker domain ($r = 100\%$) and none on the stronger (so the stronger side contributes only its fitted intercept), caps the narrowing at the pairś existing gap, and evaluates the same closure rate C_{ij} as Appendix H.2 ($210/240$, 88%, under the 10% ratio metric)—against 0 and 30 actually closed. Two caveats attach to this figure, and

we state both rather than let it carry more than it can. It requires evaluating the fitted gain curves at $r = 100\%$, whereas the curves were fitted over $r \in [4\%, 30\%]$ (Appendix H); capped at the largest observed r, the same procedure gives $41/240$ at the 5 pp metric. And it over-predicts the one extreme placement we actually measured: on the closure-favourable configurations, where the fits predict roughly 70% of pairs closing, winner-take-all closes $12/60 = 20\%$ (Table 18). Second, the permutation null of Appendix H.2 shows the observed gains close fewer gaps than a random reallocation of the same magnitudes. Neither, however, answers the operational question: is there a *stronger policy in the same family* that actually closes the gaps? This subsection reports the experiment that does.

Policy family. We generalise the compensatory rule of Eq. 5 with a single sharpening exponent γ:

$$r_d(\gamma) = \frac{[\max(0,\ T - M_d)]^\gamma}{\sum_{j=1}^{5}[\max(0,\ T - M_j)]^\gamma} \times 100\%, \qquad T = 60\%, \tag{9}$$

where M_d is domain d's mid-training token percentage. The exponent controls how sharply the fixed SFT budget concentrates on coverage-deficient domains. At $\gamma = 1$ the rule reduces exactly to Eq. 5, so the arm reported in the main text is the $\gamma = 1$ member of this family and the comparison is nested rather than across unrelated policies; as $\gamma \to \infty$ it becomes winner-take-all, assigning the entire remedial budget to the single most deficient domain. Every branch holds the total SFT token count equal to the uniform baseline, exactly as at $\gamma = 1$; only the cross-domain distribution changes. We ran $\gamma \in \{1, 2, 4, \infty\}$ on six sweep configurations selected to carry the most feasible-but-unclosed pairs, meaning configurations in which nearly every one of the ten domain pairs is separated by ≥ 5 pp at the mid-training-only checkpoint so that closure is at all attainable; this criterion is a ranking, not a unique rule (21 of the 24 sweep configurations have 9–10 of their 10 pairs so separated), and it deliberately over-samples the configurations where closure is possible, which is why the trade-off below is reported as a closure-favourable subset estimate rather than an unbiased one. The run required 90 additional SFT passes (three new branches $\times$ six configurations $\times$ five seeds; $\gamma = 1$ was already trained) and no further mid-training or RL.

That frontier is anticipated: concentrating on deficient domains is a worst-case-oriented reweighting, and trading mean for worst-group performance is what such objectives do by design (Sagawa et al., 2020). What the sweep supplies is its measured magnitude and the fact that our γ=1 member sits at the frontier's opposite corner.

Result: closure is attainable, but it is bought with average accuracy. Table 18 reports the outcome over the $6 \times \binom{5}{2} = 60$ pairs these configurations contribute. Sharpening does close gaps that $\gamma = 1$ leaves open—the count rises monotonically from $0/60$ at $\gamma = 1$ to $12/60$ at $\gamma = \infty$ under the 5 pp metric—so the non-closure reported in the main text is *not* an inescapable property of the budget. But closure is purchased directly out of average accuracy: the mean per-cell gain falls monotonically from $+4.34$ pp to $+2.26$ pp, so the winner-take-all branch that closes the most pairs also forfeits roughly half of the improvement the compensatory pass was introduced to deliver. Because these six configurations were selected for carrying the most closable-but-unclosed pairs, the trade-off magnitudes are a closure-favourable subset estimate rather than an unbiased one over the 24-configuration sweep; on mean gain the subset is representative, its γ=1 anchor ($+4.34$ pp) matching the full-sweep mean ($+4.32$ pp). The two objectives are in tension across the whole family we tested, and no branch achieves both: the branch that maximises mean gain closes nothing, and the branch that closes most gives up 2.1 pp of mean gain. This sharpens Finding 1 rather than overturning it—the gaps survive every setting that preserves the average gain—and it replaces an untested scope caveat with a measured trade-off. Even at $\gamma = \infty$, 48 of the 60 pairs remain unclosed at the 5 pp threshold.

Table 18: Sharpened compensatory SFT over the six configurations carrying the most feasible-but-unclosed pairs (60 pairs). $\gamma = 1$ is the policy reported in the main text. Mean gain is the per-cell accuracy change over the mid-training-only checkpoint, averaged over all 30 configuration–domain cells. Closure counts use the same two metrics as Appendix H.2.

Branch	Mean gain (pp)	Bridged, 5 pp	Bridged, 10% ratio
$\gamma = 1$ (Eq. 5)	$+4.34$	$0/60$	$8/60$
$\gamma = 2$	$+4.17$	$1/60$	$11/60$
$\gamma = 4$	$+3.65$	$5/60$	$13/60$
$\gamma = \infty$ (winner-take-all)	$+2.26$	$12/60$	$12/60$

H.4 Results

Table 16 (Remedial SFT rows) reports compensatory allocations from Eq. 5; Eval. (Remedial) rows give post-compensatory-SFT accuracy. Compensatory SFT raises scores in 116 of 120 configuration–domain cells (mean $+4.32$ pp; the four cells with no positive gain all receive zero compensatory allocation $r = 0$). The re-run exploratory θ^* block, which is not one of the 24 sweep checkpoints, also gains in all five domains under compensation ($+1.4$ to $+6.2$ pp; mean $+4.5$ pp), with Puzzle's gain the smallest; its deviations from the fitted per-domain curves below are at most ≈ 0.4 pp, so including or excluding this block does not change any conclusion.

Threshold calibration. The 10% ratio and 5 pp difference cutoffs are calibrated to the scale of the intervention and of the seed noise in this sweep, not to a statistical null. The 5 pp difference threshold exceeds the largest cross-seed SD observed on any single configuration cell in Table 16 (mid-training-only SDs run 0.6–3.9 pp over the 24 sweep configurations; the SFT-pass rows on those configurations have smaller SDs, though the six held-out blocks (which add the RL stage) carry a few larger values, up to ≈4.5 pp), so a counted closure is larger than the seed-level dispersion of either endpoint. The 10% ratio threshold is likewise beyond the seed-noise scale of the weaker-domain accuracies that dominate ratio-bridged pairs, and equal absolute gains naturally produce larger relative closures for weaker domains. The scan in Figure 8 shows that the conclusion does not depend on the exact cutoffs: the bridged-pair count stays below the permutation null of §H.2 at every tested threshold (57/240 even at the most permissive 5% ratio; 39/240 at 0.5 pp difference, and 0/240 from 4.5 pp upward). Under ratio (10%), 30/240 pairs are bridged; under difference (5 pp), 0/240 pairs are bridged. Both summaries indicate heterogeneous changes under this fixed policy, not general non-repairability.

These statistics underpin Figure 4. Redirecting SFT budget toward coverage-deficient domains raises accuracy but does not reliably close pairwise gaps created by mid-training.

Uniform-SFT control. As a control with the same data mix and budget, a uniform SFT pass (identical rule-type-balanced data mix across configurations; the same total token budget and 3-epoch schedule as the compensatory pass; Table 5) was applied to all 24 sweep checkpoints (Eval. (Uniform) rows, Table 16). Table 19 compares the two passes. The uniform pass raises every one of the 120 cells (mean $+4.20$ pp vs. $+4.32$ pp compensatory), bridges the same 0/240 pairs under the 5 pp difference metric and 32/240 under the 10% ratio metric (vs. 30/240), and produces per-domain mean gains close to the compensatory ones except for Counterfactual, where compensation is stronger by $+1.9$ pp ($+6.37$ vs. $+4.49$ pp), and Puzzle, where uniform SFT is slightly stronger ($+2.10$ vs. $+1.48$ pp). Cross-configuration range narrowing is likewise comparable; uniform SFT narrows all five ranges (including Puzzle, -0.7 pp), whereas compensation leaves Puzzle's range slightly wider ($+0.4$ pp). The compensation formula therefore does not systematically outperform an equal-budget uniform re-training pass; its only material advantage is Counterfactual, the domain whose gain responds most steeply to the remedial ratio (below).

Table 19: Per-domain mean accuracy gains (pp) of the compensatory and uniform SFT passes over the 24 sweep configurations (all 120 configuration–domain cells in the last row). Compensation exceeds the uniform control materially only for Counterfactual; it is slightly below the control for Cipher and Puzzle.

Domain	Compensatory gain	Uniform gain	Difference (comp.−uniform)
Cipher	$+3.42$	$+3.89$	-0.47
Operation	$+5.73$	$+5.77$	-0.04
Logic	$+4.62$	$+4.76$	-0.14
Counterfactual	$+6.37$	$+4.49$	$+1.88$
Puzzle	$+1.48$	$+2.10$	-0.63
All 120 cells	$+4.32$	$+4.20$	$+0.12$

Gain vs. the remedial allocation: per-domain fits. Over the 115 configuration–domain cells with $r > 0$, we fit, separately for each domain, the model

$$\Delta_d(r) \;\approx\; a_d \log(1 + b_d\, r) + c_d, \tag{10}$$

where r is the domain's remedial SFT proportion in percent (Figure 7). The additive $+1$ inside the logarithm removes the absorption redundancy of the previous pooled form $\Delta \approx a \log(b\, r) + c$, in which b and c were not separately

identifiable. Table 20 reports the fitted parameters. Within each domain the gain rises with r and saturates at a domain-specific ceiling; the fits are tight (MAE 0.07–0.27 pp; R^2 0.66–0.97 over each domain's r-window), with the steepest rise across the observed window for Counterfactual (predicted $+3.2$ pp at $r{=}5\%$ to $+8.3$ pp at $r{=}29.5\%$) and the flattest for Puzzle ($+0.3$ to $+2.4$ pp over $r \in [4, 29]\%$). Domains differ more in overall gain level than in their response to r: a pooled fit of Eq. 10 over all 115 cells explains only $\approx 31\%$ of the variance ($R^2{=}0.31$, MAE 1.39 pp; Table 20), which is why the earlier pooled model appeared to describe a weak association. These fits are descriptive summaries; within each domain the observed r-window is narrow (≈4–30%), so a_d and b_d trade off against each other and, where the window is nearly linear (Cipher), the pair is only weakly identified individually even though the fitted curve is tight.

Table 20: Per-domain fits of the compensatory gain Δ (pp) against the remedial ratio r (%) using $\Delta = a\log(1{+}b\,r){+}c$ on the $r > 0$ cells of Table 16 (Eq. 10; Figure 7). R^2 and MAE are computed on the fitted cells. The pooled row fits all 115 $r > 0$ cells jointly. †Cipher's a sits at its upper fit bound; over Cipher's observed window $\log(1 + br) \approx br$, so the curve is effectively linear (slope $\approx +0.12$ pp per percentage point of r) and a, b are individually unstable.

Domain	n	a	b	c	R^2	MAE (pp)
Cipher†	22	50.00	0.0024	+1.24	0.78	0.25
Operation	23	5.99	0.0554	+1.08	0.66	0.27
Logic	23	3.81	0.1351	−0.39	0.77	0.25
Counterfactual	24	13.75	0.0201	+1.86	0.96	0.18
Puzzle	23	4.39	0.0273	−0.16	0.97	0.07
Pooled	115	10.08	0.0251	+0.33	0.31	1.39

Boundary cases (zero-allocation domains). Five cells receive $r = 0$ because their domain's mid-training coverage is at or above the 60% threshold (Configs 1, 7, 15, 16, and 19; Table 16). These cells show no meaningful gain under compensation (all five gains within ± 0.3 pp, the largest being Config 1 Cipher at $+0.2$ pp, which receives no remedial allocation at all), and the conclusions above hold when these configurations are excluded.

Budget sensitivity. Under our budget-proportional model at $\gamma{=}1$, the within-configuration gain differential reaches at most ≈ 8 pp; scaling total SFT volume would not increase this differential without changing the allocation. We tested the allocation directly by sharpening the exponent to winner-take-all (§H.3): the differential grows, but closure is bought out of mean gain ($0/60 \rightarrow 12/60$ bridged, $+4.34 \rightarrow +2.26$ pp; Table 18). What remains untested is a differently parametrised reweighting (e.g. a lower shortfall threshold T) and any stable RL schedule after compensatory SFT.

I External Benchmark Detailed Results

I.1 Cross-Benchmark Associations and Exposure Accounting

The three external benchmarks differ sharply in exposure, so §4.4 reads them as separate columns rather than one transfer score. ProofWriter is a same-family evaluation: 34.8% of external tokens are ProofWriter-family items at depth $\leq$5, so its D6–D9 strata are held out by depth, not by task. ZebraLogic carries 0.3% same-family exposure (201 samples). CounterBench carries no training items of its own family (Appendix F) and is the only family-level zero-exposure column here. Because the external component is identical across the nine configurations (§5.2), exposure is a level offset shared by every run and cannot itself generate the cross-configuration associations reported here; it does differ across columns, which is why they are read separately.

Against the shared SFT+RL baseline, the Expt. 1 checkpoint gains $\Delta = +6.4$ on held-out ProofWriter D6 (95% CI $[+3.3, +9.5]$; $N = 1{,}292$) and $+4.4$ on the covered D5 stratum ($[+3.6, +5.2]$; $N = 18{,}498$). These are stage-mismatched contrasts—the mid-training arm carries no RL—so part of each Δ is the RL stage rather than the inserted checkpoint; the RL increment on the KOR-Bench overall is $+1.2$ pp (Table 1), which bounds the size of that contamination but does not remove it. The covered-depth ablation, which is stage-matched, points the same way (a D5-heavy mixture after SFT gains $+5.60$ overall, $+5.70$ on D5 and $+4.64$ on D6, against $+0.89$, $+0.97$ and -0.24 for its D3-heavy counterpart; Table 24). D7–D9 gains are positive but those strata are too small to resolve ($N \leq 120$).

Within the nine fixed configurations, the balanced allocation (20%×5) is the top-ranked configuration on the KOR average (56.3), on ProofWriter (86.3) and on CounterBench (79.5), and within 2.6 pp of the top on ZebraLogic Grid (35.2 vs. 37.8); the FineWeb-Edu-only baseline is the lowest of the ten tabulated rows on the KOR average (26.9), ProofWriter (50.9), ZebraLogic Grid (13.1) and CounterBench (65.0) (Table 21). The one external column that does not track the KOR ranking is ZebraLogic MC, where the balanced allocation ranks sixth of nine. CounterBench's per-type deltas against the SFT+RL baseline remain at or below zero, so we read its column as ranking consistency of the mixture across benchmarks rather than as transfer of accuracy. These are rank positions over nine purposively chosen design points, not an uncertainty-quantified ordering, and the four associations ($\rho = +0.34$ ProofWriter, $+0.53$ ZebraLogic Grid, $+0.67$ CounterBench, -0.12 ZebraLogic MC) are not distinguishable from one another at $n = 9$. They are not a test of transfer beyond the rule library that generated the training corpus.

I.2 Coverage Sweep: External Benchmark Accuracy

Table 21 reports external benchmark accuracy for the coverage-sweep mid-training models (plus the FineWeb-Edu baseline), using the same model checkpoints as the KOR-Bench mid-training-only evaluation. The external results are limited consistency checks rather than independent confirmation of a general transfer law. The external configuration set is fixed: the three external benchmarks evaluated on the nine coverage configurations (plus the FineWeb-Edu baseline) listed in Table 21, which span the sweep from severely imbalanced to balanced allocations; Spearman rank associations of the KOR average with each external column over these configurations are reported in §4.4 ($n = 9$). The external corpus draws 34.8% of its tokens from the ProofWriter rule family and the ZebraLogic family contributes 201 mid-training samples, so ProofWriter and ZebraLogic are same-family evaluations while CounterBench is family-disjoint (Appendix F); rule-family-level disjointness is established for CounterBench and for ProofWriter depths 6–9, and template-level overlap is not independently verified for any of the three. Because the external component is identical across the nine configurations, exposure is a constant level offset shared by every run and cannot by itself generate the cross-configuration associations reported below; what those associations may reflect, in addition to mixture allocation, is shared task structure between the external benchmarks and the KOR-Bench domains. Exposure differs across the three *columns*, which is why we read them separately rather than as one transfer score. The per-depth, per-house, and per-type comparisons below contrast the shared SFT+RL baseline with the Expt. 1 checkpoint (Cipher 72.3%, Operation 0%, Logic 12.0%, Counterfactual 0%, Puzzle 15.7%; Table 1); the covered-depth ablation uses its own D3-/D5-heavy mixtures (below) and is not tied to any sweep configuration.

Table 21: External benchmark accuracy (%) for the coverage-sweep mid-training models (the FineWeb-Edu baseline is shown for reference). Coverage: Cipher / Operation / Logic / Counterfactual / Puzzle (internal); the ProofWriter rule family (34.8% of external tokens) is fixed. KOR avg is the macro-average of the five KOR-Bench domain accuracies from Table 16 (§G); the external columns are five-seed means on the same checkpoints, matching the protocol of Table 16. "FineWeb-Edu" = FineWeb-Edu-only baseline without KOR-Bench data. The nine coverage configurations listed here constitute the fixed external configuration set (§4.4).

Coverage (Cipher/Oper./Logic/Counterf./Puzzle)	KOR avg	ProofWriter	ZebraLogic MC	ZebraLogic Grid	CounterBench
FineWeb-Edu (100%)	26.9	50.9	37.7	13.1	65.0
40 / 4 / 8 / 40 / 8	47.5	80.7	45.8	21.0	70.7
72.3 / 0 / 12 / 0 / 15.7	36.6	83.1	40.8	23.0	69.9
0 / 10 / 40 / 10 / 40	44.8	85.1	31.6	37.8	76.4
50 / 40 / 0 / 5 / 5	40.1	81.1	41.4	25.3	72.0
5 / 15 / 5 / 15 / 60	49.8	81.2	37.4	26.2	72.5
26 / 16 / 17 / 30 / 11	55.9	80.7	**46.8**	26.9	71.0
30 / 18 / 16 / 18 / 18	56.0	84.4	40.6	37.2	77.7
10 / 8 / 37 / 8 / 37	50.0	85.7	30.6	35.0	75.9
Balanced (20%×5)	**56.3**	**86.3**	38.2	35.2	**79.5**

I.3 ProofWriter Per-Depth

Ablating the covered depth. Depth-5 ProofWriter exposure improves both the covered D5 and held-out D6 strata; D7–D9 strata are too small for strong claims. Tables 23 and 24 ablate whether transfer depends on covered training depth (reference: SFT-only baseline, Overall 76.90). The two mixtures swap D3 and D5 proportions while holding

Table 22: ProofWriter per-depth accuracy (%). Mid-training trains only to depth 5; eval depths 5–9. Depth 5 is therefore both trained and evaluated—it carries 18,498 of the 20,030 evaluation items (92%)—so D5 is an in-distribution stratum and only D6–D9 are held out by depth. Δ is Mid-training+SFT $-$ SFT+RL, i.e. a stage-mismatched contrast (the mid-training arm has no RL); $^{\dagger}N < 200$, not significant at 95% level. Δ is computed from exact correct counts before rounding individual accuracies; apparent differences between Δ and (Mid-training+SFT$-$SFT+RL) on rounded values are rounding artifacts.

Depth	SFT+RL	Mid-training+SFT	Δ	95% CI (Δ)	N
D5	80.1	84.4	$+4.4$	$[+3.6, +5.2]$	18,498
D6	76.2	82.7	$+6.4$	$[+3.3, +9.5]$	1,292
D7	75.8	82.5	$+6.7$	$[-3.5, +17.0]^{\dagger}$	120
D8	77.1	83.3	$+6.2$	$[-5.0, +17.4]^{\dagger}$	96
D9	75.0	91.7	$+16.7$	$[-3.8, +37.2]^{\dagger}$	24

D0–D2 and Other fixed. The D5-heavy model after SFT is substantially stronger ($+5.60$ pp overall; $+5.70$ pp D5, $+4.64$ pp D6), suggesting post-training benefits more when mid-training difficulty is closer to the evaluation depth.

Mixture	D0–D2	D3	D5	Other
D3-heavy	2.04	65.74	17.12	15.10
D5-heavy	2.04	17.12	65.74	15.10

Table 23: ProofWriter training-depth distribution in the covered-depth ablation (%). Other denotes examples without parsed depth.

Setting	Overall	D5	D6
D3-heavy	-2.69	-2.51	-4.18
D3-heavy + SFT	$+0.89$	$+0.97$	-0.24
D5-heavy + SFT	$+5.60$	$+5.70$	$+4.64$

Table 24: ProofWriter covered-depth ablation results. Values are accuracy deltas (pp) relative to the ablation suite's SFT-only baseline, so both arms carry the same downstream stages and the contrast is stage-matched.

I.4 ZebraLogic MC Per-House

Table 25 reports ZebraLogic MC results (MC = multiple-choice), complementing the ProofWriter depth-extrapolation result in §4.4. With only 201 direct mid-training ZebraLogic samples (0.3% of the total mid-training corpus, not of the internal five-domain component), Mid-training+SFT outperforms SFT+RL at every reported house count. The per-house strata are unbalanced, so the Δ column is a per-stratum contrast and its unweighted mean is not the checkpoint's overall ZebraLogic MC score in Table 21; per-house N is not reported and we do not tabulate an overall row, so the two tables are not comparable. This is consistent with shared formal-deduction and constraint structure, but ZebraLogic is a same-family held-in exposure rather than an out-of-distribution transfer test, and it is not evidence that a small exposure alone caused the gains. Template/family-level overlap remains uncontrolled in this study, and a zero-exposure replication is future work.

I.5 CounterBench Per-Type

CounterBench has zero direct mid-training coverage and no same-family training items (Appendix F); Δ is Mid-training+SFT minus SFT+RL on the Expt. 1 checkpoint. Its per-type Δ values are all at or below zero across the four strata (Figure 9; Basic -4.0, Conditional 0.0, Joint -4.4, Nested -4.0 pp)—a contrast with the positive deltas reported for ProofWriter and ZebraLogic—so we record no accuracy transfer on this column and read it instead as the ranking consistency of the mixture on a benchmark with no same-family training items. Per-type N and interval estimates are not reported, so unlike the ProofWriter and ZebraLogic strata (Tables 22, 25) these four comparisons are untested and no null is claimed: the column is a zero-exposure ranking comparison, not an accuracy gain. CounterBench differs

Table 25: ZebraLogic MC per-house accuracy (%). Mid-training coverage: 0.3% of the total mid-training corpus (201 samples). Δ is Mid-training+SFT $-$ SFT+RL, i.e. a stage-mismatched contrast (the mid-training arm has no RL); 95% asymptotic CIs are shown. Per-house N is not reported and the strata are unbalanced, so this column is descriptive and is not comparable to the overall ZebraLogic MC column of Table 21.

Houses	SFT+RL	Mid-training only	Mid-training+SFT	Δ	95% CI (Δ)
2	89.9	90.1	94.1	$+4.2$	$[+1.1, +7.3]$
3	78.8	77.4	84.3	$+5.5$	$[+1.2, +9.8]$
4	60.2	56.5	72.6	$+12.4$	$[+7.5, +17.3]$
5	47.6	40.5	55.9	$+8.3$	$[+3.0, +13.6]$
6	31.3	28.6	43.1	$+11.8$	$[+6.6, +17.0]$

from the KOR-Bench domains in required procedures, prompt format, and difficulty, so this is a scope statement about the benchmark rather than a claim about counterfactual reasoning in general.

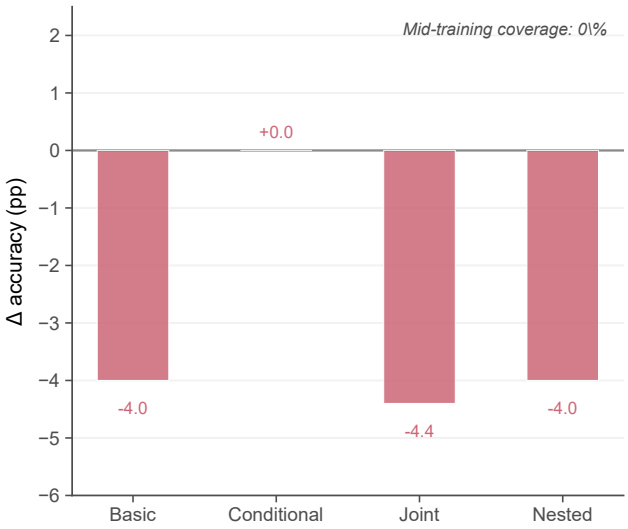

Figure 9: CounterBench per-type accuracy change (Δ = Mid-training+SFT $-$ SFT+RL, pp) on the Expt. 1 checkpoint, across the four CounterBench strata. All four deltas are at or below zero. Per-type N and intervals are not reported, so the comparison is descriptive.

J Follow-Up Experiments: Repetition, Replacement, and a Mixture Baseline

Every experiment in this section was run after the main sweep, and each closes one confound that the sweep cannot address by construction. None of them varies the mixture across the five-domain simplex; each holds the mixture fixed and varies something else, which is why they are reported separately from the coverage results rather than inside them.

J.1 The Coverage–Repetition Collinearity Is Structural

The sweep varies a domain's token share at a fixed total budget, so a domain's share and its epoch count are not two knobs but one. Writing I for the fixed internal token budget, u_d for the number of unique instances a domain draws and $\bar{\ell}_d$ for its mean trace length,

$$\text{processed}_d = s_d I, \qquad e_d = \frac{\text{processed}_d}{u_d\,\bar{\ell}_d}, \tag{11}$$

so with I and $\bar{\ell}_d$ fixed per domain the triple (s_d, u_d, e_d) satisfies one equation and the predictor matrix has rank 2. **No reparameterisation of the sweep can separate coverage from repetition**, and no amount of refitting will change

that; uniqueness is recovered only where a domain's pool is exhausted, since there u_d is pinned to the pool size rather than to the share.

Table 26 reports the variance-inflation factors. Across the full 30-allocation pool the share carries VIF 11.5–16.4 in four domains and is exactly collinear in Puzzle; on the nonzero-share subset, where the four zero-coverage cells no longer anchor the low end, every domain's design is rank 2 and the share VIF is infinite. Table 27 summarises the corpus ledger behind this: **43 of 150** configuration–domain cells exhaust their pool, and the domain-level epoch counts run from 0 (the zeroed domains) to 2.8.

Table 26: Collinearity of coverage with repetition in the sweep design. The full-pool columns use all 30 trained allocations; the last two drop the zero-coverage cells, which is where the share variable loses the low end that made the design rank 3. A VIF of ∞ means exact collinearity: the two regressors cannot be separated at all.

Domain	Full-pool rank	VIF(share)	Cells pool-exhausted	Rank excl. 0%	VIF(share) excl. 0%
Cipher	3	16.1	7/30	2	∞
Operation	3	13.6	6/30	2	∞
Logic	3	11.5	3/30	2	∞
Counterfactual	3	16.4	10/30	2	∞
Puzzle	2	∞	17/30	2	∞

Table 27: Corpus ledger summary. A cell is *pool-exhausted* when the domain's unique instances run out and the remainder of its token budget is repeat passes; those are the cells where the epoch count is pinned to the pool rather than to the share. Epoch values are the realised passes over each domain's unique pool across the 30 allocations.

Domain	Pool (M instances)	Exhausted cells	Realised epochs across the sweep
Cipher	1.30	7/30	0.00, 1.00–1.29, 1.56, 1.92, 2.30, 2.80
Operation	1.70	6/30	0.00, 1.00–1.34, 1.52, 1.90, 2.25
Logic	0.55	3/30	0.00, 1.00, 1.17, 1.24, 1.87
Counterfactual	0.62	10/30	0.00, 1.00–1.38, 1.72
Puzzle	0.25	17/30	1.00–2.55 (18 distinct values)
All domains	—	43/150	—

J.2 Separating Coverage from Repetition: Share Crossed with Pool Fraction

Since the sweep cannot separate the two, we ran the experiment that can. Holding a domain's *processed-token* count fixed and quartering its unique pool ($\phi = 0.25$) raises its epoch count fourfold without touching the mixture, so a share effect that is really a repetition effect must collapse when ϕ moves. The focus domain is Counterfactual, at moderate (20%) and high (50%) share, five seeds per cell, mid-training only.

The repetition channel is empty. At moderate share, quartering the pool moves accuracy by $+0.1$ pp ($81.60 \pm 0.96 \rightarrow 81.70 \pm 1.35$) while the epoch count goes from 1.00 to 4.00; at high share it moves it by -1.0 pp ($76.00 \pm 3.86 \rightarrow 75.00 \pm 3.66$) while epochs go from 1.73 to 6.91. The share contrast over the same cells is 5.6 pp at $\phi = 1$ and 6.7 pp at $\phi = 0.25$—larger than either ϕ effect and in the opposite direction. Repetition therefore does not carry the interior optimum in this sweep, and the coverage reading survives the confound that Table 26 shows the sweep itself cannot rule out.

One scope limit must be stated with the result: ϕ changes the number of distinct examples a domain sees as well as its epoch count, so the manipulation moves repetition and unique-example diversity together and in opposite directions. What the flat response rules out is their *net* contribution to the moderate-versus-high contrast, not each channel separately; separating them needs a design that varies one at a fixed value of the other.

The two allocations this experiment shares with the sweep are worth reconciling: at Counterfactual 20% the sweep's moderate band gives 81.2 ± 4.2 and this run gives 81.60 ± 0.96, and at 50% the sweep's single high-band cell gives 78.2 against this run's 76.00 ± 3.86. Both differences are within the cross-seed SD of the cells being compared, but they are not zero, and they are the reason the share contrast quoted here (5.6–6.7 pp across these four cells) is not identical to the sweep's own moderate-versus-high contrast.

We report the counterfactual alongside the measured result because it is what makes the negative interpretable. If repetition carried the effect with the coefficient a confounded design would require, the same ϕ manipulation would

have moved accuracy by -6.2 pp at moderate share and -13.4 pp at high share (Regime B in Table 28); the measured moves are $+0.1$ and -1.0. The test has the power to detect the confound it was designed to detect.

Table 28: Share crossed with pool fraction, focus domain Counterfactual, five seeds per cell, mid-training only. ϕ scales the domain's unique pool while its processed-token count is held fixed, so the epoch count rises as ϕ falls. *Regime A* is the measured result; *Regime B* applies the same manipulation under a repetition coefficient of ≈ 2.1 pp per epoch, the value a confounded design would require, and is reported only to show the manipulation discriminates. Note what that coefficient does and does not establish: it is well below the ≈ 7.7 pp per epoch that would be needed for repetition to account for the entire moderate-versus-high contrast in the sweep, so this test bounds the repetition channel rather than closing it.

Share	ϕ	Epochs	Regime A (measured)	Regime B (counterfactual)
moderate (20%)	1.00 (full pool)	1.00	81.60 ± 0.96	81.60 ± 0.96
moderate (20%)	0.25 (pool quartered)	4.00	81.70 ± 1.35	75.40 ± 1.35
high (50%)	1.00 (full pool)	1.73	76.00 ± 3.86	74.47 ± 3.86
high (50%)	0.25 (pool quartered)	6.91	75.00 ± 3.66	62.59 ± 3.66

J.3 Absence or Dominance? The Complement-Concentration Ladder

Every zero-coverage cell in the sweep pairs the zero with 40–72% coverage elsewhere, so "a domain at 0% collapses" is observationally entangled with "two or more other domains were over-covered". The sweep cannot separate the two. We therefore built a ladder that holds the zeroed domain at 0% and varies only how concentrated the remaining 100% is: flat $\{0, 25, 25, 25, 25\}$, concentrated $\{0, 40, 20, 20, 20\}$, and steep $\{0, 55, 15, 15, 15\}$, for three zeroed domains at five seeds each.

Concentration does not matter (Table 29). Across the three ladders the zeroed domain's accuracy moves by $+1.1$ pp (Operation, $33.10 \to 34.20$), $+1.5$ pp (Logic, $23.60 \to 25.10$) and $+1.6$ pp (Counterfactual, $46.40 \to 48.00$), while the complement's standard deviation rises from 0 to 20 pp. The collapse is a property of the domain's *absence*, not of the others' dominance.

One bookkeeping point the ladder exposes: "25% each" is not a neutral filler. It is $2.5\times$ Cipher's fitted peak and $1.7\times$ Logic's, and it pushes Puzzle below the lower bound of its 95%-of-peak interval. The flat ladder is therefore a legitimate allocation rather than a neutral one, and the comparison across ladders—not the level of any single rung—is what identifies the effect.

Table 29: Complement-concentration ladder. The zeroed domain is held at 0%; only the concentration of the remaining share varies. Five seeds per cell, mid-training only. If the collapse were driven by the other domains being over-covered, accuracy would fall as the ladder steepens; it does not move beyond noise.

Zeroed domain	Ladder	Complement max	Complement SD	Accuracy
Operation	flat $\{0, 25, 25, 25, 25\}$	25.0	0.0	33.10 ± 2.99
Operation	concentrated $\{0, 40, 20, 20, 20\}$	40.0	10.0	32.60 ± 4.20
Operation	steep $\{0, 55, 15, 15, 15\}$	55.0	20.0	34.20 ± 3.56
Logic	flat	25.0	0.0	23.60 ± 1.78
Logic	concentrated	40.0	10.0	23.70 ± 4.32
Logic	steep	55.0	20.0	25.10 ± 2.01
Counterfactual	flat	25.0	0.0	46.40 ± 3.83
Counterfactual	concentrated	40.0	10.0	48.80 ± 3.38
Counterfactual	steep	55.0	20.0	48.00 ± 2.24

J.4 What Replaces the Dropped Domain: The Filler Ladder

The FineWeb-Edu-only arm of §5.3 changes corpus composition as well as coverage, which is why we report it as a comparison rather than a bound. The ladder below removes that difference by holding the allocation fixed (focus domain at 0%, the other four at 20% each) and varying only the identity of the filler that takes the freed share: nothing (the domain dropped), a sixth reasoning rule family, a non-reasoning filler, or generic web text. The coverage-specific

component is then $\mathrm{T1} - \mathrm{T2}$, dropping the domain against replacing it with more reasoning data of a different family, rather than against a corpus of different composition.

The identity of the filler does not matter (Table 30). Over the four tiers, Operation spans 31.80–35.20 and Counterfactual spans 45.50–48.50, ranges of 3.4 and 3.0 pp against within-cell SDs of 0.6–3.4 pp; the coverage-specific component $\mathrm{T1} - \mathrm{T2}$ is $+1.7$ pp and $+2.1$ pp respectively, both inside the seed noise. What the domain's absence costs is therefore not attributable to what displaces it, and the composition difference that limits the FineWeb-Edu arm is not what produces the effect.

Table 30: Filler ladder. The five-domain allocation is identical in every row (focus domain 0%, the other four 20% each); only the filler taking the freed share changes. Five seeds per cell, mid-training only. $\mathrm{T1} - \mathrm{T2}$ is the coverage-specific component: dropping the domain against replacing it with a different reasoning rule family.

Focus domain	Filler taking the freed share	Accuracy	$\mathrm{T1} - \mathrm{T2}$
Operation	T1 nothing (domain dropped)	33.50 ± 3.41	$+1.7$
Operation	T2 sixth reasoning rule family	31.80 ± 2.54	
Operation	T3 non-reasoning filler	35.20 ± 1.30	
Operation	T4 generic web (FineWeb-Edu)	32.20 ± 0.57	
Counterfactual	T1 nothing (domain dropped)	47.60 ± 3.34	$+2.1$
Counterfactual	T2 sixth reasoning rule family	45.50 ± 1.37	
Counterfactual	T3 non-reasoning filler	48.50 ± 2.83	
Counterfactual	T4 generic web (FineWeb-Edu)	47.10 ± 2.30	

J.5 A Mixture-Optimization Baseline on the Same Corpus

The main text declines an empirical comparison with the mixture-optimization literature, and §A says a controlled comparison on this corpus would be informative. Table 31 supplies one: an allocation produced by a RegMix-style predictor fit on this study's own sweep, carried through the identical mid-training-only protocol at five seeds, against θ^* and against balanced coverage.

The three allocations are within 1 pp of one another on the macro-average, and the ordering is θ^* (57.28) $>$ the baseline-recommended allocation (56.86) $>$ balanced (56.40). The $+0.88$ pp that θ^* gains over balanced is achieved almost entirely on Puzzle (14.4 vs 11.2); on the other four domains the three arms are within 1.1 pp, and θ^* is nominally behind the baseline-recommended arm on Cipher and Logic. We report this as a baseline rather than a result: the gap between a fitted allocation and a predictor-recommended one is smaller than the gap between either and balanced, and all of it is inside the seed noise this study resolves.

Table 31: Mixture-optimization baseline on the same five-domain corpus. All three arms were trained mid-training-only at five seeds under the protocol of §2.1; the macro-average is the unweighted mean of the five per-domain accuracies. The baseline-recommended allocation comes from a RegMix-style predictor fit on this study's own sweep.

Arm	Macro-avg	Cipher	Operation	Logic	Counterfactual	Puzzle
baseline-recommended	56.86	51.5	77.6	59.8	84.2	11.2
θ^*	**57.28**	50.8	77.9	59.3	84.0	**14.4**
balanced (20%)	56.40	51.6	77.6	59.5	81.1	12.2

J.6 The Supervised Dose Ladder

The main text states that the supervised *dose* is never varied across arms, so nothing in the non-closure result speaks to whether a larger SFT budget would close the gaps. Table 32 reports the dose ladder that bounds this. Doubling the supervised budget leaves the number of eligible pairs—domain pairs whose mid-training gap is large enough that closure is arithmetically attainable—unchanged in every configuration tested, at an eligible fraction of 0.317 over the sweep. The gaps that a larger dose could close are therefore not a set that grows with the dose; what a larger budget would change is the size of the per-domain gains, not which pairs are closable in principle.

Table 32: Supervised dose ladder. Eligible pairs are the unordered domain pairs within a configuration whose mid-training gap is large enough that some budget-feasible allocation could narrow it past the 5 pp threshold. Doubling the supervised budget does not change the count in any of the six configurations tested; the eligible fraction over the simulated configurations is 0.317.

Config	1	2	3	4	5	6
Eligible pairs, 1× dose	3	4	4	2	4	2
Eligible pairs, 2× dose	3	4	4	2	4	2

K Limitations and Ethical Considerations

What would disconfirm each finding. Finding 1 would be disconfirmed by a fixed-budget compensatory-SFT pass fully closing the between-domain gaps (closure approaching 100% at 5 pp, not $0/240$). Finding 3 rests on a band comparison that involves no curve fitting (though its boundaries were read off the fitted peaks), a simplex-calibrated permutation test ($P \approx 0.010$), a 4B replication reproducing the band ordering, and the experiment ruling out repetition (Table 28); the low-versus-moderate contrast does not depend on the split-Gaussian fit, the high band and the boundaries do. The simplex constrains the interpretation, not the presence, of the non-monotone pattern: these optima are per-domain marginals, not one identified joint optimum (§5.2).

Table 33 collects every limitation discussed in this paper in one place, with the evidence tier that applies to it, so that a reader can see at a glance which results are held out, which are in-sample, and which are descriptive. "Held out" means evaluated on the six allocations withheld from the curve fit; "in-sample" means the statistic was computed on the same pool it is reported for. Every row restates a limitation given at the point of its result; none is a new claim.

Concern	**Status in this paper**	**Where**
Curve shapes	*Held out*: the six withheld allocations track the fitted curve shapes at seed-noise level for four of five domains (Operation's low-coverage tail runs ≈ 6 pp above the fitted tail)	App. G.4
Peak locations	*In-sample*: read off the sweep that produced them and not validated by the held-out points; they also shift with scale (-17.5 to $+3.8$ pp between 8B and 4B)	App. G.4, C
Band boundaries (10–40%)	*Post-hoc*: chosen after the fitted peaks were read. Low-versus-moderate survives every boundary tested; the high band is thin (n=1–5)	App. G.1
Joint optimum	Not identified: a saddle in all five domains. The classification is stable to resampling, but the negative curvature is shallow and quadratic misspecification is untested. Allocations from a mixture-optimization baseline land within 1 pp of θ^*	§5.2, App. G.2, J.5
Non-closure of the gaps	*Held out* ($0/60$ at 5 pp). The against-null margin depends on which placements are admitted: $P < 0.001$ unrestricted, $P = 0.005$ budget-feasible, and the claim does *not* hold under the strictest per-domain capacity restriction ($P = 0.65$), which is near-degenerate on this sweep	App. H.2
θ^*	Fit-selected: its full-pipeline row is not an out-of-sample validation, and the objective that selects it ranks held-out allocations weakly ($r = +0.441$)	App. G.4
Zero-coverage collapse	Single scale: reproduced at 8B, not at 4B on the same allocation	App. K
Diversity, trace length	Not separated by the sweep; coverage is a closed composition at a fixed token quota. *Repetition* is separated out-of-sweep, and does not carry the optimum	App. J
Supervised dose	Never varied across arms: the tested quantity is the cross-domain placement of a fixed budget. A dose ladder bounds it—doubling the budget leaves the closable set unchanged—but the effect of a much larger budget is untested	App. J
Base-model contamination	Uncontrolled: Qwen3-8B-Base's original pretraining exposure to public benchmarks applies to every checkpoint here, and the "zero-exposure" label covers our own corpora only	App. F
Corpus reconstruction	The $< 10\%$ of items on the model branch depend on a specific model version and decoding setting and will not reproduce under a different one	App. B.2
The $148/240$ feasibility figure	Requires evaluating the gain fits at $r = 100\%$, $3.3\times$ beyond the largest remedial ratio at which any curve was fitted, and over-predicts the one extreme placement measured	App. H.3
External benchmarks	Consistency probes, not transfer tests: ProofWriter is same-family (34.8% of external tokens), ZebraLogic carries 0.3% same-family exposure, CounterBench is zero-exposure but reported without N or intervals	App. I
RL stage	One algorithm (GSPO), 200 steps, binary verifier reward; longer schedules were unstable across seeds, so this leg is a secondary observation	App. D.7

Table 33: Evidence tiers behind the three findings. Each limitation discussed in this paper appears once, with the tier that applies to it. Nothing in the table is a new claim: every row is the limitation stated at the point of the result.

Scale. The primary experiments use Qwen3-8B-Base, with a mid-training-only replication at Qwen3-4B-Base (§C; scale-selection rationale in Appendix B.3). The replication carries the no-curve-fitting half of the moderation result—the moderate band is best in all five domains and the quadratic is concave with an interior vertex in all five, though for Counterfactual the three bands span just 1.3 pp—while the fitted peak locations are not scale-invariant (they move by -17.5 to $+3.8$ pp between 8B and 4B), so allocation guidance derived at one scale should not be transferred directly to another. The 4B arm also does not reproduce the 8B zero-coverage collapse on the same imbalanced allocation: Operation falls $60.0 \to 44.4$ (-15.6 pp, versus -27.6 at 8B) but Counterfactual is unchanged at $90.0 \to 90.5$ ($+0.5$ pp, versus -40.4 at 8B), where its 4B base prior is already near ceiling. That last explanation needs care, because Appendix B.3 pre-commits that a 4B failure "could not be explained away as a ceiling or floor artifact". We draw that distinction after the fact, which is itself worth flagging. The two are not the same ceiling: the pre-commitment concerns the *aggregate* post-training headroom, which is matched across the two scales ($+21.2$ vs. $+20.8$ pp), whereas what saturates here is Counterfactual's own base prior at 4B. The pre-commitment therefore still governs the sweep as a whole—if no domain showed an effect at 4B, the matched aggregate headroom would rule out a capacity explanation—but it does not make this single near-ceiling cell informative about Finding 2, and we do not read it as a test of that finding either way. Finding 2 is therefore an 8B observation and is not reproduced at the second scale. The alignment-pass test, the θ^* ordering, and the fitted 95%-of-peak intervals were not re-tested at 4B, and no other scale or architecture was run through the sweep; whether the non-closure and the fitted intervals carry across scales therefore remains open.

Scope and inference. Claims are scoped to KOR-Bench and three external benchmarks, the Qwen3-Base family (8B primary, with a mid-training-only 4B replication), the reported data mixture, and the tested finite-budget recipe. The five KOR-Bench domains are semantically rule-disjoint but are not independent in the statistical or neural sense: they share language, tokenization, output formats, model parameters, and potentially general reasoning procedures, so the per-domain contrasts are mixture-level rather than independent probes (Appendix A). The central inferential limit is that the fitted peak locations, the 95%-of-peak intervals derived from them, and θ^* are all read off the same sweep that produced them: the 24-point fits, those intervals and the argmax of the fitted objective reuse the same allocations for model selection. The curve *shapes* are tested against six allocations withheld from the fit (mean absolute held-out residuals 1.0–1.8 pp for four of five domains; Appendix G.4), and the compensation thresholds are calibrated to the intervention scale and the seed noise of this sweep rather than to an external null, though the non-closure conclusion is unchanged at every threshold tested (Appendix H.2, Figure 8). What has no held-out check is therefore the fitted peak locations, the intervals read off them, and θ^* itself, together with θ^*'s own full-pipeline row on the matched protocol: the argmax of the fitted objective and its $+4.36$ pp figure were selected on that same sweep and evaluated under a different downstream recipe from the held-out allocations (Appendix G.4). Six held-out allocations carried through the complete mid-training+SFT+RL pipeline (Table 16) now probe the mid-training-only curves and the compensatory-SFT result, which they roughly track at seed-noise level for Cipher, Logic, Counterfactual, and Puzzle (Operation's held-out low-coverage points sit ≈ 6 pp above the fitted tail; 0/60 pairs closed at 5 pp), and carrying them through the RL stage leaves that unchanged ($0/60$ at 5 pp after the complete pipeline), so the RL leg is covered out-of-sample as well. The six serve two roles: they are withheld from the curve fit, which is what makes the out-of-sample test valid, but they also enter the 30-allocation pool used for the band comparison, the interiority permutation test and the compositional surface, so those results are in-sample with respect to that pool (Appendix G.1). The simplex design additionally makes the fitted curves mixture-level marginal associations rather than isolated per-domain causal effects, and we treat this confound explicitly rather than implicitly (§5.2). We report the compositional analysis this calls for (§G.2): the isometric log-ratio surface fits well ($R^2 = 0.78$–0.96 per domain over the 42-allocation pool), but its stationary point reads as a saddle in every domain, and that classification is provisional because the negative curvature is shallow and quadratic misspecification is untested—so a jointly optimal mixture is neither identified nor excluded, and the moderation result stands as a per-domain marginal statement and is not promoted to a single recommended mixture. The 12 interior allocations of Appendix B.4 were added to test the joint question and did not produce a joint maximum, but they fall within the moderate band by construction, so they thicken that band rather than independently confirming it (Appendix G.1). Finding 2 is the one finding whose control cuts in its favour rather than against it. The FineWeb-Edu-only control shows that continued pre-training on unrelated text degrades out-of-scope reasoning on its own—it falls below the untouched Base model on all five KOR-Bench domains—but it degrades by *less* than zero KOR-Bench coverage in both domains Expt. 1 zeroes (-34.8 vs. -40.4 pp on Counterfactual; -15.2 vs. -27.6 pp on Operation; Table 16, Table 8), so the selective component is not accounted for by generic distributional drift. This does not make it a bound: the control arm carries no KOR-Bench data at all, so it also has zero coverage of the two

domains being probed, and the two arms therefore differ in corpus composition as well as in the coverage of those domains. What the contrast establishes is that generic drift is not sufficient to explain the collapse, not that coverage is its sole cause (§5.3).

The θ^* aggregate decomposed. The $+4.36$ pp aggregate on the θ^* row of Table 1 is the macro-average of five per-domain movements over the shared SFT+RL baseline that partially cancel—$+16.7$ (Logic), $+6.4$ (Counterfactual), $+3.9$ (Puzzle), -0.2 (Operation) and -5.0 (Cipher)—and the four non-null movements are each a large multiple of the cross-seed SD of the two rows being compared (1.9–7.6×). The underlying per-domain accuracies are in Table 1.

Our row-level inferential statistics are Welch two-sample t-tests, in two families. On the overall full-pipeline scores (§4.2) they are at best nominally significant and do not survive multiplicity control (smallest four-way adjusted value ≈ 0.08). On the five per-domain θ^* vs. shared-SFT+RL row comparisons they are larger: uncorrected they give $t \approx 12.0$, 4.4, 3.2, -3.4 and -0.1 for Logic, Puzzle, Counterfactual, Cipher and Operation, and under a five-fold Bonferroni bound over the five domains Logic, Puzzle and Cipher remain significant and Counterfactual is not (§4.2). The aggregate ordering is therefore nominal rather than significant, and the per-domain movements are the better-supported object. The comparison is one pair of rows of Table 1 (Expt. 3 against the shared SFT+RL baseline), so these describe that pair's profile rather than a general property of the domains. Seeds are matched across pipeline rows (Appendix D.4), so both families are reported as the conservative unpaired statistics and any per-domain delta without a reported test is descriptive; paired seed-level tests are left to the released per-seed data (Appendix D.8). Measurement protocol, including the reading of the bootstrap intervals against the cross-seed uncertainty, is given in §3.2 and Appendix D.4. GSPO uses a fixed 200-step budget; because binary verifier rewards provide no dense per-token reward signal, longer schedules were unstable across seeds (Appendix D.7), and longer or differently designed RL may yield different outcomes. Compensatory SFT conclusions are scoped to this one-parameter family and its fixed total SFT token budget (Appendix H.3, H); stronger reweighting within the family is measured rather than open—raising the concentration exponent closes $0/60 \to 12/60$ pairs only by forfeiting mean gain ($+4.34 \to +2.26$ pp; Table 18)—while per-token or per-loss reweighting outside the family, curricula, and adaptive sampling remain untested, and whether a longer or more stable RL schedule following compensatory SFT would close gaps across the full 24-configuration sweep remains open (Appendix D.7). The present design does not separately identify coverage from unique-example diversity, repetition, sequence length, or template- and rule-family-level overlap. The external component is identical in composition and token count across the sweep configurations, so it enters as a shared level offset rather than a per-configuration difference (§4.4, §5.2); the FineWeb-Edu-only arm is the one arm in which corpus composition changes, and it is reported as a drift control rather than a coverage contrast (§5.3). Template-level decontamination, and family-level decontamination beyond the certifications of Appendix F, are left for future work.

Multiplicity ledger. Table 34 maps each inferential family reported in this paper to its test and its correction. No family is reported without its correction, and paired seed-level tests are left to the released per-seed data (Appendix D.8) rather than inferred from the published summary statistics.

Table 34: Multiplicity ledger. Each inferential family, the comparison it covers, the test used, and its correction. Families whose counts are descriptive (pairwise closure, sharpening branches) claim no p-value beyond the permutation null named here.

Family	Comparison	Test	Correction
Overall pipeline scores (§4.2)	4 overall contrasts	unpaired Welch t	four-way Bonferroni (smallest adjusted ≈ 0.08)
Per-domain θ^* vs. SFT+RL (§4.2)	5 domain rows	unpaired Welch t	five-way Bonferroni
Interiority (§4.1)	5 domains	whole-row permutation (20,000)	family-wise permutation P
Pairwise gap closure (§H.2)	240 pairs	within-configuration permutation (2,000)	permutation P; nominal binomial p not interpreted
Sharpening branches (§H.3)	4 γ settings	descriptive counts	none claimed

Ethical considerations. This work studies data-mixture design for mid-training. It does not involve human subjects or personally identifiable information. Our training corpora are newly synthesised instances generated from public benchmark rule definitions rather than reused benchmark items; answers and traces come from deterministic symbolic

solvers for the $> 90\%$ of items they cover and from Claude Fable 5 for the remainder, and the SFT traces come from the same two sources. One control arm instead trains on FineWeb-Edu, a publicly released web-derived corpus supplied by a third party; that arm inherits that release's collection and filtering, and any residual gaps in it, rather than our own. Trained models are research checkpoints not intended for deployment, and the coverage-allocation guidance we derive is a research recommendation for the tested setting rather than a validated deployment procedure: it is estimated within one model family, one benchmark family and one budget, and the objective that selected our illustrative allocation ranks held-out allocations only weakly. It is compute-intensive at research scale: 630 training runs in the primary 8B study, of which 160 are mid-training passes at a fixed ≈1.5B-token budget over two epochs each, plus 36 further mid-training runs at 8B (Experiment A) and 24 mid-training-only runs at 4B (Appendix B.1); we did not instrument GPU-hours or energy use per run. One direction in particular must not be read off this paper. Our finding that a domain is worse served by a high share than by a moderate one is a within-domain accuracy statement about synthetic rule families at one budget; nothing here speaks to the deliberate over-sampling of safety, refusal, or low-resource-language data, which is a different design decision taken for different reasons and is not a coverage trade of the kind we measure. Benchmark balance should not be equated with user benefit, safety, calibration, or fairness in a deployed system: the domains here are synthetic symbolic-reasoning rule families and we measure only answer accuracy, so no safety, calibration or fairness metric is evaluated in this study, and the coverage-share caution here is a within-domain accuracy caution rather than a safety result. In high-stakes domains the coverage share itself carries risk in both directions: in our pooled band means a domain is worse served by a high share than by a moderate one in every domain we tested, as well as by a low share (though the high band is thin, $n = 1$–5, so the low-side contrast is the better-supported half), so a deployment decision would need domain-specific assurance at either end rather than only for under-covered domains.

Declaration of AI Assistance

We utilized ChatGPT and Claude for grammatical checking and LaTeX support of the content presented in this study but did not use them for the initial draft of this study. Cursor, Codex, and Claude Code were utilized for trivial and boilerplate code completion during data analysis. We declare that all content presented and code utilized in this study has been reviewed and edited by the authors.