LONGER ANSWERS, HEAVIER QUESTIONS: HOW LOSS WEIGHTING SHAPES COMPARISONS OF SYNTHETIC REASONING DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

When a language model is fine-tuned on synthetic reasoning data with a token-mean loss, each question counts in proportion to the length of its target. We call this *implicit question weighting*. A recipe that sets how much reasoning each question receives therefore also sets how much each question counts in training. With targets of similar length the weights are close to uniform; in the mixed-depth corpus we study, direct-answer targets are 42% of the questions but receive under 6% of the loss weight. We measure how strongly a comparison between two recipes depends on this weighting. Both recipes come from one teacher pool on the same $11{,}853$ questions: one gives every question a brief explanation, the other mixes direct answers with deeper reasoning. With the targets held fixed, replacing the token mean by an equal weight per example causes no detectable change for the fixed-depth corpus but costs the mixed-depth corpus 8.6 and 5.5 points of four-benchmark accuracy on Qwen3-4B-Base and OLMo-3-7B (three seeds per arm), so the lead of the fixed-depth corpus grows from 1.6 to 9.9 points on one student and from 3.1 to 8.3 on the other. Training the fixed-depth targets with the mixture's own question weights does not reproduce this response, so the weights alone do not explain it. With each source dataset's share of the loss held fixed, moving the direct-answer targets to their share of the examples costs more than equalizing the examples within the direct-answer and the reasoning groups. Comparisons of reasoning datasets should therefore report their loss objective and the loss share of their shortest targets.

1 INTRODUCTION

A synthetic reasoning dataset pairs each question with a target, which may be a bare answer, a brief explanation or a long reasoning trace, and recipes differ in which questions receive which kind of target (Guha et al., 2026; Zhang et al., 2026; Yu et al., 2025a; Li et al., 2025a). Training adds a choice that the dataset does not specify: how the loss weights targets whose lengths differ by three orders of magnitude (Bai et al., 2024; Yao et al., 2025). Under the token mean, an example's coefficient is proportional to its number of target tokens, so each question is weighted by the length of its target; we call this *implicit question weighting*. In the mixed-depth corpus of our main comparison ($11{,}853$ questions; Section 4), the targets generated as direct answers (T1, the shallowest of five prompted depths T1–T5) are 42.1% of the questions but receive 5.6% of the loss coefficients on Qwen3, while the 0.7% of questions whose targets come from the teacher's thinking mode receive 13.0% (Figure 1). Choosing a target for a question therefore also chooses how much the question counts in training, and a comparison of two recipes that leaves the weighting unstated cannot separate the two choices.

That the averaging convention matters is established: token and example averaging are contrasted, switched or blended in Tülu 3, Nemotron 3 Super, LongAlign, Hierarchical Balance Packing, GritLM and on-policy supervised fine-tuning (Section 2). We introduce no normalizer and claim no general ranking of them. Our question is narrower: with the training rows and their targets held fixed, how much of the measured difference between two synthetic reasoning corpora is decided by their implicit question weighting, and which specified redistribution of weight inside a corpus carries that effect?

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.

We answer it with one teacher pool of prompted depths, two student families (Qwen3-4B-Base and OLMo-3-7B, base checkpoints), three seeds per arm and four benchmarks. Every objective comparison holds the training rows, their targets, their order, the seeds and the optimization settings fixed and changes the loss coefficients; the token-mean arms reused from the registered comparison ran with cuDNN attention and the later arms without it (Appendix O). The recipe-by-objective comparison and its shared-weight control run on the $11{,}853$ questions for which both recipes can be built; the interventions of Section 5 and the checks of Section 6 run on the full $12{,}108$-question mixture unless stated. The study was staged (Section 3): a broader recipe comparison was registered before any student was trained (Appendix B), and each later experiment was specified after earlier results had been seen, with its predictions fixed before its own runs. The teacher pool and one studied recipe, a reliability-guided mixture of depths, come from prior work (Maqsood et al., 2026); the hypotheses registered for that recipe are reported in full, including those not supported (Section 6).

What we find. On the same questions, switching from the token mean to an equal-example objective costs the mixed-depth corpus 8.64 and 5.46 points on the two students, while for the fixed-depth corpus no penalty is detected. The difference between the two penalties is 8.31 and 5.23 points; its sign was predicted before these runs (Section 4; Appendix O, item (x)). Trained with the mixture's own question weights, the fixed-depth targets do not gain (-2.24 and -0.43 points), so the interaction is not explained by the question weights alone. The effect does not need redistribution of loss between datasets: with every dataset's loss share held fixed, equalizing the examples inside the datasets still costs 9.28 and 6.24 points. Inside a dataset the switch does two things at once, and a post hoc panel separates them with the targets and the dataset shares fixed. Moving the T1 targets to their example share of the loss costs 7.74 and 5.22 points; equalizing the examples within the T1 and the T2–T5 groups costs less (1.81, and on OLMo 0.20, not distinguishable from zero), although it moves more weight away from the longest targets (Section 5). These are the measured costs of two specified interventions in this setting; they order the interventions.

Contributions. (i) We quantify implicit question weighting, how strongly the token mean concentrates the loss of a mixed-depth corpus on few questions, with diagnostics that need no training (Section 3). (ii) A paired recipe comparison showing that the measured difference between fixed-depth and mixed-depth supervision depends strongly on the weighting objective, with the targets held fixed inside each objective comparison, and a shared-weight control showing that this is not explained by the question weights alone. (iii) Fixed-target interventions that separate redistribution between datasets, between the T1 and the T2–T5 group inside a dataset, and within those groups, and that order the tested redistributions by their accuracy cost. Supporting and negative results are reported with them: the objective gap survives separately selected learning rates, a reliability-guided allocation of depths shows no consistent advantage over the matched-random control (random placement of depths, matched in questions, depth mix and token budget), and the registered non-inferiority hypothesis for the guided mixture is not supported (Section 6). The supplementary material contains the code, the protocols, the training targets (withholding rule in Appendix N) and the summary result files, with benchmark questions identified by source, identifier and hash (Appendix M).

2 Related work

Loss normalization. Prior work shows that loss aggregation affects supervised fine-tuning and output length (Lambert et al., 2025; Muennighoff et al., 2025; Bai et al., 2024; Yao et al., 2025), and BLOOMZ already normalized each target's loss by its length so that short and long targets count equally, reporting longer generations and lower validation accuracy without that normalization (Muennighoff et al., 2023); on-policy SFT reports a normalization-related collapse of reasoning length and accuracy (Zhao et al., 2026), and Nemotron uses token and conversation averaging in successive stages (NVIDIA, 2026a;b). GritLM also studies aggregation over groups of examples (Muennighoff et al., 2025). Closest to our group-level result, TLDR changes the sampling proportions of short and long reasoning data during training, guided by development accuracy and response length (Li et al., 2026); in reinforcement learning, DAPO replaces GRPO's sample-level loss reduction with token-level policy-gradient loss aggregation (Yu et al., 2025b), and Dr. GRPO removes response-length-dependent loss normalization (Liu et al., 2025). Relative to this evidence, we hold the training rows and target strings fixed, show that the objective alone moves a recipe

comparison by five to eight points, separate the direct-answer targets' share of the loss from the weighting inside the groups with interventions on dataset shares, group shares and within-group weights, and give diagnostics that need no training (Appendix A).

Reasoning-data construction and mixtures. OpenThoughts studies data recipes for reasoning models (Guha et al., 2026), DC-CoT benchmarks data-centric choices for reasoning distillation (Zhang et al., 2026), Chen et al. (2025) examine reasoning granularity, and compression studies shorten or restructure long teacher traces (Griot et al., 2026; Lyu et al., 2026). LS-Mixture combines long traces with shorter rewrites (Yu et al., 2025a), NaturalThoughts evaluates mixtures of direct and reasoned responses (Li et al., 2025a), and MI-Distillation selects among teacher trajectories for the student (Lan et al., 2026). Our depths are complete generation policies, not compressed versions of one trace, and our allocation experiment asks a narrower question than those selectors (Appendix A); the recipe comparison itself (Appendix B) is reported as context.

3 SETUP AND OBJECTIVES

Teacher pool and recipes. We build on the candidate pool and the reliability-guided selector ("Minimum Sufficient Reasoning", here GUIDED-MIX) of Maqsood et al. (2026). In that pool a Qwen3.6-35B-A3B teacher (Qwen Team, 2026) answers 14,870 training problems from GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), ARC-Challenge (Clark et al., 2018) and StrategyQA (Geva et al., 2021) at five prompted depths. T1 asks for a direct answer (64-token cap), T2 for at most two or three sentences of reasoning (300), T3 for concise step-by-step reasoning (900), T4 for a short plan and its execution (1,400), and T5 enables the teacher's thinking mode (8,192). Each depth is sampled $k{=}8$ times on yes/no and multiple-choice questions and $k{=}4$ times on free-form mathematics, and every sample is checked against the gold answer. The depths are complete prompting configurations, not a pure manipulation of length. Two recipes carry this paper: GUIDED-MIX takes, per question, the shallowest depth whose within-depth pass rate clears a per-format reliability bar and then its shortest correct sample, so its corpus mixes depths; FIXED-T2 always takes the shortest correct T2 sample. The other recipes are compared in Appendix B: ANSWER-ONLY (the gold answer), FIXED-T3 and FIXED-T4 (the shortest correct T3 or T4 sample), SHORTEST-CORRECT (the shortest correct sample at any depth), LONG (a random correct T5 trace) and BON-SHORT (the shortest correct of 15 separately generated long samples; cf. Munkhbat et al., 2025).

Populations. 12,708 problems have a target under each of the four recipes of the registered comparison (LONG, BON-SHORT, SHORTEST-CORRECT, GUIDED-MIX); we hold out 600 of them (150 per dataset) for development and train on the remaining 12,108 (GSM8K 7,073, MATH 2,425, ARC 1,225, StrategyQA 1,385). 11,853 of these have a correct T2 sample, and the recipe-by-objective comparison of Section 4 uses exactly those, with GUIDED-MIX-T2SUB denoting the restriction of the mixture to them; 37 of them lack a correct sample at T3, at T4 or at both and keep their T2 target in FIXED-T3 and FIXED-T4. All interventions of Section 5 and the checks of Section 6 use the full 12,108-question mixture. A comparison between a recipe trained on 11,853 questions and one trained on 12,108 is descriptive and labelled as such. Within each population, recipes share questions, prompts and the three-epoch schedule, but not token budgets. Targets were kept verbatim; Appendix C counts the instruction phrases and bare answers they contain.

Students, training and evaluation. We fully fine-tune Qwen3-4B-Base (Yang et al., 2025) and OLMo-3-7B (Team Olmo et al., 2025) (base checkpoints) for three epochs with a configured global batch of 64 examples, AdamW ($\beta{=}(0.9, 0.95)$, no weight decay), cosine decay with 3% warm-up, gradient clipping at 1.0 and bfloat16, without packing; 12,108 questions give 570 optimizer steps and 11,853 give 558. The learning rate was selected once per student from $\{5{\times}10^{-6}, 10^{-5}, 2{\times}10^{-5}\}$ on the development accuracy of the long recipe and then shared by all recipes ($2{\times}10^{-5}$ for both students); Section 6 selects it separately for each objective. Students are evaluated zero-shot on GSM8K test (1,319), MATH-500 (Lightman et al., 2024), ARC-Challenge test (1,172), a StrategyQA split of 300 questions held out before any teacher generation, and, as a secondary out-of-distribution check, eight BBH tasks (Suzgun et al., 2023) (1,937 questions; Appendix M). The primary metric is the equal-weight mean accuracy over the first four. We report one sampled

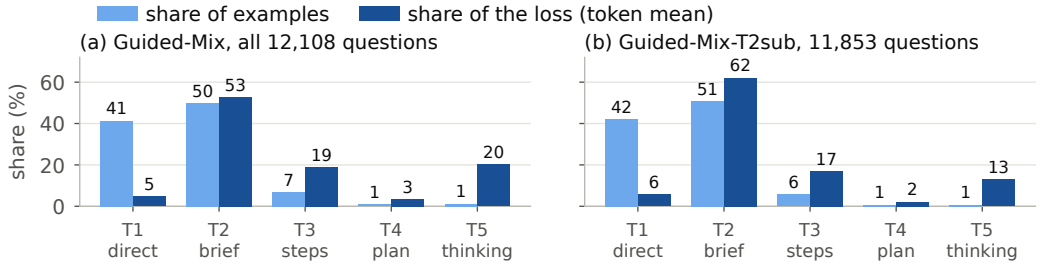

Figure 1: Implicit question weighting: share of the training examples (light) and of the accumulated token-mean loss coefficients (dark) per prompted depth, for the full GUIDED-MIX set and its 11,853-question restriction (Qwen3 tokenizer; OLMo within one point, Appendix D). Direct answers (T1) are 41 to 42% of the questions but carry about 5% of the loss; the 1% of thinking-mode targets carry 13 to 20%.

completion per question (temperature 0.6, top-p 0.95, top-k 20, at most 8,192 tokens, a fixed seed per question) and greedy decoding as a sensitivity check. An unextractable answer counts as incorrect; reaching the token cap is reported but not forced incorrect. We call a response of at most 16 generated tokens a *direct answer* (Appendix L).

The two objectives. With targets that range from a bare answer to thousands of tokens, the averaging convention of the loss decides how much each example counts. For a global accumulation window of n examples, let t_i count the supervised tokens of example i and $\bar{\ell}_i$ be their mean cross-entropy; we compare

$$L_{\text{tok}} = \frac{\sum_i t_i \bar{\ell}_i}{\sum_i t_i}, \qquad L_{\text{ex}} = \frac{1}{n} \sum_i \bar{\ell}_i.$$

Unless stated otherwise students are trained with L_{tok}, so an example's weight is proportional to its supervised tokens (the loss of a final partial window is normalized by the examples and tokens actually processed, as in the gradient-accumulation corrections of Debut et al., 2024; Han & Han, 2024). Under the token mean the T1 examples (targets generated at depth T1, as direct answers), 41% of GUIDED-MIX's full 12,108-question training set, carry about 5% of the accumulated loss coefficients (each example's coefficients w_i summed over training; Figure 1, and Appendix D for every depth's share), and the 1% of examples with thinking-mode targets about 20%; the equal-example objective L_{ex} gives every example the same coefficient. We call this dependence of a question's weight on the length of its target *implicit question weighting*. On the 11,853 questions of Section 4 it gives GUIDED-MIX-T2SUB an effective number of 2,917 examples ($1/\sum_i c_i^2$ for accumulated coefficient shares c_i that sum to one), against 10,356 for FIXED-T2 and 11,849 under equal-example weighting (Qwen3; OLMo: 2,779, 10,383 and 11,849). Both conventions are established (Bai et al., 2024; Yao et al., 2025; NVIDIA, 2026b). Unlike changing the proportion of short examples (Wang et al., 2025), this intervention holds the training set fixed and changes only the coefficients. We verified both objectives, and the diagnostic objectives of Section 5, against their definitions through the training code, including partial accumulation windows and, on small random models, the distributed path (Appendix D).

Statistics and staging. Contrasts use a paired bootstrap that resamples questions within each dataset (10,000 resamples) on seed-averaged correctness, with percentile intervals, two-sided p-values ($2\min\{P(\Delta^* \le 0), P(\Delta^* \ge 0)\}$; the registered non-inferiority test is one-sided) and Holm correction at level 0.05 inside families of hypotheses that were fixed before the corresponding runs. Intervals are unadjusted, conditional on the trained checkpoints, and resample questions, not seeds; Appendix I adds 95% t-intervals over the three seed-wise differences, which, among the 55 contrasts listed there, exclude zero wherever the bootstrap intervals do except for four, each marked where it is used. When no resample falls beyond zero, p is below the bootstrap's resolution of 2×10^{-4}, and the Holm-adjusted value is given as a bound. The study was staged; we use four stage labels. *Registered*: fixed in a written protocol whose hash was recorded before any student was trained; this covers the recipe comparison of Appendix B and its six contrasts. *Pre-specified*:

Table 1: Scope of the experiments. "Fixed" names what is identical between the arms of a comparison; the objective comparisons additionally keep the target strings, the question order and the seeds. Stage: the labels defined in Section 3 (pre-spec.: pre-specified). The allocation control matches each dataset's supervised-token total within 0.5%.

Experiment	Questions	Fixed	Changed	Stage	Sec.
Recipe $\times$ objective	11,853	questions, schedule	recipe, objective	pre-spec.	4
Shared question weights	11,853	questions, question weights	targets	post hoc	4
Between or within datasets	12,108	targets	where loss is shared	pre-spec.	5
Depth groups in a dataset	12,108	targets, dataset shares	group, within-group	post hoc	5
Learning rate per objective	12,108	targets	objective, rate	pre-spec.	6
Allocation of depths	12,108	questions, depths, tokens	placement of depths	pre-spec.	6
Recipe comparison	12,108 / 11,853	prompt, schedule	complete recipes	registered	B

fixed in a later protocol, written after the results of the registered comparison had been analysed and before any of its own runs, with predictions and hypothesis families. It covers the experiments so labelled in Table 1, the depth ladder and the OLMo runs of ANSWER-ONLY, FIXED-T2 and GUIDED-MIX-T2SUB; the Qwen3 token-mean cells of the recipe-by-objective comparison are the registered FIXED-T2 versus GUIDED-MIX-T2SUB contrast. *Post hoc*: designed in response to results of the pre-specified experiments themselves, with objectives, tests and claim rules fixed before their own runs; this covers the panel of Section 5 that splits the within-dataset switch and the shared-weight control of Section 4, and descriptive comparisons added after results were seen carry the same label without a test. *Descriptive*: comparisons without a test, such as the fixed depths against long traces. Appendix O lists the protocol amendments and deviations. Table 1 lists what each experiment holds fixed and what it changes.

4 THE COMPARISON BETWEEN TWO RECIPES DEPENDS ON THE WEIGHTING

Holding targets, question order, seeds and optimization settings fixed (attention kernel excepted, Appendix O), the equal-example objective lowers four-benchmark accuracy on the full mixture (descriptive; Appendix H) by 9.15 points $[7.75, 10.54]$ for GUIDED-MIX and 9.67 $[8.37, 10.97]$ for the matched-random control of Section 6 on Qwen3, and by 5.25 $[3.93, 6.60]$ and 6.36 $[5.13, 7.61]$ on OLMo. The lost accuracy is concentrated in mathematics and in the BBH subset, where part of it is a format effect: bare labels that the grader does not accept (Appendix L). Unextractable answers do not explain it, since in every loss-weighting arm, under sampled decoding, at most 3.5% of the MATH-500 responses and 0.3% of those on the other primary benchmarks lack an extractable answer; instead, students answer directly more often and less accurately (Appendix J). Is this a property of loss weighting in general, or does it depend on what the corpus contains?

The same switch, two recipes, one set of questions. On the $11{,}853$ questions that have a correct T2 sample we trained both the fixed brief depth and the restricted mixture GUIDED-MIX-T2SUB under both objectives, with identical questions, order, seeds and optimization settings, the attention kernel excepted (Appendix O). The two recipes share questions and, where the mixture chose depth T2, targets; each objective contrast holds its own targets fixed. For the fixed-depth corpus we detect no penalty (Figure 2a): token mean minus equal example is $+0.33$ points $[-0.60, +1.26]$ on Qwen3 and $+0.23$ $[-0.63, +1.10]$ on OLMo. The mixed-depth corpus loses 8.64 $[7.31, 10.01]$ and 5.46 $[4.20, 6.74]$. The pre-specified interaction, the difference between the two penalties, is $+8.31$ points $[6.77, 9.87]$ on Qwen3 and $+5.23$ $[3.70, 6.74]$ on OLMo, positive on every seed, with Holm-adjusted $p \leq 4 \times 10^{-4}$ in both students (family of two); greedy decoding gives $+7.36$ and $+5.10$.

What this does and does not show. Under the shared optimizer and schedule, how these two recipes compare depends on the objective. The direction does not change: the fixed-depth corpus is ahead of the mixture under both objectives, by $+1.56$ points $[0.24, 2.86]$ on Qwen3 (positive on every seed; seed-level interval $[-0.67, +3.80]$, Appendix I) and $+3.10$ $[1.70, 4.52]$ on OLMo under the token mean (Appendix B); what the weighting changes is the size of that difference, which grows to 9.88 and 8.33 points under equal-example weighting. The same data thus show the two recipes as close or as far apart depending only on the loss objective. The two corpora differ in reasoning

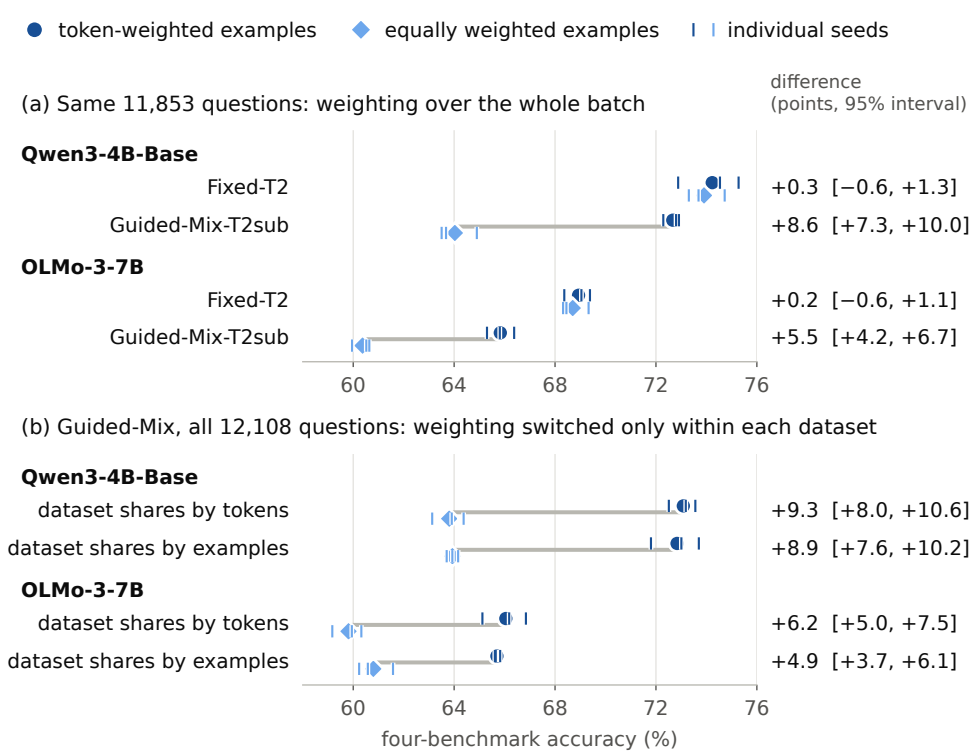

Figure 2: Sensitivity to loss weighting. Four-benchmark accuracy under sampled decoding (markers: mean of three seeds; ticks: individual seeds) with token-weighted and with equally weighted examples; right: token-weighted minus equally weighted, in points, with a 95% paired-bootstrap interval. (a) Two recipes trained on the same 11,853 questions, weighted over the whole batch (token mean versus equal example): no penalty is detected for the fixed-depth corpus, whereas the mixed-depth corpus loses five to nine points. (b) GUIDED-MIX targets on all 12,108 questions, with the weighting switched only *within* each dataset while the datasets' loss shares stay at their token shares (upper row, starting at the token mean) or at their example shares (lower row, ending at the equal-example objective); the other two points are the diagnostic objectives of Section 5. The penalty appears in both rows, so rebalancing datasets is not needed for it.

content and in the distribution of target lengths, so the comparison does not by itself identify which difference is responsible.

The same question weights, other targets. The switch moves the mixture's question weights far more than those of the fixed-depth corpus (effective number of examples 2,917 against 10,356 on Qwen3; Section 3), so the interaction could reflect the size of the weight change rather than the targets. A post hoc control, with its main estimate, reading rule and analysis code fixed before its runs, removes that difference: the fixed-depth targets are trained with the mixture's own token-mean question weights, window by window (Appendix F). Relative to equal-example weighting, these weights change the fixed-depth corpus by -2.24 points $[-3.44, -1.05]$ on Qwen3 and -0.43 $[-1.68, +0.83]$ on OLMo, while they raise the mixture by $+7.91$ and $+5.46$ points (on Qwen3 with the mixture arm retrained under the same attention kernel; 8.64 with the registered runs, Appendix F). By the reading rule fixed before the runs, the question weights alone reproduce less than half of the mixture's response on both students, also under greedy decoding and under the scoring sensitivity of Appendix L, and the test of the difference between the two responses, fixed in the control's protocol, is rejected on both. What the switch does to the mixture is therefore not explained by the question weights alone. The rest of this paper asks which redistribution inside the mixed corpus carries the penalty.

Table 2: Between or within the depth groups (post hoc), with the targets and every dataset's loss share fixed. Upper block: what each condition does; the code labels are those of the released analysis files (the suffix '-mode' marks this panel's objectives); their first letter says how the T1 and T2–T5 groups share a dataset's loss, the second how the examples within a group are weighted (T: by supervised tokens, E: equally). Lower block: two intervention penalties, each token-mean accuracy minus accuracy after one change, and their difference (sampled decoding, three seeds; unadjusted 95% intervals conditional on the trained checkpoints; * rejected after Holm correction in a family of six; Appendix E).

What the condition does	Share of the two groups	Weight within a group	Label
The training objective of the study	by supervised tokens	by supervised tokens	TT
Equalize examples within the groups	by supervised tokens	equal per example	TE-mode
Move the groups to example shares	by example counts	by supervised tokens	ET-mode
Both changes (the within-dataset switch)	by example counts	equal per example	EE-mode

Change inside every dataset	Qwen3-4B-Base	OLMo-3-7B
Between groups: example shares (TT − ET-mode)	7.74^{*} $[6.47, 9.01]$	5.22^{*} $[3.90, 6.53]$
Within groups: equal examples (TT − TE-mode)	1.81^{*} $[0.74, 2.85]$	0.20 $[-0.87, +1.27]$
Difference (TE-mode − ET-mode)	5.93^{*} $[4.63, 7.27]$	5.02^{*} $[3.69, 6.39]$

5 Which redistribution of weight carries the penalty?

Between datasets or inside them? The two objectives differ in how the loss is shared between the four training datasets and in how it is shared between the examples of one dataset. We crossed the two ways on the GUIDED-MIX targets with two diagnostic objectives: one keeps every dataset's token-mean share and gives equal weight to the examples within a dataset; the other gives every dataset its share of examples and keeps token weighting within it (definitions and gradient checks in Appendix D). Equalizing examples within datasets, with the dataset shares untouched (Figure 2b), lowers accuracy by 9.28 points $[7.95, 10.62]$ on Qwen3 and 6.24 $[4.98, 7.46]$ on OLMo, on every seed (Holm-adjusted $p \leq 1.2 \times 10^{-3}$ in a pre-specified family of six). Changing only the dataset shares has an estimated penalty of 0.25 points $[-0.82, +1.26]$ and 0.35 $[-0.78, +1.49]$, and the interaction of the two switches is $+0.37$ $[-1.09, +1.84]$ and $+1.34$ $[-0.09, +2.73]$; none of these four contrasts is rejected, and greedy decoding agrees: the same Holm step, applied afterwards to the greedy values, gives the same decisions (Appendix H). Redistribution of loss between datasets is therefore not necessary for the penalty: under the tested settings it appears with the dataset shares held fixed, when short and long examples of the same dataset receive equal weight.

Inside a dataset: two changes in one. The within-dataset switch does two things at once. It moves the T1 targets (a label of the generation policy rather than of content) from their token share to their example share of the dataset's loss, and it equalizes the examples within the T1 and the T2–T5 groups. A post hoc panel of twelve runs, specified after the within-dataset result had been seen, applies each change alone to token-mean training, with the targets and every dataset's loss share fixed; its objectives, tests and claim rules were fixed before those runs (Appendix E). Table 2 names the four conditions and reports the two penalties and their difference. The within-dataset switch moves the estimated accumulated loss-coefficient share of T1 examples from 4.8% to 36.5% on Qwen3 and from 4.5% to 36.8% on OLMo; the dataset-share switch alone moves it to 10.4% and 10.1%. Students trained with the within-dataset switch answer directly more often, and less accurately, on mathematics (Appendix J).

Moving the groups costs more than equalizing within them. On both students, giving the two groups their example shares costs more than equalizing the examples inside them (Table 2): 7.74 and 5.22 points against 1.81 and 0.20, a difference of 5.93 $[4.63, 7.27]$ and 5.02 $[3.69, 6.39]$ that is positive on every seed; its sign and all six Holm decisions are unchanged under greedy decoding and under the scoring sensitivity of Appendix L. Both penalties compare the reused token-mean arm, trained with cuDNN attention, with a later arm trained without it; their difference compares two arms trained under the same setting (Appendix O). Equalizing within the groups still lowers accuracy on Qwen3 (1.81 $[0.74, 2.85]$), so the group shares alone do not explain the penalty there; on OLMo

its interval includes zero (0.20 $[-0.87, +1.27]$). With targets and dataset shares fixed, the specified redistribution between T1 and T2–T5 is thus more damaging than the specified redistribution within them. The two interventions differ in size, and which is larger depends on the measure, the T1 share or the longest tenth's share of the loss (Table 4, Appendix E). The two single changes are not an additive split of the within-dataset penalty, their interaction is exploratory, and raising the share of the T1 group necessarily lowers the share of the T2–T5 group, which these experiments cannot separate.

6 Robustness and supporting results

Does the gap survive a learning rate chosen for each objective? The shared learning rate was selected for the token-mean objective on long targets, so the equal-example objective might simply need a different rate. For each student and objective we trained the GUIDED-MIX corpus at three rates (10^{-5}, 2×10^{-5}, 4×10^{-5}) and selected one on the development set only (Appendix K). The gap between the objectives remains at the selected rates: $+8.50$ points $[7.18, 9.89]$ on Qwen3 and $+6.35$ $[5.01, 7.72]$ on OLMo, positive on every seed (Holm-adjusted $p \leq 8 \times 10^{-4}$ in the pre-specified family of four replication and tuning contrasts), and $+7.93$ and $+6.03$ under greedy decoding. In all four cells the lowest rate was selected, and students trained at it are more accurate under both objectives than at the shared rate (by 0.91 to 2.64 points), so a lower rate suits this corpus under both objectives; the penalty nevertheless persists after separate development-set selection among the three tested rates. Every selected rate lies at the lower edge of the tested grid (Appendix K).

Does it matter which questions receive explanations? GUIDED-MIX's distinctive ingredient is an allocation rule: the teacher's repeated success decides which questions receive a direct answer and which receive deeper reasoning. Its advantage over shortest-correct selection ($+5.72$ and $+4.82$ points; Appendix B) does not test that rule, because the two recipes also differ in how many questions receive explanations and in total tokens. We therefore built a control that keeps the questions, the depth histogram of every dataset and, approximately, the token budget, and randomizes the placement. Depths are assigned at random among those at which the teacher produced a correct sample (all but 12 questions have more than one), and the target is the shortest correct sample at the assigned depth; random swaps then match the supervised-token total of each dataset to GUIDED-MIX's within 0.5% under both student tokenizers. The overlap with GUIDED-MIX's own choices was left at its natural value (66%); we drew three allocations, paired each with a training seed, and reused them under both objectives. Under the training objective used throughout, GUIDED-MIX minus the control is $+0.73$ points $[-0.44, +1.92]$ on Qwen3 and -0.98 $[-2.13, +0.18]$ on OLMo; under the equal-example objective it is $+1.24$ $[+0.05, +2.47]$ (Holm-adjusted p=0.12; seed-level interval $[-0.42, +2.91]$) and $+0.13$ $[-0.96, +1.22]$ (Figure 4 in Appendix H). No contrast is rejected after correction, and under the token mean the signs disagree between students, so we find no consistent advantage of reliability-guided placement at matched depth mixture and token budget. This is not an equivalence result: no margin was set, the control keeps correctness filtering and most of GUIDED-MIX's own choices, and the intervals admit effects of up to about 2.5 points. Our predictions of a positive allocation effect, made before these runs, are not supported (the second holds in direction only; Appendix J), so the recipe's advantage over shortest-correct selection cannot be read as evidence for the placement rule.

Registered tests of GUIDED-MIX. The registered hypothesis that GUIDED-MIX would be non-inferior to BON-SHORT within one point is not supported on either student (-7.35 points $[-8.78, -5.90]$ on Qwen3 and -12.58 $[-14.07, -11.07]$ on OLMo), and on OLMo GUIDED-MIX is close to the untrained starting model on the four-benchmark mean, above it under sampling and below it under greedy decoding (66.0 against 64.2 sampled; 66.7 against 67.3 greedy). On OLMo, FIXED-T2, LONG and BON-SHORT gain 4.7 to 14.4 points over the same starting model, while SHORTEST-CORRECT falls 3.0 points below it and ANSWER-ONLY 18.9. A registered control that spends GUIDED-MIX's token budget on long traces ($1{,}085$ examples, 51 optimizer steps, Qwen3 only) is reported in Appendix B; its contrast depends on decoding and on how capped responses are read, so we draw no accuracy conclusion from it.

Context from the broader comparison. Targets that contain reasoning matter (GUIDED-MIX exceeds ANSWER-ONLY by 20.82 and 20.71 points), and on the same $11{,}853$ questions accuracy

rises with the depth of a fixed target at every pre-specified step; Appendix B reports the full comparison, in which no recipe dominates every benchmark.

7 Discussion and limitations

What a practitioner can take from this. Choosing the targets of a synthetic reasoning corpus also chooses its implicit question weights; the setting is common, since hybrid think/no-think corpora mix bare answers with traces (Wang et al., 2025) and training codebases differ in their default loss averaging (Lambert et al., 2025; NVIDIA, 2026b). A comparison of corpora should therefore state how the loss weights the targets: in the tested setting the measured gap between two recipes moved by 8.31 and 5.23 points with the data unchanged. Reporting the example shares and the loss-coefficient shares of the important groups (here T1 against T2–T5) makes that visible. A difference under another objective can be traced before it is credited to the selection policy: hold every dataset's loss share fixed, then separate group shares from within-group weights (Section 5).

One teacher pool, one question set. Targets come from one teacher (ANSWER-ONLY uses gold labels) and $12{,}108$ benchmark-derived questions in four datasets; the matched-random control is another policy on the same pool, not an independent corpus, so we make no claim about other synthetic reasoning datasets. **Two students, three seeds.** Both students are base models of 4 and 7 billion parameters, fully fine-tuned for three epochs; larger students, other families and parameter-efficient tuning are untested. Bootstrap intervals resample questions only, and seed-level intervals over three seeds are wide (Appendix I); the three matched-random allocations are paired one-to-one with the training seeds, so allocation and seed variation are not separated.

Implementation. The token-mean arms reused from the registered comparison ran with cuDNN attention, which was disabled for all later runs (Appendix O); retraining the three Qwen3 GUIDED-MIX-T2SUB runs without it changes their accuracy by -0.7 (sampled) and $+0.5$ (greedy) points (Appendix F). **Unequal interventions.** The tested redistributions differ in size and are not an additive split of the within-dataset penalty (Section 5). **Labels and mechanism.** T1 versus T2–T5 labels the generation policy, not the content of a target; the experiments isolate no mechanism (Appendix J). **Optimization.** The learning rate was selected once per student on the long recipe and shared by all recipes; the check of Section 6 selects a lower rate for the mixed-depth corpus under both objectives, so recipe comparisons, including the interaction of Section 4, hold under the shared settings and not at each recipe's own optimum. That check covers three rates on one corpus, and every selected rate lies at the lower edge of its grid.

Content differences. Targets were kept verbatim, so the two recipes differ in reasoning content and in the distribution of target lengths; the shared-weight control removes the difference in how the switch moves their question weights, not these, and it also concentrates the fixed-depth corpus on fewer questions (Appendix F). The comparison does not isolate an intrinsic property of mixed depth, and the small fixed-depth estimates are not evidence of an exactly zero effect. **Scoring and decoding.** Removing quoted format instructions from the responses before grading (Appendix L) changes no test decision and, outside the token-matched control, moves no earlier contrast by more than 0.26 points. On BBH, part of the equal-example penalty is a format effect of the grader, the same grader that filtered the teacher samples (Appendix L). Greedy decoding is a sensitivity throughout, Holm-corrected as a separate family where stated.

8 Conclusion

Under the token mean, a synthetic reasoning corpus weights its questions by the length of their targets. When this weighting changed, the measured difference between a fixed-depth and a mixed-depth corpus on the same questions moved by 8.31 and 5.23 points, and the mixture's own question weights did not help the fixed-depth targets. With targets and dataset loss shares held fixed, shifting weight towards T1 targets cost more accuracy than equalizing the examples within the T1 and T2–T5 groups, on both students. The penalty on the full mixed corpus survived learning rates selected for each objective, and a reliability-guided allocation of depths showed no consistent advantage over a matched-random control. We recommend that comparisons of data-construction recipes state their loss normalization and the share of the loss that their shortest targets receive.

AI USE STATEMENT

We used AI assistants for literature search, for writing and debugging the training, evaluation and analysis code and its checks, for the audits of stored responses reported in Appendix L, and for drafting and editing the manuscript. Teacher-model outputs form the synthetic supervision targets described in Section 3; the setup and the appendix describe the teacher, the generation settings, target selection and verification. The authors reviewed all AI-assisted work, approved every protocol, amendment and launch, made the final decisions on the experiments, the claims and the submitted version, and are responsible for the final text, citations, experimental claims, code and artefacts, including material produced with AI assistance.

ETHICS STATEMENT

The study uses public benchmarks (GSM8K, MATH, MATH-500, ARC-Challenge, StrategyQA and eight BBH tasks) and openly released models; it involves no human subjects and no personal data. The training targets are outputs of an openly licensed teacher model. Because benchmark licences vary, the supplementary material identifies questions by source, identifier and hash instead of redistributing their text, and withholds the training targets that reproduce eight or more consecutive words of a benchmark question (Appendices M and N). The contamination audit of Appendix M checks the training questions against the evaluation and development sets. The students are research checkpoints trained for this comparison and are not released for deployment.

REPRODUCIBILITY STATEMENT

Appendix C gives the teacher prompts, sampling settings and target construction; Appendix D the exact objectives, their verification and the micro-batch sizes; Appendices M and N the data sources, evaluation splits, prompts, serving settings, model revisions and software versions; Appendix O the staging of the study, its protocol amendments and every deviation. All contrasts, intervals, seed-wise values and Holm decisions are listed in Appendices E, F, H and I. The supplementary material contains the protocols, the training, evaluation and analysis code, the training targets keyed by question identifier (targets that reproduce eight or more consecutive words of a benchmark question are withheld and listed by identifier), question manifests with hashes, and the summary result files underlying every reported number, table and figure; the withheld targets and the per-question responses with their grades (2.4 GB) will be released with the final version.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. LongAlign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1376–1395. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.74. URL https://aclanthology.org/2024.findings-emnlp.74/.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Prasoon Varshney, Makesh Narsimhan, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi Mahabadi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf,

Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Shaona Ghosh, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika Katariya, Chris Alexiuk, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, and Eric Chung. Llama-Nemotron: Efficient reasoning models, 2025. URL https://arxiv.org/abs/2505.00949.

Xinghao Chen, Zhijing Sun, Wenjin Guo, Miaoran Zhang, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, and Xiaoyu Shen. Unveiling the key factors for distilling chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15094–15119. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.782. URL https://aclanthology.org/2025.findings-acl.782/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Lysandre Debut, Arthur Zucker, Zachary Mueller, Yih-Dar Shieh, Benjamin Bossan, and Pedro Cuenca. Fixing gradient accumulation. Hugging Face Blog, October 2024. URL https://huggingface.co/blog/gradient_accumulation. Published October 16, 2024.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21/.

Maxime Griot, Paul Steven Scotti, and Tanishq Mathew Abraham. Compress-Distill: Reasoning trace compression for efficient knowledge distillation, 2026. URL https://arxiv.org/abs/2606.05988.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Rea Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah M Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. OpenThoughts: Data recipes for reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7xjoTuaNmN.

Daniel Han and Michael Han. Bug fixes in LLM training – gradient accumulation. Unsloth Blog, October 2024. URL https://unsloth.ai/blog/gradient. Published October 15, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https:

//datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Mathew Huerta-Enochian and Seung Yong Ko. Instruction fine-tuning: Does prompt loss matter? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22771–22795, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1267. URL https://aclanthology.org/2024.emnlp-main.1267/.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH.

Yangsong Lan, Renkai Hu, HongKai Zheng, Bo Zhang, Renzhi Wang, Hongliang Dai, and Piji Li. MI-Distillation: Selecting from model-interpolated instruct-reasoning data spectrum for chain-of-thought distillation, 2026. URL https://arxiv.org/abs/2608.29623.

Yang Li, Youssef Emad, Karthik Padthe, Jack Lanchantin, Weizhe Yuan, Thao Nguyen, Jason Weston, Shang-Wen Li, Dong Wang, Ilia Kulikov, and Xian Li. NaturalThoughts: Selecting and distilling reasoning traces for general reasoning tasks, 2025a. URL https://arxiv.org/abs/2507.01921.

Yuan Li, Zhengzhong Liu, and Eric Xing. Data mixing optimization for supervised fine-tuning of large language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35419–35437. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/li25bh.html.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25366–25394, Vienna, Austria, July 2025c. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1301. URL https://aclanthology.org/2025.findings-acl.1301/.

Zhong-Zhi Li, Xiao Liang, Zihao Tang, Lei Ji, Peijie Wang, Haotian Xu, Xing W, Haizhen Huang, Weiwei Deng, Yeyun Gong, Zhijiang Guo, Xiao Liu, Fei Yin, and Cheng-Lin Liu. Too long, do re-weighting for efficient LLM reasoning compression. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39943–39962, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1856. URL https://aclanthology.org/2026.acl-long.1856/.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations*, 2024. URL https://mlanthology.org/iclr/2024/lightman2024iclr-let/.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Siyang Lyu, Xinghao Chen, Zhijing Sun, Tong Liu, Dawei Zhu, and Xiaoyu Shen. When compression helps and when it hurts: Condition-aware analysis of chain-of-thought distillation, 2026. URL https://arxiv.org/abs/2606.21704v2.

Maaz Maqsood, Nikhil Ranjan, Owais Aijaz, Ahmer Tabassum, Haonan Li, Hector Xuguang Ren, Zhengzhong Liu, and Eric P. Xing. Minimum sufficient reasoning: How much chain-of-thought does training data need? In *COLM 2026 Workshop on Efficient Reasoning*, 2026. URL https://openreview.net/forum?id=ali6mfCRJl.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL https://aclanthology.org/2023.acl-long.891/.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BC4lIvfSzv.

Tergel Munkhbat, Namgyu Ho, Seo Hyun Kim, Yongjin Yang, Yujin Kim, and Se-Young Yun. Self-training elicits concise reasoning in large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25127–25152, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1289. URL https://aclanthology.org/2025.findings-acl.1289/.

NVIDIA. Nemotron 3 Super: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for agentic reasoning, 2026a. URL https://arxiv.org/abs/2604.12374.

NVIDIA. Stage 1: Supervised fine-tuning (SFT). Nemotron documentation, Super3 SFT, 2026b. URL https://docs.nvidia.com/nemotron/latest/nemotron/super3/sft.html. Accessed September 20, 2026.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Hongjin Song, Runwu Shi, Weiqiao Shan, Jiale Luo, Yujin Wang, Yifei Wu, and Chunxiang Jin. Rethinking length-based training: Batch composition and loss normalization in speech token language models, 2026. URL https://arxiv.org/abs/2609.25890.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner,

Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Shouren Wang, Wang Yang, Xianxuan Long, Qifan Wang, Vipin Chaudhary, and Xiaotian Han. Demystifying hybrid thinking: Can LLMs truly switch between think and no-think?, 2025. URL https://arxiv.org/abs/2510.12680.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yongqiang Yao, Jingru Tan, Kaihuan Liang, Feizhao Zhang, Jiahao Hu, Shuo Wu, Yazhe Niu, Ruihao Gong, Dahua Lin, and Ningyi Xu. Hierarchical balance packing: Towards efficient supervised fine-tuning for long-context LLM. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 45509–45536. Curran Associates, Inc., 2025. doi: 10.52202/085713-1517. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/410bbba8388369d8bb5875544d1d4428-Paper-Conference.pdf.

Bin Yu, Hang Yuan, Haotian Li, Xueyin Xu, Yuliang Wei, Bailing Wang, Weizhen Qi, and Kai Chen. Long-short chain-of-thought mixture supervised fine-tuning eliciting efficient reasoning in large language models, 2025a. URL https://arxiv.org/abs/2505.03469v2.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244. Curran Associates, Inc., 2025b. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Ruichen Zhang, Rana Shahroz, Zhen Tan, Dawei Li, Song Wang, and Tianlong Chen. The quest for efficient reasoning: A data-centric benchmark to CoT distillation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Dud8FtScW7.

Anhao Zhao, Ziyang Chen, Junlong Tong, Yingqi Fan, Fanghua Ye, Shuhao Li, Yunpu Ma, Wenjie Li, and Xiaoyu Shen. On-policy supervised fine-tuning for efficient reasoning, 2026. URL https://arxiv.org/abs/2602.13407.

A Related work in detail

Loss aggregation and its scope. Several precedents show that the reduction applied to token losses can change training outcomes. LongAlign reweights examples inside packed sequences to recover equal sequence weighting and reports gains over its pack-local baseline; that baseline is not our global token mean (Bai et al., 2024). HBP compares Sample-Mean, rank-local Token-Mean,

Sum and Ave-Token normalization; its Ave-Token construction is the closest reported comparator to our global token mean and outperforms Sample-Mean on its reported general and long-context metrics (Yao et al., 2025). Tülu 3 illustrates token and example aggregation under gradient accumulation but selects a summed loss with a corresponding learning-rate adjustment, rather than our normalized global-token objective (Lambert et al., 2025). BLOOMZ normalizes the loss of each target by its length, so that short and long targets receive the same weight, the equal-example convention (Muennighoff et al., 2023). GritLM associates sample-level aggregation with shorter generations and token-level aggregation with longer generations in a joint embedding–generation setting, including intermediate grouped objectives (Muennighoff et al., 2025). DeepSeekMath formulates SFT as an expectation of the per-response mean token log-likelihood, without a comparative normalizer ablation (Shao et al., 2024). On-Policy SFT replaces each response's own length denominator with a shared batch-maximum denominator and reports stable training where the former collapses; its data are dynamically generated and correctness-filtered (Zhao et al., 2026). Nemotron 3 Super uses the same two endpoint formulas as successive stages while also changing packing length, batch size and data composition, so it is practical precedent rather than an isolated objective comparison (NVIDIA, 2026a;b). Whether prompt tokens enter the loss is a separate masking decision studied in instruction tuning (Huerta-Enochian & Ko, 2024). Closest to our group-level result, TLDR changes the sampling proportions of short and long reasoning data during training, guided by development accuracy and response length (Li et al., 2026); our panel instead keeps the training rows and every dataset's loss share fixed and separates the change of group shares from the change within groups. In reinforcement learning for reasoning, DAPO replaces GRPO's sample-level loss reduction with token-level policy-gradient loss aggregation (Yu et al., 2025b), and Dr. GRPO removes response-length-dependent loss normalization and reward-standard-deviation normalization, two optimization biases that it identifies in GRPO (Liu et al., 2025); their signed, on-policy objectives differ from supervised fine-tuning. Llama-Nemotron connects the higher learning rates it uses for long reasoning targets to sequence-length-dependent token-loss averaging (Bercovich et al., 2025, Section 4.1), which bears on our learning-rate check. Concurrent work on speech token language models separates length-based batching from the token weights of a batch-mean loss and reports, for one tokenizer, a perplexity gain from length grouping that is not observed under a token-balanced loss (Song et al., 2026); our experiments concern comparisons of reasoning-data recipes rather than batch composition.

Composition, allocation and dataset balance. Mixtures of long and short reasoning, and of thinking and non-thinking targets, are established data-construction strategies (Yu et al., 2025a; Li et al., 2025a; Wang et al., 2025). Shortest-correct selection among sampled traces is also used for self-training towards concise reasoning (Munkhbat et al., 2025), and the difficulty of distilling long traces from strong reasoners into small students has been studied separately (Li et al., 2025c). Data Mixing Optimization compares equal-token and equal-instance allocations across domains, but changes the sampled dataset mixture rather than the coefficients on an identical set of rows (Li et al., 2025b). Our experiments address a narrower question: with the target strings of each recipe fixed, how do objective coefficients change the comparison between matched fixed-depth and mixed-depth recipes (which share questions but only part of their targets), does the mixed-corpus penalty remain when dataset shares are held fixed, and, with those shares fixed, does moving the T1 group to its example share cost more than equalizing the examples within the groups? The contribution is the controlled empirical comparison under one teacher pool.

B Broader recipe comparisons

This appendix reports the comparison of the supervision recipes themselves. It contains the recipe comparison registered before any student was trained, together with later additions listed in Section 3 (the OLMo runs of the ANSWER-ONLY, FIXED-T2 and GUIDED-MIX-T2SUB recipes, the depth ladder, and the descriptive comparison with long traces, which was added post hoc); the main text uses it as context, because its recipes differ in several respects at once, while the experiments of Sections 4 and 5 change one thing at a time.

The recipes. All recipes share the same student prompt, which contains the question and a final-answer format instruction but no depth instruction. Six recipes select from the five-depth pool of Section 3: LONG uses a random correct T5 trace; SHORTEST-CORRECT the shortest correct

sample at any depth; GUIDED-MIX the shallowest depth whose within-depth pass rate clears a per-format reliability bar ($8/8$ on yes/no, $\geq 6/8$ on ARC multiple-choice, $\geq 3/4$ on free-form), then the shortest correct sample at that depth; FIXED-T2, FIXED-T3 and FIXED-T4 the shortest correct sample at that depth. Two recipes do not use the pool: BON-SHORT takes the shortest correct of 15 separately generated, unconstrained long samples per question, and ANSWER-ONLY the gold answer in the final-answer format. The fixed-depth recipes and GUIDED-MIX-T2SUB are trained on the 11,853 questions with a correct T2 sample, the others on all 12,108; comparisons across the two populations are descriptive and labelled as such.

Table 3: Training questions (n), four-benchmark accuracy (Acc., sampled, mean of three seeds), supervised target tokens per epoch in millions (Tgt.), mean generated tokens over the four benchmarks (Out.), and mean hours of the training loop per run on one eight-GPU node (Hours; excludes generation, learning-rate selection, evaluation, model loading). Rows name the recipes in words: the reliability-guided mixture is GUIDED-MIX, its restriction to the T2-eligible questions GUIDED-MIX-T2SUB, shortest correct at any depth SHORTEST-CORRECT, shortest of 15 long samples BON-SHORT, random correct long trace LONG.

		Qwen3-4B-Base				OLMo-3-7B			
Supervision recipe	n	Tgt. (M)	Acc.	Out.	Hours	Tgt. (M)	Acc.	Out.	Hours
Answer only	12,108	0.09	52.3	7	0.26	0.08	45.3	6	0.24
Shortest correct (any depth)	12,108	0.55	67.4	73	0.27	0.49	61.2	108	0.28
Reliability-guided mixture	12,108	0.98	73.1	147	0.32	0.91	66.0	203	0.36
on the T2-eligible questions	11,853	0.80	72.7	125	0.28	0.74	65.8	174	0.28
Fixed depth T2 (brief)	11,853	0.86	74.2	103	0.26	0.79	68.9	107	0.24
Fixed depth T3	11,853	1.61	79.2	212	0.23	1.49	76.4	251	0.24
Fixed depth T4	11,853	2.57	82.5	413	0.24	2.43	78.6	432	0.26
Shortest of 15 long samples	12,108	14.95	80.4	2,000	0.86	14.15	78.6	2,172	1.12
Random correct long trace	12,108	24.17	79.5	2,677	1.25	23.05	77.9	2,777	1.58

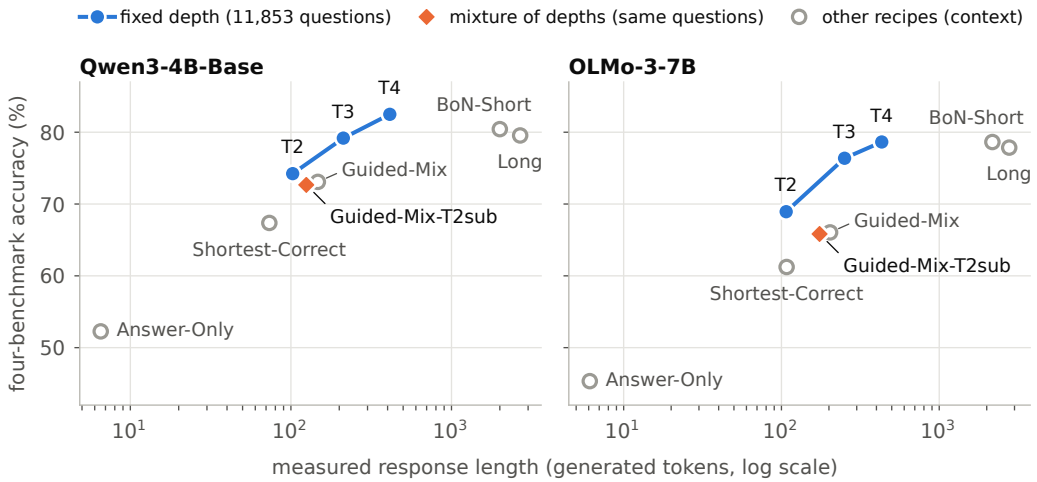

Figure 3: Observed accuracy–length trade-off among the tested recipes (sampled decoding, mean of three seeds). Blue: fixed depths T2, T3 and T4 trained on the same 11,853 questions; orange: the mixture of depths restricted to those questions; grey: recipes trained on all 12,108 questions, shown as context and not as matched endpoints. Response length is the mean number of generated tokens over the four benchmarks, with an 8,192-token limit.

Table 3 and Figure 3 give the measured trade-off. **Reasoning targets matter.** Supervision with bare answers is far below every recipe of Table 3 that contains reasoning: GUIDED-MIX exceeds ANSWER-ONLY by 20.82 points $[19.31, 22.31]$ on Qwen3 and 20.71 $[19.21, 22.20]$ on OLMo, and ANSWER-ONLY leaves mathematics far below the starting model on both students.

Accuracy rises with the depth of a fixed target. On the same 11,853 questions, moving from two or three sentences (T2) to concise steps (T3) raises four-benchmark accuracy by 4.95 points [3.63, 6.29] on Qwen3 and 7.44 [6.11, 8.77] on OLMo, and moving on to a short plan followed by its execution (T4) by a further 3.31 [2.17, 4.46] and 2.27 [1.00, 3.54]; all four pre-specified steps are positive on every seed and are rejected after Holm correction (adjusted $p \leq 8 \times 10^{-4}$). Students respond with about 103, 212 and 413 tokens on Qwen3 and 107, 251 and 432 on OLMo, while the hours of the training loop are almost flat across the three depths although the target tokens triple (Table 3); these costs describe this implementation and its fixed three-epoch recipe.

Against long traces, FIXED-T4 (about 420 generated tokens) has higher observed accuracy on Qwen3 and similar observed accuracy on OLMo, under our generation budget. The two long recipes were trained on all 12,108 questions, 2% more than the fixed depths, so the following comparison is descriptive and was added after the ladder results were seen. FIXED-T4 is +2.94 points [+1.63, +4.26] against LONG and +2.06 [+0.82, +3.34] against BON-SHORT on Qwen3 (the seed-level interval of the latter, [−1.03, +5.15], contains zero; Appendix I), and +0.78 [−0.48, +2.05] and +0.01 [−1.28, +1.28] on OLMo (the OLMo intervals include zero), with responses 4.8 to 6.5 times shorter and 16 to 28% of the training-loop hours. The composition of this result matters. The long recipes remain ahead on GSM8K and ARC, and LONG on the held-out BBH tasks, while FIXED-T4 leads on MATH-500 (per-dataset values in Appendix G), where students of the long recipes reach the 8,192-token limit on 31 to 43% of the questions and students of FIXED-T4 on 4 to 6%. Counting every capped response as incorrect lowers the long recipes by 0.4 to 1.9 points, because most capped responses already lack an answer, and leaves the fixed depths unchanged; under that rescoring LONG also falls below FIXED-T3 on both students (Appendix L). The statement therefore holds for an 8,192-token budget, and we make no claim about an optimal depth or about saturation.

Mixing depths. GUIDED-MIX exceeds SHORTEST-CORRECT by 5.72 points [4.48, 6.95] on Qwen3 and 4.82 [3.55, 6.05] on OLMo, positive on every seed; this compares two complete policies, and GUIDED-MIX's targets contain about 1.8 times as many tokens, so it does not isolate the reliability bar. Section 6 compares the placement rule with the matched-random control. On identical questions FIXED-T2 is more accurate than GUIDED-MIX-T2SUB on the four-benchmark mean: FIXED-T2 exceeds GUIDED-MIX-T2SUB by +1.56 [0.24, 2.86] on Qwen3 (Holm-adjusted p=0.04; seed-level interval [−0.67, +3.80]) and +3.10 [1.70, 4.52] on OLMo (adjusted $p \leq 8 \times 10^{-4}$), on every seed, while the mixture is ahead or level on the BBH tasks (Appendix G); neither recipe dominates every endpoint. The registered hypothesis that GUIDED-MIX would be non-inferior to BON-SHORT within one point is not supported on either student (−7.35 points [−8.78, −5.90] on Qwen3; −12.58 [−14.07, −11.07] on OLMo), and on OLMo GUIDED-MIX is close to the untrained starting model (Section 6), while SHORTEST-CORRECT is below it. Per-dataset results, including the starting models, are in Appendix G.

Measurement notes. A registered control that spends GUIDED-MIX's token budget on long traces (Qwen3; 1,085 examples, 51 optimizer steps) is −18.92 points [−20.64, −17.17] against GUIDED-MIX under sampling, where its students reach the token limit on 63 to 78% of the questions, and +1.60 [+0.06, +3.08] under greedy decoding (adjusted p=0.084, not rejected). Without quoted instructions (Appendix L) the values are −11.19 and +0.48 [−1.05, +1.97]; the contrast depends on how capped responses are read, so this control supports no accuracy conclusion.

C TEACHER PROMPTS AND TARGET CONSTRUCTION

The teacher receives the question, a dataset-specific answer-format hint and one of five depth instructions. T1: "Answer the question directly. Do not show any reasoning or explanation." (at most 64 generated tokens). T2: "Reason briefly in at most 2-3 sentences, then answer." (300). T3: "Reason step by step, keeping it concise (about 4-8 short steps), then answer." (900). T4: "First write a short plan (2-4 bullet points), then carry out the plan concisely, then answer." (1,400). T5: "Think carefully and thoroughly before answering." with the teacher's thinking mode enabled (8,192). Depths T1 to T4 are sampled at temperature 0.7 and T5 at 0.6 (top-p 0.95, top-k 20), four times per question on GSM8K and MATH and eight times on ARC-Challenge and StrategyQA. BON-SHORT's 15 samples per question use the teacher's thinking mode without a depth instruction, a 6,000-token cap and the T5 sampling settings. The student prompt is

identical for every recipe and contains no depth instruction. Supervised tokens are the tokens of the target under the student's tokenizer, including the end-of-turn token; prompt tokens carry no loss. **Target content.** Targets were kept verbatim. Of the 12,108 LONG targets, 5,217 (43.1%) contain the T5 instruction phrase "Think carefully and thoroughly before answering" (5,003 of them with its final period), and a broader detector flags 6,243 that contain any depth-instruction string. Outside the long-trace family the same detector flags at most 61 targets per training file: 49 in GUIDED-MIX, 39 in GUIDED-MIX-T2SUB, 60, 61 and 57 in the three random allocations, one each in SHORTEST-CORRECT and BON-SHORT, and none in ANSWER-ONLY or the three fixed depths; the token-matched long control draws its targets from LONG and carries 544 such targets among its 1,085 (Qwen3 draw; the OLMo draw, built but not trained, has 499 of 1,041). These are exact-string counts on every training file of both students, not a semantic annotation; bare answers without the "Final Answer:" line occur in every file that contains direct-answer targets: 1,284 SHORTEST-CORRECT, 792 GUIDED-MIX, 785 GUIDED-MIX-T2SUB and 697, 670 and 719 targets of the three random allocations; all targets of the other recipes carry the line (one FIXED-T4 target ends with a "Final Answer" heading and the answer on the next line instead). Three targets passed the correctness filter without containing a final answer: two LONG StrategyQA traces and one BON-SHORT GSM8K trace were cut off at the teacher's generation cap after quoting the format instruction of the prompt, and the grader read the quoted instruction as the answer (Appendix L); one of the two is also in the OLMo draw of the token-matched long control, which was built but not trained. The targets were kept as trained; they are the samples the automatic grader accepted, and the three listed above lack a final answer. Two teacher-side checks, the first similar to one in Maqsood et al. (2026), qualify the reliability rule of GUIDED-MIX: on fresh samples its chosen depth clears the bar again for 92.0% of the 12,708 questions, against 75.4% for SHORTEST-CORRECT; and regenerating T1 with a 256-token cap leaves 13,115 of 14,870 depth decisions unchanged; these were independent new draws with the other depths retained, so changed decisions combine sampling variation with the effect of the cap.

D LOSS OBJECTIVES

Let one accumulation window over all devices contain n examples with T supervised tokens, of which n_d examples and T_d tokens belong to dataset d; let t_i be the number of supervised tokens of example i and $\bar{\ell}_i$ its mean token cross-entropy. Every objective has the form $\sum_i w_i \bar{\ell}_i$. Token mean: $w_i = t_i/T$. Equal example: $w_i = 1/n$. The two diagnostic objectives cross the two ways in which these differ. Dataset token shares with equal examples within a dataset: $w_i = (T_d/T)(1/n_d)$. Dataset example shares with token weighting within a dataset: $w_i = (n_d/n)(t_i/T_d)$. We verified all four in a single-process CPU test that calls the trainer's real entry points on full and on partial windows, including a window from which one dataset is absent; in that single-process test the largest relative gradient error against the definition is 1.1×10^{-6}, and the token-mean and equal-example paths reproduce the implementations used in the earlier experiments exactly. The distributed runs use the same entry points. After the experiments we also checked the distributed path itself on eight GPUs under ZeRO stage 3: for a small, randomly initialized model of each student architecture and each of the four objectives, two steps of plain stochastic gradient descent through the training loop were compared with a single-process reference gradient of the objective as defined above, on a full window of 32 examples and on a partial window of 16 examples, 4 of them repeated by the data loader and one dataset absent. In 32-bit precision the applied gradient equals the reference to a relative error of at most 2.6×10^{-6} on every parameter tensor in all 8 runs. In bfloat16, the precision of the experiments, the relative error over all parameters was below 2.5%, with ratios of applied to reference gradient norm between 0.997 and 1.003; the reference is in 32-bit precision, so these differences are consistent with reduced-precision effects and show no gross scaling discrepancy. The check uses small random models of each student architecture and plain gradient descent. The micro-batch size is a memory setting that does not change any of the objectives; the value used is stored with every run (1 for every LONG run, for OLMo BON-SHORT and for the token-matched control; 2 for Qwen3 BON-SHORT and for two Qwen3 matched-random token-mean runs, seeds 0 and 1 (Appendix O, item xii); 4 for all other runs). The configured global batch is 64. In the final partial window of an epoch the data loader (Accelerate's `even_batches`) fills the last round with repeated examples, and how many depends on the micro-batch size (for 12,108 examples, 20 repeated rows at micro-batch 4 and 4 at micro-batch 2, in 3 of 570 steps); the loss is normalized by the examples and tokens actually processed. Every arm of the recipe-by-objective experiment, of the

normalization decomposition (the between- and within-dataset experiment of Section 5, test family N), of the learning-rate check and of the fixed-depth ladder uses micro-batch 4. Loss-coefficient shares quoted in the main text are estimated accumulated shares over 20 seeded random orders; they are neither corpus token shares nor gradient influence. Per depth, on the full GUIDED-MIX set (Qwen3), T1 to T5 are 41.3, 49.8, 6.8, 0.9 and 1.1% of the examples and carry 4.8, 52.7, 18.9, 3.4 and 20.2% of the token-mean coefficients (4.5, 52.2, 19.0, 3.5 and 20.8% on OLMo); on the 11,853-question GUIDED-MIX-T2SUB set they are 42.1, 50.9, 5.7, 0.7 and 0.7% of the examples and carry 5.6, 62.1, 17.1, 2.2 and 13.0% (5.3, 61.7, 17.1, 2.3 and 13.5% on OLMo). Under the dataset-share switch alone (dataset shares by examples, token weighting within them) T1 carries 10.4% on Qwen3 and 10.1% on OLMo. The token mean is defined per accumulation window, as in the gradient-accumulation loss corrections of October 2024 (Debut et al., 2024; Han & Han, 2024). As an exploratory summary of the normalization decomposition, the within-dataset switch alone reproduces 1.014 $[0.901, 1.138]$ of the full gap between the two objectives on Qwen3 and 1.187 $[0.999, 1.447]$ on OLMo.

E BETWEEN AND WITHIN THE DEPTH GROUPS

Design. This panel was specified after the within-dataset result of Section 5 had been seen, so it is post hoc; its objectives, arms, test family, prediction and claim rules were fixed in a protocol before its twelve new runs. Every training example carries its dataset d and a group g: T1, if the selected target was generated at depth T1, or T2–T5 otherwise. For an accumulation window, let n and T count its examples and supervised tokens, n_d and T_d those of dataset d, n_{dg} and T_{dg} those of the cell (d, g), and t_i the tokens of example i. The four arms weight the mean token loss of example i by

$$\text{TT: } \frac{t_i}{T}, \qquad \text{TE-mode: } \frac{T_{dg}/T}{n_{dg}}, \qquad \text{ET-mode: } \frac{T_d}{T}\,\frac{n_{dg}}{n_d}\,\frac{t_i}{T_{dg}}, \qquad \text{EE-mode: } \frac{T_d/T}{n_d}.$$

Every arm gives dataset d its token share T_d/T, and an absent cell contributes nothing. TT is the token mean L_{tok}. EE-mode equalizes the examples within each dataset, the diagnostic objective of Section 5; it is *not* the global equal-example objective L_{ex}, which also changes the dataset shares. TT and EE-mode are the existing runs of that section; TE-mode and ET-mode were trained for this panel with the same recipe, seeds and data, with the attention setting of Appendix O (item viii). Before their runs, a separate eight-GPU job repeated the gradient check of Appendix D for these two objectives, with the harness extended to them and to a partial window from which one (dataset, group) cell is absent: in its four 32-bit runs every parameter tensor agrees with the reference to a relative error of at most 2.6×10^{-6}, and in its four bfloat16 runs the relative error over all parameters is below 2.9%, with norm ratios from 0.999 to 1.003. Table 4 gives the arm means and the estimated accumulated loss-coefficient shares.

Tests and claims. The primary family holds P_{between} (TT minus ET-mode), P_{within} (TT minus TE-mode) and S (TE-mode minus ET-mode) for both students: sampled decoding, two-sided paired bootstrap over questions, Holm over the six tests; greedy decoding is a separately corrected sensitivity family. The protocol allowed the statement of Section 5 only if P_{between} and S were positive and their null hypotheses of no difference were rejected after Holm correction, for both students under both the original and the rescored grades; both conditions hold (Table 5 and the next paragraph). P_{within} is rejected on Qwen3 and not on OLMo; a significant effect on one student and a non-significant one on the other is not a difference between the student families. Their interaction is exploratory (Table 6). Intervals are unadjusted and conditional on the trained checkpoints.

Scoring sensitivity. The quotation-removing rescoring of Appendix L was applied to all four arms by re-grading all 250,944 of their responses (Table 7): 31 responses change their extracted answer or grade, and 16 of these flip correctness; every contrast moves by less than 0.09 points, and every Holm decision is unchanged in both families. The 125,472 responses of the two reused arms were already among the 1,206,728 responses of the quoted-instruction rescoring of Appendix L, so together with the 125,472 new responses the rescoring covers 1,332,200 distinct responses.

Table 4: Between or within the depth groups: four-benchmark accuracy of the four arms (three seeds), and the estimated accumulated loss-coefficient shares of T1 targets and of the longest tenth of targets under each objective (simulated orders with the trainer's accumulation windows).

Student	Quantity	TT	TE-mode	ET-mode	EE-mode
Qwen3	accuracy, sampled	73.08	71.28	65.34	63.80
Qwen3	accuracy, greedy	73.72	71.94	66.01	64.66
OLMo	accuracy, sampled	66.05	65.85	60.83	59.81
OLMo	accuracy, greedy	66.67	66.78	61.23	60.00
Qwen3	T1 loss share (%)	4.75	4.75	36.48	36.48
Qwen3	longest tenth, loss share (%)	45.73	22.89	28.84	13.80
OLMo	T1 loss share (%)	4.50	4.50	36.83	36.83
OLMo	longest tenth, loss share (%)	46.52	23.65	29.07	14.10

Table 5: Between or within the depth groups: the six tests of the panel's primary family (sampled decoding) and of its separately corrected greedy family, in points (* rejected after Holm correction within its family of six; Holm p adjusted within the family). Seed-wise: the contrast computed with seed k of both arms.

Student	Decoding	Test	Estimate	95% interval	Holm p	Seed-wise
Qwen3	sampled	P_{between}	7.74^*	$[6.47, 9.01]$	$\leq 1.2\times10^{-3}$	$+8.98, +7.44, +6.81$
Qwen3	sampled	P_{within}	1.81^*	$[0.74, 2.85]$	1.2×10^{-3}	$+2.16, +1.65, +1.62$
Qwen3	sampled	S	5.93^*	$[4.63, 7.27]$	$\leq 1.2\times10^{-3}$	$+6.82, +5.80, +5.19$
Qwen3	greedy	P_{between}	7.71^*	$[6.42, 9.01]$	$\leq 1.2\times10^{-3}$	$+8.50, +8.30, +6.33$
Qwen3	greedy	P_{within}	1.78^*	$[0.72, 2.82]$	3.2×10^{-3}	$+1.35, +2.76, +1.23$
Qwen3	greedy	S	5.93^*	$[4.65, 7.20]$	$\leq 1.2\times10^{-3}$	$+7.15, +5.54, +5.10$
OLMo	sampled	P_{between}	5.22^*	$[3.90, 6.53]$	$\leq 1.2\times10^{-3}$	$+3.93, +6.73, +5.00$
OLMo	sampled	P_{within}	0.20	$[-0.87, +1.27]$	0.71	$-0.42, +2.15, -1.14$
OLMo	sampled	S	5.02^*	$[3.69, 6.39]$	$\leq 1.2\times10^{-3}$	$+4.36, +4.57, +6.14$
OLMo	greedy	P_{between}	5.44^*	$[4.13, 6.75]$	$\leq 1.2\times10^{-3}$	$+4.21, +6.00, +6.10$
OLMo	greedy	P_{within}	-0.11	$[-1.21, +0.99]$	0.84	$-1.02, +0.15, +0.53$
OLMo	greedy	S	5.55^*	$[4.20, 6.89]$	$\leq 1.2\times10^{-3}$	$+5.23, +5.85, +5.57$

Table 6: Between or within the depth groups: exploratory contrasts under sampled decoding, in no hypothesis family, in points. The last one per student repeats the within-dataset penalty of Table 13 with the same estimate and an interval from this panel's own bootstrap stream.

Student	Contrast	Estimate	95% interval	Seed-wise
Qwen3	(TT − TE-mode) − (ET-mode − EE-mode)	0.27	$[-1.15, +1.70]$	$+0.70, -0.19, +0.30$
Qwen3	ET-mode − EE-mode	1.54	$[0.51, 2.55]$	$+1.45, +1.83, +1.32$
Qwen3	TE-mode − EE-mode	7.47	$[6.13, 8.81]$	$+8.27, +7.63, +6.52$
Qwen3	TT − EE-mode (within-dataset penalty)	9.28	$[7.94, 10.63]$	$+10.43, +9.28, +8.13$
OLMo	(TT − TE-mode) − (ET-mode − EE-mode)	-0.82	$[-2.28, +0.63]$	$-1.29, +1.20, -2.38$
OLMo	ET-mode − EE-mode	1.02	$[0.05, 2.00]$	$+0.87, +0.95, +1.24$
OLMo	TE-mode − EE-mode	6.04	$[4.78, 7.33]$	$+5.22, +5.52, +7.38$
OLMo	TT − EE-mode (within-dataset penalty)	6.24	$[4.99, 7.49]$	$+4.80, +7.67, +6.24$

Table 7: Rescoring of the four arms with the quotation-removing rule of Appendix L: responses whose extracted answer or grade changed, and those among them whose correctness flipped; responses of the reused arms are counted once.

Arms	Responses	Changed	Flipped
TT, EE-mode (reused; in the quoted-instruction rescoring)	125,472	23	14
TE-mode, ET-mode (new)	125,472	8	2
All four arms	250,944	31	16

F SHARED QUESTION WEIGHTS

Design. This control was designed after the recipe-by-objective result had been seen and discussed, so it is post hoc; its test, main estimate and reading rule were fixed in a protocol amendment before its runs, and the analysis code was frozen, with its hash recorded, before any of its results existed. It asks whether the two target sets of Section 4 respond differently to the same change of question weights. Every FIXED-T2 training row carries r_i, the number of supervised tokens that the GUIDED-MIX-T2SUB target of the same question has under the same tokenizer, and the loss of an accumulation window is $\sum_i w_i \bar{\ell}_i$ with $w_i = r_i / \sum_j r_j$ over the window's rows. The two training files hold the same questions in the same order with the same prompts, so with the same seed every window holds the same questions in both arms and the new arm F, w_G receives exactly the coefficients of the mixture's token-mean arm G, w_G. With r_i set to the example's own length the objective is the token mean, with $r_i = 1$ the equal-example objective (checked through the training code); an eight-GPU gradient check against a single-process reference gave a worst relative tensor error below 3×10^{-6} for both students. The comparison arms are the existing runs of Section 4 (G, u, F, u and, as context, F, w_F, the fixed-depth corpus under its own token mean), three seeds each.

What the mixture's weights do to the fixed-depth targets. Under w_G, the FIXED-T2 targets of the questions whose mixture target is a direct answer (42.07% of the questions) receive 5.63% of the coefficients instead of 35.20% under their own token mean on Qwen3 (5.34% instead of 35.99% on OLMo); those whose mixture target is T3–T5 (7.06%) receive 32.24% instead of 10.29% (32.92% and 10.27%); the effective number of examples falls from 10,356 to 2,917 (10,383 to 2,779), and the per-token multiplier relative to the fixed-depth token mean reaches 30.9 (30.4).

Attention setting of the Qwen3 mixture arm. The registered Qwen3 runs were trained with PyTorch's default attention settings, before cuDNN attention was disabled. A diagnostic on 40 real micro-batches per student shows that the default settings select cuDNN attention for both students, padded or not, and that disabling it changes the loss of a micro-batch by up to 1.3% and its gradient by up to 42% in relative norm on Qwen3 (0.4% and 52% on OLMo). The three Qwen3 G, w_G runs were therefore retrained with the current setting and are the primary arm; their accuracy is 71.94 (sampled) and 72.98 (greedy) against 72.67 and 72.49 for the registered runs, and with the registered runs J is 10.88 $[+9.16, +12.63]$.

Tests and reading. $D_G = A(G, w_G) - A(G, u)$ is the mixture's response and $K_1 = A(F, w_G) - A(F, u)$ the response of the fixed-depth targets to the same weight change, the main estimate fixed in the protocol; $K_2 = A(F, w_G) - A(F, w_F)$, against the fixed-depth corpus under its own token mean, is descriptive. It is read through $Q = K_1 - D_G/2$ on the same question resamples: if Q's 95% interval lies below zero, the question weights alone reproduce less than half of the mixture's response (a); above zero, at least half (b); otherwise the share is not determined (c). The statement of Section 4 required reading (a) for both students under sampled and greedy decoding and under the quoted-instruction rescoring; all eight readings are (a); the six of Table 8 and the two greedy readings under the rescoring, which are in the analysis file. The test fixed in the protocol, $J = D_G - K_1$ (Holm over the two students), is rejected on both students (10.15 and 5.89 points, Holm-adjusted $p \leq 4.0\times10^{-4}$); $J \leq 0$ would have required the fixed-depth corpus to reach 81.81 points on Qwen3 and 74.16 on OLMo under w_G, so Q is the informative summary. The quoted-instruction rescoring changes 10 responses of the new and retrained arms.

What the fixed-depth targets do under the mixture's weights. They do not gain: K_1 is negative for both students and both decodings, with a bootstrap interval below zero for Qwen3 under sampling, where all three seed-wise values are negative (analysis file), and for OLMo under greedy decoding (-1.51 $[-2.73, -0.28]$; point estimate in Table 8, interval in the analysis file); on Qwen3 the lost accuracy is concentrated in MATH-500 and StrategyQA (Table 10). The seed-level interval of K_2 on Qwen3 contains zero (Table 8). Under w_G the fixed-depth runs clip their gradient norm at 1.0 on 100% of the logged steps on Qwen3 and 78% on OLMo, against 93% and 41% under equal-example weighting (Table 9). This control does not separate the effect of concentrating the weights from that of the targets.

Table 8: Shared question weights: estimates in points with 95% paired-bootstrap intervals (sampled), seed-level t-intervals (sampled), greedy estimates and the quotation-removing rescoring of Appendix L (EAE, echo-aware extraction; sampled). * rejected after Holm correction over the two students. (a): the question weights alone reproduce less than half of the mixture's response. The Qwen3 mixture arm is the retrained one except in the last row; "–": not computed under the rescoring.

Quantity	Student	Sampled	Seed-level	Greedy	EAE, sampled
J (test fixed in the protocol)	Qwen3	$+10.15^{*}$ $[+8.38, +11.95]$	$[+7.86, +12.44]$	$+9.49^{*}$	$+10.04$
	OLMo	$+5.89^{*}$ $[+4.04, +7.68]$	$[+3.42, +8.37]$	$+7.27^{*}$	$+5.81$
D_G: mixture, $w_G - u$	Qwen3	$+7.91$ $[+6.54, +9.28]$	$[+5.45, +10.37]$	$+8.36$	$+7.80$
	OLMo	$+5.46$ $[+4.19, +6.73]$	$[+4.85, +6.07]$	$+5.76$	$+5.38$
K_1: FIXED-T2, $w_G - u$	Qwen3	-2.24 $[-3.44, -1.05]$	$[-2.99, -1.49]$	-1.14	-2.24
	OLMo	-0.43 $[-1.68, +0.83]$	$[-2.42, +1.55]$	-1.51	-0.43
$Q = K_1 - D_G/2$	Qwen3	-6.20 $[-7.54, -4.87]$ (a)	$[-7.38, -5.02]$	-5.32 (a)	-6.14 (a)
	OLMo	-3.16 $[-4.61, -1.78]$ (a)	$[-5.39, -0.94]$	-4.39 (a)	-3.12 (a)
K_2: FIXED-T2, $w_G - w_F$	Qwen3	-2.57 $[-3.73, -1.44]$	$[-5.86, +0.73]$	-1.63	–
	OLMo	-0.67 $[-1.85, +0.55]$	$[-1.93, +0.59]$	-2.17	–
J, registered token-mean arm	Qwen3	$+10.88$ $[+9.16, +12.63]$	$[+9.19, +12.57]$	$+9.00$	–

Table 9: Shared question weights: four-benchmark accuracy of each arm (mean of three seeds), responses of at most 16 tokens (mean over the four benchmarks, sampled), median logged gradient norm before clipping and the share of logged steps above the clipping threshold 1.0 (means over seeds). The Qwen3 G, w_G arm is retrained; the last row gives the registered runs.

Arm	Student	Sampled	Greedy	Direct (%)	Grad. norm	Clipped (%)
G, w_G: mixture, own token mean	Qwen3	71.94	72.98	48.2	1.76	97
G, w_G: mixture, own token mean	OLMo	65.83	66.82	48.3	1.01	51
G, u: mixture, equal example	Qwen3	64.03	64.63	67.3	3.28	100
G, u: mixture, equal example	OLMo	60.37	61.06	62.1	2.36	86
F, w_G: fixed depth, mixture's weights (new)	Qwen3	71.66	73.10	0.0	2.26	100
F, w_G: fixed depth, mixture's weights (new)	OLMo	68.27	67.49	0.0	1.34	78
F, u: fixed depth, equal example	Qwen3	73.91	74.24	0.0	1.70	93
F, u: fixed depth, equal example	OLMo	68.70	69.00	0.0	0.95	41
F, w_F: fixed depth, own token mean	Qwen3	74.23	74.73	0.0	1.58	89
F, w_F: fixed depth, own token mean	OLMo	68.93	69.66	0.0	0.91	37
G, w_G with the registered runs	Qwen3	72.67	72.49	48.6	1.75	97

Table 10: Shared question weights: K_1 and D_G per benchmark (sampled), in points with 95% paired-bootstrap intervals.

Quantity	Student	GSM8K	MATH-500	ARC	StrategyQA
K_1	Qwen3	-0.56 $[-1.84, +0.78]$	-4.00 $[-6.40, -1.60]$	$+0.26$ $[-0.88, +1.37]$	-4.67 $[-8.44, -0.89]$
	OLMo	-0.18 $[-1.47, +1.14]$	-0.53 $[-3.00, +2.07]$	-0.03 $[-1.45, +1.34]$	-1.00 $[-4.89, +3.00]$
D_G	Qwen3	$+10.84$ $[+9.05, +12.71]$	$+15.20$ $[+12.00, +18.40]$	$+1.71$ $[+0.65, +2.76]$	$+3.89$ $[+0.00, +7.89]$
	OLMo	$+7.43$ $[+5.59, +9.33]$	$+11.07$ $[+8.00, +14.13]$	$+1.68$ $[+0.37, +3.01]$	$+1.67$ $[-1.67, +5.00]$

G PER-DATASET RESULTS

Table 11 gives the accuracy of every recipe on each benchmark, and Table 12 the same for the loss-weighting experiments. "Cap %" is the share of MATH-500 responses that reach the 8,192-token limit. The starting models are evaluated with the same prompt; their responses are long and often reach the token limit (Table 11); under greedy decoding they score 36.25 (Qwen3) and 67.29 (OLMo) on the four-benchmark mean. Answer-only supervision raises multiple-choice and yes/no accuracy but lowers mathematics far below the starting model on both students (GSM8K 29.1 against 61.0 on Qwen3 and 15.8 against 72.5 on OLMo).

H ALL CONTRASTS UNDER BOTH DECODING SETTINGS

Figure 4 plots the allocation contrasts of Section 6 under both objectives. Table 13 lists the contrasts of the study before the panel of Appendix E under sampled decoding (primary) and greedy decoding (sensitivity), and Table 14 their three seed-wise differences; it omits two descriptive contrasts per

Table 11: Per-dataset accuracy (sampled decoding, mean of three seeds; starting models: one run), BBH task-macro accuracy, mean generated tokens over the four primary benchmarks, and share of MATH-500 responses at the token cap.

Qwen3-4B-Base Recipe	GSM8K	MATH-500	ARC	StrategyQA	Mean	BBH	Out. tok.	Cap %
Starting model	61.0	34.4	25.4	28.7	37.4	39.0	3,176	31.8
ANSWER-ONLY	29.1	20.6	87.8	71.6	52.3	58.9	7	0.0
SHORTEST-CORRECT	65.7	41.9	89.2	72.8	67.4	56.5	73	1.0
GUIDED-MIX	76.6	52.5	90.0	73.2	73.1	68.7	147	1.3
GUIDED-MIX-T2SUB	75.8	51.7	89.9	73.3	72.7	67.3	125	1.1
FIXED-T2	78.8	52.4	89.5	76.2	74.2	65.1	103	1.0
FIXED-T3	88.4	66.1	88.7	73.6	79.2	68.0	212	1.9
FIXED-T4	90.0	73.6	89.0	77.3	82.5	68.1	413	5.7
BON-SHORT	93.2	62.0	91.2	75.3	80.4	73.9	2,000	31.4
LONG	93.0	59.5	92.1	73.7	79.5	76.2	2,677	37.6
LONG, token-matched	58.3	43.5	60.8	54.1	54.2	51.1	6,240	78.4
OLMo-3-7B Recipe	GSM8K	MATH-500	ARC	StrategyQA	Mean	BBH	Out. tok.	Cap %
Starting model	72.5	58.2	67.9	58.3	64.2	36.5	954	12.8
ANSWER-ONLY	15.8	13.0	81.8	70.8	45.3	52.2	6	0.0
SHORTEST-CORRECT	58.4	35.7	83.8	67.0	61.2	52.3	108	2.5
GUIDED-MIX	67.3	41.9	84.3	70.7	66.0	64.3	203	3.9
GUIDED-MIX-T2SUB	67.5	41.2	83.6	71.1	65.8	62.8	174	2.7
FIXED-T2	74.1	45.5	83.3	72.9	68.9	62.4	107	1.3
FIXED-T3	85.6	61.1	83.7	75.1	76.4	63.3	251	3.1
FIXED-T4	86.7	70.4	85.1	72.3	78.6	66.7	432	3.7
BON-SHORT	91.4	58.9	87.3	77.0	78.6	62.5	2,172	41.1
LONG	92.7	59.0	87.2	72.6	77.9	71.0	2,777	43.3

Table 12: Loss-weighting experiments: per-dataset accuracy under each objective (sampled decoding, mean of three seeds). "Token shares" and "example shares" state how the loss is divided between datasets, "equal within" and "token within" how it is divided inside a dataset. Columns as in Table 11.

Qwen3-4B-Base Recipe	GSM8K	MATH-500	ARC	StrategyQA	Mean	BBH	Out. tok.	Cap %
GUIDED-MIX, token mean	76.6	52.5	90.0	73.2	73.1	68.7	147	1.3
GUIDED-MIX, equal example	63.2	36.9	88.0	67.7	63.9	56.5	55	0.7
GUIDED-MIX, token shares, equal within	62.7	35.3	88.1	69.2	63.8	55.1	51	0.5
GUIDED-MIX, example shares, token within	76.7	55.1	86.8	72.8	72.8	63.9	133	2.0
Matched-random, token mean	75.3	52.9	89.4	71.8	72.4	66.0	156	2.3
Matched-random, equal example	59.1	32.8	87.1	71.7	62.7	46.3	55	0.9
GUIDED-MIX-T2SUB, token mean	75.8	51.7	89.9	73.3	72.7	67.3	125	1.1
GUIDED-MIX-T2SUB, equal example	63.9	35.9	87.6	68.7	64.0	54.8	48	0.5
FIXED-T2, token mean	78.8	52.4	89.5	76.2	74.2	65.1	103	1.0
FIXED-T2, equal example	79.0	52.3	88.9	75.4	73.9	63.9	117	1.8
OLMo-3-7B Recipe	GSM8K	MATH-500	ARC	StrategyQA	Mean	BBH	Out. tok.	Cap %
GUIDED-MIX, token mean	67.3	41.9	84.3	70.7	66.0	64.3	203	3.9
GUIDED-MIX, equal example	60.5	31.9	82.8	68.0	60.8	47.2	92	2.1
GUIDED-MIX, token shares, equal within	60.4	29.2	82.6	67.1	59.8	49.0	84	1.8
GUIDED-MIX, example shares, token within	67.3	44.0	81.8	69.7	65.7	60.6	165	3.2
Matched-random, token mean	70.0	45.1	83.8	69.1	67.0	59.6	177	3.4
Matched-random, equal example	60.6	31.5	82.3	68.3	60.7	48.2	92	2.1
GUIDED-MIX-T2SUB, token mean	67.5	41.2	83.6	71.1	65.8	62.8	174	2.7
GUIDED-MIX-T2SUB, equal example	60.0	30.1	81.9	69.4	60.4	48.8	62	0.9
FIXED-T2, token mean	74.1	45.5	83.3	72.9	68.9	62.4	107	1.3
FIXED-T2, equal example	73.8	45.9	83.3	71.9	68.7	60.9	108	1.5

student of the normalization decomposition (each diagnostic objective minus equal example), which are in the released analysis files, and the panel's own contrasts are in Tables 5 and 6. The column "Test" identifies each test and "Holm p" gives its adjusted p-value for the sampled value ("n.r.": not rejected): C1 to C6 are the six registered contrasts of the recipe comparison on Qwen3 (on OLMo the three contrasts that its four trained recipes allow form their own family; Appendix O, item xi); E1 to E3 the allocation-by-objective tests, one family per student; B1 and B2 the two recipe-by-objective interactions; R1 to R4 the two OLMo controls and the two tuned objective gaps; N the six tests of the normalization decomposition (Section 5); F1 to F4 the four ladder steps. C2,

the registered non-inferiority test of GUIDED-MIX against BON-SHORT, is one-sided with a margin of one point: its p-value is for the null hypothesis that the difference is at most -1 point, so a large negative difference gives $p = 1$; the claim needed both the lower end of the 95% interval above -1 and rejection after Holm correction, and neither holds. Rows marked "descr." are descriptive and belong to no hypothesis family, which includes the four rows "Objective effect" (token mean minus equal example on the named targets); those marked post hoc († in Table 13) were added after the corresponding means had been seen. The tests of the paper are those on the sampled values. For the registered comparison and the allocation-by-objective experiment the same Holm procedure was also run on the greedy values, as a separately corrected sensitivity. Before the panel there are 46 Holm decisions: 15 sampled and 15 greedy for these two analyses, and 16 sampled for the extension families (B, R, N and F), which were tested on sampled values only; the panel adds six sampled and six greedy decisions. The follow-up is the allocation-by-objective experiment, and the extension comprises the experiments specified after it (Appendix O). For family N the same Holm procedure, applied afterwards to the greedy values, gives the same six decisions; this descriptive check is in the released analysis files and is not counted among the 46 decisions. A penalty is token mean minus the switched objective; the within-dataset switch keeps the dataset shares fixed and the dataset-share switch keeps the within-dataset weighting fixed. Intervals are 95% paired-bootstrap intervals over questions, unadjusted; Holm corrections are applied within the families stated in the text, on the sampled values.

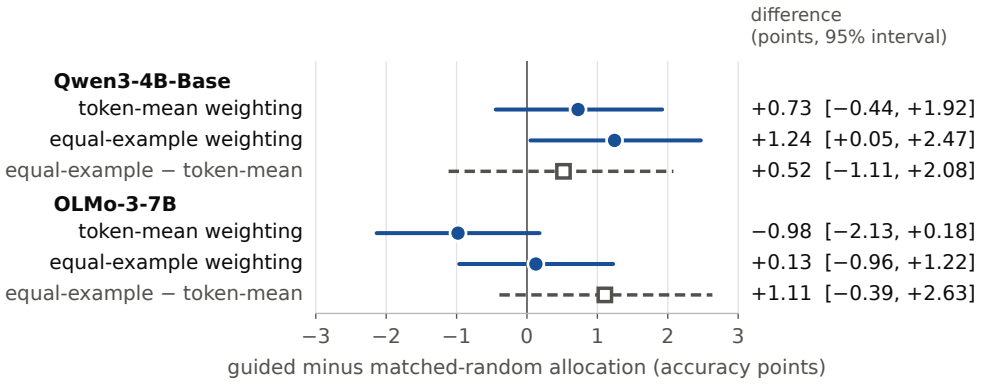

Figure 4: Does it matter which questions receive explanations? Four-benchmark accuracy of GUIDED-MIX minus the matched-random control (same questions, same histogram of depths per dataset, supervised tokens within 0.5%), under each loss weighting, with 95% paired-bootstrap intervals (unadjusted; sampled decoding; three seeds per arm, each control seed with its own random allocation). Grey rows: the equal-example row minus the token-mean row (E3 = E2 − E1). No contrast is rejected after Holm correction (Section 6).

I SEED-LEVEL INTERVALS

The bootstrap intervals of the paper resample questions and hold the trained checkpoints fixed, so they do not include seed-to-seed variation. Table 15 adds, for every contrast of Tables 13 and 5 under sampled decoding, a 95% t-interval over its three seed-wise differences (mean $\pm$ 4.30 standard errors, two degrees of freedom). With three seeds these intervals are wide, and they treat the seeds as the only source of variation. They exclude zero wherever the bootstrap interval does, except for the four contrasts marked ‡; no test decision of the paper is based on them.

J STUDENT BEHAVIOUR UNDER THE OBJECTIVES AND THE TWO ALLOCATIONS

Where the equal-example penalty falls. On Qwen3 the equal-example objective lowers GSM8K accuracy from 76.6 to 63.2, MATH-500 from 52.5 to 36.9 and the BBH subset from 68.7 to 56.5

Table 13: Contrasts in four-benchmark accuracy points with 95% intervals. "–": not run (the token-matched long control was trained on Qwen3 only). Test: test identifier (families C, E, B, R, N and F, see text); Holm p: adjusted p-value of the sampled value, a bound when marked $\leq$ (no resample beyond zero), "n.r." not rejected; †: added after the corresponding means had been seen; "Guided": GUIDED-MIX.

Contrast	Student	Sampled	Greedy	Test	Holm p
GUIDED-MIX − SHORTEST-CORRECT	Qwen3	$+5.72$ [$+4.48$, $+6.95$]	$+5.72$ [$+4.55$, $+6.89$]	C1	$\leq$0.0012
	OLMo	$+4.82$ [$+3.55$, $+6.05$]	$+5.06$ [$+3.87$, $+6.26$]	C1	$\leq$0.0006
GUIDED-MIX − BON-SHORT	Qwen3	-7.35 [-8.78, -5.90]	-6.31 [-7.85, -4.78]	C2	1.000 n.r.
	OLMo	-12.58 [-14.07, -11.07]	-11.36 [-12.90, -9.85]	C2	1.000 n.r.
GUIDED-MIX − LONG	Qwen3	-6.46 [-7.91, -4.98]	-6.27 [-7.80, -4.75]	C5	$\leq$0.0012
	OLMo	-11.81 [-13.28, -10.31]	-11.69 [-13.18, -10.17]	C5	$\leq$0.0006
LONG, token-matched − GUIDED-MIX	Qwen3	-18.92 [-20.64, -17.17]	$+1.60$ [$+0.06$, $+3.08$]	C6	$\leq$0.0012
	OLMo	–	–	–	–
GUIDED-MIX − ANSWER-ONLY	Qwen3	$+20.82$ [$+19.31$, $+22.31$]	$+21.32$ [$+19.90$, $+22.76$]	C3	$\leq$0.0012
	OLMo	$+20.71$ [$+19.21$, $+22.20$]	$+20.92$ [$+19.46$, $+22.38$]	R1	$\leq$0.0008
FIXED-T2 − GUIDED-MIX-T2SUB	Qwen3	$+1.56$ [$+0.24$, $+2.86$]	$+2.25$ [$+0.88$, $+3.66$]	C4	0.040
	OLMo	$+3.10$ [$+1.70$, $+4.52$]	$+2.84$ [$+1.46$, $+4.23$]	R2	$\leq$0.0008
Guided − matched-random, token mean	Qwen3	$+0.73$ [-0.44, $+1.92$]	$+0.22$ [-0.88, $+1.34$]	E1	0.455 n.r.
	OLMo	-0.98 [-2.13, $+0.18$]	-1.06 [-2.17, $+0.02$]	E1	0.274 n.r.
Guided − matched-random, equal example	Qwen3	$+1.24$ [$+0.05$, $+2.47$]	$+1.29$ [$+0.10$, $+2.48$]	E2	0.116 n.r.
	OLMo	$+0.13$ [-0.96, $+1.22$]	-0.15 [-1.21, $+0.89$]	E2	0.816 n.r.
Allocation × objective interaction	Qwen3	$+0.52$ [-1.11, $+2.08$]	$+1.07$ [-0.45, $+2.62$]	E3	0.517 n.r.
	OLMo	$+1.11$ [-0.39, $+2.63$]	$+0.91$ [-0.51, $+2.33$]	E3	0.292 n.r.
Objective effect, GUIDED-MIX	Qwen3	$+9.15$ [$+7.75$, $+10.54$]	$+9.22$ [$+7.90$, $+10.55$]	descr.	–
	OLMo	$+5.25$ [$+3.93$, $+6.60$]	$+5.88$ [$+4.63$, $+7.20$]	descr.	–
Objective effect, matched-random	Qwen3	$+9.67$ [$+8.37$, $+10.97$]	$+10.29$ [$+9.03$, $+11.56$]	descr.	–
	OLMo	$+6.36$ [$+5.13$, $+7.61$]	$+6.79$ [$+5.56$, $+8.02$]	descr.	–
Objective effect, GUIDED-MIX-T2SUB	Qwen3	$+8.64$ [$+7.31$, $+10.01$]	$+7.86$ [$+6.58$, $+9.18$]	descr.	–
	OLMo	$+5.46$ [$+4.20$, $+6.74$]	$+5.76$ [$+4.48$, $+7.04$]	descr.	–
Objective effect, FIXED-T2	Qwen3	$+0.33$ [-0.60, $+1.26$]	$+0.49$ [-0.43, $+1.41$]	descr.	–
	OLMo	$+0.23$ [-0.63, $+1.10$]	$+0.67$ [-0.28, $+1.61$]	descr.	–
Recipe × objective interaction	Qwen3	$+8.31$ [$+6.77$, $+9.87$]	$+7.36$ [$+5.81$, $+8.96$]	B1	$\leq$0.0004
	OLMo	$+5.23$ [$+3.70$, $+6.74$]	$+5.10$ [$+3.56$, $+6.62$]	B2	$\leq$0.0004
Within-dataset switch, penalty	Qwen3	$+9.28$ [$+7.95$, $+10.62$]	$+9.06$ [$+7.73$, $+10.37$]	N	$\leq$0.0012
	OLMo	$+6.24$ [$+4.98$, $+7.46$]	$+6.67$ [$+5.44$, $+7.94$]	N	$\leq$0.0012
Dataset-share switch, penalty	Qwen3	$+0.25$ [-0.82, $+1.26$]	$+0.76$ [-0.29, $+1.81$]	N	1.000 n.r.
	OLMo	$+0.35$ [-0.78, $+1.49$]	-0.25 [-1.35, $+0.83$]	N	1.000 n.r.
Interaction of the two switches	Qwen3	$+0.37$ [-1.09, $+1.84$]	$+0.60$ [-0.87, $+2.05$]	N	1.000 n.r.
	OLMo	$+1.34$ [-0.09, $+2.73$]	$+0.54$ [-0.88, $+1.95$]	N	0.261 n.r.
Tuned token mean − tuned equal example	Qwen3	$+8.50$ [$+7.18$, $+9.89$]	$+7.93$ [$+6.67$, $+9.20$]	R3	$\leq$0.0008
	OLMo	$+6.35$ [$+5.01$, $+7.72$]	$+6.03$ [$+4.72$, $+7.36$]	R4	$\leq$0.0008
FIXED-T3 − FIXED-T2	Qwen3	$+4.95$ [$+3.63$, $+6.29$]	$+4.89$ [$+3.61$, $+6.18$]	F1	$\leq$0.0008
	OLMo	$+7.44$ [$+6.11$, $+8.77$]	$+6.59$ [$+5.22$, $+7.94$]	F3	$\leq$0.0008
FIXED-T4 − FIXED-T3	Qwen3	$+3.31$ [$+2.17$, $+4.46$]	$+2.39$ [$+1.14$, $+3.64$]	F2	$\leq$0.0008
	OLMo	$+2.27$ [$+1.00$, $+3.54$]	$+2.26$ [$+0.93$, $+3.57$]	F4	0.0008
FIXED-T4 − FIXED-T2	Qwen3	$+8.26$ [$+7.00$, $+9.51$]	$+7.28$ [$+5.99$, $+8.60$]	descr.	–
	OLMo	$+9.71$ [$+8.37$, $+11.08$]	$+8.85$ [$+7.43$, $+10.29$]	descr.	–
FIXED-T3 − LONG	Qwen3	-0.37 [-1.67, $+0.91$]	-0.37 [-1.71, $+0.96$]	descr.†	–
	OLMo	-1.48 [-2.77, -0.16]	-2.10 [-3.41, -0.73]	descr.†	–
FIXED-T3 − BON-SHORT	Qwen3	-1.25 [-2.53, $+0.02$]	-0.41 [-1.70, $+0.89$]	descr.†	–
	OLMo	-2.26 [-3.60, -0.92]	-1.78 [-3.10, -0.44]	descr.†	–
FIXED-T4 − LONG	Qwen3	$+2.94$ [$+1.63$, $+4.26$]	$+2.03$ [$+0.75$, $+3.29$]	descr.†	–
	OLMo	$+0.78$ [-0.48, $+2.05$]	$+0.15$ [-1.16, $+1.46$]	descr.†	–
FIXED-T4 − BON-SHORT	Qwen3	$+2.06$ [$+0.82$, $+3.34$]	$+1.98$ [$+0.73$, $+3.24$]	descr.†	–
	OLMo	$+0.01$ [-1.28, $+1.28$]	$+0.48$ [-0.84, $+1.76$]	descr.†	–

(GUIDED-MIX; 68.9 to 61.0 under the format-tolerant rescoring of Appendix L). Qwen3 students answer directly on 35% instead of 23% of GSM8K questions and on 47% instead of 21% of MATH-500 questions, and the accuracy of those direct answers falls from 73% to 49% and from 55% to 31%. The same pattern holds for the matched-random control and for OLMo. In the recipe-by-objective comparison (Section 4; trained on 11,853 questions), students of the fixed-depth corpus never answer directly on the four benchmarks under either objective, whereas students of the mixture answer directly more often under equal-example weighting (Qwen3: 34% instead of 23% of GSM8K questions, 46% instead of 22% of MATH-500 questions) and their mean response length falls from 125 to 48 tokens on Qwen3 (OLMo: 174 to 62). In the normalization decomposition, students trained with equal examples within datasets answer directly far more often on mathematics (Qwen3, MATH-500: 47% of questions instead of 21 to 22%, with an accuracy of those direct answers of 26 to 31% instead of 55 to 56%) and respond with 51 to 55 tokens instead of 133 to 147

Table 14: Seed-wise differences (sampled decoding) of the contrasts of Table 13: the contrast computed with seed k of both arms. "–": not run.

Contrast	Student	Seed 0	Seed 1	Seed 2
GUIDED-MIX − SHORTEST-CORRECT	Qwen3	+6.05	+7.18	+3.93
	OLMo	+3.83	+5.69	+4.94
GUIDED-MIX − BON-SHORT	Qwen3	−6.41	−7.61	−8.01
	OLMo	−13.54	−12.37	−11.84
GUIDED-MIX − LONG	Qwen3	−5.80	−6.41	−7.19
	OLMo	−13.40	−10.88	−11.15
LONG, token-matched − GUIDED-MIX	Qwen3	−17.83	−19.47	−19.46
	OLMo	–	–	–
GUIDED-MIX − ANSWER-ONLY	Qwen3	+21.44	+20.84	+20.17
	OLMo	+19.13	+21.63	+21.39
FIXED-T2 − GUIDED-MIX-T2SUB	Qwen3	+1.74	+0.59	+2.36
	OLMo	+3.01	+3.74	+2.56
Guided − matched-random, token mean	Qwen3	+1.04	+1.37	−0.23
	OLMo	−1.61	−0.21	−1.12
Guided − matched-random, equal example	Qwen3	+0.50	+1.43	+1.80
	OLMo	+0.19	−0.21	+0.40
Allocation × objective interaction	Qwen3	−0.54	+0.06	+2.04
	OLMo	+1.80	0.00	+1.52
Objective effect, GUIDED-MIX	Qwen3	+9.86	+9.02	+8.58
	OLMo	+4.54	+6.61	+4.61
Objective effect, matched-random	Qwen3	+9.32	+9.08	+10.62
	OLMo	+6.34	+6.61	+6.13
Objective effect, GUIDED-MIX-T2SUB	Qwen3	+9.29	+8.61	+8.01
	OLMo	+5.74	+5.35	+5.29
Objective effect, FIXED-T2	Qwen3	−0.19	−0.42	+1.59
	OLMo	+0.05	+0.73	−0.08
Recipe × objective interaction	Qwen3	+9.48	+9.04	+6.42
	OLMo	+5.69	+4.62	+5.38
Within-dataset switch, penalty	Qwen3	+10.43	+9.28	+8.13
	OLMo	+4.80	+7.67	+6.24
Dataset-share switch, penalty	Qwen3	+1.76	+0.17	−1.19
	OLMo	−0.53	+1.23	+0.36
Interaction of the two switches	Qwen3	+2.33	+0.43	−1.64
	OLMo	−0.27	+2.30	+1.99
Tuned token mean − tuned equal example	Qwen3	+7.82	+8.33	+9.36
	OLMo	+6.35	+6.53	+6.17
FIXED-T3 − FIXED-T2	Qwen3	+5.53	+5.24	+4.07
	OLMo	+6.91	+7.58	+7.84
FIXED-T4 − FIXED-T3	Qwen3	+3.16	+3.45	+3.33
	OLMo	+2.15	+2.39	+2.27
FIXED-T4 − FIXED-T2 (descr.)	Qwen3	+8.70	+8.69	+7.39
	OLMo	+9.05	+9.97	+10.10
FIXED-T3 − LONG (descr., post hoc)	Qwen3	+0.71	−1.46	−0.36
	OLMo	−2.23	−1.09	−1.13
FIXED-T3 − BON-SHORT (descr., post hoc)	Qwen3	+0.10	−2.67	−1.18
	OLMo	−2.37	−2.59	−1.82
FIXED-T4 − LONG (descr., post hoc)	Qwen3	+3.87	+1.98	+2.97
	OLMo	−0.08	+1.29	+1.14
FIXED-T4 − BON-SHORT (descr., post hoc)	Qwen3	+3.26	+0.78	+2.15
	OLMo	−0.23	−0.20	+0.45

on Qwen3 (OLMo: 84 to 92 instead of 165 to 203); the dataset-share switch alone changes direct answering on mathematics by at most 1.1 percentage points but raises it where direct answers are usually correct (StrategyQA 65% to 78% and ARC 85% to 88% on Qwen3; 69% to 79% and 84% to 89% on OLMo).

Allocation by objective. The difference between the two allocation effects of Section 6 and Figure 4 (the interaction with the objective) has the same sign in all four analyses ($+0.52$ to $+1.11$ points) with every interval containing zero. The protocol of this experiment fixed four predictions before its runs: a small positive allocation effect under the token mean (P1), a larger one under equal-example weighting (P2), more direct answers under equal-example weighting for both allocations with more accurate direct answers for GUIDED-MIX (P3), and lower mathematics accuracy under equal-example weighting for both allocations (P4). P1 is not supported, P2 holds in direction only, P3 holds for the rates of direct answers only, and P4 holds on both students.

Direct answering under the two allocations. Students trained on GUIDED-MIX and on the matched-random control answer some questions directly and reason on others without being told

Table 15: Seed-level 95% intervals (t, two degrees of freedom) next to the paired question-bootstrap intervals, sampled decoding, in points. Seed-wise: the contrast computed with seed k of both arms. ‡ the seed-level interval contains zero while the bootstrap interval excludes it.

Contrast	Student	Seed-wise	Mean	Seed-level 95%	Bootstrap 95%
GUIDED-MIX − SHORTEST-CORRECT	Qwen3	+6.05, +7.18, +3.93	+5.72	[+1.61, +9.83]	[+4.48, +6.95]
	OLMo	+3.83, +5.69, +4.94	+4.82	[+2.49, +7.14]	[+3.55, +6.05]
GUIDED-MIX − BON-SHORT	Qwen3	−6.41, −7.61, −8.01	−7.35	[−9.42, −5.28]	[−8.78, −5.90]
	OLMo	−13.54, −12.37, −11.84	−12.58	[−14.75, −10.42]	[−14.07, −11.07]
GUIDED-MIX − LONG	Qwen3	−5.80, −6.41, −7.19	−6.46	[−8.20, −4.73]	[−7.91, −4.98]
	OLMo	−13.40, −10.88, −11.15	−11.81	[−15.24, −8.37]	[−13.28, −10.31]
LONG, token-matched − GUIDED-MIX	Qwen3	−17.83, −19.47, −19.46	−18.92	[−21.27, −16.57]	[−20.64, −17.17]
GUIDED-MIX − ANSWER-ONLY	Qwen3	+21.44, +20.84, +20.17	+20.82	[+19.24, +22.40]	[+19.31, +22.31]
	OLMo	+19.13, +21.63, +21.39	+20.71	[+17.29, +24.14]	[+19.21, +22.20]
FIXED-T2 − GUIDED-MIX-T2SUB	Qwen3	+1.74, +0.59, +2.36	+1.56	[−0.67, +3.80]‡	[+0.24, +2.86]
	OLMo	+3.01, +3.74, +2.56	+3.10	[+1.61, +4.59]	[+1.70, +4.52]
Guided − matched-random, token mean	Qwen3	+1.04, +1.37, −0.23	+0.73	[−1.38, +2.83]	[−0.44, +1.92]
	OLMo	−1.61, −0.21, −1.12	−0.98	[−2.75, +0.79]	[−2.13, +0.18]
Guided − matched-random, equal example	Qwen3	+0.50, +1.43, +1.80	+1.24	[−0.42, +2.91]‡	[+0.05, +2.47]
	OLMo	+0.19, −0.21, +0.40	+0.13	[−0.64, +0.89]	[−0.96, +1.22]
Allocation × objective interaction	Qwen3	−0.54, +0.06, +2.04	+0.52	[−2.83, +3.87]	[−1.11, +2.08]
	OLMo	+1.80, +0.00, +1.52	+1.11	[−1.30, +3.52]	[−0.39, +2.63]
Objective effect, GUIDED-MIX	Qwen3	+9.86, +9.02, +8.58	+9.15	[+7.55, +10.76]	[+7.75, +10.54]
	OLMo	+4.54, +6.61, +4.61	+5.25	[+2.33, +8.18]	[+3.93, +6.60]
Objective effect, matched-random	Qwen3	+9.32, +9.08, +10.62	+9.67	[+7.61, +11.73]	[+8.37, +10.97]
	OLMo	+6.34, +6.61, +6.13	+6.36	[+5.76, +6.96]	[+5.13, +7.61]
Objective effect, GUIDED-MIX-T2SUB	Qwen3	+9.29, +8.61, +8.01	+8.64	[+7.05, +10.23]	[+7.31, +10.01]
	OLMo	+5.74, +5.35, +5.29	+5.46	[+4.85, +6.07]	[+4.20, +6.74]
Objective effect, FIXED-T2	Qwen3	−0.19, −0.42, +1.59	+0.33	[−2.41, +3.06]	[−0.60, +1.26]
	OLMo	+0.05, +0.73, −0.08	+0.23	[−0.85, +1.31]	[−0.63, +1.10]
Recipe × objective interaction	Qwen3	+9.48, +9.04, +6.42	+8.31	[+4.21, +12.42]	[+6.77, +9.87]
	OLMo	+5.69, +4.62, +5.38	+5.23	[+3.87, +6.59]	[+3.70, +6.74]
Within-dataset switch, penalty	Qwen3	+10.43, +9.28, +8.13	+9.28	[+6.43, +12.13]	[+7.95, +10.62]
	OLMo	+4.80, +7.67, +6.24	+6.24	[+2.66, +9.81]	[+4.98, +7.46]
Dataset-share switch, penalty	Qwen3	+1.76, +0.17, −1.19	+0.25	[−3.42, +3.91]	[−0.82, +1.26]
	OLMo	−0.53, +1.23, +0.36	+0.35	[−1.83, +2.54]	[−0.78, +1.49]
Interaction of the two switches	Qwen3	+2.33, +0.43, −1.64	+0.37	[−4.56, +5.31]	[−1.09, +1.84]
	OLMo	−0.27, +2.30, +1.99	+1.34	[−2.15, +4.82]	[−0.09, +2.73]
Tuned token mean − tuned equal example	Qwen3	+7.82, +8.33, +9.36	+8.50	[+6.56, +10.45]	[+7.18, +9.89]
	OLMo	+6.35, +6.53, +6.17	+6.35	[+5.91, +6.79]	[+5.01, +7.72]
FIXED-T3 − FIXED-T2	Qwen3	+5.53, +5.24, +4.07	+4.95	[+3.02, +6.88]	[+3.63, +6.29]
	OLMo	+6.91, +7.58, +7.84	+7.44	[+6.25, +8.64]	[+6.11, +8.77]
FIXED-T4 − FIXED-T3	Qwen3	+3.16, +3.45, +3.33	+3.31	[+2.95, +3.67]	[+2.17, +4.46]
	OLMo	+2.15, +2.39, +2.27	+2.27	[+1.97, +2.57]	[+1.00, +3.54]
FIXED-T4 − FIXED-T2 (descr.)	Qwen3	+8.70, +8.69, +7.39	+8.26	[+6.40, +10.12]	[+7.00, +9.51]
	OLMo	+9.05, +9.97, +10.10	+9.71	[+8.29, +11.13]	[+8.37, +11.08]
FIXED-T3 − LONG (descr., post hoc)	Qwen3	+0.71, −1.46, −0.36	−0.37	[−3.08, +2.33]	[−1.67, +0.91]
	OLMo	−2.23, −1.09, −1.13	−1.48	[−3.09, +0.12]‡	[−2.77, −0.16]
FIXED-T3 − BON-SHORT (descr., post hoc)	Qwen3	+0.10, −2.67, −1.18	−1.25	[−4.70, +2.20]	[−2.53, +0.02]
	OLMo	−2.37, −2.59, −1.82	−2.26	[−3.25, −1.27]	[−3.60, −0.92]
FIXED-T4 − LONG (descr., post hoc)	Qwen3	+3.87, +1.98, +2.97	+2.94	[+0.59, +5.29]	[+1.63, +4.26]
	OLMo	−0.08, +1.29, +1.14	+0.78	[−1.09, +2.65]	[−0.48, +2.05]
FIXED-T4 − BON-SHORT (descr., post hoc)	Qwen3	+3.26, +0.78, +2.15	+2.06	[−1.03, +5.15]‡	[+0.82, +3.34]
	OLMo	−0.23, −0.20, +0.45	+0.01	[−0.95, +0.96]	[−1.28, +1.28]
Between groups: example shares (Appendix E)	Qwen3	+8.98, +7.44, +6.81	+7.74	[+4.98, +10.51]	[+6.47, +9.01]
	OLMo	+3.93, +6.73, +5.00	+5.22	[+1.72, +8.72]	[+3.90, +6.53]
Within groups: equal examples (Appendix E)	Qwen3	+2.16, +1.65, +1.62	+1.81	[+1.05, +2.56]	[+0.74, +2.85]
	OLMo	−0.42, +2.15, −1.14	+0.20	[−4.11, +4.50]	[−0.87, +1.27]
Difference of the two (Appendix E)	Qwen3	+6.82, +5.80, +5.19	+5.93	[+3.89, +7.98]	[+4.63, +7.27]
	OLMo	+4.36, +4.57, +6.14	+5.02	[+2.60, +7.44]	[+3.69, +6.39]

to. Under token-mean training and sampled decoding, the students of the matched-random control show a similar aggregate difficulty profile to those of GUIDED-MIX: on MATH-500 the share of direct answers falls from 45% (GUIDED-MIX) and 44% (matched-random) at level 1 to 5% at level 5 on Qwen3, and from 36% to about 10% for both on OLMo. At the intermediate levels the two profiles differ by less than 6.4 percentage points. Under equal-example training the profiles differ more, and the matched-random control's is flatter (OLMo: 47% at level 1 and 38% at level 5, not monotone in between): at level 5 the matched-random control's students answer directly on 48% (Qwen3) and 38% (OLMo) of the questions, against 33% and 29%. Difficulty-graded short answering thus appears under both allocations; question-by-question decisions were not compared.

K LEARNING-RATE CHECK

For the check of Section 6, for each student and objective we trained the GUIDED-MIX corpus with seed 0 at three learning rates (half, equal to and double the shared rate) and selected one on the 600-question development set only (highest sampled accuracy; ties by greedy accuracy, then the smaller rate). Table 16 shows all twelve candidates. Every selected rate lies at the lower edge of the grid. Among the newly trained candidate rates, only the selected configurations received additional seeds and a test evaluation; the existing three-seed results at the shared rate were reused, and no test result entered the selection. On the test sets, the students trained at the selected rate reach 75.1 (Qwen3) and 68.1 (OLMo) under the token-mean objective and 66.6 and 61.7 under the equal-example objective, with three seeds per arm.

Table 16: Development accuracy (sampled / greedy) of the learning-rate candidates; used for selection only.

Student	Objective	10^{-5}	2×10^{-5}	4×10^{-5}	Selected
Qwen3	token mean	83.5 / 84.3	80.7 / 80.5	74.0 / 75.3	1×10^{-5}
Qwen3	equal example	76.0 / 75.7	72.7 / 71.7	66.5 / 66.8	1×10^{-5}
OLMo	token mean	76.8 / 76.0	76.5 / 74.0	67.2 / 69.2	1×10^{-5}
OLMo	equal example	71.3 / 71.7	69.5 / 67.2	62.2 / 64.0	1×10^{-5}

L SCORING SENSITIVITY

A response is graded from its last "Final Answer:" line (matched case-insensitively on the four primary benchmarks and case-sensitively on BBH). Without that line, a response that reached the token limit fails extraction even if it contains a boxed answer, while a complete mathematics response falls back to its last boxed expression or last line; mathematical equivalence is checked with Math-Verify 0.9.0 (latex2sympy2_extended 1.11.0, SymPy 1.14.0). In this environment every one of the 2,119 mathematics gold answers of the evaluation and development sets verifies against itself. Reaching the token cap is reported but not by itself counted as an error in the reported results. Answers that cannot be extracted are rare in the loss-weighting experiments: the largest share in any of their arms is 0.20% of the responses on GSM8K, 3.47% on MATH-500, 0.03% on ARC-Challenge and 0.22% on StrategyQA (sampled decoding, mean over seeds), against 58.62% for a starting model on ARC-Challenge and 35.2% on MATH-500 for a long-trace recipe. Table 18 counts every capped response as incorrect. Recipes with short targets change by at most 0.1 points and the three fixed depths not at all, because their capped responses were already incorrect. BON-SHORT falls by 0.4 and 0.6 points and LONG by 1.2 and 1.9, which moves LONG below FIXED-T3 on both students; the order of the other recipes of Table 3 is unchanged. The token-matched long control falls from 54.2 to 28.0, below ANSWER-ONLY. The BBH grader requires the final-answer format of the training targets; Table 19 rescores BBH with a rule that also accepts a response consisting of exactly one valid label. The rule matters for SHORTEST-CORRECT and for the mixture arms trained with equal examples inside the datasets, whose students often reply with a bare label (FIXED-T2 under equal examples is unaffected): for GUIDED-MIX under the equal-example objective, BBH rises from 56.5 to 61.0 on Qwen3 and from 47.2 to 53.0 on OLMo, so its BBH penalty shrinks from 12.2 to 7.9 and from 17.2 to 11.8 points. Among the token-mean arms the rule moves SHORTEST-CORRECT by 5.7 (Qwen3) and 5.6 (OLMo) points and every other by less than a point except the OLMo matched-random arm (1.7). Two pairs of BBH causal-judgement questions have identical text and opposite gold labels in the benchmark source; omitting these four questions changes no entry's BBH task-macro accuracy by more than 0.17 points, and the four-benchmark accuracy is unaffected. A response counts as a *direct answer* when it has at most 16 generated tokens; on a fixed sample of 60 such responses (Qwen3, GUIDED-MIX-T2SUB, seed 0, sampled; read with an AI assistant) all 60 were bare answers without a reasoning step.

Quoted instructions. After the registered comparison, the follow-up and the extension (Appendix O) were finished (before the panel of Appendix E, whose own rescoring is reported there), a later audit of the stored responses, performed with an AI assistant (see the AI use statement), found that the grader can read a quoted format instruction as the answer. It searches for "Final

Answer:" anywhere in a line and, for yes/no and multiple-choice questions, takes the first valid label that follows. A response that quotes the format instruction of the prompt (*"Final Answer: ⟨answer⟩" where ⟨answer⟩ is Yes or No*) therefore produces an answer line of its own: it is read as "Yes", or as the example option "(A)", it is credited whenever that is the gold label, and as the last such line it can hide a real answer line above it. Among the 1,206,728 stored responses, 11,763 have the placeholder inside the selected answer span and 2,043 of those are credited, 1,592 of them in the token-matched long control; 24 credited responses of the untrained OLMo model consist only of a copy of the prompt. The registered scores remain the primary results. As a labelled sensitivity, whose rule and test cases were fixed before any contrast was computed under it, we delete literal copies of the prompt text and every header that merely quotes the placeholder, and apply the unchanged grader (Table 17); where the placeholder is used as a tag around an answer or is directly followed by one, the header is kept and only the placeholder token is removed. This changes the extracted answer or the grade of 9,165 responses (2,408 become correct, 1,386 incorrect). No contrast changes sign, none of the 46 Holm decisions changes, and the four learning rates selected on the development set stay the same; 94 of the 106 contrasts (the 53 contrasts before the panel, each under both decodings) move by less than 0.1 points and, outside the token-matched control, none by more than 0.26. The token-matched long control is the exception. Its sampled accuracy rises from 54.17 to 61.81, because quoted instructions had hidden real answer lines, and its contrast with GUIDED-MIX becomes -11.19 $[-12.77, -9.64]$ under sampling and $+0.48$ $[-1.05, +1.97]$ under greedy decoding. In addition, 1,565 sampled and 662 greedy final lines of this control carry the correct label, option letter or number directly followed by an underscore or other characters ("Yes_Pods", "A_Podsman"), a form that the grader parses for no recipe and that occurs in no other entry; because its contrast depends on how such lines are read, we draw no accuracy conclusion from this control. The rule removes literal prompt text only: paraphrased quotations remain; in 41 responses the grade depends on which of two headers on one line is read, 40 of which weigh two formats (*"Final Answer: Yes" or "Final Answer: No"*, 38 of them in traces cut off at the token limit) and one of which corrects itself after its first header; one MATH-500 response loses its credit only because deleting the placeholder changes how the checker parses the remaining text; and responses were read in stratified samples (149 flagged responses before the rule was written; 22 that gained and 24 that lost credit afterwards). **Replay.** Re-running the registered scorer on all stored responses in the original environment reproduces every extracted answer. The grades of 20 predictions longer than 10,000 characters (18 from the token-matched long control) depend on the checker's five-second time limit, and two Math-Verify parsing limits affect eight further mathematics responses; any one of these grades moves a three-seed four-benchmark mean by at most 0.02 points, and the registered grades are kept.

M DATA SOURCES, CONTAMINATION AUDIT, PROMPTS AND SERVING

Sources. Training questions come from the training split of GSM8K (`openai/gsm8k`), a level-stratified sample of 3,999 MATH training problems (`EleutherAI/hendrycks_math`), ARC-Challenge train and validation (`allenai/ai2_arc`) and the labelled StrategyQA mirror `ChilleD/StrategyQA`, its train and test splits pooled. Before any generation, a normalized exact-match screen against all evaluation questions removed 10 of the 14,880 pooled training questions (8 from ARC-Challenge, one from MATH and one from StrategyQA), leaving the 14,870 of the teacher pool: 7,473 GSM8K, 3,998 MATH, 1,410 ARC-Challenge and 1,989 StrategyQA. Evaluation uses the GSM8K test split (1,319), MATH-500 (`HuggingFaceH4/MATH-500`, the representative subset of the MATH test set of Lightman et al., 2024), the ARC-Challenge test split (1,172), 300 questions held out from the pooled StrategyQA set with a fixed shuffle before any teacher generation (not the official test set), and eight BBH tasks (`lukaemon/bbh`; 1,937 questions): date_understanding, logical_deduction_three_objects, temporal_sequences, disambiguation_qa, navigate, web_of_lies and sports_understanding with 250 questions each, and causal_judgement with 187. We report BBH as the task-macro accuracy of this subset, not of the full benchmark. MATH-500 and BBH were never used for generation. **Contamination audit.** Between the 12,108 training questions and the 5,228 evaluation questions, a normalized exact-match check finds no overlap. Among the training targets, 131 (in 69 training questions) share an eight-word run with one of 27 evaluation questions; they were trained on and are withheld only from the released targets (Appendix N). A five-word-shingle check at Jaccard similarity ≥ 0.8 flags four training problems (three from counting and probability, one from number theory) as template siblings of two MATH-500 problems (`math500-0115`, `math500-0268`); they differ in the word whose letters are arranged or in one number, and were retained. The same check flags four pairs between training

Table 17: Registered scores and the quoted-instruction sensitivity: literal copies of the prompt and headers followed by the placeholder are removed (where the placeholder tags or directly precedes an answer, the header is kept and only the placeholder token is removed), then the unchanged grader is applied. Contrasts in four-benchmark accuracy points with 95% intervals (sampled decoding unless stated); "–": not run.

Contrast	Student	Original grader	Quoted instructions removed
Recipe × objective interaction	Qwen3	$+8.31$ [$+6.77$, $+9.87$]	$+8.29$ [$+6.74$, $+9.84$]
	OLMo	$+5.23$ [$+3.70$, $+6.74$]	$+5.14$ [$+3.61$, $+6.66$]
Within-dataset switch, penalty	Qwen3	$+9.28$ [$+7.95$, $+10.62$]	$+9.20$ [$+7.86$, $+10.55$]
	OLMo	$+6.24$ [$+4.98$, $+7.46$]	$+6.23$ [$+4.96$, $+7.45$]
Tuned token mean − tuned equal example	Qwen3	$+8.50$ [$+7.18$, $+9.89$]	$+8.45$ [$+7.12$, $+9.84$]
	OLMo	$+6.35$ [$+5.01$, $+7.72$]	$+6.29$ [$+4.96$, $+7.66$]
GUIDED-MIX − SHORTEST-CORRECT	Qwen3	$+5.72$ [$+4.48$, $+6.95$]	$+5.64$ [$+4.38$, $+6.88$]
	OLMo	$+4.82$ [$+3.55$, $+6.05$]	$+4.81$ [$+3.54$, $+6.05$]
GUIDED-MIX − BON-SHORT	Qwen3	-7.35 [-8.78, -5.90]	-7.42 [-8.87, -5.97]
	OLMo	-12.58 [-14.07, -11.07]	-12.46 [-13.94, -10.93]
GUIDED-MIX − LONG	Qwen3	-6.46 [-7.91, -4.98]	-6.72 [-8.16, -5.25]
	OLMo	-11.81 [-13.28, -10.31]	-11.82 [-13.32, -10.31]
Guided − matched-random, token mean	Qwen3	$+0.73$ [-0.44, $+1.92$]	$+0.64$ [-0.53, $+1.84$]
	OLMo	-0.98 [-2.13, $+0.18$]	-0.99 [-2.15, $+0.17$]
FIXED-T4 − LONG (descr., post hoc)	Qwen3	$+2.94$ [$+1.63$, $+4.26$]	$+2.77$ [$+1.44$, $+4.09$]
	OLMo	$+0.78$ [-0.48, $+2.05$]	$+0.78$ [-0.48, $+2.06$]
LONG, token-matched − GUIDED-MIX, sampled	Qwen3	-18.92 [-20.64, -17.17]	-11.19 [-12.77, -9.64]
	OLMo	–	–
LONG, token-matched − GUIDED-MIX, greedy	Qwen3	$+1.60$ [$+0.06$, $+3.08$]	$+0.48$ [-1.05, $+1.97$]
	OLMo	–	–
Accuracy (four benchmarks / BBH task macro), sampled			
LONG	Qwen3	79.55 / 76.15	79.72 / 75.77
	OLMo	77.86 / 71.03	77.86 / 70.63
BON-SHORT	Qwen3	80.43 / 73.93	80.42 / 73.83
	OLMo	78.63 / 62.55	78.50 / 62.34
GUIDED-MIX	Qwen3	73.08 / 68.70	73.00 / 68.68
	OLMo	66.05 / 64.35	66.04 / 64.33
LONG, token-matched	Qwen3	54.17 / 51.09	61.81 / 53.87
	OLMo	–	–
Untrained starting model	Qwen3	37.36 / 39.01	36.98 / 37.46
	OLMo	64.23 / 36.51	64.22 / 35.99

Table 18: Four-benchmark accuracy as scored and with every response at the token cap counted as incorrect (sampled decoding).

Student	Recipe	As scored	Cap hits counted incorrect
Qwen3	ANSWER-ONLY	52.27	52.27
Qwen3	SHORTEST-CORRECT	67.37	67.37
Qwen3	GUIDED-MIX	73.08	72.98
Qwen3	GUIDED-MIX-T2SUB	72.67	72.64
Qwen3	FIXED-T2	74.23	74.23
Qwen3	FIXED-T3	79.18	79.18
Qwen3	FIXED-T4	82.49	82.49
Qwen3	BON-SHORT	80.43	80.02
Qwen3	LONG	79.55	78.40
Qwen3	LONG, token-matched	54.17	28.01
OLMo	ANSWER-ONLY	45.34	45.34
OLMo	SHORTEST-CORRECT	61.23	61.23
OLMo	GUIDED-MIX	66.05	65.99
OLMo	GUIDED-MIX-T2SUB	65.83	65.75
OLMo	FIXED-T2	68.93	68.93
OLMo	FIXED-T3	76.37	76.37
OLMo	FIXED-T4	78.64	78.64
OLMo	BON-SHORT	78.63	78.06
OLMo	LONG	77.86	75.97

Table 19: BBH task-macro accuracy under the frozen grader and under format-tolerant rescoring (sampled decoding, mean of three seeds; starting models: one run; 1,937 questions); the lower block lists the arms of the loss-weighting experiments. Frozen: the grader of the study; Tolerant: the format-tolerant rescoring; Bare labels: mean number of bare-label responses.

Student	Recipe	Frozen	Tolerant	Bare labels
Qwen3	Starting model	39.0	39.0	0
Qwen3	ANSWER-ONLY	58.9	58.9	0
Qwen3	SHORTEST-CORRECT	56.5	62.1	205
Qwen3	GUIDED-MIX	68.7	68.9	5
Qwen3	GUIDED-MIX-T2SUB	67.3	67.3	2
Qwen3	FIXED-T2	65.1	65.1	0
Qwen3	FIXED-T3	68.0	68.0	0
Qwen3	FIXED-T4	68.1	68.1	0
Qwen3	BON-SHORT	73.9	73.9	0
Qwen3	LONG	76.2	76.2	0
Qwen3	LONG, token-matched	51.1	51.1	0
OLMo	Starting model	36.5	36.6	5
OLMo	ANSWER-ONLY	52.2	52.2	0
OLMo	SHORTEST-CORRECT	52.3	57.9	212
OLMo	GUIDED-MIX	64.3	64.8	16
OLMo	GUIDED-MIX-T2SUB	62.8	63.1	9
OLMo	FIXED-T2	62.4	62.4	0
OLMo	FIXED-T3	63.3	63.3	0
OLMo	FIXED-T4	66.7	66.7	0
OLMo	BON-SHORT	62.5	62.5	0
OLMo	LONG	71.0	71.0	0
Qwen3	GUIDED-MIX, equal example	56.5	61.0	172
Qwen3	GUIDED-MIX, token shares, equal within	55.1	59.9	162
Qwen3	GUIDED-MIX, example shares, token within	63.9	64.2	10
Qwen3	Matched-random, token mean	66.0	66.6	16
Qwen3	Matched-random, equal example	46.3	56.4	373
Qwen3	GUIDED-MIX-T2SUB, equal example	54.8	60.0	194
Qwen3	FIXED-T2, equal example	63.9	63.9	0
OLMo	GUIDED-MIX, equal example	47.2	53.0	230
OLMo	GUIDED-MIX, token shares, equal within	49.0	53.6	187
OLMo	GUIDED-MIX, example shares, token within	60.6	61.2	21
OLMo	Matched-random, token mean	59.6	61.3	68
OLMo	Matched-random, equal example	48.2	54.0	232
OLMo	GUIDED-MIX-T2SUB, equal example	48.8	54.1	208
OLMo	FIXED-T2, equal example	60.9	60.9	0

and development problems (three of the same counting problems against one development problem, and one further pair), and one pair between a development problem and `math500-0115`; all are template siblings, and no pair of question sets shares an exact match: the counting problems differ in the word whose letters are arranged, and one training–development pair has identical quantities and asks for two socks of different colours in one problem and of the same colour in the other. Two development problems are thus template siblings of training problems; the development set is used only to select learning rates. **Student prompt and serialization.** The prompt is the question (for ARC followed by one line per option, `A. <option text>`), an empty line, and one line that ends with a dataset-specific hint:

```
End your response with exactly one line: "Final Answer: <answer>" <hint>
```

where `<hint>` is `where <answer> is the final number or mathematical expression.` for GSM8K and MATH-500, `where <answer> is the letter of the chosen option.` for ARC and `where <answer> is Yes or No.` for StrategyQA. BBH uses the same line zero-shot, with `where <answer> is the letter of the chosen option in parentheses, for example (A).` for its four multiple-choice tasks and `where <answer> is Yes or No.` for the other four. Prompts and targets are serialized with a ChatML template: the Qwen3 tokenizer's own template with thinking disabled,

and a fixed ChatML template for OLMo; the end-of-turn token is appended once to every target and is supervised. At evaluation, generation stops at the end-of-turn token <|im_end|> (token 151645 for Qwen3, 100265 for OLMo), which we pass explicitly because the server otherwise ignores the checkpoint's end-of-sequence list. Before training we checked on 20 examples per student that targets survive serialization, that the end-of-turn token occurs once at the end and that the loss mask covers exactly the target: 20 of 20 on OLMo, and on Qwen3 20 of 20 for the first two checks and 19 of 20 for the mask (the exception is one long-trace target). We also checked that the training tokenizer and the server produce identical prompt token IDs for the 20 Qwen3 examples; this comparison was not run for OLMo. **Serving.** Checkpoints are served with vLLM. For OLMo the server needs the rotary-embedding parameters in a flat form; we pass the checkpoint's own values (YaRN with factor 8, base 500,000, original length 8,192), a compatibility translation and not a change of the model. **Licences.** The model cards declare Apache-2.0 for the two students and the teacher; the dataset repositories declare MIT for GSM8K and CC-BY-SA-4.0 for ARC. For MATH, MATH-500, StrategyQA and BBH the declared repository licences do not settle the rights to every underlying question, so the release identifies questions by source, identifier and hash.

N MODELS, SOFTWARE AND COMPUTE

Teacher: `Qwen/Qwen3.6-35B-A3B`, revision `995ad96`. Students: `Qwen/Qwen3-4B-Base`, revision `906bfd4`, and `allenai/Olmo-3-1025-7B`, revision `a81bae4` (base checkpoints; full revision identifiers are in the supplementary material). Training uses PyTorch 2.11.0, Transformers 5.14.1, DeepSpeed 0.19.6 (ZeRO stage 3) and Accelerate 1.15.0 on one node with eight H100 GPUs per run; evaluation serves checkpoints with vLLM 0.25.1 on one node with eight GPUs, either one checkpoint per GPU or one checkpoint on all eight GPUs, with the same decoding settings. Every run stores the micro-batch size actually used and its training loss. Runs of the extension additionally log the hash of each executed source file and of their training file; five early runs predate that record and nine entries carry an evaluation-time record; the run and source records give the per-run provenance. Across the records of these runs, all before the panel of Appendix E, the trainer, the evaluator, the data builder, the serving script and the DeepSpeed configuration have one hash each; the job wrapper has three versions (cache relocation, checkpoint cleanup), each matching the hash that the jobs recorded. The twelve runs of that panel use a later trainer, which adds its two objectives, a new job wrapper, and training files that add the group label to every row without changing rows, order or messages; the evaluator, the data builder, the serving script and the DeepSpeed configuration are unchanged. Each of these runs was trained and evaluated in a single job that wrote its own source record when it started and its own evaluation record; before launch, every hash that the panel's analysis requires was checked against the files. The supplementary material attached to the submission holds the code, the protocols, the training targets (those that reproduce eight or more consecutive words of a benchmark question are withheld and listed by identifier), the question manifests and the summary results.

O STAGING, PROTOCOL AMENDMENTS AND DEVIATIONS

The recipe comparison, its six contrasts and the non-inferiority margin were registered before any student was trained (the protocol's hash and freeze time are recorded in the environment file of the supplementary material). The allocation-by-objective experiment (the follow-up), and later the experiments of the extension, namely the OLMo controls (ANSWER-ONLY, FIXED-T2 and GUIDED-MIX-T2SUB on OLMo), the recipe-by-objective experiment (whose Qwen3 token-mean cells are the registered contrast C4), the normalization decomposition and the learning-rate check, and the fixed-depth ladder, were each specified after earlier results had been seen, with predictions, hypothesis families and selection rules fixed before the corresponding runs. The descriptive comparison of the fixed depths with long traces was added after the ladder results had been seen. The split of the within-dataset switch (Appendix E) is post hoc: it was specified after the within-dataset result, with its protocol frozen before its twelve runs. The shared-weight control (Appendix F) is post hoc in the same sense: designed after the recipe-by-objective result, with its main estimate, reading rule and analysis code fixed before its nine runs. Run records with every job are part of the supplementary material. Deviations: (i) an interim report of the allocation-by-objective experiment on OLMo used the opposite sign convention for the interaction and had been read

before the convention was corrected; all reports were regenerated with the pre-specified convention, which changed neither the two-sided significance nor any Holm decision. (ii) Jobs interrupted by storage-quota limits were restarted from the same seed (training) or resumed by question (evaluation), and every affected entry passed the same completeness gate as the others (all files present, row counts and question identifiers equal to the gold splits). (iii) Every contrast of the extension draws its bootstrap resamples from its own seeded stream; an earlier shared stream could have moved an interval by up to 0.08 points, with point estimates and all test decisions unchanged, and the paper reports the final values. The registered analysis is restricted to the arms it was registered with and reproduces the published reports. (iv) The registered token-matched long control ($1{,}085$ examples, 51 optimizer steps; Qwen3 only) is reported in Table 13 and Appendix B. (v) The quoted-instruction rescoring of Appendix L was added after the earlier experiments and before the panel of Appendix E; the registered scores are unchanged, its rule was fixed before any contrast was computed under it, and a draft step that read the last of several headers on one line was dropped at that stage (41 responses; Appendix L). (vi) The analysis halts when a registered seed or question identifier is missing. (vii) The protocol named Qwen3.5-4B-Base as the primary student and Qwen3-4B-Base as a pre-specified fallback; before any training the untrained Qwen3.5-4B-Base scored 93.2% on the development set (greedy), leaving no room to learn, and a recorded amendment adopted the fallback; OLMo-3-7B was unchanged. (viii) After three OLMo BoN-SHORT runs diverged and one OLMo LONG run terminated with an error in PyTorch's cuDNN attention kernel, that kernel was disabled for all later runs, and the four runs were retrained with their seeds. The other registered runs, including the token-mean GUIDED-MIX, FIXED-T2 and GUIDED-MIX-T2SUB arms reused later, were trained with it enabled and kept after their training curves were checked; for OLMo the kernel was in use, and a later diagnostic shows that the default settings select it for Qwen3 as well (Appendix F), where retraining the Qwen3 GUIDED-MIX-T2SUB token-mean arm without it is also reported. Contrasts between a reused token-mean arm and a later arm (E1 and E3, the objective effect of GUIDED-MIX, B1, the three tests per student of the normalization decomposition, P_{between} and P_{within}, R1, F1, and on OLMo C2 and C5) therefore also differ in the attention kernel; contrasts whose arms were trained under the same setting (S, B2, C4, R2 to R4, F2 to F4) do not. Its effect was measured on one arm: retraining the Qwen3 GUIDED-MIX-T2SUB token-mean runs without the kernel changes their accuracy by -0.7 (sampled) and $+0.5$ (greedy) points (Appendix F); the measured effect is reported for that arm. (ix) The development evaluation of the Qwen3 learning-rate candidates was repeated after a server stop-token setting was corrected, before any rate was selected. (x) The sign of the recipe-by-objective interaction was predicted after the objective penalty of the allocation-by-objective experiment on the full mixture had been seen. (xi) The protocol named one hypothesis family for the six registered contrasts; the analysis code tests each student separately, so on OLMo the three contrasts that its four trained recipes allow form their own family (Appendix H); pooling them into the Qwen3 family (nine tests) would change one decision: the adjusted p of C4 would rise from 0.04 to 0.06 and C4 would no longer be rejected; the other eight decisions are unchanged. (xii) Two matched-random runs on Qwen3 (allocations 0 and 1, seeds 0 and 1) were trained at micro-batch 2 after an out-of-memory stop at micro-batch 4, with the same seed and recipe; the long-trace recipes used micro-batch 1 or 2 throughout. The objectives do not depend on the micro-batch (Appendix D).