Strong Post-Training from Permissive, Reasoning-Dominant, Web-Scale Pretraining

Ali Elganzory[1,4]* **Harsh Raj**[2,3,4,5]* **Marianna Nezhurina**[2,3,4,5]*
Victor May[4,6]* **Van Khue Nguyen**[4,7] **David Salinas**[1,4,5,8]
Huu Nguyen[3,4,5,6] **Jenia Jitsev**[2,3,4,5]

[1]University of Freiburg [2]LAION
[3]Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ)
[4]Open-Ψ (Open-Sci) Collective [5]OpenEuroLLM team
[6]Ontocord [7]École Polytechnique [8]ELLIS Institute Tübingen

Abstract

Can front-loading instruction and reasoning data into pre-training make permissive-first pretraining competitive with traditional non-permissive mixtures after post-training? We compare 1.7B-parameter models pretrained on 300B tokens under a fixed recipe across four pre-training datasets, of which two are permissive. We first look at Tulu3 SFT and DPO, where such enrichment via front-loading improves reasoning and instruction scores in every pre-training dataset. At matched pretraining compute, both enriched permissive mixtures outperform the tested unenriched non-permissive references on the reasoning-and-instruction suite. At 16K context, they also score above both enriched non-permissive references. We then perform further reasoning SFT on the Tulu3 SFT checkpoints using the same 30K samples OpenThoughts3 dataset. Enriched models from both permissive and non-permissive pre-training datasets achieve substantial further gains on math/code/reasoning. Without front-loading, reasoning SFT outcome strongly drops for all pre-training datasets. Additional controls on MixtureVitae as a permissive mixture show that excluding the OpenThoughts3 reasoning traces data from pretraining reduces strongly subsequent post-training gain on reasoning evals. The beneficial front-loading effect is not restricted to OpenThoughts3 - similar gains are obtained when using other reasoning corpora for post-training. While front-loading reasoning/instruct boosts math/code/reasoning capabilities in post-training for both permissive and non-permissive pre-training mixtures, the permissive mixtures still show gap on general language understanding evals. These results establish front-loading reasoning and instruction data as a route to strong post-training under permissive data constraint and show that pretraining composition strongly determines the effectiveness of successive post-training. Code, evaluation records, and checkpoints are public.[1]

1 Introduction

Open alternatives to closed frontier foundation models have improved steadily, in released weights (Llama [1], Qwen [2]) and in open pretraining datasets such as RedPajama [3] and DCLM [4], and fully reproducible pipelines now combine open pretraining (Olmo [5]) with open post-training

*Equal contribution, listed in alphabetical order of first name.

[1]Code: `https://github.com/MixtureVitae/permissive-post-train`. The Reproducibility Statement at the start of the appendix links the evaluation records and the checkpoints.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

(Tulu [6]). Efforts such as CommonPile [7] build pretraining corpora from public-domain and openly licensed text. Recent work, including CommonPile, MixtureVitae [8] and SYNTH [9], reports increasingly competitive base model performance, but whether permissive base models keep up after post-training has not been measured.

We answer this question with a controlled comparison in which the pretraining recipe and subsequent training procedures are held fixed across mixtures. We use *enrichment* to mean including a designated package of data containing instruction and reasoning examples in a pretraining mixture. We refer to sampling from that package alongside the other sources throughout pretraining as *front-loading*. Our main experiments use 1.7B-parameter models pretrained on 300B tokens, comparing the permissive-first MixtureVitae mixture, which enriches pretraining with instruction and reasoning data [10], with reference corpora under the same architecture, tokenizer, and training budget. Phase 1 evaluates instruction alignment with Tulu3 supervised fine-tuning and DPO. In the main Phase 2 comparison, each model undergoes reasoning SFT on the same 30K-example OpenThoughts3 dataset, starting from its own Tulu3 SFT checkpoint, to measure how much it gains from the same small post-training budget. More heavily pretrained open-weight models provide additional, uncontrolled references.

After Phase 1, MixtureVitae outperforms the original reference mixtures, which lack this enrichment, on the reasoning and instruction-following suite. Phase 2 reveals a sharper difference in their response to further training: the same reasoning SFT procedure produces large mathematical gains for MixtureVitae but little or no mathematical improvement for these reference mixtures. The comparison therefore concerns not only the quality of the post-trained models, but also how much their different pretraining histories allow them to benefit from a subsequent reasoning SFT run.

We investigate this response with complementary interventions on pretraining and post-training. Removing the OpenThoughts3 source from MixtureVitae pretraining substantially reduces the gain from the otherwise identical reasoning SFT procedure. Conversely, replacing OpenThoughts3 at post-training with alternative reasoning corpora still produces large mathematical gains on the intact MixtureVitae model. The response is thus sensitive to pretraining composition, but is not confined to using OpenThoughts3 again at post-training.

MixtureVitae is both reasoning-enriched and permissively sourced, however, so these comparisons alone do not establish how competitive it remains against enriched references. We therefore cross four web pre-training datasets, two of them permissive, with the inclusion or exclusion of the same instruction and reasoning subset, at the full web-scale 300B-token pretraining budget (larger than usual scales used for comparisons). Enrichment improves post-trained reasoning in all four pre-training datasets, and with the subset in every mixture both permissive mixtures reach competitive reasoning scores after instruction alignment, although a gap in general language understanding remains for both. MixtureVitae is therefore not the only permissive mixture that supports strong post-training. Under the same reasoning SFT procedure, the three enriched reference mixtures also gain substantially, and in each pre-training dataset enrichment turns a decline into a large gain.

These results extend prior work on how reasoning-rich pretraining and mid-training shape subsequent learning [10, 11, 12] by testing strong post-training under a permissive-first sourcing constraint: **(i) Strong post-training under permissive-first sourcing.** At matched pretraining compute, MixtureVitae outperforms the original reference mixtures after post-training on Tulu3 and expands gains further after a 30K-example reasoning SFT. **(ii) Front-loading across pre-training datasets.** A full-budget factorial shows that front-loading same instruction and reasoning subset improves reasoning post-training in all four pre-training datasets, with competitive reasoning obtained also when starting from permissive pre-training mixtures. Through reasoning SFT, the three enriched reference mixtures gain substantially, and in each pre-training dataset enrichment turns a decline into a large gain. **(iii) Pre-training composition determines post-training receptivity.** Excluding the OpenThoughts3 source from pretraining reduces MixtureVitae's gain from the SFT run, while three other reasoning corpora also produce large broad gains in post-training if OpenThoughts 3 is front-loaded.

The base, long-context, SFT, DPO, and reasoning post-trained checkpoints are public, with the evaluation records and the pipeline that regenerates every table and figure (Reproducibility Statement).

2 Methods and experimental setup

2.1 Pretraining recipe and base models

The bases of our main comparisons are 1.7B decoder transformers pretrained on 300B tokens under the `open-sci-ref` protocol [13], an open and reproducible reference recipe. It fixes the architecture, tokenizer, optimizer, learning rate schedule, batch size, and total training tokens, so that bases trained under it differ only in their pretraining data (hyperparameters in Appendix A.2). Each base also has a 16K context variant (Section 2.2), for example MV and MV-16K, unless stated otherwise. The bases enter three comparisons.

Original mixture comparison. Our default base, MV, is pretrained on MixtureVitae [8]. Its matched-compute references, the original reference mixtures, are pretrained on four widely used pretraining corpora without the instruction and reasoning subset of MixtureVitae: Nemotron-CC-HQ [14] (the strongest pretraining dataset identified in the `open-sci-ref` evaluation), FineWeb-Edu [15], DCLM [4] and CommonPile v0.1 [7] (denoted Comma0.1, which is permissively licensed like MV but has no instruction and reasoning subset of its own). All of them also have a 16K variant.

Front-loading factorial. MixtureVitae is not the only reasoning-enriched mixture in this study. To separate what the mixture contains from where its web text comes from, we train a 4×2 factorial of bases. The first factor is the pre-training dataset supplying the web text, at four levels: two that carry the permissive constraint, the curated web portion of MixtureVitae and CommonPile v0.1 (Comma0.1), and the two non-permissive references Nemotron-CC-HQ and FineWeb-Edu. The second factor is front-loading, namely whether the same MixtureVitae instruction and reasoning subset is present in the pre-training mixture (as in [10]). Pretraining is a single stage in which every example is sampled at random from the mixture, so the subset has no curriculum of its own. The three enriched reference mixtures all give the subset 42.1% of their tokens, close to its 45.0% in MixtureVitae, which makes the front-loading effect comparable between them. It holds 175B tokens of mathematics, code, reasoning traces and other instruction data (Appendix A.1). Every arm is trained for the full 300B tokens, so the MixtureVitae arm without the subset repeats part of its remaining data. All eight bases pass through the same long context extension and Phase 1. MixtureVitae's enriched arm is the main MV model, so the study adds three enriched reference mixtures: the permissive Comma0.1 and the non-permissive Nemotron-CC-HQ and FineWeb-Edu. The main Phase 2 comparison (Section 3.3) uses MV and the original reference mixtures. The three enriched reference mixtures, and a Nemotron-CC-HQ control without front-loading, also go through Phase 2.

OpenThoughts3 exclusion. To measure the contribution of the OpenThoughts3 source to subsequent gains from reasoning SFT, we compare MV with a variant pretrained without that source and nothing else changed (OT3-excluded MV, which keeps the rest of its instruction and reasoning data), and give both bases the same Phase 2 reasoning SFT procedure.

External open weights references. SmolLM2 1.7B [16], Qwen2.5 1.5B [17] and Qwen3 1.7B [18] are released models with their own recipes, pretrained on 37, 60 and 120 times as many tokens as our bases. They are higher-compute references outside the controlled comparison. SmolLM2 (8K native context) is passed through our 16K extension before post-training, while Qwen2.5 and Qwen3 (both 32K) are post-trained as released.

Two auxiliary sets of MV bases, a 0.4B model compared with Baguettotron [9] and three 1.7B variants at 100B tokens for mixture composition ablations, are described in Appendices F.3 and F.1.

2.2 Post-training pipeline

Post-training has two phases, with data and recipe identical across models within each phase. Phase 1 covers the bases of the original comparison, the factorial and the external references, and Phase 2 starts from selected 16K checkpoints (Section 3.3). Figure 1 gives an overview.

Long context extension. Each base is extended from 4K to 16K tokens with YaRN [19] and one epoch of full fine-tuning on long context instruction data (SeaLong [20] and LongAlign [21]) mixed with short conversational data (SmolTalk [16]). Appendix A.2 gives the settings. The same recipe is applied identically to MV and to every 1.7B reference base before Phase 1 post-training, so that long context behavior is attributable to the base and not to recipe differences.

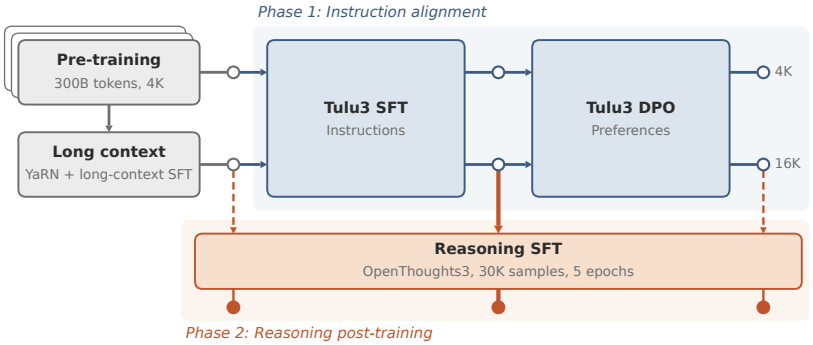


Figure 1: **Post-training pipeline**. Circles denote checkpoints. Each base is extended from 4K to 16K context with YaRN and one epoch of fine-tuning. Phase 1 applies Tulu3 SFT and then DPO to the 4K base and to its 16K extension, which gives six checkpoints: base, SFT and DPO at 4K and 16K. Phase 2 applies reasoning SFT on a 30K-sample subset of OpenThoughts3, only from 16K checkpoints. The solid line is the main comparison, which starts from the SFT checkpoint. The dashed lines are two controls that start from the 16K extended checkpoint (without Tulu3) and from the DPO checkpoint.

Phase 1: Instruction alignment (Tulu3 SFT + DPO). We apply the first two stages of the Tulu3 [6] recipe, SFT and DPO, and omit its third stage, reinforcement learning with verifiable rewards (RLVR). The supervised fine-tuning stage uses the Tulu3 SFT mixture, which combines public and synthetic instruction, math, code, and safety data. The preference optimization stage uses DPO [22] on Tulu3's preference data. Both stages use the learning rates, batch sizes and epochs of the Tulu3 recipe without model-specific tuning, and DPO uses $\beta = 5$ with the standard sigmoid loss, without Tulu3's length normalization (Appendix Table 8). The 16K checkpoints are post-trained with the same 4,096-token SFT and 2,048-token DPO sequences as the 4K ones. Tulu3 SFT computes the loss on prompt and response tokens, unlike the response-only loss of Phase 2, and Appendix F.4 compares response-only masking for Phase 1. We use the historically filtered Tulu3 datasets, and Section 2.4 documents the residual overlap and the filtering coverage.

Phase 2: Reasoning post-training (OpenThoughts3). Starting from each model's Phase 1 SFT checkpoint, we further train for 5 epochs on a uniformly sampled 30K sample subset of OpenThoughts3 (OT3) [23] following the OT3 medium hyperparameter set. For MV, we also run the same reasoning SFT procedure from the DPO checkpoint and, without Tulu3, from the 16K extended checkpoint, as controls.

2.3 Benchmarks and evaluation

Reasoning and instruction-following suite (primary). This suite targets the capabilities that post-training is meant to improve. We report accuracy on AIME24 [24], AIME25 [25], AMC23 [26], GSM8K [27] and MATH500 [28] for math reasoning; HumanEval [29], LiveCodeBench [30], and MBPP [31] for code; GPQA-Diamond [32] and JEEBench [33] for science and knowledge-intensive question answering; and IFEval [34] for instruction following. Decoding and few-shot settings follow the OpenThoughts3 evaluation protocol [23] and are summarized in Table 15.

Language understanding suite (secondary). We assess whether post-training degrades general language capabilities, following the evaluation strategy of MixtureVitae [8]. All standard language evaluations are run with the LM Evaluation Harness [35] on a fixed suite of recognized downstream benchmarks: MMLU [36], HellaSwag [37], CommonsenseQA [38], ARC Challenge and Easy [39], PIQA [40], BoolQ [41], Winogrande [42], OpenBookQA [43], COPA [44], and LAMBADA [45]. Few-shot counts for each task are given in Table 16.

Evaluation protocols and uncertainty. Phase 1 and the factorial are evaluated under one fixed generation budget with three decoding seeds (42, 1234, 2025), reported as mean$_{\pm \mathrm{SD}}$; Appendix Table 15 gives the settings. The seeds measure how much repeated evaluation of one checkpoint varies, not how much a result varies between training runs, which these results do not estimate (Appendix B.2). The reasoning post-training experiments of Sections 3.3 and 3.4 use the reasoning

generation budget of Appendix Table 15, sampled at temperature 0.7 with three runs and reported as $\text{mean}_{\pm\text{SD}}$ over the runs, so their scores are changes measured against a starting checkpoint re-evaluated under the same protocol rather than values directly comparable with Phase 1. For comparisons on the suite average, we report pointwise 95% paired item-level bootstrap confidence intervals and adjust two-sided p-values within each stated comparison family using Benjamini–Hochberg. Appendix B.2 gives the method and the number of bootstrap replicates. The suite average is the unweighted mean of the eleven tasks, and reweighting its domains changes neither MV's lead after Phase 1 nor the front-loading effect (Tables 19 and 21), and Phase 2's changes are reported by domain (Appendix Table 34). AIME24, AIME25 and AMC23 contain 30 to 40 items, so single-benchmark differences on them are descriptive and we do not draw conclusions from them alone.

2.4 Benchmark overlap and robustness

We assess training–evaluation overlap using 13-gram and exact matching against the benchmark items used in our evaluations, auditing the post-training datasets and MV's pretraining reasoning sources within the coverage specified in Appendix A.3. The audit identifies substantial overlap on several benchmarks: 650/1,221 CommonsenseQA items occur in the Tulu3 SFT data, and MV's pretraining reasoning pool matches 211/500 MATH500, 25/30 AIME24 and 199/369 LiveCodeBench problems. We therefore complement scores on complete benchmarks with sensitivity analyses that exclude a common set of identified items from every compared checkpoint wherever item-level outcomes are available. On MATH500, excluding 213 identified items leaves 287 problems: MV's SFT→OT3 gain remains $+41.9$ points $[+37.6, +46.5]$, compared with $+41.8$ on the full benchmark. The OT3-source exclusion still reduces the subsequent gain, and all three alternative reasoning corpora retain substantial mathematical gains on the same subset. Appendix A.3 reports the overlap counts, the audit coverage and every comparison, including changes in significance.

3 Experiments and results

We organize the results around the two post-training phases of Section 2.2. Section 3.1 reports Phase 1, where the Tulu3 recipe is applied to every base. Section 3.2 then reports the central result of this work, the controlled comparison in which the same instruction and reasoning subset is added to, or left out of, four pre-training datasets. Section 3.3 reports Phase 2, the reasoning post-training stage that separates the mixtures most sharply, and Section 3.4 measures how sensitive MV's response is to pretraining composition and to the data for reasoning SFT.

3.1 Phase 1: Tulu3 SFT and DPO post-training

Math, code, and instruction following. Table 1 summarizes Phase 1 on the primary reasoning and instruction-following suite of Section 2.3. *In the original mixture comparison, MV-16K has the highest matched-compute suite average at every stage*: it averages 29.1 as a base model, 26.6 after Tulu3 SFT, and 26.9 after DPO. The original reference mixtures stay far behind after the identical pipeline, at 12.0 for Comma0.1, 8.3 for Nemotron-CC-HQ, 8.2 for FineWeb-Edu, and 8.2 for DCLM. Comma0.1 is the informative reference among the matched-compute models, since it is permissively licensed but is not enriched with instruction and reasoning data, and it lands with the other references rather than with MV. Enriched with the same subset it does not: it is then ahead of MV at 4K and level with MV-16K after SFT and DPO (Section 3.2).

The 16K Tulu3 reasoning drop. Table 1 also shows a phenomenon that Phase 2 returns to. MV-16K is strongest before Phase 1, at 29.1 against 17.0 at 4K, and Tulu3 SFT lowers the reasoning suite average to 26.6; DPO recovers part of the loss to 26.9 but does not restore the base level. The jump to 16K comes from the extension recipe of Section 2.2: besides YaRN, it trains one epoch on instruction data: SmolTalk conversations and long-context instruction examples. Every base shown at both context lengths gains from it, MV the most (Appendix Table 17), so the 16K checkpoint is already lightly instruction-tuned. The Tulu3 drop is therefore a drop from that checkpoint rather than from an untouched base, and the resulting MV checkpoints still outperform every original reference mixture by more than a factor of two. Phase 2 tests whether the lost performance can be recovered (Section 3.3), which does not by itself identify the cause of the decline.

Table 1: **Phase 1 results for all models.** Benchmark columns are the endpoint after Tulu3 SFT→DPO, and the last three columns give the eleven-task suite average before Tulu3 and after each stage. MixtureVitae and Comma0.1 are permissive. Matched-compute models and SmolLM2 are at 16K context, Qwen2.5 and Qwen3 at their released context. Scores are on the full benchmarks, which include items identified in the training data (Section 2.4), so they are not uniformly held-out estimates. Protocol A (Appendix Table 15 defines the abbreviations), means over three decoding seeds, with suite average SD ≤ 0.7. Appendix Table 17 gives SDs and both context lengths.

	Mathematics					Code			Science/QA		IF	Suite average		
Model	A24	A25	AMC	M500	GSM8K	HE	LCB	MBPP	GPQA	JEE	IFE	Base	SFT	DPO
Matched compute, 300B tokens														
MixtureVitae	3.3	1.1	17.5	37.0	52.6	52.6	9.4	40.3	23.4	6.2	52.6	29.1	26.6	26.9
Comma0.1	0.0	0.0	2.5	2.1	6.7	17.1	0.0	25.9	23.4	5.1	48.8	11.1	11.7	12.0
Nemotron-CC-HQ	0.0	0.0	2.5	2.5	7.7	5.3	0.0	5.3	23.4	1.7	43.3	7.8	8.1	8.3
FineWeb-Edu	0.0	3.3	0.0	3.3	9.2	4.9	0.1	2.5	23.9	2.7	40.7	8.3	7.7	8.2
DCLM	0.0	0.0	0.0	1.4	7.1	6.1	0.0	4.9	25.8	1.1	44.1	7.5	8.1	8.2
External references, 11–36T tokens, uncontrolled														
SmolLM2 1.7B	0.0	0.0	3.3	9.6	33.8	32.3	3.2	36.9	26.3	5.4	54.9	16.8	18.2	18.7
Qwen2.5 1.5B	2.2	0.0	10.0	21.2	40.4	55.7	3.3	44.5	27.9	8.0	58.6	9.9	23.4	24.7
Qwen3 1.7B	0.0	0.0	8.3	31.2	69.8	67.7	4.1	54.7	31.3	13.3	64.9	30.7	31.5	31.4

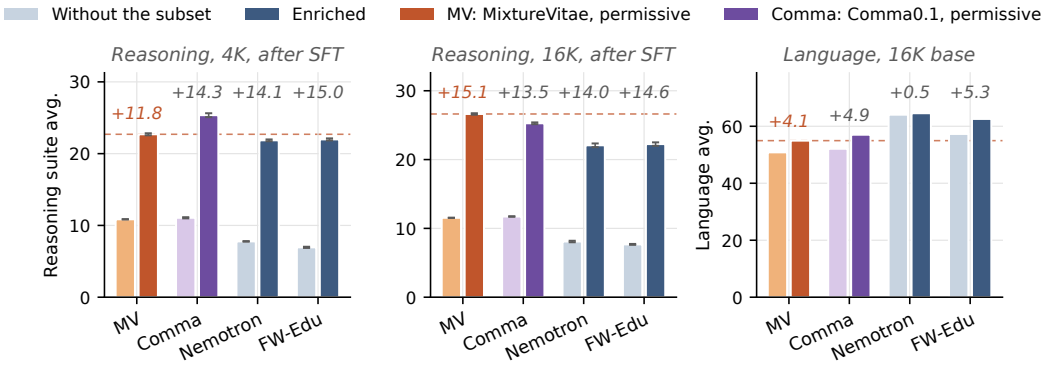


Figure 2: **Front-loading instruction and reasoning data drives the post-training reasoning gain for every pre-training dataset.** Each pair of bars is one pre-training dataset without and with the instruction and reasoning subset in pretraining, and the number above it is the front-loading effect. Nemotron is Nemotron-CC-HQ and FW-Edu is FineWeb-Edu, both non-permissive. The dashed line marks MV's enriched value. The reasoning panels give the suite average after Tulu3 SFT, with the standard deviation over three decoding seeds. The language panel gives the 16K extended checkpoints before Tulu3, at one seed for selecting prompt examples (intervals in Appendix Table 23). Appendix Table 20 gives every stage, and Appendix A.3 the contrasts without the identified overlaps.

3.2 Front-loading across pre-training datasets

Front-loading instruction and reasoning data improves post-training reasoning for every pre-training dataset. Phase 1 compares mixtures that differ in two ways at once: MixtureVitae is permissively sourced *and* enriched with instruction and reasoning data. The 4×2 design of Section 2.1 varies the two factors separately (Figure 2, Appendix Table 20). Adding the instruction and reasoning subset to the pretraining mixture raises the reasoning suite average in all twenty-four combinations of pre-training dataset, context length and stage, by 9.8 to 20.1 points, with every interval excluding zero ($q < 0.001$, Appendix Table 22). Tulu3 contains instruction, mathematics, and code data of its own, but it does not substitute for the subset in pretraining: Phase 1 narrows the effect in every case but one and closes it in none (Appendix Table 22). A 100B-token ablation points the same way: removing the subset from MixtureVitae leaves the post-trained model well below the full mixture, while removing its curated web subset leaves it level with or ahead of the full mixture (Appendix F.1). The benefit of enrichment persists after excluding identified benchmark overlaps (Appendix A.3).

At matched enrichment, the permissive datasets reach competitive reasoning scores after post-training. With the subset in every mixture, both permissive pre-training datasets lead both non-permissive ones at 16K after SFT and after DPO, by 3.0 to 5.1 points ($q < 0.001$), also without the identified overlaps, under which some other pre-training dataset contrasts and interactions change significance. The only comparison a non-permissive pre-training dataset wins is against MixtureVitae's 4K pretrained base, before post-training. Appendix Table 22 gives every estimate, interval and q-value on the full benchmarks, and Appendix A.3 reports both analyses.

Front-loading gains are broad, while the language gap persists. Averaged over the four pre-training datasets at 16K after SFT, enrichment is worth $+21.8$ points on code, $+15.7$ on mathematics, $+9.0$ on instruction following, and $+2.2$ on science and knowledge-intensive question answering. Instruction following improves in all sixteen post-training combinations. At 16K after SFT, the front-loading effects and the pre-training dataset contrasts stay positive under an equal-domain average and with any one domain left out (Appendix Table 21). General language understanding behaves differently. On the 16K extended checkpoints, before Tulu3, enrichment raises the language average of every pre-training dataset except Nemotron-CC-HQ, whose change cannot be told from zero, and at matched front-loading the permissive pre-training datasets remain 5.5 to 9.6 points behind ($q < 0.001$, Figure 2, right, and Appendix Table 23). Seven of its eight arms go through the reasoning SFT on OpenThoughts3 of Phase 2, all but MixtureVitae without the subset (Section 3.3), and its language comparisons cover the base checkpoints only.

3.3 Phase 2: OpenThoughts3 reasoning post-training

MV gains under the controlled reasoning SFT on OpenThoughts3. Every model undergoes SFT on the same 30K-sample OpenThoughts3 dataset, starting from its own Tulu3 SFT checkpoint (Table 2, Appendix Table 27). MV's suite average rises from 23.4 to 34.1, a gain of $+10.7$ points $[+8.2, +13.2]$, while the original reference mixtures change by -4.0 to -1.3 points, every interval below zero. MV's MATH500 gain holds without the identified items (Section 2.4).

Front-loading reasoning data into pre-training strongly and broadly improves reasoning post-training. Starting from the factorial's Tulu3 SFT checkpoints, reasoning SFT on OpenThoughts3 raises the suite average of all three enriched reference mixtures, by $+13.3$ to $+15.8$ points with every interval above zero (Table 2). For each pre-training dataset, front-loading adds $+14.6$ to $+19.8$ points to the gain, every interval above zero (Appendix E.5). Front-loading instruction and reasoning data enables large gains from reasoning SFT across all three reference pre-training datasets. Under the same SFT recipe, their front-loaded models improve on the reasoning suite average, whereas their models pretrained without front-loading decline.

The original reference mixtures' declines coincide with very long outputs and frequent missing final answers: after reasoning SFT, 53.2 to 100.0% of their answers are near the generation limit on every task but GSM8K, against at most 12.2% after SFT (Appendix Table 38). In the audited OT3 sample, 62.4% of training targets are cut at 16,384 tokens without a supervised end-of-turn token, which may contribute to this noncompletion. Under AceReason [46] and Nemotron-Cascade [47], whose targets end in 95.5% and 96.8% of samples (Appendix Table 7), the original reference mixtures still gain at most 1.6 MATH500 points (Table 4). Limited MATH500 improvement therefore also occurs with training data with substantially less truncation, and the contribution of truncation to the OT3-specific declines remains unisolated. These declines establish failure under the tested recipe without separating a deficit in mathematical competence from termination or answer-format failures.

Phase 2 recovers what Phase 1 erodes, and also works without Tulu3. The same reasoning SFT procedure also recovers the long-context mathematics performance that Tulu3 erodes in Section 3.1, so that erosion is not a limit of the base, although the recovery does not identify its cause. Applying OpenThoughts3 directly to MV's 16K extended checkpoint, with no Tulu3 stage, reaches 35.6 on the suite average and 69.8 on MATH500, close to the 71.0 of the main route through Tulu3 SFT, so for MV the SFT run works without Tulu3. The 16K extension itself trains on instruction data, so this route is not free of it.

Reasoning SFT leaves MV's language understanding nearly unchanged. Measured on the same questions, with the identified overlap items excluded, the reasoning SFT on OpenThoughts3 changes

Table 2: **Reasoning post-training.** Every model undergoes SFT on the same 30K-sample OpenThoughts3 dataset. Benchmark columns and the suite average are the endpoint, and Δ is the change in the suite average from each row's starting checkpoint. Evaluation uses Protocol B rather than Table 1's Protocol A. Scores are on the full benchmarks, which include items identified in the training data (Section 2.4), so they are not uniformly held-out estimates. Starting scores, SDs and the external references are in Appendix Table 26, and the interval method in Appendix B.2.

	Mathematics					Code			Science/QA		IF	Suite average	
Model	A24	A25	AMC	M500	GSM8K	HE	LCB	MBPP	GPQA	JEE	IFE	Avg	Δ [95% CI]
(a) Main comparison: Tulu3 SFT → OT3-30K													
Matched compute, 300B tokens													
MixtureVitae	11.1	16.7	46.7	71.0	49.0	51.0	11.5	28.5	28.8	18.8	42.1	34.1	$+10.7$ $[+8.2, +13.2]$
Comma0.1	0.0	0.0	0.0	1.3	13.4	1.4	0.0	0.1	26.8	2.6	36.8	7.5	-4.0 $[-5.1, -3.0]$
Nemotron-CC-HQ	0.0	0.0	1.7	1.5	6.3	0.2	0.0	0.0	24.2	3.7	34.7	6.6	-1.3 $[-2.0, -0.6]$
FineWeb-Edu	0.0	0.0	1.7	0.6	6.2	0.0	0.0	0.0	23.9	1.9	33.4	6.2	-1.3 $[-2.3, -0.5]$
Enriched, 300B tokens													
Comma0.1	15.6	23.3	50.8	74.7	49.5	62.2	13.6	30.1	27.9	25.8	44.3	38.0	$+15.8$ $[+13.2, +18.4]$
Nemotron-CC-HQ	16.7	12.2	46.7	72.1	48.2	57.7	12.7	28.6	31.3	23.0	39.4	35.3	$+14.2$ $[+11.6, +16.7]$
FineWeb-Edu	16.7	18.9	40.8	69.4	48.0	53.5	11.3	17.1	30.1	19.1	41.0	33.3	$+13.3$ $[+10.7, +16.0]$
(b) MixtureVitae starting point controls													
16K extended	8.9	25.6	42.5	69.8	58.8	54.7	13.4	28.8	30.1	18.4	40.9	35.6	$+7.6$ $[+5.2, +10.0]$
Tulu3 SFT + DPO	0.0	4.4	17.5	34.4	44.4	46.1	8.6	34.5	27.1	11.5	49.2	25.3	$+0.5$ $[-1.5, +2.7]$

its 11-task language average by $+0.20$ points from the SFT checkpoint, the main Phase 2 route (95% interval -0.21 to $+0.68$), and by $+0.23$ on all items (Appendix Tables 37 and 11).

3.4 Sensitivity to pretraining composition and data for reasoning SFT

Two controls test how sensitive MV's response to reasoning SFT is to its pretraining data and to the corpus used for reasoning SFT, without identifying a mechanism.

Excluding the OpenThoughts3 source from pretraining reduces the subsequent gain. Table 3 compares MV and OT3-excluded MV, both pretrained for 300B tokens, with the same reasoning SFT applied directly to their 16K extended checkpoints. Excluding OpenThoughts3 reduces the suite average gain by $+5.9$ points $[+2.2, +9.9]$ and the MATH500 gain by $+14.5$ $[+10.9, +18.4]$. The mathematical gain remains positive without OT3, and the contrast persists after excluding identified benchmark overlaps. This ablation removes the OT3 source while most of its mathematical questions remain elsewhere in pretraining; it does not isolate question exposure, response content or repetition. Appendix E.2 gives the detailed estimates and reruns with a separate seed for each question.

The gains are not confined to OpenThoughts3 only. We next substitute three alternative reasoning corpora for OpenThoughts3 (Table 4). The runs match 30K examples, five epochs, and optimization settings, but supervised token exposure ranges from 0.88B to 2.20B (Appendix Table 7). Gains occur again broadly on math/code/reasoning evals. On the common 287-problem MATH500 subset, MV gains $+41.9$ points from its Tulu3 SFT checkpoint under OpenThoughts3, and $+38.7$, $+38.4$ and $+26.6$ under AceReason, Nemotron-Cascade and Superior-Reasoning [48], with every interval excluding zero. Large gains thus occur with all four reasoning corpora (Appendix Table 33), while 46.0 to 66.4% of the alternative corpora's sampled questions match MV's pretraining reasoning pool (Appendix Table 32). These experiments therefore test robustness across various reasoning datasets used in post-training, while not excluding amplification effects in post-training due to repetition of reasoning data already included in pre-training. The size and profile of the gain depend on the corpus and the starting checkpoint (Appendix Table 33).

Drop on instruction following alongside reasoning gains. From the Tulu3 SFT checkpoint, the four reasoning corpora raise MV's mathematics average by $+13.5$ to $+25.4$ points and change its IFEval score by -9.9 to -12.8 (Appendix Table 34). IFEval falls significantly after thirty of the thirty-one reasoning SFT runs in that table. IFEval decrease can have multiple causes: degree of specialization on reasoning, response style and answer termination change together after post-training.

Table 3: **Effect of excluding the OpenThoughts3 (OT3) source from pretraining.** Two 1.7B models pretrained for 300B tokens, with and without the OT3 source, undergo SFT on the same 30K-sample OpenThoughts3 dataset directly from their 16K extended checkpoints, without Tulu3. Cells show the mean over three sampled evaluations, with SD below. Avg is the unweighted mean of all eleven benchmarks. Evaluation follows Protocol B.

	Mathematics					Code			Science/QA			
Stage	**A24**	**A25**	**AMC**	**M500**	**GSM8K**	**HE**	**LCB**	**MBPP**	**GPQA**	**JEE**	**IFE**	**Avg**
With OT3 source												
Before	3.3 ±3.33	8.9 ±3.85	25.8 ±12.58	46.5 ±11.29	55.0 ±1.78	43.7 ±0.93	8.9 ±0.27	31.3 ±1.47	26.9 ±0.29	10.7 ±1.58	46.6 ±0.30	28.0 ±2.73
After	8.9 ±5.09	25.6 ±1.92	42.5 ±6.61	69.8 ±1.25	58.8 ±1.52	54.7 ±3.36	13.4 ±0.87	28.8 ±1.91	30.1 ±4.98	18.4 ±0.41	40.9 ±1.14	35.6 ±1.15
Without OT3 source												
Before	1.1 ±1.92	0.0 ±0.00	12.5 ±4.33	29.8 ±4.33	55.7 ±0.57	36.4 ±1.86	6.6 ±2.60	32.3 ±1.40	26.1 ±0.77	11.0 ±1.89	49.4 ±0.46	23.7 ±0.91
After	2.2 ±1.92	3.3 ±0.00	14.2 ±3.82	38.5 ±1.10	58.5 ±0.22	39.8 ±3.92	5.0 ±1.74	33.5 ±1.33	27.9 ±1.05	10.4 ±1.42	46.1 ±0.46	25.4 ±0.50

Suite-average gain: with the OT3 source, $+7.6$, without it, $+1.7$ points.

Table 4: **Large MATH500 gains for MixtureVitae across reasoning corpora, small changes for the original reference mixtures.** Change in MATH500 accuracy when each base's Tulu3 SFT checkpoint receives 30K samples from one of four reasoning corpora, on the common 287-problem subset left after excluding the identified overlaps (Section 2.4). Brackets are 95% paired bootstrap intervals. **Bold** and *italic* mark significant increases and decreases after Benjamini–Hochberg correction over the 20 changes. Appendix Table 31 gives all 500 problems. Protocol B (Appendix Table 15).

Base	**OpenThoughts3**	**AceReason**	**Nemotron-Cascade**	**Superior-Reasoning**
MixtureVitae	$\mathbf{+41.9}$ [+37.6, +46.5]	$\mathbf{+38.7}$ [+34.3, +43.1]	$\mathbf{+38.4}$ [+34.0, +42.5]	$\mathbf{+26.6}$ [+22.5, +30.8]
Comma0.1	-0.1 [−1.4, +1.2]	$+0.5$ [−0.8, +1.7]	$+1.6$ [+0.3, +3.0]	$+0.5$ [−0.8, +1.7]
FineWeb-Edu	$+0.0$ [−0.9, +0.8]	$+1.0$ [+0.0, +2.2]	$+0.9$ [−0.1, +2.0]	$+1.3$ [+0.1, +2.6]
Nemotron-CC-HQ	$+0.5$ [−0.7, +1.5]	$+1.5$ [+0.0, +3.1]	$+0.9$ [−0.6, +2.3]	$+0.5$ [−0.7, +1.6]
SmolLM2	$+0.3$ [−2.2, +2.8]	$\mathbf{+7.2}$ [+4.3, +10.2]	$\mathbf{+7.2}$ [+4.2, +10.1]	$\mathbf{+7.5}$ [+4.8, +10.2]

Running DPO first limits the loss under OpenThoughts3 to -5.3, but that SFT run also gains only $+1.4$ in mathematics, and from DPO the other three corpora lower IFEval by -10.7 to -13.4.

4 Related work

Permissively-licensed Pretraining Datasets. CommonCorpus [49] collects $\approx$ 2T tokens of uncopyrighted, permissively licensed data, and MixtureVitae [8] showed that much of the gap to mixed-license datasets closes when a large share of reasoning and synthetic data enters pretraining.

Highly Performant Pretraining Datasets. Quality filtering of web data is key to recent improvements in pretraining datasets. FineWeb-Edu [15] introduced a filter for high educational value, delivering substantial gains on benchmarks such as ARC [39] and MMLU [36]. Nemotron-CC [14] combined classifier ensembles with synthetic rephrasing for a better balance of quality and size. On the evaluation side, `open-sci-ref` [13] established a protocol for evaluating pretraining mixtures, finding Nemotron-CC-HQ [14] to be the strongest pretraining dataset, followed by DCLM [4] and FineWeb-Edu, across multiple model sizes.

Instruction and Reasoning Data in Pretraining. Prior work establishes that post-training outcomes depend on earlier training. LIMA [50] and LIMO [51] demonstrate strong performance after limited supervised fine-tuning of existing models, motivating the importance of previously acquired knowledge. Front-Loading Reasoning [10] directly varies reasoning data during from-scratch pretraining and studies its interaction with subsequent SFT and RL. PRISM [11] examines how reasoning-rich mid-training affects retention and downstream RL, while Zhang et al. [12] use

controlled synthetic tasks to isolate how prior exposure and training difficulty shape post-training generalization. Our study extends these investigations to permissive-first sourcing. By including or excluding the same instruction and reasoning subset across four pre-training datasets, we measure the benefit of reasoning enrichment within each pre-training dataset and compare post-trained models at matched front-loading. Experiments that remove mixture components or vary the reasoning corpus test how pretraining composition affects the response to reasoning SFT. The distinction is therefore the controlled pre-training dataset comparison and its implications for permissive-first pretraining, rather than reasoning enrichment itself.

Open Post-Training Datasets and Recipes. OpenThoughts [23] provides a scalable pipeline for generating questions paired with responses, reaching strong results on AIME24/25 and GPQA-D [32]. Tulu-3 [6] is a complete post-training recipe with an SFT stage and preference tuning via DPO [22]. DeepSeek-R1 [52], DAPO [53] and s1 [54] induce reasoning through RLVR or test-time scaling. These papers primarily vary the post-training algorithm or reasoning data. In contrast, we largely hold post-training fixed and ask how pretraining data composition changes a base model's receptivity to the post-training procedure.

5 Discussion and conclusion

Front-loading reasoning/instruct data into pretraining supports strong post-training also in the setting where permissive data is utilized. After Tulu3, MixtureVitae achieves more than twice the reasoning and instruction suite average of every original matched-compute reference. The factorial extends this result beyond MixtureVitae: adding the same instruction and reasoning subset improves math/code/reasoning evals following post-training across all four pre-training datasets, and with it both permissive mixtures are competitive at 4K and lead both non-permissive mixtures at 16K.

Pretraining composition also strongly determines receptivity to further reasoning post-training. MixtureVitae and the three reasoning/instruct enriched reference mixtures show large gains following the reasoning SFT on OpenThoughts3, while skipping front-loading shows a clear disadvantage for all mixtures. This contrast persists after excluding identified benchmark overlaps and is not specific to post-training on OpenThoughts3, persisting for three additional reasoning corpora. A separate ablation excluding the OpenThoughts3 source from pretraining strongly reduces MixtureVitae's gains under the same subsequent reasoning SFT procedure, showing that its advantage in reasoning post-training depends on front-loading reasoning data into pretraining.

Pretraining mixtures should therefore be compared by both their post-trained endpoints and their response to further training. Base model rankings can switch: at 4K, MixtureVitae is behind both enriched non-permissive references before post-training but becomes competitive thereafter. While catching up or even outperforming on math/code/reasoning, front-loaded permissive mixtures still show gaps on general language understanding. Further closing the language gap under permissive sourcing and preserving instruction following through reasoning training are subject of follow-up work. The findings are optimistic for the open foundation model community - if front-loading reasoning/instruct data, permissive pre-training data can also provide a strong base model for an effective post-training pipeline, offering a path towards capable open models where data, recipes, and checkpoints can be shared with a reduced or none legal risk [2].

Limitations. Our main comparisons use 1.7B-parameter models pretrained on 300B tokens, scaling trends obtained from smaller scale experiments do not provide accurate predictions towars larger unseen scales. Intervals quantify evaluation uncertainty for fixed checkpoints and variability across training runs is unmeasured. We test Tulu3 SFT, DPO and reasoning SFT, while reinforcement learning and larger reasoning data scale budgets are left for follow-up work. Comparisons across reasoning corpora match sample counts and epochs, but differ in the number of tokens. The overlap audit covers only part of the pretraining data and does not establish which documents each run sampled. Parts of coding evals used in Phase 2 (LiveCodeBench and MBPP) are leaked in the pre-training data, which might exaggerate the absolute scores. Appendix G provides further details and discusses additional limitations, including truncated OpenThoughts3 training targets.

[2] This does not constitute legal advice, but provides outlook for the future of research based on permissive data.

Acknowledgments and Disclosure of Funding

We acknowledge co-funding by EU from Digital Europe Programme under grant no. 101195233 (openEuroLLM) and under grant no. 101198470 (LLMs4EU), funding by EU Horizon under grant no. 101214398 (ELLIOT), co-funding from EuroHPC Joint Undertaking programme under grant no. 101182737 (MINERVA), funding by the German Federal Ministry of Research, Technology and Space (BMFTR) under grant no. 01IS24085C (OPENHAFM) and under the grant 16HPC117K (MINERVA).

We gratefully acknowledge the Gauss Centre for Supercomputing e.V. for funding this work by providing computing time through the John von Neumann Institute for Computing (NIC) on the supercomputer JUWELS Booster at Jülich Supercomputing Centre (JSC), storage resources on JUST granted and operated by JSC and supported by Helmholtz Data Federation (HDF), computing time granted by the JARA and JSC on the supercomputer JURECA at JSC, through the JUREAP (JUPITER Early Access Program), and computing time provided through the Gauss AI Competition (reformo) on JUPITER via GCS and BMFTR.

LAION further acknowledges a public storage grant from HuggingFace, enabling broad community access to open-source research outputs through HuggingFace repositories. We also thank the teams operating the supercomputing facilities at JSC for their support, especially Björn Hagemeier for assistance with data storage and handling.

We also would like to express gratitude to all the people who are working on making code, models and data publicly available, advancing community based research and making research more reproducible. Specifically, we would like to thank Open-Ψ (Open-Sci) Collective[3], openEuroLLM Consortium[4] and all the members of the LAION Discord server[5] community for providing fruitful ground for scientific exchange and open-source development.

References

[1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[3] Maurice Weber, Daniel Y Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492, 2024.

[4] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Suchin Gururangan, Luke Zettlemoyer, Ali Farhadi, Vaishaal Shankar, Alexander Ratner, and Ludwig Schmidt. Datacomp-lm: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*, 2024.

[5] Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk,

[3] `https://discord.gg/GsKh4mBVcv`

[4] `https://openeurollm.eu`

[5] `https://discord.gg/BZqhreFazY`

Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[6] 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.

[7] Nikhil Kandpal, Brian Lester, Colin Raffel, Sebastian Majstorovic, Stella Biderman, Baber Abbasi, Luca Soldaini, Enrico Shippole, A. Feder Cooper, Aviya Skowron, Shayne Longpre, Lintang Sutawika, Alon Albalak, Zhenlin Xu, Guilherme Penedo, Loubna Ben allal, Elie Bakouch, John David Pressman, Honglu Fan, Dashiell Stander, Guangyu Song, Aaron Gokaslan, John Kirchenbauer, Tom Goldstein, Brian R. Bartoldson, Bhavya Kailkhura, and Tyler Murray. The common pile v0.1: An 8TB dataset of public domain and openly licensed text. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[8] Huu Nguyen, Victor May, Harsh Raj, Marianna Nezhurina, Yishan Wang, Yanqi Luo, Vu Minh Chien, Taishi Nakamura, Ken Tsui, Van Khue Nguyen, David Salinas, Aleksandra Krasnodębska, Christoph Schuhmann, Mats Leon Richter, Xuan-Son Vu, and Jenia Jitsev. Mixturevitae: Open web-scale pretraining dataset with high quality instruction and reasoning data built from permissive-first text sources. *Transactions on Machine Learning Research*, 2026. Featured Certification.

[9] PleIAs. Synth: The new data frontier, 2025.

[10] Syeda Nahida Akter, Shrimai Prabhumoye, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, Yejin Choi, and Bryan Catanzaro. Front-loading reasoning: The synergy between pretraining and post-training data. *arXiv preprint arXiv:2510.03264*, 2025.

[11] Bharat Runwal, Aakriti Agrawal, Abhishek Roy, and Priyadarshini Panda. PRISM: Demystifying retention and interaction in mid-training. In *International Conference on Machine Learning (ICML)*, 2026.

[12] Yuxiao Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and RL on reasoning language models. In *International Conference on Machine Learning (ICML)*, 2026.

[13] Marianna Nezhurina, Jörg Franke, Taishi Nakamura, Timur Carstensen, Niccolò Ajroldi, Ville Komulainen, David Salinas, and Jenia Jitsev. Open-sci-ref-0.01: Open and reproducible reference baselines for language model and dataset comparison. In *AAAI 2026 Workshop on Reproducible AI*, 2026.

[14] Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset, 2025.

[15] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2024.

[16] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model. In *Second Conference on Language Modeling*, 2025.

[17] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

[18] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[19] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, 2024.

[20] Siheng Li, Cheng Yang, Zesen Cheng, Lemao Liu, Mo Yu, Yujiu Yang, and Wai Lam. Large language models can self-improve in long-context reasoning. *arXiv preprint arXiv:2411.08147*, 2024.

[21] 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*, pages 1376–1395, 2024.

[22] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[23] Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Rea Sprague, et al. Openthoughts: Data recipes for reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[24] MAA. Aime 2024 problems, 2024. Accessed: 2025-05-11.

[25] MAA. Aime 2025 problems, 2025. Accessed: 2025-05-11.

[26] MAA. Amc 2023 problems, 2023. Accessed: 2025-05-11.

[27] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[28] 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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[29] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[30] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025.

[31] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[32] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[33] Daman Arora, Himanshu Singh, et al. Have llms advanced enough? a challenging problem solving benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7527–7543, 2023.

[34] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *CoRR*, 2023.

[35] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, September 2021.

[36] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

[37] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[38] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[39] 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. *arXiv preprint arXiv:1803.05457*, 2018.

[40] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

[41] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[42] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

[43] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, 2018.

[44] Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95, 2011.

[45] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany, August 2016. Association for Computational Linguistics.

[46] Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron 1.1: Advancing math and code reasoning through sft and rl synergy. *arXiv preprint arXiv:2506.13284*, 2025.

[47] Boxin Wang, Chankyu Lee, Nayeon Lee, Sheng-Chieh Lin, Wenliang Dai, Yang Chen, Yangyi Chen, Zhuolin Yang, Zihan Liu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nemotron-cascade: Scaling cascaded reinforcement learning for general-purpose reasoning models, 2026.

[48] Shaotian Yan, Kaiyuan Liu, Chen Shen, Bing Wang, Sinan Fan, Jun Zhang, Yue Wu, Zheng Wang, and Jieping Ye. Distribution-aligned sequence distillation for superior long-cot reasoning. *arXiv preprint arXiv:2601.09088*, 2026.

[49] Pierre-Carl Langlais, Pavel Chizhov, Catherine Arnett, Carlos Rosas Hinostroza, Mattia Nee, Eliot Krzysztof Jones, Irène Girard, David Mach, Anastasia Stasenko, and Ivan P. Yamshchikov. Common corpus: The largest collection of ethical data for LLM pre-training. In *The Fourteenth International Conference on Learning Representations*, 2026.

[50] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

[51] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. In *Second Conference on Language Modeling*, 2025.

[52] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[53] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[54] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims of the abstract and Sec. 1 map to Sec. 3.2, Sec. 3.1, and Sec. 3.4, and they include two negative results: both permissive pre-training datasets trail on general language understanding, and reasoning fine-tuning reduces instruction following. Scope is bounded in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Sec. 5, which discusses the scales tested, uncertainty estimates that cover fixed checkpoints but not training runs, the post-training recipes tested, and a factorial that is carried through reasoning SFT only in part: seven of its eight arms, with OpenThoughts3 only. Appendix G discusses them in full.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper contains no theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 2 gives the factorial design, the pre-training configuration, the post-training recipes, and both evaluation protocols. App. A gives the hyperparameters, App. B.1 the protocol tables, and App. A.3 the benchmark-overlap filtering, audit and robustness analyses. The checkpoints, the evaluation records and the code that regenerates every table and figure from them are public, and the Reproducibility Statement at the start of the appendix links them.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code that regenerates every table, figure and result number from the evaluation records, the records themselves, and the checkpoints are public, and the code's README gives the commands that rebuild the results. The Reproducibility Statement at the start of the appendix links all of them. Sec. 2 and App. A give the training settings.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 2 gives the factorial design, the base models and the post-training pipeline. App. A.2 gives the `open-sci-ref` architecture and training schedules, App. B.1 gives both evaluation protocols in full, and App. A.3 gives the decontamination procedure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: All Protocol A results report a mean and a sample standard deviation over three decoding seeds (42, 1234, 2025), and for the contrasts on the suite average we report pointwise 95% paired item-level bootstrap confidence intervals and adjust two-sided p-values within each stated comparison family using Benjamini–Hochberg (App. B.2 gives the replicate counts). The factors of variability, the resampling scheme, and the small-sample caveat for the contest mathematics benchmarks are stated in Sec. 2.3 and App. B.2. Protocol B results report a mean and a sample standard deviation over three sampled runs, and their changes carry paired item-level intervals, in which the code tasks enter through their aggregate scores (App. E.1). These intervals describe evaluation variability conditional on fixed checkpoints and do not estimate variability across independent pre-training runs, which App. B.2 states.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: App. A reports the hardware and single-run GPU-hours for reasoning post-training (Tab. 14) and long-context extension (App. A.4). Pre-training follows the `open-sci-ref` 1.7B configuration, whose architecture, per-token FLOPs, and token schedules are in Tab. 5 and Tab. 6. We do not report per-run pre-training GPU-hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The research conducted in this paper conforms with the NeurIPS Code of Ethics in every respect.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See the broader impact statement in Appendix G.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released artifacts are 1.7B research checkpoints trained on public corpora, with no capability beyond comparable open releases, so no specific safeguard is required. The general dual-use considerations for open model releases are noted in the broader impact statement.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Licensing is the subject of the paper. Sec. 2.1 identifies which pre-training datasets are permissive and which are not, and Sec. 4 cites the originating release for every corpus. The post-training data (Tulu3, OpenThoughts3, AceReason, Nemotron-Cascade and Superior-Reasoning), the benchmark suites, and the evaluation harnesses are cited at first use and used under their published terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets are the pre-training runs of the 300B factorial and their long-context, SFT and DPO descendants, the reasoning post-trained descendants of selected checkpoints, and the artifact generation pipeline. The code's README documents how every result is rebuilt, the records dataset has a card with its contents and license, and each checkpoint's entry in the Hugging Face collections names the table row it appears in.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The work does not involve research with human subjects, so no institutional review was required.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: Large language models are the object of study rather than a component of the research method, and no LLM was used as an important or original part of the methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

Appendix

Appendix guide

Appendices A and B give the methods, including the benchmark-overlap audit, Appendices C to E the complete results and controls of the main experiments, and Appendix F the auxiliary experiments.

	Reproducibility Statement	23
A	Training data and implementation details	23
B	Evaluation and statistical methods	28
C	Phase 1: complete results	30
D	Front-loading across pre-training datasets	34
E	Reasoning post-training: controls and diagnostics	39
F	Auxiliary experiments	51
G	Limitations and broader impact	56
H	Author contributions	57

Reproducibility Statement

What is released. The code that writes every table, figure and result number of this paper from its evaluation records is at `MixtureVitae/permissive-post-train` (MIT license). Its README gives the commands that download the records, rebuild every generated file and compare each one with the version printed here. The records, including the per-question scores that the intervals resample, are the Hugging Face dataset `permissive-post-train/evaluation-records` (CC BY 4.0). The checkpoints are in four Hugging Face collections, each listing its checkpoints in the order of the table rows with a note that names the row: Phase 1 (Section 3.1), the front-loading study (Section 3.2), Phase 2 (Sections 3.3 and 3.4) and the auxiliary experiments (Appendix F).

What the release covers.

Rebuilding the paper's results. The code and the records rebuild every table, figure and result number. The Phase 2 code tasks enter through their aggregate scores, because most Phase 2 records hold no per-question verdicts for them (Appendix B.2).

Evaluating the checkpoints again. Appendix B.1 gives both protocols and the language settings, and the evaluation code is public: `Ali-Elganzory/evalchemy` for reasoning and `OpenEuroLLM/oellm-eval` for language. Every checkpoint is on Hugging Face.

Training the models again. Section 2 and Appendix A give the recipes and the configurations of pretraining, the 16K extension and reasoning post-training. The training code is public: `ontocord/mixturevitae` for pretraining, `OpenEuroLLM/post-training` for the 16K extension, `Ali-Elganzory/alignment-handbook` for the Tulu3 stages, with their run settings, and `open-thoughts/open-thoughts` for reasoning post-training, whose 30K samples and configuration are in `marianna13/open-thoughts3-30k`. Appendix A.1 gives the share of each enriched run that the instruction and reasoning subset takes.

A Training data and implementation details

A.1 Pretraining mixtures and the instruction and reasoning subset

Every base of the main comparisons is a 1.7B model pretrained on 300B tokens under the `open-sci-ref` recipe (Section 2), so the bases differ only in their pretraining data. The subset that the front-loading study adds to or removes from a mixture is the instruction and reasoning data of MixtureVitae, 175B tokens. In the files of the public release (`ontocord/MixtureVitae-v1-decontaminated`), it holds mathematics (`dart-math`, `math`, `math_word_problems`, `prism_math`, `nemo_science_math`), code (`ling-coder-sft`, `ling-coder-dpo`, `ling-coder-syntheticqa`), reasoning traces (`glaiveai-reasoning`, `open_thoughts`, `os-q2`, `ring-lite-sft`) and other instruction data (`few-shot`, `formatted_text`, `misc_instruct`, `pyedu_reasoning`). Pretraining is a single stage with no curriculum: every example is sampled at random from the whole mixture, so the subset is

spread over the whole run. The 175B tokens are the size of the subset, not the share of a 300B-token run that it takes. The enriched Comma0.1, Nemotron-CC-HQ and FineWeb-Edu runs give it the same share, 42.1% of their tokens (126B, 0.72 passes over the subset), its share of MixtureVitae as computed from the 175B tokens. MixtureVitae's own copy of these files is larger, mainly in `formatted_text`, and takes 45.0% of MixtureVitae. MixtureVitae without the subset trains on a pool of 237.5B tokens, so about 21% of its 300B tokens repeat earlier data.

A.2 Model and training configurations

Table 5 gives the two model scales and Table 6 the pretraining schedules of the 100B and 300B runs. Pretraining is a single stage. Embeddings are tied, and FLOPs follow the $6N$ rule with the embeddings included in N, as `open-sci-ref` [13] computes them. The learning rate rises linearly to its peak over the warmup steps and decays to zero over the cooldown, the last 20% of the steps.

Table 5: Architecture and scales of the `open-sci-ref` [13] models. We used tied embedding weights in all experiments.

Parameters (B) (Non-Emb + Emb)	Layers	Hidden	Heads	FFN Hidden	Memory	FLOPs
$0.35 + 0.05 = 0.40$	22	1024	16	3840	2.88 GB	2.4×10^9
$1.61 + 0.10 = 1.71$	24	2048	32	8192	9.884 GB	1.0×10^{10}

Table 6: The training schedules used in our experiments.

Tokens	Global Batch Size (tokens)	Iterations	Learning Rate	Warmup	Cooldown (20%)
100B	4.12M	23,842	4×10^{-3}	8,334	4,768
300B	4.12M	72,661	4×10^{-3}	25,000	14,532

Long context extension To extend a base's context window from 4K to 16K tokens, we use YaRN [19], which applies frequency-dependent interpolation to rotary position embeddings, with a scaling factor of 4.0, and fine-tune all model parameters for one epoch with a learning rate of 2×10^{-4} and cosine decay. The training mixture uses a 30/70 split between long context and short context data. The long context portion combines SeaLong [20] and LongAlign [21], which provide instruction-following examples requiring reasoning over extended documents. The short context portion is SmolTalk [16], a conversational instruction dataset. The result is each base's 16K extended checkpoint.

Tulu3 SFT and DPO. Phase 1 applies the SFT and DPO stages of the Tulu3 recipe to every base, with the Tulu3 SFT mixture and preference data as filtered before training (Appendix A.3). The training code that the Reproducibility Statement links holds the recipes of both stages, and Table 8 gives their settings. They follow the Tulu3 recipe in learning rate, schedule, batch size and epochs. Three of them matter for reading the results. The 16K checkpoints are post-trained with the same maximum lengths as the 4K ones, 4,096 tokens for SFT and 2,048 for DPO, so their 16K context comes from the long context extension alone. SFT computes the loss on the whole conversation, prompt included, as TRL does for data in conversational format unless told otherwise, while the reasoning SFT of Phase 2 masks the prompt. Appendix F.4 compares SFT with the loss on the response tokens only. DPO uses $\beta = 5$ with the standard sigmoid loss, while the Tulu3 recipe pairs the same β with length normalization.

Reasoning SFT. Table 8 also gives the settings of every reasoning post-training run, whose loss covers the response tokens only. All runs share one recipe, whatever the reasoning dataset. Each trains on 30,000 samples drawn uniformly at random from one release: OpenThoughts3 (its sample is released as `marianna13/open-thoughts3-30k`), AceReason-1.1-SFT [46], the first SFT stage of Nemotron-Cascade [47] and Superior-Reasoning-SFT-gpt-oss-120b [48], the last three under CC

BY 4.0. Multi-turn samples keep all their turns. Table 7 gives each sample's domains and training tokens: the runs match examples and epochs, not supervised tokens. In the OpenThoughts3 sample, 62.4% of the training targets are cut at 16,384 tokens without a supervised end-of-turn token, and 62.7% of its answers as released never close their reasoning, before any cut.

Table 7: **The reasoning datasets: domains and training tokens.** The 30,000 samples of the MixtureVitae runs as trained. Math to Other count the samples of each domain, as the release labels them: Other is general questions for Nemotron-Cascade and instruction following for Superior-Reasoning. Tokens are counted in the tokenized data that LLaMA-Factory stored for the runs, in the tokenizer of the `open-sci-ref` models, after the chat template and the cut at 16,384 tokens. At 16,384 is the percent of samples cut at the limit, whose target ends without the end-of-turn token, and Ends the percent trained to stop. Supervised tokens count the answer tokens over the five epochs. The runs match examples and epochs, not supervised tokens. In the OpenThoughts3 sample as released, 62.7% of the answers never close their reasoning, before any cut.

	Samples				Tokens	At 16,384	Ends	Supervised
Dataset	Math	Code	Science	Other	Per sample	(%)	(%)	Tokens
OpenThoughts3	21,326	6,268	2,406	–	14,861	62.4	37.6	2.20B
Superior-Reasoning	10,386	7,742	5,659	6,213	9,763	38.8	61.3	1.42B
AceReason	20,146	9,854	–	–	7,517	4.5	95.5	1.10B
Nemotron-Cascade	14,650	7,205	1,652	6,493	6,044	3.2	96.8	0.88B

Table 8: Settings of the three post-training stages. Each reasoning SFT run trains on 30,000 samples of OT3, AceReason, Nemotron-Cascade or Superior-Reasoning, with the same settings. The 16K checkpoints are post-trained with the same maximum lengths as the 4K ones, so their 16K context comes from the long context extension alone. DPO uses the standard sigmoid loss, without the length normalization of the Tulu3 recipe. The reasoning SFT also uses the Liger kernel and groups samples by length, with FlashAttention-2 for the Llama and Qwen architectures and SDPA for the `open-sci-ref` architecture. A dash marks a setting of DPO only. The Tulu3 data are `ali-elganzory/tulu-3-sft-mixture-decontaminated` for SFT and `ali-elganzory/llama-3.1-tulu-3-8b-preference-mixture-decontaminated` for DPO.

Setting	Tulu3 SFT	Tulu3 DPO	Reasoning SFT
Framework	TRL 0.27.1 SFTTrainer	TRL DPOTrainer	LLaMA-Factory
Training data	936,509 conversations	272,585 pairs	30,000 samples
Fine-tuning	Full	Full	Full, DeepSpeed ZeRO-3
Maximum length	4,096	2,048 (prompt 512)	16,384
Packing	No	No	No
Batch per step	128	128	128 ($1 \times 4 \times 32$ H100)
Epochs (steps)	2 (14,634)	1 (2,130)	5 (1,175)
Learning rate	5×10^{-6}	5×10^{-7}	4×10^{-5}
Schedule, warmup	Linear, 3%	Linear, 10%	Cosine, 10%
Loss	Prompt and response tokens	Sigmoid DPO, $\beta = 5$	Response tokens
Length normalization	–	No	–
Optimizer	AdamW	AdamW	Fused AdamW
Weight decay, clipping	0, 1.0	0, 1.0	0, 1.0
Precision, seed	BF16, 42	BF16, 42	BF16, 42

A.3 Benchmark overlap: filtering, audit and robustness

Filtering before training. Table 9 gives how each training dataset was filtered for benchmark overlap before training. The Tulu3 SFT and preference data were processed with the MixtureVitae 13-gram filter [8], and the context extension data were used without benchmark-overlap filtering. The historical index did not cover all evaluated benchmarks and splits. Dataset and checkpoint names that contain "decontaminated" are the artifacts' historical names, not a guarantee of benchmark disjointness.

The audit. A retrospective audit looks for the evaluated benchmark items, on the splits and items the evaluations score, in the training data that can be read (Table 9). It combines 13-gram scans, with

Table 9: **Benchmark-overlap filtering before training, and the audit.** How each training dataset was filtered for benchmark overlap before training, and what the audit read. "Not audited" means that no item-level match was looked for.

Training data	Filtered before training	Audited
16K extension data	None	Every message
Tulu3 SFT mixture, as trained	MixtureVitae 13-gram filter	Every message
Tulu3 preference data, as trained	The same filter, probably on the prompts only	Prompt and both answers
Phase 2 30K samples of the four reasoning corpora	Not recorded	As trained, after the 16K cut
OpenThoughts3 source of MixtureVitae	Not recorded	Its 65,089 questions
MixtureVitae's other reasoning data (37 shards)	Filtered, index not established	Every document
Rest of MixtureVitae pretraining (web and other text)	Filtered, index not established	Not audited
Reference pretraining datasets and external models	As released	Not audited

the normalization and rule of the MixtureVitae filter, with an exact pass that matches each question verbatim after normalizing case, Unicode form and white space, and counts the choices shown near it. An item counts as identified when a training document holds at least half of its 13-grams, holds its question verbatim with its choices, or holds a multiple-choice question of 60 characters or more without all its choices. These rules were fixed before any model's answers were examined. Table 10 gives the 1,720 items identified, of seventeen benchmarks. The audit does not establish disjointness: it misses paraphrases and translations, short items without a 13-gram can only be matched verbatim, the rest of MixtureVitae's pretraining data was not audited, and whether a 300B-token run drew a given document of the pool is not recorded.

Table 10: **Benchmark items identified in the training data.** Items is the number scored, Found the items identified anywhere, and Left those kept in the comparisons without them. The last six columns count the items found in each training dataset: the 16K extension data, the Tulu3 SFT and preference data as trained, the Phase 2 30K samples as trained, the OpenThoughts3 questions and the public samples of the three other reasoning corpora, and MixtureVitae's pretraining reasoning pool other than OpenThoughts3. An item can be found in several. No item of AIME25, HumanEval, GPQA-Diamond, IFEval or LAMBADA was identified, which means no item-level match under this procedure, not absence.

Benchmark	Items	Found	Left	16K ext.	Tulu3 SFT	Tulu3 DPO	Phase 2 30K	Reasoning questions	MV pool
CommonsenseQA	1,221	670	551	15	650	204	12	4	0
MATH500	500	213	287	2	1	0	7	1	211
LiveCodeBench	369	199	170	0	0	0	0	0	199
JEEBench	515	172	343	0	0	0	3	1	169
OpenBookQA	500	162	338	0	0	0	0	0	162
MMLU	14,042	79	13,963	9	13	3	13	0	62
COPA	100	45	55	0	2	7	0	0	39
ARC-Easy	2,376	36	2,340	1	23	11	0	0	3
AMC23	40	26	14	0	0	0	2	1	26
AIME24	30	25	5	0	0	0	0	0	25
BoolQ	3,270	21	3,249	0	15	3	0	4	0
MBPP	500	18	482	0	1	0	3	1	15
WinoGrande	1,267	17	1,250	0	14	6	0	0	0
PIQA	1,838	14	1,824	0	4	3	0	0	10
ARC-Challenge	1,172	12	1,160	0	8	3	0	0	1
HellaSwag	10,042	10	10,032	0	8	1	0	0	1
GSM8K	1,319	1	1,318	0	0	0	0	0	1
All		1,720							

Comparisons without the identified items. Every comparison of Phase 2, of Table 4, of the front-loading study and of the language suite is recomputed with the identified items left out of every compared checkpoint, one common set for all, with the intervals, seeds and corrections of the full benchmarks, whose published values the recomputation reproduces. Table 11 gives the principal comparisons, Table 12 every family, and Table 13 the comparisons whose significance changes. The principal MATH500 gains from reasoning SFT persist. Some calls of the front-loading study and of the language suite change, mostly because leaving out many of a benchmark's items widens the

intervals. The Phase 2 suite averages keep the 199 LiveCodeBench and 18 MBPP items identified, because the Phase 2 records hold no per-question verdicts for the code tasks. This is a sensitivity analysis to excluding the identified items, not an evaluation on uncontaminated items. The audit's code is in the released repository, and its item identifiers and all 691 comparisons are released with the evaluation records.

Table 11: **The principal comparisons with and without the identified items.** Changes in points with 95% intervals, each from its own analysis (Sections 3.3 and 3.4), the MATH500 change from Table 4's. Without the items, MATH500 keeps 287 of its 500 problems, and every compared checkpoint is scored on the same items. The suite averages keep the LiveCodeBench and MBPP items found, because the Phase 2 records hold no per-question verdicts for the code tasks.

Comparison	All items	Without the identified items
MV SFT→OT3-30K, MATH500	$+41.8$ $[+38.5, +45.1]$	$+41.9$ $[+37.6, +46.5]$
MV SFT→OT3-30K, suite average	$+10.7$ $[+8.2, +13.2]$	$+11.0$ $[+8.1, +13.9]$
OT3-source exclusion, difference of MATH500 gains	$+14.5$ $[+10.9, +18.4]$	$+11.5$ $[+6.3, +16.8]$
OT3-source exclusion, difference of suite average gains	$+5.9$ $[+2.2, +9.9]$	$+7.8$ $[+3.5, +12.1]$
MV SFT→OT3-30K, language average	$+0.23$ $[-0.23, +0.66]$	$+0.20$ $[-0.21, +0.68]$

Table 12: **Every comparison of the overlap analysis without the identified items.** For each family of comparisons: the number of comparisons, the largest change of an estimate in points, and how many cross the significance threshold ($q < 0.05$) in either direction. Leaving out many of a benchmark's items widens its intervals, so a call can change with little change in the estimate. Table 13 lists the calls that change, and the released audit has every comparison.

Comparisons	n	Largest change	Calls changed
Table 27, changes of the suite average	11	1.39	0
Table 27, MATH500 changes	11	2.29	1
OT3-source exclusion, gains and their difference	6	3.04	0
Table 4, MATH500 changes	20	1.93	1
Factorial, front-loading effects	24	1.84	0
Factorial, pre-training dataset contrasts	24	1.61	5
Factorial, interactions	12	1.02	1
Language, changes of the average	33	0.45	8
Language, front-loading contrasts of the average	20	0.46	1
Language, changes of single tasks	330	5.64	7
Language, front-loading contrasts of single tasks	200	5.82	11

Table 13: **Comparisons whose significance changes without the identified items.** Estimates in points with 95% intervals and Benjamini–Hochberg q-values, on all items and without the identified ones, for the families of Table 12 other than single language tasks.

	All items		Without them	
Comparison	Estimate	q	Estimate	q
MV DPO→OT3-30K, MATH500	$+4.3$ $[+1.3, +7.2]$	0.008	$+2.0$ $[-1.7, +5.9]$	0.464
FineWeb-Edu SFT→AceReason, MATH500	$+1.1$ $[+0.3, +1.9]$	0.006	$+1.0$ $[+0.0, +2.2]$	0.105
Pre-training datasets, MV − Nemotron-CC-HQ, 4K, SFT	$+0.8$ $[-0.9, +2.5]$	0.406	$+2.4$ $[+0.2, +4.6]$	0.035
Pre-training datasets, MV − FineWeb-Edu, 4K, SFT	$+0.7$ $[-0.9, +2.3]$	0.407	$+2.3$ $[+0.4, +4.2]$	0.025
Pre-training datasets, MV − FineWeb-Edu, 4K, DPO	$+1.3$ $[-0.5, +3.2]$	0.223	$+2.8$ $[+0.7, +5.2]$	0.015
Pre-training datasets, MV − Nemotron-CC-HQ, 16K, 16K base	$+2.6$ $[+0.8, +4.3]$	0.006	$+2.2$ $[-0.1, +4.5]$	0.076
Pre-training datasets, MV − FineWeb-Edu, 16K, 16K base	$+3.0$ $[+1.1, +4.9]$	0.007	$+3.1$ $[+0.0, +6.4]$	0.061
Interaction, MV − non-permissive, 4K, SFT	-2.7 $[-4.5, -1.0]$	0.009	-1.8 $[-4.4, +0.6]$	0.520
MV 16K base, from 4K base, language	-0.87 $[-1.56, -0.15]$	0.042	-0.80 $[-1.71, +0.19]$	0.140
MV OT3-30K, from 16K base, language	$+0.68$ $[+0.30, +1.10]$	0.004	$+0.75$ $[+0.18, +1.34]$	0.059
MV SFT→OT3-30K, from 16K base, language	$+0.72$ $[+0.12, +1.35]$	0.042	$+0.86$ $[+0.04, +1.63]$	0.109
MV DPO→OT3-30K, from 16K base, language	$+0.78$ $[+0.19, +1.41]$	0.042	$+0.86$ $[+0.05, +1.63]$	0.109
Comma0.1 SFT, from 16K base, language	-0.95 $[-1.58, -0.29]$	0.005	-0.85 $[-1.70, +0.04]$	0.081
Comma0.1 DPO, from 16K base, language	-0.75 $[-1.37, -0.13]$	0.018	-0.73 $[-1.60, +0.15]$	0.116
FineWeb-Edu 4K DPO, from 4K SFT, language	$+0.09$ $[+0.03, +0.15]$	0.012	$+0.10$ $[+0.02, +0.19]$	0.060
Nemotron-CC-HQ DPO, from SFT, language	$+0.17$ $[+0.04, +0.30]$	0.068	$+0.28$ $[+0.13, +0.44]$	0.020
Interaction, MV − non-permissive, 16K base, language	$+1.22$ $[+0.15, +2.33]$	0.028	$+0.76$ $[-0.78, +2.27]$	0.299

Memorization probes. Probes of the CommonsenseQA items found in the Tulu3 SFT data, in their training format, find no reproduction of the choices of a training question, a small lift of the choices'

likelihood only inside the exact training text, and no consistent accuracy advantage over questions not found in the training data, with wide intervals. They are secondary diagnostics: they do not restore held-out status to questions present in training.

A.4 Compute

Table 14 gives the GPU-hours of each reasoning post-training run, by model family and starting checkpoint, and each family's pretraining tokens. Each 16K extension, of MixtureVitae, Comma0.1, DCLM, FineWeb-Edu and Nemotron-CC-HQ, took 12 GPU-hours on 8 GH200 GPUs (2 nodes of 4), for one epoch at about 16K tokens per sequence, with YaRN scaling and SDPA attention.

Table 14: Compute of one OT3-30K SFT run, in GPU-hours on 32 H100 (8 nodes $\times$ 4 GPUs), by the checkpoint it starts from: the 16K extended checkpoint (the released base for Qwen), the Tulu3 SFT or the Tulu3 DPO checkpoint. A dash marks a run that does not exist. FA2 (Llama and Qwen) runs about twice as fast as SDPA (the `open-sci-ref` architecture). The OT3-30K run from SmolLM2's released Instruct model took 48 GPU-hours, and the 18 runs 1,417 together, not the compute of the whole project. The last two columns give each family's pretraining tokens and their ratio to MV's.

		OT3-30K SFT from			Pretraining	
Model family	**Attention**	**Base**	**SFT**	**DPO**	**Tokens (T)**	**Ratio**
MixtureVitae-1.7B	SDPA	106	106	105	0.3	1.0
Comma0.1-1.7B	SDPA	–	105	158	0.3	1.0
FineWeb-Edu-1.7B	SDPA	–	106	–	0.3	1.0
Nemotron-CC-HQ-1.7B	SDPA	–	106	106	0.3	1.0
Qwen3-1.7B	FA2	60	57	57	36	120.0
Qwen2.5-1.5B	FA2	53	53	52	18	60.0
SmolLM2-1.7B	FA2	48	44	47	11	36.7

B Evaluation and statistical methods

B.1 Evaluation protocols

Two evaluation protocols are used (Table 15), and no value produced under one is ever compared against a value produced under the other. Protocol A decodes greedily on every benchmark under one fixed generation budget, which is what makes thirteen model families and three decoding seeds comparable, and it covers Phase 1, the 300B factorial, and the mixture ablations. Protocol B samples at temperature 0.7, with three runs per checkpoint, and it covers every OpenThoughts3 experiment. Its generation limit is longer on the short tasks, GSM8K, HumanEval, MBPP and IFEval, because models that have received reasoning post-training emit long chains of thought that a short budget truncates. Under Protocol B, IFEval checks the whole response, reasoning trace included. Both run LiveCodeBench (v5) on the same problems, and they differ in one task setting, MBPP's shots. The same checkpoint therefore scores differently under the two, which is why every result in Sections 3.3 and 3.4 is reported as a paired change against a starting checkpoint re-evaluated under Protocol B. Chat benchmarks are run through Evalchemy, an open evaluation framework for post-trained models, in the version that the Reproducibility Statement links, and GSM8K through the LM Evaluation Harness `gsm8k` task. The language suite (Table 16) is scored by the LM Evaluation Harness under its own settings, on neither reasoning protocol, as accuracy in percent, normalized where the harness reports it.

B.2 Statistical methodology and uncertainty

A result can vary for three reasons: evaluating the same checkpoint again, the finite number of questions in each benchmark, and training the same recipe again. The standard deviations of the results tables measure the first, the intervals the second, and no result of this paper measures the third.

Repeated evaluation of one checkpoint. Every Protocol A result is a mean over decoding seeds 42, 1234, and 2025, with the sample standard deviation over the three. Greedy decoding leaves only the

Table 15: **The two reasoning evaluation protocols.** Protocol A, used for Phase 1, the front-loading study and the mixture ablations, decodes greedily, and its results are the mean and sample SD over decoding seeds 42, 1234 and 2025. Protocol B, used for every OpenThoughts3-30K experiment, samples at temperature 0.7, with one answer per question in each of three runs, and its results are the mean and SD over the runs. A task has the same metric under both. GSM8K is scored by exact match on the last number of the answer, the LM Evaluation Harness's flexible-extract filter. The token limits are upper bounds, lower for a model with a shorter context. The suite average is the unweighted mean of the eleven tasks, which weights each domain by its number of tasks: mathematics at 45%, code at 27%, science at 18% and instruction following at 9%.

			Max. new tokens	
Task	**Metric**	**Shots A / B**	**A**	**B**
Mathematics				
AIME24 (A24) [24]	Accuracy	0 / 0	32,768	32,000
AIME25 (A25) [25]	Accuracy	0 / 0	32,768	32,000
AMC23 (AMC) [26]	Accuracy	0 / 0	32,768	32,000
MATH500 (M500) [28]	Accuracy	0 / 0	32,768	32,000
GSM8K (GSM8K) [27]	Exact match (flexible extract)	5 / 5	2,048	16,000
Code				
HumanEval (HE) [29]	Pass@1 (`python_pass@1`)	0 / 0	2,048	16,000
LiveCodeBench (v5) (LCB) [30]	Pass@1	0 / 0	32,768	16,000
MBPP (MBPP) [31]	Functional Pass@1	3 / 0	2,048	16,000
Science				
GPQA-Diamond (GPQA) [32]	Accuracy	0 / 0	32,768	32,000
JEEBench (JEE) [33]	Accuracy (JEE score)	0 / 0	32,768	32,000
Instruction following				
IFEval (IFE) [34]	Instruction-level accuracy (strict)	0 / 0	2,048	16,000

Table 16: Standard language understanding evaluation benchmarks used in our result tables. All tasks use the standard LM Evaluation Harness [35] task definitions, with the shot counts listed below, seed for selecting prompt examples 1234 and no chat template. Normalized accuracy is the harness's `acc_norm`, which scores each answer by its log-likelihood divided by its length in characters. MMLU is the harness's overall score, the mean over all 14,042 questions of its 57 subjects.

Abbr.	**Task**	**Ref.**	**Metric**	**#Shots**
MMLU	MMLU	[36]	Accuracy	5
HellaS	HellaSwag	[37]	Normalized accuracy	10
CSQA	CommonSenseQA	[38]	Accuracy	10
ARC-C	ARC-Challenge	[39]	Normalized accuracy	10
ARC-E	ARC-Easy	[39]	Normalized accuracy	10
PIQA	PIQA	[40]	Normalized accuracy	10
BoolQ	BoolQ	[41]	Accuracy	10
WinoG	Winogrande	[42]	Accuracy	0
OBQA	OpenBookQA	[43]	Normalized accuracy	0
COPA	COPA	[44]	Accuracy	0
LAMBADA	LAMBADA	[45]	Accuracy	0

batching and kernel nondeterminism of the inference engine, so these standard deviations are small, except on the 30-problem contest sets, where one problem is worth 3.3 points. Every Protocol B result is a mean over three runs sampled at temperature 0.7, so its standard deviation measures how much the sampled answers vary. On most tasks every question of a run is sampled with that run's seed, so the answers of one run are correlated, and the standard deviation includes shifts shared by a whole run, which the intervals below, resampling questions, do not cover (Appendix E.2). For the language suite, the seeds change only the few-shot examples of the prompt.

The finite number of questions. Intervals are pointwise 95% percentile intervals from a paired item-level bootstrap with 2,000 replicates (20,000 for Tables 19 and 41). Each item is first scored as its mean over the seeds or runs. Each replicate resamples items with replacement within each benchmark, which keeps the composition of the eleven-task average, and recomputes the contrast on the resampled items. IFEval is scored by instruction, and the instructions of one prompt share one response, so its replicates resample the prompts, keep all instructions of each drawn prompt,

and recompute the share of instructions followed. Every checkpoint in a contrast answers the same items, so the interval is for the contrast itself. JEEBench awards partial credit, so its item scores lie in $[0, 1]$. Two-sided p-values come from the bootstrap distribution and are adjusted with the Benjamini–Hochberg procedure within each family that a caption names. The intervals themselves are not adjusted. The language changes resample the questions of each language task in the same way.

The code tasks. The Protocol A records hold a verdict for every code problem, so the Phase 1 and enrichment intervals resample code problems like any other item. Most Phase 2 records hold no per-question verdicts for HumanEval, LiveCodeBench and MBPP, so in every Phase 2 interval each of these tasks adds its observed contrast plus a normal term whose standard deviation is $\sum_k |w_k| \sqrt{p_k(1-p_k)/n}$ over the checkpoints k in the contrast, with p_k the checkpoint's score, n the number of problems and w_k the checkpoint's weight: $+1$ and -1 for a change, and $+1$, -1, -1 and $+1$ for a difference between gains (each model's endpoint, then its start). This is the largest standard deviation that resampling the problems could give the contrast, whatever the correlation between the checkpoints, so these intervals are conservative for the code tasks, up to the normal approximation.

Differences between training runs. Every comparison in this paper is between single trained checkpoints, so no standard deviation or interval includes the variation between independent training runs of one recipe. Estimating it would require repeating the 300B-token pretraining runs. A significant difference therefore shows that two checkpoints differ on these benchmarks, not that every training run of the two recipes would.

Reporting conventions. Scores are reported as mean±SD over repeated evaluations. A change is a checkpoint minus the checkpoint it was trained from. One arrow (→SFT) is one training stage, and a chain of arrows (→SFT→OT3-30K) counts every training stage from the first checkpoint of its block. An front-loading effect is a model with the instruction and reasoning subset minus the same pre-training dataset without it, at the same stage and context length, and a difference of gains subtracts one model's change from another's. Brackets are 95% intervals and q is the Benjamini–Hochberg q-value. A dash marks a cell without records.

B.3 Sample sizes and power

The minimum detectable effect is the smallest difference that a paired comparison of two models detects in 80% of repeated experiments, at a two-sided level of 0.05. We compute it by simulation for one outcome per item and a sign-flip test, with a correlation of 0.3 between the two models' outcomes on an item, at the references' own scores in Table 19. This describes a single run of each model, so it is an approximation for Table 19, which averages each item over three seeds and uses the bootstrap. A smaller difference can still come out significant, as AMC23 does there. It is about 26 points on AIME24 and AIME25 (n=30), 20 on AMC23 (n=40), 12 on GPQA-Diamond (n=198), 11 on HumanEval (n=164), 7 on MBPP (n=500), 7 on IFEval (n=769), 4 on MATH500 (n=500), 4 on JEEBench (n=515), 4 on GSM8K (n=1319) and 2 on LiveCodeBench (n=369). It grows with the baseline score up to 50%, so it is larger for comparisons between stronger models. A non-significant result on a small benchmark may reflect limited power and does not establish equivalence, and the suite average and the large mathematics and code benchmarks carry the statistical support for the claims of Section 3.1.

The contest mathematics benchmarks. AIME24 and AIME25 contain 30 problems each and AMC23 contains 40, so their percentage-point estimates are intrinsically imprecise and we do not base claims about effect size on them. Where we quote them, as in the pretraining ablation of Table 3, where AIME24 and AIME25 fall to 2.2 and 3.3, the claim is whether a model clears the benchmark floor at all.

C Phase 1: complete results

This appendix gives every model after the Tulu3 SFT and DPO pipeline of Section 3.1: the reasoning suite under Protocol A (Table 17), the language suite (Table 18), and the matched-compute contrasts with intervals (Table 19). The matched-compute models are pretrained on 300B tokens under the

same `open-sci-ref` recipe, so they differ only in their data. SmolLM2, Qwen2.5 and Qwen3 are released models with 37 to 120 times as many pretraining tokens (Table 14) and other architectures, so comparisons with them attribute differences to data and compute jointly. Table 17 gives both context variants of every model except Qwen2.5 and Qwen3, which are evaluated at their released context only, and Table 1 in the main text shows one per family.

C.1 Reasoning and language for all models

Table 17: Phase 1 (Tulu3 SFT and DPO) for all models at the 1.7B scale, both context variants where both were evaluated. This is the complete version of Table 1. Protocol A (Appendix Table 15). Values are mean±SD over decoding seeds 42, 1234, and 2025. The table has two panels: (a) mathematics, (b) code, science and instruction following, and the suite average.

(a) Mathematics

Model	A24	A25	AMC	M500	GSM8K
Matched compute, 300B tokens, controlled comparison					
MixtureVitae	0.0±0.00	7.8±5.09	5.0±0.00	17.8±0.72	54.1±0.00
→ SFT	0.0±0.00	0.0±0.00	12.5±0.00	29.5±0.23	49.9±0.00
→ SFT → DPO	0.0±0.00	3.3±0.00	10.8±1.44	30.5±0.12	50.2±0.00
MixtureVitae-16K	1.1±1.92	3.3±0.00	23.3±8.04	47.0±0.53	60.0±0.00
→ SFT	3.3±0.00	3.3±0.00	15.0±0.00	35.0±0.53	51.6±0.00
→ SFT → DPO	3.3±0.00	1.1±1.92	17.5±0.00	37.0±0.80	52.6±0.00
Nemotron-CC-HQ	0.0±0.00	0.0±0.00	0.0±0.00	0.2±0.00	4.1±0.00
→ SFT	0.0±0.00	0.0±0.00	5.0±0.00	3.1±0.23	6.8±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	2.5±0.00	3.5±0.42	8.0±0.00
Nemotron-CC-HQ-16K	0.0±0.00	0.0±0.00	0.0±0.00	2.8±0.35	8.2±0.00
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	2.9±0.31	8.6±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	2.5±0.00	2.5±0.12	7.7±0.00
FineWeb-Edu	0.0±0.00	0.0±0.00	0.0±0.00	0.0±0.00	3.2±0.00
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	2.1±0.12	7.1±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	2.5±0.00	1.7±0.31	6.7±0.00
FineWeb-Edu-16K	0.0±0.00	0.0±0.00	5.0±0.00	1.8±0.20	8.1±0.00
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	3.1±0.12	8.7±0.00
→ SFT → DPO	0.0±0.00	3.3±0.00	0.0±0.00	3.3±0.12	9.2±0.00
DCLM	0.0±0.00	0.0±0.00	0.0±0.00	0.2±0.00	3.0±0.00
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	0.8±0.00	5.2±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	1.7±1.44	1.1±0.23	4.6±0.00
DCLM-16K	0.0±0.00	0.0±0.00	1.7±1.44	1.5±0.12	5.8±0.00
→ SFT	0.0±0.00	0.0±0.00	0.8±1.44	1.3±0.12	7.2±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	0.0±0.00	1.4±0.00	7.1±0.00
Comma0.1	0.0±0.00	0.0±0.00	0.0±0.00	0.0±0.00	5.5±0.00
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	2.5±0.23	9.3±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	0.0±0.00	2.5±0.31	9.7±0.00
Comma0.1-16K	0.0±0.00	0.0±0.00	1.7±1.44	2.8±0.35	12.2±0.00
→ SFT	0.0±0.00	0.0±0.00	2.5±0.00	2.5±0.12	6.9±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	2.5±0.00	2.1±0.12	6.7±0.00
External open-weights references, 11–36T tokens, uncontrolled					
SmolLM2	0.0±0.00	0.0±0.00	0.0±0.00	0.0±0.00	29.3±0.00
→ SFT	0.0±0.00	0.0±0.00	2.5±2.50	6.3±0.12	33.3±0.00
→ SFT → DPO	1.1±1.92	0.0±0.00	1.7±1.44	8.3±0.46	33.0±0.00
SmolLM2-16K	0.0±0.00	0.0±0.00	0.8±1.44	11.8±0.92	31.9±0.00
→ SFT	0.0±0.00	0.0±0.00	1.7±1.44	9.6±0.20	32.8±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	3.3±1.44	9.6±0.53	33.8±0.00
Qwen2.5	0.0±0.00	0.0±0.00	0.0±0.00	0.2±0.35	61.9±0.00
→ SFT	3.3±0.00	0.0±0.00	8.3±1.44	18.8±0.69	39.0±0.00
→ SFT → DPO	2.2±1.92	0.0±0.00	10.0±2.50	21.2±1.00	40.4±0.00
Qwen3	3.3±0.00	2.2±1.92	20.8±3.82	48.0±0.35	71.6±0.00
→ SFT	0.0±0.00	0.0±0.00	16.7±2.89	28.2±0.35	69.4±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	8.3±1.44	31.2±1.06	69.8±0.00

Table 17: (Continued) Panel (b), code, science and instruction following, and the suite average.

(b) Code, science and instruction following, and the suite average

Model	HE	LCB	MBPP	GPQA	JEE	IFE	Avg
Matched compute, 300B tokens, controlled comparison							
MixtureVitae	13.6±0.93	7.7±0.16	19.5±0.12	24.4±0.77	2.9±0.51	34.7±0.13	17.0±0.38
→ SFT	44.3±0.70	1.2±0.16	39.9±0.23	20.7±0.87	3.7±0.55	48.0±0.08	22.7±0.14
→ SFT → DPO	47.4±0.35	1.0±0.16	41.2±0.20	21.2±0.00	4.5±0.34	49.8±0.08	23.6±0.17
MixtureVitae-16K	54.1±0.35	11.5±0.16	40.1±0.23	23.4±0.29	8.0±0.24	47.7±0.34	29.1±0.69
→ SFT	54.1±0.35	9.7±0.16	40.4±0.35	23.4±0.29	5.1±0.27	51.9±0.46	26.6±0.11
→ SFT → DPO	52.6±0.70	9.4±0.31	40.3±0.12	23.4±0.29	6.2±0.20	52.6±0.60	26.9±0.04
Nemotron-CC-HQ	0.0±0.00	0.0±0.00	2.4±0.00	5.7±0.29	0.0±0.00	29.2±0.08	3.8±0.02
→ SFT	4.9±0.00	0.0±0.00	5.5±0.23	20.2±0.00	2.0±0.17	38.1±0.33	7.8±0.04
→ SFT → DPO	6.1±0.00	0.0±0.00	5.1±0.12	21.2±0.00	2.7±0.66	39.3±0.40	8.0±0.07
Nemotron-CC-HQ-16K	1.8±0.00	0.0±0.00	5.7±0.31	22.7±0.51	5.6±1.04	39.0±0.13	7.8±0.07
→ SFT	5.3±0.35	0.0±0.00	5.9±0.12	22.2±0.87	1.2±0.08	42.9±0.13	8.1±0.11
→ SFT → DPO	5.3±0.35	0.0±0.00	5.3±0.12	23.4±1.17	1.7±0.37	43.3±0.42	8.3±0.06
FineWeb-Edu	0.0±0.00	0.0±0.00	0.0±0.00	25.1±0.58	0.0±0.00	34.7±0.00	5.7±0.05
→ SFT	2.4±0.00	0.0±0.00	2.3±0.23	26.9±0.58	1.6±0.46	34.1±0.00	6.9±0.08
→ SFT → DPO	2.4±0.00	0.0±0.00	1.9±0.12	25.9±0.29	1.6±0.17	33.9±0.45	7.0±0.06
FineWeb-Edu-16K	3.7±0.00	0.0±0.00	1.7±0.23	25.4±0.29	7.9±1.02	38.1±0.65	8.3±0.10
→ SFT	4.9±0.00	0.0±0.00	2.5±0.23	23.6±0.58	2.4±0.40	39.2±0.27	7.7±0.07
→ SFT → DPO	4.9±0.00	0.1±0.16	2.5±0.12	23.9±0.29	2.7±0.54	40.7±0.53	8.2±0.07
DCLM	0.0±0.00	0.0±0.00	0.0±0.00	19.5±0.77	0.0±0.00	37.0±0.08	5.4±0.08
→ SFT	6.7±0.00	0.0±0.00	5.4±0.35	21.7±0.00	1.1±0.49	39.5±0.23	7.3±0.02
→ SFT → DPO	6.7±0.00	0.0±0.00	5.6±0.00	20.7±0.00	0.5±0.03	41.3±0.08	7.5±0.15
DCLM-16K	0.4±0.35	0.3±0.00	4.4±0.00	23.7±0.00	5.0±0.16	39.2±0.15	7.5±0.13
→ SFT	6.3±0.35	0.0±0.00	4.9±0.12	25.4±0.29	1.1±0.12	41.9±0.53	8.1±0.19
→ SFT → DPO	6.1±0.00	0.0±0.00	4.9±0.12	25.8±0.51	1.1±0.56	44.1±0.45	8.2±0.08
Comma0.1	7.9±0.00	0.0±0.00	8.8±0.20	20.2±1.34	0.0±0.00	38.4±0.00	7.3±0.12
→ SFT	15.2±0.00	0.2±0.16	22.5±0.46	25.6±0.29	3.9±0.25	42.4±0.08	11.1±0.08
→ SFT → DPO	15.2±0.00	0.6±0.16	22.7±0.12	26.8±1.34	3.9±0.49	44.7±0.53	11.5±0.07
Comma0.1-16K	13.4±0.00	0.0±0.00	16.0±0.20	25.8±1.01	7.8±0.19	41.9±0.64	11.1±0.22
→ SFT	16.5±0.00	0.2±0.31	25.2±0.35	23.2±0.00	4.8±0.66	47.2±0.08	11.7±0.04
→ SFT → DPO	17.1±0.00	0.0±0.00	25.9±0.31	23.4±0.29	5.1±0.12	48.8±0.68	12.0±0.04
External open-weights references, 11–36T tokens, uncontrolled							
SmolLM2	0.0±0.00	0.0±0.00	1.9±0.12	15.7±1.01	0.0±0.00	31.5±0.23	7.1±0.08
→ SFT	30.9±0.70	1.6±0.00	37.2±0.20	24.7±1.01	3.7±0.37	49.5±0.40	17.2±0.34
→ SFT → DPO	30.9±1.76	1.7±0.16	37.7±0.42	25.8±0.87	4.4±0.30	50.4±0.46	17.7±0.19
SmolLM2-16K	24.0±0.35	0.1±0.16	36.5±0.12	24.6±0.29	7.6±0.39	47.4±0.15	16.8±0.06
→ SFT	32.9±0.61	3.1±0.31	37.5±0.23	24.9±1.91	4.7±0.72	52.9±0.39	18.2±0.03
→ SFT → DPO	32.3±1.06	3.2±0.16	36.9±0.50	26.3±0.00	5.4±0.69	54.9±0.08	18.7±0.18
Qwen2.5	5.7±0.70	0.0±0.00	0.1±0.12	11.1±1.82	0.3±0.10	29.9±0.34	9.9±0.15
→ SFT	51.0±0.35	3.6±0.16	44.3±0.23	29.0±0.77	6.1±0.54	54.0±0.33	23.4±0.08
→ SFT → DPO	55.7±0.70	3.3±0.00	44.5±0.23	27.9±0.58	8.0±0.64	58.6±0.67	24.7±0.44
Qwen3	64.2±0.93	1.8±1.03	46.5±0.81	25.1±0.77	18.0±0.76	36.4±0.26	30.7±0.28
→ SFT	67.1±0.61	4.4±0.56	54.3±0.58	33.0±1.05	11.9±0.54	61.4±0.26	31.5±0.33
→ SFT → DPO	67.7±0.61	4.1±0.47	54.7±0.23	31.3±0.87	13.3±1.25	64.9±1.47	31.4±0.23

On the reasoning suite, MV-16K leads every matched-compute reference at every stage. After DPO it is ahead of SmolLM2-16K and Qwen2.5 and below Qwen3, which has 120 times its pretraining tokens. On the language suite, Phase 1 barely moves any model: MV stays at 55.7 to 55.8 mean accuracy across its three stages and MV-16K at 55.0 to 55.5, above Comma0.1 (50.0 to 51.4) and below FineWeb-Edu (56.7 to 57.5), DCLM-16K (58.5 to 58.8), Nemotron-CC-HQ (64.2 to 64.5, and 64.1 to 64.5 at 16K) and the external references. MV is ahead of FineWeb-Edu and DCLM-16K on BoolQ, CommonsenseQA and MMLU and behind them and Nemotron-CC-HQ on HellaSwag, COPA and LAMBADA.

Table 18: Full language suite results for the main 1.7B comparison. Phase 1 Tulu3 SFT/DPO leaves language scores largely unchanged, with MixtureVitae matching FineWeb-Edu, trailing Nemotron, and remaining below the more heavily pretrained external baselines. Each value is one run, printed with one decimal, and Avg is the mean of the eleven tasks.

Model	ARC-C	ARC-E	BoolQ	CSQA	COPA	HellaS	LAMBADA	MMLU	OBQA	PIQA	WinoG	Avg
Comma0.1	36.2	63.1	63.2	20.4	70.0	53.5	54.5	26.5	32.6	71.2	59.3	50.0
→SFT	36.5	64.6	64.1	20.3	74.0	54.0	57.0	27.2	33.2	71.9	60.9	51.3
→SFT →DPO	36.7	64.7	64.7	20.5	74.0	54.2	57.3	27.2	33.2	71.9	60.9	51.4
MV	40.3	70.9	73.5	51.4	71.0	54.5	48.1	41.9	34.2	70.4	57.8	55.8
→SFT	38.8	70.2	73.2	49.6	70.0	55.7	49.6	42.7	34.4	70.0	58.2	55.7
→SFT →DPO	38.8	70.2	73.2	50.0	71.0	55.8	49.8	42.7	34.4	70.0	58.2	55.8
MV-16k	41.6	71.4	72.6	48.8	67.0	53.1	49.4	41.1	31.8	69.7	57.9	55.0
→SFT	39.2	70.2	72.9	50.4	69.0	54.9	50.0	41.9	34.2	69.9	57.5	55.5
→SFT →DPO	39.2	70.5	72.9	50.3	69.0	55.0	50.2	42.0	34.4	69.8	57.4	55.5
Qwen2.5-1.5B	53.8	80.5	78.4	75.8	83.0	68.0	63.0	61.0	40.8	77.1	63.4	67.7
→SFT	51.2	79.0	69.0	76.6	82.0	68.0	60.7	60.5	39.4	77.0	62.4	66.0
→SFT →DPO	51.6	79.1	69.3	76.6	82.0	68.4	60.6	60.6	39.2	77.0	62.4	66.1
Qwen3-1.7B	54.0	82.2	82.6	75.8	82.0	67.1	63.9	62.7	39.0	76.6	64.4	68.2
→SFT	53.0	80.7	82.6	76.5	80.0	67.7	61.9	62.9	38.6	76.1	63.9	67.6
→SFT →DPO	53.5	80.9	82.7	76.4	81.0	68.3	61.9	63.0	39.0	76.3	63.6	67.9
SmolLM2-1.7B	53.6	79.6	74.3	60.5	83.0	73.1	67.7	50.4	44.6	78.9	66.1	66.5
→SFT	52.4	80.5	74.9	60.2	85.0	73.8	69.7	49.2	43.8	78.7	67.1	66.8
→SFT →DPO	53.0	80.7	74.8	60.5	85.0	74.1	69.8	49.3	43.6	78.7	67.0	67.0
FineWeb-Edu	48.5	76.9	64.9	21.1	82.0	67.1	55.7	26.1	42.4	77.1	62.4	56.7
→SFT	48.1	78.5	65.9	19.7	83.0	67.6	59.2	27.0	41.8	77.0	63.2	57.4
→SFT →DPO	48.4	78.5	65.9	19.7	83.0	67.8	59.3	26.9	42.0	77.1	63.4	57.5
Nemotron	52.4	80.9	77.4	48.6	85.0	71.5	58.0	46.4	42.0	78.7	65.0	64.2
→SFT	52.3	80.8	78.3	47.8	84.0	72.6	61.3	45.1	41.8	79.1	65.4	64.4
→SFT →DPO	52.5	80.9	78.4	47.8	84.0	72.8	61.4	45.1	41.8	79.1	65.4	64.5
Nemotron-16k	52.0	80.7	77.1	50.6	83.0	70.0	60.1	46.5	41.8	78.1	64.6	64.1
→SFT	52.0	80.0	78.1	52.2	83.0	71.9	60.2	46.8	41.4	78.5	64.2	64.4
→SFT →DPO	52.0	80.3	78.0	52.7	83.0	72.3	60.5	46.7	41.4	78.2	64.8	64.5
DCLM-16k	45.2	75.3	72.3	20.6	88.0	68.7	68.9	23.4	39.8	78.1	66.1	58.8
→SFT	41.6	73.4	70.5	20.1	89.0	71.1	69.5	24.8	39.0	78.5	66.2	58.5
→SFT →DPO	41.4	73.3	70.3	20.6	88.0	71.2	69.7	25.0	39.2	78.6	66.1	58.5

C.2 Matched-compute contrasts after Phase 1

Table 19 tests MV-16K against the two closest matched-compute references, Comma0.1-16K and Nemotron-CC-HQ-16K, benchmark by benchmark after Tulu3 SFT and on the suite average after SFT and after DPO. On the suite average MV is ahead by $+14.9$ [$+13.2$, $+16.7$] and $+18.5$ [$+16.7$, $+20.4$] points after SFT, and with the four domains weighted equally, or with any one of them left out, by $+11.2$ to $+22.1$ points. The lead comes from mathematics and code: MATH500, GSM8K, HumanEval, MBPP and LiveCodeBench differ significantly against both references, while GPQA-Diamond does not differ, and on AIME24 and AIME25 the models are one problem apart. These contrasts compare single checkpoints, so they do not include the variation between pretraining runs (Appendix B.2).

Table 19: Phase 1 matched-compute contrasts under Protocol A at 16K context. Differences are MV minus the reference, in points, so positive values favour MV. Each item is scored as its mean over the three decoding seeds, and the 95% intervals come from the paired item-level bootstrap of Appendix B.2, with items resampled within each benchmark and IFEval's prompts with all their instructions. This table uses 20,000 replicates rather than 2,000, so that its q-values are not limited by the smallest p-value the bootstrap can give. q is the Benjamini–Hochberg adjusted q-value, computed within each column for the per-task rows, over the four suite average rows, and over the ten reweighted rows. The reweighted rows recompute the suite average after SFT with the four domains of Table 15 weighted equally, or with one domain left out, as Table 21 does for the front-loading study. n is the number of scored items: HumanEval counts its 164 Python problems and IFEval the 769 instructions of its 500 prompts. A dash marks a benchmark on which both models score zero on every item, which is not tested.

	MV − Comma0.1			MV − Nemotron-CC-HQ		
Task	Δ pp	95% CI	q	Δ pp	95% CI	q
Per task, after Tulu3 SFT						
AIME24 (n=30)	+3.3	[+0.0, +10.0]	0.793	+3.3	[+0.0, +10.0]	0.792
AIME25 (n=30)	+3.3	[+0.0, +10.0]	0.793	+3.3	[+0.0, +10.0]	0.792
AMC23 (n=40)	+12.5	[+2.5, +22.5]	0.017	+15.0	[+5.0, +27.5]	0.003
MATH500 (n=500)	+32.5	[+28.3, +36.7]	< 0.001	+32.1	[+28.0, +36.1]	< 0.001
GSM8K (n=1319)	+44.7	[+41.9, +47.5]	< 0.001	+43.0	[+40.1, +45.9]	< 0.001
HumanEval (n=164)	+37.6	[+29.9, +45.3]	< 0.001	+48.8	[+41.1, +56.7]	< 0.001
LiveCodeBench (n=369)	+9.5	[+6.6, +12.6]	< 0.001	+9.7	[+6.8, +12.7]	< 0.001
MBPP (n=500)	+15.2	[+11.0, +19.5]	< 0.001	+34.5	[+30.2, +38.9]	< 0.001
GPQA-Diamond (n=198)	+0.2	[−7.9, +8.2]	0.971	+1.2	[−7.2, +9.6]	0.792
JEEBench (n=515)	+0.3	[−1.3, +2.0]	0.793	+3.9	[+2.4, +5.6]	< 0.001
IFEval (n=769)	+4.7	[+0.7, +8.6]	0.033	+9.0	[+5.3, +12.8]	< 0.001
Eleven-task suite average						
Tulu3 SFT, 16K	+14.9	[+13.2, +16.7]	< 0.001	+18.5	[+16.7, +20.4]	< 0.001
SFT→DPO, 16K	+15.0	[+13.3, +16.7]	< 0.001	+18.6	[+16.8, +20.4]	< 0.001
Suite average reweighted, after Tulu3 SFT						
Equal-domain average	+11.2	[+9.5, +13.0]	< 0.001	+15.5	[+13.7, +17.3]	< 0.001
Drop math domain	+11.2	[+9.1, +13.4]	< 0.001	+17.9	[+15.7, +20.0]	< 0.001
Drop code domain	+12.7	[+10.6, +15.0]	< 0.001	+13.9	[+11.7, +16.2]	< 0.001
Drop science domain	+18.2	[+16.2, +20.2]	< 0.001	+22.1	[+20.2, +24.2]	< 0.001
Drop instruction domain	+15.9	[+14.1, +17.9]	< 0.001	+19.5	[+17.6, +21.5]	< 0.001

D Front-loading across pre-training datasets

The front-loading study of Section 3.2 pretrains each of four pre-training datasets, the web part of MixtureVitae, Comma0.1, Nemotron-CC-HQ and FineWeb-Edu, with and without the instruction and reasoning subset of Appendix A.1, and passes every base through the same 16K extension and Phase 1. Comma0.1 is the second permissive pre-training dataset. Four cells are models of Table 1: MixtureVitae with enrichment, which is the main MV model, and Comma0.1, Nemotron-CC-HQ and FineWeb-Edu without it. Every reasoning result of this appendix is under Protocol A, as the mean and sample SD over decoding seeds 42, 1234 and 2025. Table 20 summarizes the study: the reasoning suite average of every model at every stage, and the language average of its 4K pretrained base and 16K extended checkpoint. Figure 2 in the main text shows its SFT columns.

Robustness to how the suite is weighted. The eleven-task average weights mathematics at 45% and code at 27%. Table 21 recomputes the enrichment effects and the pre-training dataset contrasts at 16K after SFT under an equal-domain average and with each domain left out in turn. Under every view, each front-loading effect and each pre-training dataset contrast stays positive, although their sizes change.

D.1 Contrasts

Table 22 gives every contrast of the study with its interval, in three families corrected separately: the front-loading effects, the pre-training dataset contrasts at matched enrichment, and the interaction. The intervals come from the paired item-level bootstrap of Appendix B.2 with 2,000 replicates, with

Table 20: **Front-loading across pre-training datasets at 300B pretraining tokens.** Each row is one 1.7B model trained under the identical `open-sci-ref` configuration, differing only in the pre-training dataset and in whether the MixtureVitae instruction and reasoning subset is included. Base, SFT and DPO: the eleven-task reasoning suite average under Protocol A (Appendix Table 15), mean±SD over decoding seeds 42, 1234 and 2025. Language: the eleven-task understanding suite on the base checkpoint of the block, one seed. Tables 22 and 23 give the front-loading effects with 95% intervals.

Pre-training dataset	Enriched	Base	SFT	DPO	Language
4K context					
MixtureVitae *(permissive)*	No	7.3±0.03	10.9±0.02	11.2±0.08	51.0
	Yes	17.0±0.38	22.7±0.14	23.6±0.17	55.8
Comma0.1 *(permissive)*	No	7.3±0.12	11.1±0.08	11.5±0.07	50.0
	Yes	27.4±0.14	25.4±0.27	26.7±0.30	56.6
Nemotron-CC-HQ	No	3.8±0.02	7.8±0.04	8.0±0.07	64.2
	Yes	22.1±0.22	21.9±0.13	22.8±0.25	64.6
FineWeb-Edu	No	5.7±0.05	6.9±0.08	7.0±0.06	56.7
	Yes	24.4±0.05	22.0±0.14	22.4±0.08	62.8
16K context					
MixtureVitae *(permissive)*	No	11.3±0.03	11.5±0.02	11.7±0.13	50.8
	Yes	29.1±0.69	26.6±0.11	26.9±0.04	55.0
Comma0.1 *(permissive)*	No	11.1±0.22	11.7±0.04	12.0±0.04	52.1
	Yes	27.5±0.34	25.3±0.13	25.4±0.27	57.0
Nemotron-CC-HQ	No	7.8±0.07	8.1±0.11	8.3±0.06	64.1
	Yes	26.5±0.39	22.1±0.27	22.4±0.26	64.6
FineWeb-Edu	No	8.3±0.10	7.7±0.07	8.2±0.07	57.3
	Yes	26.0±0.06	22.2±0.25	21.8±0.07	62.6

Table 21: The front-loading effects and pre-training dataset contrasts at 16K after Tulu3 SFT, under other weightings of the suite average, whose eleven tasks weight math at 45% and code at 27%. Under every view each of them stays positive. Each cell is a point estimate, the difference of two scores averaged over the three decoding seeds.

	Front-loading effect				Permissive − non-permissive			
Averaging view	MV	Comma	Nemotron	FineWeb	MV vs. Nemotron	MV vs. FineWeb	Comma vs. Nemotron	Comma vs. FineWeb
Suite average (11 tasks)	+15.1	+13.5	+14.0	+14.6	+4.6	+4.4	+3.2	+3.0
Equal-domain average	+11.6	+10.9	+12.8	+13.4	+2.6	+3.0	+2.3	+2.7
Drop math domain	+10.8	+11.5	+15.1	+15.2	+2.7	+3.5	+3.0	+3.7
Drop code domain	+13.5	+11.2	+10.3	+11.0	+3.6	+3.0	+2.1	+1.5
Drop science domain	+18.4	+16.2	+16.1	+17.1	+5.9	+5.8	+4.0	+3.8
Drop instruction domain	+15.9	+14.3	+14.3	+14.8	+5.2	+4.8	+3.5	+3.1

items resampled within each benchmark and IFEval's prompts with all their instructions, and the q-values are Benjamini–Hochberg adjusted within each family. Comma0.1's contrasts are in the pre-training dataset and interaction families. Except for MixtureVitae at 4K, the front-loading effect is largest at the base and shrinks under Phase 1 without disappearing.

It does not compare the two permissive pre-training datasets with each other. With front-loading, Comma0.1 is ahead of MixtureVitae at every 4K stage, where MV minus Comma0.1 is -10.4 to -2.7 points ($q \leq 0.003$). At 16K the two cannot be told apart once these six contrasts are corrected as a family of their own, at the base ($+1.52$ points, $q = 0.111$) or after SFT and DPO ($+1.35$ and $+1.53$ points, $q = 0.123$ and $q = 0.104$).

The language contrasts. Table 23 gives the same three families on the 11-task language average. Enrichment raises the language average of MixtureVitae, Comma0.1 and FineWeb-Edu at both context lengths, while the change for Nemotron-CC-HQ, $+0.38$ points at 4K and $+0.51$ at 16K, cannot be told from zero ($q = 0.377$ and $q = 0.258$). At matched front-loading, the permissive

pre-training datasets trail the non-permissive ones in all eight comparisons, by -9.62 to -5.54 points at 16K (each $q < 0.001$).

Table 22: Contrasts of the front-loading study on the reasoning suite average, in points, with 95% intervals and Benjamini–Hochberg q-values adjusted within each of the three families. Positive values favour enrichment in the first block and the permissive pre-training dataset in the second and third.

		4K context			16K context		
Contrast	**Stage**	**Estimate**	**95% CI**	q	**Estimate**	**95% CI**	q
Front-loading effect, within pre-training dataset							
MixtureVitae	Base	$+9.78$	$[+8.40, +11.14]$	< 0.001	$+17.73$	$[+16.08, +19.45]$	< 0.001
	SFT	$+11.84$	$[+10.17, +13.41]$	< 0.001	$+15.09$	$[+13.31, +16.95]$	< 0.001
	DPO	$+12.38$	$[+10.68, +14.15]$	< 0.001	$+15.23$	$[+13.52, +17.01]$	< 0.001
Comma0.1	Base	$+20.08$	$[+18.22, +22.03]$	< 0.001	$+16.48$	$[+14.81, +18.30]$	< 0.001
	SFT	$+14.29$	$[+12.88, +15.79]$	< 0.001	$+13.55$	$[+11.98, +15.11]$	< 0.001
	DPO	$+15.20$	$[+13.61, +16.92]$	< 0.001	$+13.42$	$[+11.87, +14.95]$	< 0.001
Nemotron-CC-HQ	Base	$+18.28$	$[+16.55, +20.05]$	< 0.001	$+18.68$	$[+17.09, +20.35]$	< 0.001
	SFT	$+14.08$	$[+12.61, +15.51]$	< 0.001	$+13.98$	$[+12.67, +15.28]$	< 0.001
	DPO	$+14.76$	$[+13.37, +16.22]$	< 0.001	$+14.05$	$[+12.69, +15.43]$	< 0.001
FineWeb-Edu	Base	$+18.67$	$[+16.89, +20.62]$	< 0.001	$+17.71$	$[+15.91, +19.50]$	< 0.001
	SFT	$+15.04$	$[+13.67, +16.49]$	< 0.001	$+14.58$	$[+13.21, +16.07]$	< 0.001
	DPO	$+15.38$	$[+13.91, +16.81]$	< 0.001	$+13.58$	$[+12.11, +15.02]$	< 0.001
Permissive minus non-permissive pre-training dataset, both enriched							
MV − Nemotron	Base	-5.02	$[-6.94, -3.06]$	< 0.001	$+2.58$	$[+0.83, +4.29]$	0.006
	SFT	$+0.84$	$[-0.93, +2.54]$	0.406	$+4.55$	$[+2.83, +6.39]$	< 0.001
	DPO	$+0.83$	$[-0.82, +2.56]$	0.350	$+4.53$	$[+2.55, +6.34]$	< 0.001
MV − FineWeb	Base	-7.35	$[-9.29, -5.49]$	< 0.001	$+3.01$	$[+1.09, +4.90]$	0.007
	SFT	$+0.71$	$[-0.93, +2.28]$	0.407	$+4.38$	$[+2.59, +6.16]$	< 0.001
	DPO	$+1.26$	$[-0.51, +3.15]$	0.223	$+5.11$	$[+3.40, +6.84]$	< 0.001
Comma − Nemotron	Base	$+5.36$	$[+3.59, +7.22]$	< 0.001	$+1.05$	$[-0.73, +2.95]$	0.284
	SFT	$+3.50$	$[+1.80, +5.30]$	< 0.001	$+3.20$	$[+1.74, +4.60]$	< 0.001
	DPO	$+3.89$	$[+2.26, +5.49]$	< 0.001	$+3.00$	$[+1.46, +4.43]$	< 0.001
Comma − FineWeb	Base	$+3.03$	$[+1.34, +4.80]$	0.002	$+1.49$	$[-0.16, +3.20]$	0.102
	SFT	$+3.37$	$[+1.68, +5.04]$	< 0.001	$+3.02$	$[+1.56, +4.49]$	< 0.001
	DPO	$+4.32$	$[+2.67, +6.08]$	< 0.001	$+3.58$	$[+2.21, +5.02]$	< 0.001
Interaction: front-loading effect, permissive minus non-permissive							
MV − non-permissive	Base	-8.70	$[-10.54, -6.88]$	0.006	-0.47	$[-2.43, +1.38]$	0.791
	SFT	-2.72	$[-4.49, -1.03]$	0.009	$+0.81$	$[-1.13, +2.79]$	0.674
	DPO	-2.69	$[-4.48, -0.77]$	0.050	$+1.42$	$[-0.65, +3.45]$	0.339
Comma − non-permissive	Base	$+1.60$	$[-0.03, +3.33]$	0.148	-1.72	$[-3.49, +0.08]$	0.148
	SFT	-0.26	$[-1.93, +1.45]$	0.836	-0.73	$[-2.35, +0.86]$	0.613
	DPO	$+0.13$	$[-1.53, +1.88]$	0.860	-0.39	$[-2.16, +1.24]$	0.791

Table 23: The contrasts of Table 22 on the 11-task language average of the sixteen checkpoints of Table 25, the 4K pretrained bases and their 16K extended checkpoints, before Tulu3. The intervals resample the questions within each task. Averaging each question over the three seeds for selecting prompt examples moves no estimate by more than 0.38 points and no contrast across $q = 0.05$.

	4K			16K		
Contrast	**Estimate**	**95% CI**	q	**Estimate**	**95% CI**	q
Front-loading effect, within pre-training dataset						
MixtureVitae	$+4.78$	$[+3.94, +5.60]$	< 0.001	$+4.12$	$[+3.26, +4.99]$	< 0.001
Comma0.1	$+6.57$	$[+5.76, +7.39]$	< 0.001	$+4.93$	$[+4.09, +5.75]$	< 0.001
Nemotron-CC-HQ	$+0.38$	$[-0.42, +1.21]$	0.377	$+0.51$	$[-0.38, +1.34]$	0.258
FineWeb-Edu	$+6.10$	$[+5.31, +6.84]$	< 0.001	$+5.27$	$[+4.51, +6.01]$	< 0.001
Permissive minus non-permissive pre-training dataset, both enriched						
MV − Nemotron	-8.73	$[-9.65, -7.83]$	< 0.001	-9.62	$[-10.56, -8.65]$	< 0.001
MV − FineWeb	-7.02	$[-7.99, -6.05]$	< 0.001	-7.62	$[-8.53, -6.73]$	< 0.001
Comma − Nemotron	-7.94	$[-8.97, -6.93]$	< 0.001	-7.55	$[-8.50, -6.63]$	< 0.001
Comma − FineWeb	-6.23	$[-7.18, -5.26]$	< 0.001	-5.54	$[-6.42, -4.67]$	< 0.001
Interaction: front-loading effect, permissive minus non-permissive						
MV − non-permissive	$+1.53$	$[+0.47, +2.50]$	0.010	$+1.22$	$[+0.15, +2.33]$	0.028
Comma − non-permissive	$+3.33$	$[+2.33, +4.35]$	0.001	$+2.04$	$[+1.06, +3.09]$	0.001

D.2 Complete results

Table 24 gives every benchmark of the four arms that Table 17 does not hold, and Table 25 the language suite of the sixteen pretrained bases and 16K extended checkpoints. Without front-loading, FineWeb-Edu and both permissive pre-training datasets score at chance on 5-shot MMLU and 10-shot CommonsenseQA, and enrichment moves them above it. Nemotron-CC-HQ scores 46.4 on MMLU and 48.6 on CommonsenseQA without front-loading. Figure 3 splits the front-loading effect by domain.

Table 24: Per-benchmark results for the 300B front-loading $\times$ pre-training dataset factorial. Enriched models include the MixtureVitae instruction and reasoning subset in their pretraining mixture, and the models marked not enriched do not. Protocol A (Appendix Table 15). Values are mean$\pm$SD over decoding seeds 42, 1234, and 2025. The table gives the four arms that are not models of Table 17: MixtureVitae without enrichment and the three enriched reference mixtures. The other four, the main MixtureVitae model and Comma0.1, Nemotron-CC-HQ and FineWeb-Edu without enrichment, are rows of Table 17 with the same values. The table has two panels: (a) mathematics, (b) code, science and instruction following, and the suite average.

(a) Mathematics

Model	A24	A25	AMC	M500	GSM8K
MV (not enriched)	0.0±0.00	0.0±0.00	0.0±0.00	0.1±0.12	3.3±0.00
→ SFT	0.0±0.00	0.0±0.00	2.5±0.00	1.9±0.12	4.4±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	2.5±0.00	1.7±0.23	4.0±0.00
Comma0.1 (enriched)	0.0±0.00	12.2±1.92	21.7±1.44	52.9±0.31	57.7±0.00
→ SFT	0.0±0.00	0.0±0.00	8.3±2.89	29.2±1.04	56.5±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	16.7±2.89	30.7±0.31	56.6±0.00
Nemotron (enriched)	1.1±1.92	5.6±1.92	25.0±0.00	44.6±0.69	52.7±0.00
→ SFT	0.0±0.00	0.0±0.00	7.5±0.00	19.4±0.40	50.9±0.00
→ SFT → DPO	1.1±1.92	0.0±0.00	7.5±2.50	19.9±0.99	51.5±0.00
FineWeb-Edu (enriched)	0.0±0.00	6.7±0.00	22.5±0.00	47.0±1.11	53.5±0.00
→ SFT	0.0±0.00	0.0±0.00	9.2±1.44	17.9±0.42	50.1±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	7.5±0.00	18.4±0.20	50.2±0.00
MV-16K (not enriched)	0.0±0.00	0.0±0.00	2.5±0.00	2.6±0.20	5.4±0.00
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	2.3±0.12	5.0±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	0.8±1.44	2.6±0.20	4.6±0.00
Comma0.1-16K (enriched)	1.1±1.92	0.0±0.00	23.3±2.89	41.8±0.00	59.6±0.00
→ SFT	0.0±0.00	0.0±0.00	10.0±0.00	26.7±0.31	55.4±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	9.2±1.44	27.7±0.12	55.9±0.00
Nemotron-16K (enriched)	0.0±0.00	1.1±1.92	20.0±2.50	39.9±0.50	59.8±0.00
→ SFT	0.0±0.00	0.0±0.00	6.7±1.44	18.5±0.46	49.6±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	5.8±1.44	20.3±0.50	50.7±0.00
FineWeb-Edu-16K (enriched)	0.0±0.00	0.0±0.00	13.3±1.44	38.0±0.35	57.2±0.00
→ SFT	0.0±0.00	2.2±1.92	5.8±1.44	21.9±0.64	51.1±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	2.5±0.00	23.1±0.31	51.1±0.00

Table 24: (Continued) Panel (b), code, science and instruction following, and the suite average.

(b) Code, science and instruction following, and the suite average

Model	HE	LCB	MBPP	GPQA	JEE	IFE	Avg
MV (not enriched)	0.0±0.00	0.0±0.00	22.0±0.20	23.6±0.29	0.0±0.00	30.9±0.33	7.3±0.03
→ SFT	19.1±0.35	1.1±0.00	25.1±0.50	25.1±0.29	2.3±0.30	37.9±0.20	10.9±0.02
→ SFT → DPO	19.7±0.35	1.1±0.00	25.4±0.00	25.1±0.29	2.5±0.53	41.7±0.15	11.2±0.08
Comma0.1 (enriched)	44.9±0.35	9.4±0.63	37.1±0.50	24.7±2.31	4.3±0.26	36.8±0.26	27.4±0.14
→ SFT	56.7±0.00	6.8±0.00	42.7±0.31	24.2±0.00	5.2±0.33	49.2±0.49	25.4±0.27
→ SFT → DPO	57.9±0.00	6.7±0.31	42.1±0.76	25.6±0.29	5.2±0.33	52.1±0.61	26.7±0.30
Nemotron (enriched)	40.2±0.61	7.7±0.95	8.7±0.12	20.4±1.27	3.3±0.42	33.5±0.20	22.1±0.22
→ SFT	47.8±0.70	2.6±0.16	35.1±0.12	24.2±0.00	4.3±0.15	48.5±0.53	21.9±0.13
→ SFT → DPO	50.0±0.00	3.3±0.27	35.5±0.23	26.1±0.29	4.8±0.78	51.1±0.53	22.8±0.25
FineWeb-Edu (enriched)	38.4±0.00	9.8±0.54	28.5±0.64	23.1±0.29	5.0±0.39	33.9±0.45	24.4±0.05
→ SFT	51.2±0.00	2.9±0.16	34.4±0.20	23.6±0.29	4.4±0.14	48.1±0.42	22.0±0.14
→ SFT → DPO	53.0±0.00	4.1±0.00	35.2±0.00	23.7±0.00	4.4±0.18	49.5±0.53	22.4±0.08
MV-16K (not enriched)	19.3±0.35	0.0±0.00	22.3±0.23	23.2±0.00	6.5±0.07	42.7±0.08	11.3±0.03
→ SFT	19.5±0.00	0.5±0.00	25.9±0.42	25.1±0.29	3.4±0.18	45.1±0.23	11.5±0.02
→ SFT → DPO	19.1±0.35	0.4±0.16	25.6±0.00	25.1±0.58	3.9±0.53	46.4±0.23	11.7±0.13
Comma0.1-16K (enriched)	45.3±0.35	2.4±0.27	41.4±0.20	24.2±0.00	10.8±1.26	52.8±0.20	27.5±0.34
→ SFT	54.3±0.00	6.1±0.16	40.8±0.40	24.1±0.29	7.2±0.83	53.4±0.34	25.3±0.13
→ SFT → DPO	54.9±0.61	6.1±0.31	41.8±0.20	23.9±0.77	6.5±0.51	53.4±0.65	25.4±0.27
Nemotron-16K (enriched)	46.5±0.35	6.7±0.16	34.7±0.12	27.3±0.00	8.3±0.83	46.8±0.15	26.5±0.39
→ SFT	49.8±0.35	0.3±0.00	32.5±0.23	25.9±0.29	6.0±0.29	53.6±0.34	22.1±0.27
→ SFT → DPO	49.4±0.61	0.3±0.00	32.9±0.12	24.7±0.00	6.8±0.50	55.4±0.72	22.4±0.26
FineWeb-Edu-16K (enriched)	43.9±0.00	8.2±0.31	35.3±0.12	29.0±0.29	9.1±0.97	52.4±0.34	26.0±0.06
→ SFT	45.5±0.35	2.7±0.00	31.7±0.23	26.8±0.51	5.4±0.30	51.5±0.08	22.2±0.25
→ SFT → DPO	45.1±0.61	2.6±0.16	31.4±0.00	25.3±0.00	5.1±1.05	53.7±0.00	21.8±0.07

Table 25: Language understanding for the sixteen factorial checkpoints, on the 4K pretrained bases and their 16K extended checkpoints. This suite is scored by the LM Evaluation Harness under its own settings and is not on either reasoning protocol. Single seed. Accuracies in percent, `acc_norm` where the harness reports it, and shot counts are in Table 16. Without enrichment, FineWeb-Edu and the MixtureVitae pre-training dataset sit at chance on MMLU and CommonsenseQA. Comma0.1 without enrichment reuses the Comma0.1 model of Table 1, and it also sits at chance on MMLU and CommonsenseQA. The enriched MixtureVitae rows reuse the main MixtureVitae model, as in Table 18. Table 23 gives 95% intervals for its contrasts.

Model	Enriched	MMLU	HellaS	CSQA	ARC-C	ARC-E	PIQA	BoolQ	WinoG	OBQA	COPA	LAMBADA	Avg
MixtureVitae	Yes	41.9	54.5	51.4	40.3	70.9	70.4	73.5	57.8	34.2	71.0	48.1	55.8
MixtureVitae	No	25.8	52.2	22.4	38.2	68.8	70.5	64.8	58.2	36.4	73.0	51.2	51.0
Comma0.1	Yes	40.4	55.4	52.1	39.8	69.4	71.2	75.3	59.5	36.2	70.0	53.4	56.6
Comma0.1	No	26.5	53.5	20.4	36.2	63.1	71.2	63.2	59.3	32.6	70.0	54.5	50.0
Nemotron	Yes	50.5	68.9	62.1	52.3	80.3	77.6	79.3	62.9	38.6	80.0	57.7	64.6
Nemotron	No	46.4	71.5	48.6	52.4	80.9	78.7	77.4	65.0	42.0	85.0	58.0	64.2
FineWeb-Edu	Yes	47.5	65.3	55.7	49.5	79.9	76.9	77.3	62.6	43.0	79.0	54.6	62.8
FineWeb-Edu	No	26.1	67.1	21.1	48.5	76.9	77.1	64.9	62.4	42.4	82.0	55.7	56.7
MixtureVitae-16K	Yes	41.1	53.1	48.8	41.6	71.4	69.7	72.6	57.9	31.8	67.0	49.4	55.0
MixtureVitae-16K	No	26.1	51.0	22.5	38.9	68.6	69.9	61.9	57.9	34.6	76.0	51.8	50.8
Comma0.1-16K	Yes	40.4	53.7	49.0	41.9	71.9	70.7	78.3	59.4	35.6	73.0	53.6	57.0
Comma0.1-16K	No	26.8	52.5	22.8	39.7	68.1	71.8	63.9	60.9	33.4	76.0	57.2	52.1
Nemotron-16K	Yes	49.5	67.7	61.7	52.3	81.4	77.9	79.5	62.8	38.2	81.0	58.5	64.6
Nemotron-16K	No	46.5	70.0	50.6	52.0	80.7	78.1	77.1	64.6	41.8	83.0	60.1	64.1
FineWeb-Edu-16K	Yes	46.4	62.5	56.4	50.2	79.6	76.7	77.6	62.1	41.8	79.0	56.1	62.6
FineWeb-Edu-16K	No	26.8	64.7	20.9	48.1	78.8	76.6	70.0	63.1	42.8	81.0	57.5	57.3

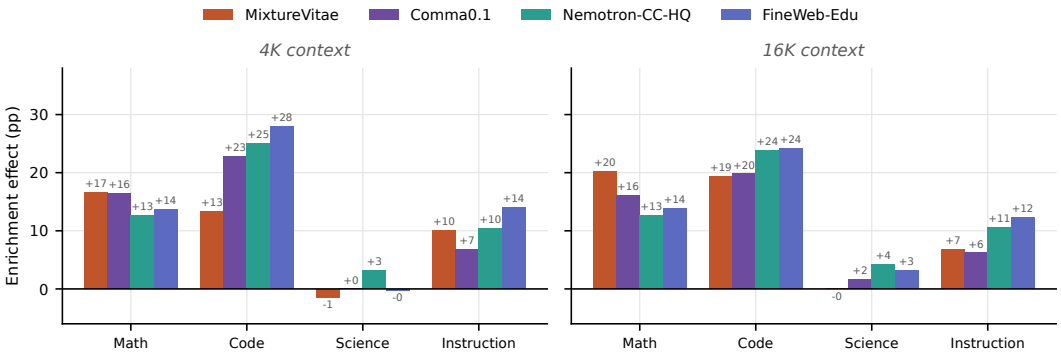


Figure 3: Front-loading effect by capability domain after Tulu3 SFT. The domains are math (AIME24, AIME25, AMC23, MATH500 and GSM8K), code (HumanEval, LiveCodeBench and MBPP), science (GPQA-Diamond and JEEBench) and instruction following (IFEval). The gain is not confined to mathematics, and instruction following improves in every pre-training dataset. Protocol A.

E Reasoning post-training: controls and diagnostics

Every change in this appendix is a checkpoint minus the checkpoint its run was trained from, under Protocol B. Brackets are 95% paired bootstrap intervals over the questions, each question scored as its mean over the three runs, and they do not include differences between training runs. The code tasks have no per-question verdicts in the Phase 2 records and enter with the conservative approximation of Appendix B.2. Bold marks a significant increase and italic a significant decrease after Benjamini–Hochberg correction, within the family that each caption names. Table 2 in the main text gives the endpoints of each reasoning SFT run, and Table 26 every checkpoint before and after it, on every task and with its standard deviation.

E.1 Main reasoning SFT comparisons

Under the reasoning SFT on OpenThoughts3 from the Tulu3 SFT checkpoint (Table 27), the suite average of the three matched-compute references changes by -1.3 $[-2.3, -0.5]$ points for FineWeb-Edu, -1.3 $[-2.0, -0.6]$ for Nemotron-CC-HQ and -4.0 $[-5.1, -3.0]$ for Comma0.1, and MV's by $+10.7$ $[+8.2, +13.2]$. SmolLM2 and Qwen3 decline as well, and the change of Qwen2.5, -0.7 $[-2.6, +1.3]$, cannot be told from zero: it improves on MATH500, from 21.0 to 44.9, but loses enough elsewhere to offset that gain. From its DPO checkpoint, MV's change of $+0.5$ $[-1.5, +2.7]$ cannot be told from zero either.

Table 28 compares MV with each reference directly, on the macro average of the five mathematics tasks and of the three contest tasks (100 problems). MV's mathematics lead over the matched-compute references is $+13.7$ to $+15.1$ points at the Tulu3 SFT checkpoints and $+35.9$ to $+37.2$ after OT3-30K, and its contest lead $+1.9$ to $+2.8$ and $+24.3$ to $+24.8$, so the post-training widens both, with every interval of the widening above zero. With each mathematics task left out in turn (Table 29), all thirty remaining gaps after OT3-30K lie above zero. On IFEval after OT3-30K, MV leads the matched-compute references by $+5.3$ to $+8.7$ points and is level with SmolLM2 ($+0.4$ $[-2.7, +3.6]$), although its own IFEval score falls under the post-training (Appendix E.4).

E.2 Excluding the OpenThoughts3 source from pretraining

Both models of Table 3 are pretrained for 300B tokens and receive the same reasoning SFT directly from their 16K extended checkpoints, rather than the Tulu3 SFT checkpoints of the main comparison. Their starting suite averages differ by 4.3 points, so the exclusion is measured as the difference between their gains (Table 30). With OpenThoughts3 in pretraining the suite average gains $+7.6$ $[+5.2, +10.0]$ points, without it $+1.7$ $[-0.2, +3.6]$, and the difference is $+5.9$ $[+2.2, +9.9]$. On MATH500 the gains are $+23.3$ and $+8.7$ points, a difference of $+14.5$ $[+10.9, +18.4]$.

The suite average gain without OpenThoughts3 cannot be distinguished from zero. Without OpenThoughts3, the MATH500 gain remains positive at $+8.7$ points $[+6.0, +11.6]$, against $+23.3$ $[+20.3, +26.0]$ with it. On the 287 MATH500 problems left after excluding the identified over-

Table 26: **Reasoning post-training: every checkpoint of Table 2.** Before and After are the checkpoints before and after each OT3-30K SFT run, not before and after Tulu3. Starting checkpoints are evaluated again here, so these values are paired within the table and must not be compared with the Phase 1 tables. Protocol B (Appendix Table 15). Values are mean±SD over three sampled runs, and Avg is the mean over the eleven tasks shown.

(a) Mathematics

Model	Stage	A24	A25	AMC	M500	GSM8K
Main comparison: Tulu3 SFT → OT3-30K						
MixtureVitae	Before	1.1±1.92	0.0±0.00	8.3±3.82	29.2±6.41	46.3±0.88
	After	11.1±3.85	16.7±3.33	46.7±6.29	71.0±1.06	49.0±1.75
Comma0.1	Before	0.0±0.00	1.1±1.92	2.5±2.50	1.7±0.61	11.0±0.77
	After	0.0±0.00	0.0±0.00	0.0±0.00	1.3±0.58	13.4±0.80
Nemotron-CC-HQ	Before	0.0±0.00	0.0±0.00	2.5±2.50	1.5±0.83	6.6±0.85
	After	0.0±0.00	0.0±0.00	1.7±2.89	1.5±0.23	6.3±0.39
FineWeb-Edu	Before	0.0±0.00	1.1±1.92	0.0±0.00	0.8±0.53	7.3±0.76
	After	0.0±0.00	0.0±0.00	1.7±1.44	0.6±0.00	6.2±0.72
SmolLM2 1.7B	Before	0.0±0.00	0.0±0.00	1.7±2.89	6.7±0.70	25.9±1.45
	After	0.0±0.00	0.0±0.00	1.7±1.44	6.6±0.92	23.1±1.34
Qwen2.5 1.5B	Before	0.0±0.00	0.0±0.00	6.7±1.44	21.0±2.78	55.1±0.39
	After	4.4±5.09	1.1±1.92	26.7±7.22	44.9±2.00	57.6±0.34
Qwen3 1.7B	Before	1.1±1.92	0.0±0.00	12.5±2.50	24.7±2.50	64.7±0.80
	After	0.0±0.00	0.0±0.00	1.7±1.44	10.9±0.42	25.7±0.38
Enriched, 300B tokens						
Comma0.1	Before	0.0±0.00	0.0±0.00	5.8±2.89	22.2±0.87	51.5±1.65
	After	15.6±8.39	23.3±3.33	50.8±5.77	74.7±0.61	49.5±0.74
Nemotron-CC-HQ	Before	0.0±0.00	0.0±0.00	15.0±4.33	19.1±2.69	46.3±1.23
	After	16.7±5.77	12.2±3.85	46.7±3.82	72.1±0.81	48.2±0.64
FineWeb-Edu	Before	1.1±1.92	0.0±0.00	8.3±5.77	17.3±3.78	47.1±0.52
	After	16.7±5.77	18.9±1.92	40.8±5.77	69.4±0.80	48.0±0.66
MixtureVitae starting point controls						
16K extended	Before	3.3±3.33	8.9±3.85	25.8±12.58	46.5±11.29	55.0±1.78
	After	8.9±5.09	25.6±1.92	42.5±6.61	69.8±1.25	58.8±1.52
Tulu3 SFT + DPO	Before	1.1±1.92	1.1±1.92	14.2±6.29	30.1±3.31	47.4±1.20
	After	0.0±0.00	4.4±1.92	17.5±4.33	34.4±1.78	44.4±1.52

laps, the difference between the gains is $+11.5$ points $[+6.3, +16.8]$. The exclusion removes the OpenThoughts3 source, but 97.8% of the mathematical question instances in the Phase 2 OT3 sample still match the remaining pretraining reasoning data (Table 32). It therefore measures dependence on this source's traces and training allocation, without separately identifying the effects of question exposure, response content or repetition.

The evaluation software samples all MATH500 questions of a run with one seed, given by the run's number, so the answers of a run are correlated, and one run can sit far from another. For MixtureVitae's 16K extended checkpoint the three released runs score 33.6, 51.6 and 54.4, and the first run's answers are much shorter than the other two runs'. An earlier single evaluation of the same checkpoint was such a first run, with the same seed, which is why nearly all of the first run's answers are the text of that earlier evaluation. To tell which runs are typical, the four checkpoints of the exclusion were evaluated again on MATH500 in three runs that give each question its own seed, with the settings of the released runs otherwise (last column of Table 30). MixtureVitae's 16K extended checkpoint scores 36.0 to 39.2 in them, a mean of 37.5, close to its first released run, and each of the other three checkpoints moves by at most 0.7 points from its released mean. The MATH500 gains are then $+32.7$ and $+8.6$ points, and their difference is $+24.1$ $[+20.3, +28.0]$, so the released runs understate it. One such rerun of each of 100 sampled MATH500 cells of Phase 2 lies within two standard errors of the released mean in 94 of them, with a median difference of -0.3 points. The reruns cover MATH500 only, so the tables report the released runs, and the suite average is not recomputed. Panel (c) of Table 34 splits the gain without OpenThoughts3 by domain, and Table 38 shows how the answers of both models end.

Table 26: (Continued) Panel (b), code, science and instruction following, and the suite average.

(b) Code, science and instruction following, and the average

Model	Stage	HE	LCB	MBPP	GPQA	JEE	IFE	Avg
Main comparison: Tulu3 SFT → OT3-30K								
MixtureVitae	Before	44.9±1.76	7.0±2.26	31.9±0.46	27.4±0.58	7.8±1.33	54.0±0.15	23.4±0.31
	After	51.0±4.97	11.5±0.87	28.5±3.29	28.8±1.75	18.8±1.34	42.1±0.20	34.1±0.93
Comma0.1	Before	14.2±0.93	0.4±0.31	16.5±2.08	26.4±5.40	5.5±0.72	46.8±0.52	11.5±0.82
	After	1.4±0.93	0.0±0.00	0.1±0.12	26.8±2.81	2.6±0.32	36.8±0.08	7.5±0.29
Nemotron-CC-HQ	Before	2.2±1.27	0.0±0.00	1.6±0.35	25.3±1.34	2.3±0.45	44.7±1.69	7.9±0.08
	After	0.2±0.35	0.0±0.00	0.0±0.00	24.2±2.62	3.7±0.40	34.7±1.40	6.6±0.59
FineWeb-Edu	Before	2.2±1.96	0.1±0.16	1.0±0.20	28.6±1.27	3.1±0.59	38.1±0.98	7.5±0.37
	After	0.0±0.00	0.0±0.00	0.0±0.00	23.9±0.77	1.9±0.44	33.4±0.67	6.2±0.08
SmolLM2 1.7B	Before	23.6±2.46	0.9±0.31	23.3±0.46	25.1±0.77	8.1±1.92	53.8±0.72	15.4±0.53
	After	14.2±3.01	0.2±0.31	10.5±0.92	26.3±1.01	3.6±0.63	41.7±0.72	11.6±0.16
Qwen2.5 1.5B	Before	46.5±0.70	3.1±0.31	39.9±1.01	25.9±3.58	6.9±2.23	54.9±0.47	23.6±0.32
	After	21.7±6.23	2.8±0.68	12.3±1.79	25.9±0.77	14.1±0.82	40.6±0.98	22.9±0.70
Qwen3 1.7B	Before	59.8±1.22	3.4±0.41	46.5±1.86	28.5±2.39	12.7±1.00	59.1±1.23	28.5±0.95
	After	8.5±1.06	0.5±0.27	6.7±0.76	29.0±2.04	2.0±0.47	38.8±0.98	11.3±0.31
Enriched, 300B tokens								
Comma0.1	Before	43.3±1.22	3.3±0.83	30.1±0.61	26.1±2.10	8.3±0.82	53.3±1.89	22.2±0.16
	After	62.2±1.06	13.6±0.54	30.1±0.64	27.9±0.29	25.8±1.53	44.3±0.65	38.0±1.08
Nemotron-CC-HQ	Before	39.0±1.61	1.4±2.12	25.6±1.31	26.8±1.75	9.4±0.73	50.5±1.35	21.2±0.37
	After	57.7±1.27	12.7±1.69	28.6±0.35	31.3±2.20	23.0±1.85	39.4±0.67	35.3±0.26
FineWeb-Edu	Before	35.8±4.58	2.9±1.03	21.8±1.97	26.4±0.58	6.8±1.19	52.4±1.11	20.0±1.22
	After	53.5±0.93	11.3±1.22	17.1±1.01	30.1±3.65	19.1±2.47	41.0±0.59	33.3±0.95
MixtureVitae starting point controls								
16K extended	Before	43.7±0.93	8.9±0.27	31.3±1.47	26.9±0.29	10.7±1.58	46.6±0.30	28.0±2.73
	After	54.7±3.36	13.4±0.87	28.8±1.91	30.1±4.98	18.4±0.41	40.9±1.14	35.6±1.15
Tulu3 SFT + DPO	Before	47.2±4.74	6.7±2.67	34.6±1.06	27.3±0.87	8.2±0.75	54.6±1.38	24.8±1.13
	After	46.1±1.41	8.6±1.49	34.5±2.44	27.1±1.27	11.5±0.51	49.2±0.61	25.3±0.80

Table 27: **Phase 2 changes on the suite average, with intervals.** The change of every run between the checkpoints of Table 26, each against the checkpoint it was trained from: the reasoning SFT runs of Table 2 other than its front-loading controls, and MixtureVitae's two Tulu3 steps. Bold and italic marks are corrected over the 11 changes. The front-loading controls, with their differences of gains, are a family of their own (Appendix E.5).

Run	Change, 95% interval	q
MixtureVitae-16K		
→ OT3-30K (without Tulu3)	$\mathbf{+7.6}$ $[+5.2, +10.0]$	< 0.001
→ SFT	$\mathit{-4.5}$ $[-6.6, -2.5]$	< 0.001
→ SFT → OT3-30K	$\mathbf{+10.7}$ $[+8.2, +13.2]$	< 0.001
→ SFT → DPO	$+1.3$ $[-0.5, +3.2]$	0.194
→ SFT → DPO → OT3-30K	$+0.5$ $[-1.5, +2.7]$	0.644
OT3-30K from each reference's Tulu3 SFT checkpoint		
FineWeb-Edu-16K	$\mathit{-1.3}$ $[-2.3, -0.5]$	0.005
Nemotron-CC-HQ-16K	$\mathit{-1.3}$ $[-2.0, -0.6]$	< 0.001
Comma0.1-16K	$\mathit{-4.0}$ $[-5.1, -3.0]$	< 0.001
SmolLM2-16K	$\mathit{-3.8}$ $[-5.1, -2.4]$	< 0.001
Qwen2.5-1.5B	-0.7 $[-2.6, +1.3]$	0.536
Qwen3-1.7B	$\mathit{-17.2}$ $[-18.7, -15.7]$	< 0.001

Table 28: **MixtureVitae against each reference before and after reasoning SFT.** MV minus the reference, in points, on the macro average of the five mathematics tasks (AIME24, AIME25, AMC23, MATH500, GSM8K), on the macro average of the three contest tasks (AIME24, AIME25, AMC23, 100 problems), and on IFEval, with both models at their Tulu3 SFT checkpoints and after OT3-30K from them. The widening is MV's gain under OT3-30K minus the reference's, so it says whether the gap after the post-training is larger than before it. Bold marks a significant difference in MV's favour and italic one in the reference's, corrected within each of mathematics, contest mathematics and IFEval.

	Mathematics			**Contest mathematics**		**IFEval**
Reference	**Tulu3 SFT**	**After OT3-30K**	**Widening**	**Tulu3 SFT**	**After OT3-30K**	**After OT3-30K**
FineWeb-Edu	**+15.1** [+13.7, +16.6]	**+37.2** [+33.3, +41.5]	**+22.0** [+18.2, +26.3]	**+2.8** [+0.6, +5.0]	**+24.3** [+17.5, +31.1]	**+8.7** [+7.0, +10.4]
Nemotron-CC-HQ	**+14.9** [+13.5, +16.3]	**+37.0** [+32.8, +41.1]	**+22.1** [+18.3, +26.5]	**+2.3** [+0.3, +4.5]	**+24.3** [+17.6, +31.1]	**+7.4** [+5.6, +9.2]
Comma0.1	**+13.7** [+12.2, +15.4]	**+35.9** [+31.9, +40.2]	**+22.2** [+18.5, +26.3]	+1.9 [−0.3, +4.3]	**+24.8** [+18.1, +31.7]	**+5.3** [+3.8, +6.9]
SmolLM2	**+10.1** [+8.8, +11.6]	**+32.6** [+28.6, +36.8]	**+22.5** [+18.6, +26.7]	**+2.6** [+0.6, +4.8]	**+24.3** [+17.3, +31.3]	+0.4 [−2.7, +3.6]
Qwen2.5	+0.4 [−1.1, +2.0]	**+11.9** [+8.7, +15.6]	**+11.5** [+7.6, +15.7]	+0.9 [−1.4, +3.1]	**+14.1** [+8.6, +19.8]	+1.5 [−0.3, +3.4]
Qwen3	*−3.6* [−5.2, −2.0]	**+31.2** [+27.3, +35.4]	**+34.9** [+30.5, +39.2]	−1.4 [−3.8, +1.1]	**+24.3** [+18.3, +31.0]	**+3.3** [+0.7, +5.9]

Table 29: **The mathematics gap after OT3-30K with each task left out.** MV minus the reference after OT3-30K on the macro average of the mathematics tasks of Table 28, with the task named in the column left out. The intervals are robustness diagnostics, chosen after seeing every removal, and are not corrected for multiple comparisons.

Reference	**Without A24**	**Without A25**	**Without AMC**	**Without M500**	**Without GSM8K**
FineWeb-Edu	+43.7 [+39.1, +48.3]	+42.3 [+38.2, +46.6]	+35.2 [+31.8, +39.6]	+28.9 [+23.7, +34.2]	+35.8 [+30.6, +41.0]
Nemotron-CC-HQ	+43.5 [+38.9, +48.4]	+42.1 [+38.0, +46.3]	+35.0 [+31.3, +39.4]	+28.9 [+23.8, +34.3]	+35.6 [+30.4, +40.9]
Comma0.1	+42.2 [+37.8, +46.8]	+40.8 [+36.8, +44.9]	+33.3 [+29.6, +37.6]	+27.5 [+22.8, +32.6]	+36.0 [+31.1, +41.4]
SmolLM2	+38.0 [+33.2, +42.8]	+36.6 [+32.5, +41.0]	+29.5 [+26.0, +33.8]	+24.7 [+19.8, +29.9]	+34.3 [+29.3, +39.7]
Qwen2.5	+13.3 [+9.3, +17.4]	+11.0 [+7.7, +14.7]	+9.9 [+6.4, +13.6]	+8.4 [+4.2, +12.9]	+17.1 [+12.9, +21.6]
Qwen3	+36.3 [+31.6, +40.8]	+34.9 [+31.0, +39.1]	+27.8 [+24.1, +31.7]	+24.0 [+19.2, +29.1]	+33.2 [+28.1, +38.4]

Table 30: **The OT3 exclusion: the two gains and their difference.** The change of each model of Table 3 from its own 16K extended checkpoint under the identical OT3-30K SFT run, and the first change minus the second, which is how much more the model with OpenThoughts3 in pretraining gains. The two bases differ before reasoning SFT, so the difference of the gains is not the difference of the two endpoints. Bold and italic marks are corrected over the six results of the first two columns. The last column evaluates the four checkpoints again on MATH500, in three runs that give each question its own seed (Appendix E.2), corrected over its three results.

	Released runs		**Reruns**
	Suite average	**MATH500**	**MATH500**
OT3 in pretraining, gain	**+7.6** [+5.2, +10.0]	**+23.3** [+20.3, +26.0]	**+32.7** [+29.3, +35.8]
OT3 removed from pretraining, gain	+1.7 [−0.2, +3.6]	**+8.7** [+6.0, +11.6]	**+8.6** [+6.1, +10.9]
Difference of the two gains	**+5.9** [+2.2, +9.9]	**+14.5** [+10.9, +18.4]	**+24.1** [+20.3, +28.0]

E.3 Alternative corpora and starting checkpoints

Table 4 in the main text gives the MATH500 change of every base from its Tulu3 SFT checkpoint under each of the four reasoning datasets, without the problems the benchmark-overlap audit identified, and Table 31 gives the same changes on all 500 problems. Table 32 gives how many of each 30K sample's questions match MixtureVitae's pretraining reasoning pool. The OT3-source exclusion keeps most of the Phase 2 OpenThoughts3 mathematics problems in other sources, with other solutions, so it removes the source, not the questions.

Table 33 gives MixtureVitae's whole suite for the three other datasets from each of its three starting checkpoints, with the recipe, the sample budget and the number of epochs held fixed. The OpenThoughts3 runs from the same checkpoints are rows of Table 26. Superior-Reasoning gives a mixed result: it raises mathematics and lowers code and instruction following, enough to lower the suite average from the 16K extended checkpoint (-1.0) but not from the SFT and DPO checkpoints ($+2.7$ and $+1.4$).

Table 31: **MATH500 changes on all 500 problems.** Change in MATH500 accuracy when each base's Tulu3 SFT checkpoint receives 30K samples from one of four reasoning corpora, on every problem, including those identified in the training data. Table 4 gives the same changes without them. Each problem is scored as its mean over three sampled runs. Brackets are 95% paired bootstrap intervals over the same 500 problems. **Bold** marks a significant increase and *italic* a significant decrease after Benjamini–Hochberg correction over the 20 changes. Protocol B (Appendix Table 15).

Base	**OpenThoughts3**	**AceReason**	**Nemotron-Cascade**	**Superior-Reasoning**
MixtureVitae	$\mathbf{+41.8}$ [+38.5, +45.1]	$\mathbf{+38.0}$ [+34.5, +41.4]	$\mathbf{+38.9}$ [+35.6, +42.1]	$\mathbf{+24.9}$ [+21.7, +28.2]
Comma0.1	-0.3 [−1.2, +0.5]	$+0.1$ [−0.7, +1.1]	$+0.9$ [−0.1, +2.0]	$+0.0$ [−0.9, +0.9]
FineWeb-Edu	-0.2 [−0.8, +0.5]	$\mathbf{+1.1}$ [+0.3, +1.9]	$+0.7$ [+0.0, +1.5]	$+1.0$ [+0.1, +1.9]
Nemotron-CC-HQ	$+0.0$ [−0.9, +0.9]	$+1.3$ [+0.1, +2.4]	$+1.0$ [−0.1, +2.1]	$+0.0$ [−0.9, +0.9]
SmolLM2	-0.1 [−1.9, +1.6]	$\mathbf{+5.9}$ [+3.9, +8.1]	$\mathbf{+5.3}$ [+3.3, +7.3]	$\mathbf{+6.3}$ [+4.3, +8.3]

Table 32: **The Phase 2 samples' questions in MixtureVitae's reasoning pool.** Percent of the questions of each 30K sample trained on that match a document of MixtureVitae's pretraining reasoning data other than OpenThoughts3, overall and by domain (13-gram and exact matching, Appendix A.3). Other is general questions for Nemotron-Cascade and instruction following for Superior-Reasoning. The model pretrained without the OpenThoughts3 source keeps this data, so the OT3-source exclusion keeps most of the Phase 2 OpenThoughts3 mathematics problems, with other solutions. A match is a document of the pool, not evidence that a 300B-token run drew it.

Corpus	Questions	All	Math	Code	Science	Other
OpenThoughts3	30,000	80.6	97.8	51.0	5.8	–
AceReason	30,000	66.4	65.8	67.7	–	–
Nemotron-Cascade	30,000	51.0	65.9	68.0	12.5	8.2
Superior-Reasoning	30,000	46.0	37.8	97.1	41.4	0.3

Table 33: The three other reasoning corpora from each of MixtureVitae's three starting checkpoints, with recipe, sample budget and epoch count held fixed, so that only the reasoning corpus changes. The change in parentheses is against the corresponding starting checkpoint in Table 26, which also gives their reasoning SFT runs on OpenThoughts3. Protocol B (Appendix Table 15). Values are mean±SD over three sampled runs, and Avg is the mean over the eleven tasks shown.

(a) Mathematics

Reasoning corpus	A24	A25	AMC	M500	GSM8K
Starting from the 16K extended checkpoint					
AceReason	6.7±6.67	13.3±0.00	45.0±2.50	68.0±1.71	59.7±0.65
Nemotron-Cascade	11.1±8.39	14.4±1.92	44.2±1.44	67.0±1.11	59.9±0.31
Superior-Reasoning	4.4±1.92	6.7±0.00	32.5±2.50	51.4±0.53	63.0±0.74
Starting from the Tulu3 SFT checkpoint					
AceReason	10.0±0.00	15.6±1.92	45.0±4.33	67.2±0.60	57.2±0.62
Nemotron-Cascade	11.1±1.92	16.7±3.33	44.2±1.44	68.1±0.81	72.0±0.18
Superior-Reasoning	4.4±1.92	7.8±5.09	28.3±5.20	54.1±1.30	57.8±0.55
Starting from the Tulu3 SFT + DPO checkpoint					
AceReason	12.2±1.92	16.7±0.00	40.8±6.29	67.1±1.33	57.3±0.26
Nemotron-Cascade	7.8±1.92	18.9±1.92	43.3±7.64	68.7±1.51	71.8±0.36
Superior-Reasoning	6.7±3.33	6.7±0.00	29.2±3.82	53.5±1.81	57.6±0.38

(b) Code, science and instruction following, and the average

Reasoning corpus	HE	LCB	MBPP	GPQA	JEE	IFE	Avg
Starting from the 16K extended checkpoint							
AceReason	58.3±0.70	13.8±0.27	37.5±1.89	29.8±3.15	20.6±0.88	40.2±1.24	35.7±0.72 (+7.7)
Nemotron-Cascade	54.7±2.82	12.6±0.78	37.7±1.63	28.6±1.27	20.6±0.60	43.1±0.87	35.8±0.68 (+7.8)
Superior-Reasoning	31.5±5.67	8.6±0.41	20.6±1.06	25.9±1.62	10.4±1.03	41.6±0.61	27.0±0.34 (−1.0)
Starting from the Tulu3 SFT checkpoint							
AceReason	54.7±1.41	14.2±1.10	36.1±0.99	26.1±1.17	20.9±1.05	41.6±0.34	35.3±0.47 (+11.9)
Nemotron-Cascade	55.1±0.93	13.6±1.76	37.2±1.59	28.1±0.58	20.9±1.14	44.1±0.98	37.4±0.34 (+13.9)
Superior-Reasoning	29.5±1.86	7.8±0.56	21.8±0.35	25.3±2.31	10.2±1.06	41.2±0.61	26.2±0.48 (+2.7)
Starting from the Tulu3 SFT + DPO checkpoint							
AceReason	52.8±7.74	14.8±0.87	36.7±0.50	29.0±2.59	20.7±0.86	42.1±1.66	35.5±0.71 (+10.7)
Nemotron-Cascade	52.8±4.06	14.4±0.72	37.7±1.62	28.6±0.77	22.2±0.44	43.9±1.43	37.3±1.03 (+12.5)
Superior-Reasoning	29.7±1.41	6.2±0.00	22.9±1.33	24.6±2.28	10.1±1.77	41.2±0.15	26.2±1.18 (+1.4)

E.4 Capability trade-offs

Table 34 splits each change by domain. From the Tulu3 SFT checkpoint, OpenThoughts3 raises MV's mathematics average by $+21.9$ points and changes its IFEval score by -11.9, and the three other reasoning datasets move both in the same directions, so MV's gain comes mostly from mathematics and outweighs a loss in instruction following. IFEval falls significantly under thirty of the thirty-one reasoning SFT runs in the table. From the DPO checkpoint, OpenThoughts3 lowers IFEval by less (-5.3) and gains little in mathematics ($+1.4$).

Table 34: **Phase 2 changes by domain.** Every run of Tables 26, 3, 4 and 33 other than the front-loading controls of Appendix E.5, against the checkpoint it was trained from, on the mean of each domain's tasks (Table 15 gives each task's domain) and of every task but mathematics, in three panels. Brackets are 95% intervals. Bold and italic marks are corrected within each column, over all three panels. A dash marks a domain with a task that the run or its parent was not evaluated on.

(a) Every base from its Tulu3 SFT checkpoint

Training stage	Math	Code	Sci. and QA	IFEval	All but math
MixtureVitae-16K					
OpenThoughts3	**+21.9** [+18.0, +26.2]	+2.4 [−3.8, +8.6]	**+6.2** [+2.2, +10.0]	*−11.9* [−16.1, −8.1]	+1.3 [−2.2, +4.5]
AceReason	**+22.0** [+18.1, +26.0]	+7.1 [+0.6, +13.0]	**+5.9** [+1.9, +9.8]	*−12.4* [−16.3, −8.4]	+3.4 [+0.0, +6.8]
Nemotron-Cascade	**+25.4** [+21.5, +29.7]	**+7.4** [+1.3, +13.5]	**+6.9** [+3.0, +10.7]	*−9.9* [−13.6, −6.2]	**+4.4** [+0.8, +7.6]
Superior-Reasoning	**+13.5** [+10.9, +16.7]	*−8.2* [−13.9, −2.5]	+0.1 [−4.0, +4.3]	*−12.8* [−16.9, −8.5]	*−6.2* [−9.3, −2.9]
FineWeb-Edu-16K					
OpenThoughts3	−0.1 [−0.8, +0.5]	*−1.1* [−1.9, −0.3]	−3.0 [−7.2, +1.2]	*−4.7* [−8.1, −1.3]	*−2.3* [−4.0, −0.9]
AceReason	−0.1 [−0.7, +0.3]	−0.6 [−2.0, +0.9]	−0.5 [−3.8, +2.8]	*−5.2* [−8.5, −2.1]	−1.3 [−2.9, +0.1]
Nemotron-Cascade	+0.1 [−0.7, +0.9]	−0.8 [−2.1, +0.4]	+0.5 [−3.4, +4.3]	*−4.0* [−7.2, −0.9]	−0.9 [−2.5, +0.6]
Superior-Reasoning	−0.5 [−1.1, +0.0]	*−1.1* [−1.9, −0.2]	−2.1 [−5.7, +1.7]	+0.9 [−2.3, +4.5]	−1.1 [−2.5, +0.3]
Nemotron-CC-HQ-16K					
OpenThoughts3	−0.2 [−0.9, +0.4]	*−1.2* [−2.2, −0.2]	+0.2 [−2.3, +2.8]	*−10.0* [−13.5, −6.5]	*−2.2* [−3.4, −1.1]
AceReason	+0.0 [−0.9, +0.8]	−0.5 [−2.1, +1.1]	+2.1 [−0.5, +4.8]	*−10.8* [−14.0, −7.5]	−1.4 [−2.7, −0.1]
Nemotron-Cascade	+0.2 [−0.7, +1.1]	−0.7 [−2.1, +0.9]	**+3.8** [+1.1, +6.2]	*−8.2* [−11.5, −4.9]	−0.4 [−1.7, +0.8]
Superior-Reasoning	−0.1 [−0.7, +0.6]	−1.1 [−2.3, +0.2]	−0.7 [−3.2, +2.0]	*−4.5* [−7.9, −1.1]	*−1.5* [−2.8, −0.2]
Comma0.1-16K					
OpenThoughts3	−0.3 [−1.2, +0.4]	*−9.9* [−12.6, −7.2]	−1.3 [−4.4, +1.9]	*−10.1* [−13.5, −6.6]	*−7.0* [−8.8, −5.2]
AceReason	−0.4 [−1.2, +0.4]	*−6.7* [−10.3, −3.2]	−0.6 [−3.4, +2.2]	*−11.5* [−14.9, −8.0]	*−5.5* [−7.6, −3.5]
Nemotron-Cascade	+0.0 [−0.9, +0.8]	*−5.4* [−9.2, −1.8]	+0.9 [−2.1, +3.8]	*−7.8* [−11.1, −4.4]	*−3.7* [−5.8, −1.4]
Superior-Reasoning	−0.5 [−1.3, +0.2]	*−9.1* [−12.0, −6.2]	−2.0 [−4.8, +0.9]	*−6.2* [−9.9, −2.8]	*−6.3* [−8.1, −4.5]
SmolLM2-16K					
OpenThoughts3	−0.6 [−1.3, +0.1]	*−7.6* [−12.1, −3.0]	−1.7 [−3.3, +0.0]	*−12.2* [−14.9, −9.6]	*−6.4* [−8.8, −4.0]
AceReason	–	−2.1 [−6.9, +2.8]	+1.1 [−1.0, +3.3]	*−7.2* [−9.7, −4.6]	−1.9 [−4.6, +0.8]
Nemotron-Cascade	**+3.5** [+2.4, +4.5]	+1.9 [−3.2, +6.8]	−0.4 [−3.6, +2.5]	*−11.4* [−14.7, −8.3]	−1.1 [−3.9, +1.7]
Superior-Reasoning	**+3.1** [+2.2, +3.9]	*−8.7* [−13.0, −4.2]	−0.3 [−2.3, +1.7]	*−12.5* [−15.8, −9.4]	*−6.5* [−9.0, −4.1]
Qwen2.5-1.5B					
OpenThoughts3	**+10.4** [+8.1, +13.0]	*−17.6* [−22.8, −12.1]	**+3.6** [+1.1, +6.1]	*−14.3* [−17.8, −10.7]	*−9.9* [−12.8, −7.1]
Qwen3-1.7B					
OpenThoughts3	*−13.0* [−14.5, −11.5]	*−31.3* [−35.6, −26.9]	*−5.1* [−7.8, −2.3]	*−20.3* [−23.6, −17.1]	*−20.7* [−23.2, −18.1]

Table 34: (Continued) Panels (b) and (c).

(b) MixtureVitae from its 16K extended checkpoint and its DPO checkpoint

Training stage	Math	Code	Sci. and QA	IFEval	All but math
MixtureVitae-16K, from its 16K extended checkpoint					
OpenThoughts3	**+13.2** [+9.8, +16.7]	+4.3 [−1.7, +10.6]	**+5.4** [+1.4, +9.3]	*−5.7* [−8.8, −2.9]	+3.0 [−0.4, +6.6]
AceReason	**+10.6** [+8.2, +13.1]	**+8.5** [+2.4, +14.7]	**+6.4** [+2.6, +10.2]	*−6.4* [−9.3, −3.6]	**+5.3** [+1.9, +8.7]
Nemotron-Cascade	**+11.4** [+8.7, +14.3]	**+7.0** [+1.2, +13.2]	**+5.8** [+1.9, +9.8]	*−3.5* [−6.2, −0.7]	**+4.9** [+1.4, +8.2]
Superior-Reasoning	**+3.7** [+1.1, +6.1]	*−7.8* [−13.7, −1.7]	−0.6 [−4.9, +3.7]	*−5.0* [−7.9, −2.2]	*−4.9* [−8.2, −1.5]
MixtureVitae-16K, from its Tulu3 SFT + DPO checkpoint					
OpenThoughts3	+1.4 [−0.8, +3.8]	+0.3 [−5.7, +6.4]	+1.6 [−0.5, +3.6]	*−5.3* [−7.9, −2.9]	−0.2 [−3.3, +3.0]
AceReason	**+20.0** [+16.1, +24.3]	+5.3 [−0.9, +11.3]	**+7.1** [+3.2, +11.0]	*−12.5* [−16.1, −9.0]	+2.9 [−0.5, +6.3]
Nemotron-Cascade	**+23.3** [+19.6, +27.3]	+5.5 [−0.5, +11.8]	**+7.7** [+3.7, +11.4]	*−10.7* [−14.1, −7.5]	+3.5 [+0.1, +6.8]
Superior-Reasoning	**+11.9** [+9.4, +14.8]	*−9.9* [−15.5, −4.2]	−0.4 [−4.4, +3.6]	*−13.4* [−17.4, −9.6]	*−7.3* [−10.6, −3.9]

(c) Controls: the Tulu3 steps, and MixtureVitae without OpenThoughts3

Training stage	Math	Code	Sci. and QA	IFEval	All but math
MixtureVitae-16K, from its 16K extended checkpoint					
Tulu3 SFT	*−10.9* [−13.7, −8.2]	−0.1 [−6.1, +6.1]	−1.2 [−3.4, +1.0]	**+7.4** [+4.3, +10.6]	+0.8 [−2.2, +3.9]
MixtureVitae-16K, from its Tulu3 SFT checkpoint					
Tulu3 DPO	**+1.8** [+0.2, +3.4]	+1.6 [−4.4, +7.4]	+0.1 [−0.7, +0.9]	+0.6 [−1.9, +3.2]	+0.9 [−2.2, +3.9]
OT3-excluded MixtureVitae-16K, from its 16K extended checkpoint					
OpenThoughts3	**+3.5** [+1.8, +5.6]	+1.0 [−5.2, +6.9]	+0.6 [−1.1, +2.3]	*−3.3* [−5.3, −1.2]	+0.2 [−2.9, +3.2]

E.5 Enriched reference mixtures through reasoning SFT

The three enriched reference mixtures of the factorial undergo the reasoning SFT on OpenThoughts3 described in Section 3.3, each from its Tulu3 SFT checkpoint of Section 3.2, with the 30K examples, five epochs and settings of MixtureVitae's reasoning SFT. Each pairs with its pre-training dataset's matched-compute row of Table 2: the same YaRN extension of the 4K base, from the factorial's own Tulu3 SFT checkpoint. FineWeb-Edu's two Tulu3 SFT runs, with and without front-loading, ran on different numbers of GPUs with the same global batch of 128 samples per step and otherwise the same settings.

The eight checkpoints were evaluated together under Protocol B. Table 26 gives every task before and after the post-training, with SDs. The three changes, and in each pre-training dataset the difference between the gains with and without enrichment on the suite average and on MATH500, form one Benjamini–Hochberg family of nine contrasts, apart from the family of Table 27. Enrichment adds $+19.8$ points $[+16.5, +23.0]$ to the gain for Comma0.1, $+15.5$ $[+12.8, +18.3]$ for Nemotron-CC-HQ and $+14.6$ $[+11.8, +17.5]$ for FineWeb-Edu, and $+52.3$ to $+52.9$ points to the MATH500 gain, each with $q < 0.001$. These contrasts were not analyzed without the identified overlap items (Table 12).

The arms' copy of the subset is close to, but not the same as, MixtureVitae's. It holds less `formatted_text` (Appendix A.1) and 58,310 more OpenThoughts2 documents (0.46B tokens), which the overlap audit of Section 2.4 did not cover. The items that the audit identified in the subset are in these arms' pretraining data as in MixtureVitae's. AIME25 and GSM8K pass no sampling seed to the inference engine, and in these eight checkpoints' records two of their runs can share every answer, so their SDs understate how much sampled answers vary. The other Phase 2 records were not checked for this.

E.6 Language understanding across training routes

Table 35 gives the language suite after reasoning post-training, Table 37 the changes with intervals, and Table 36 MixtureVitae's changes by task. The scores are percent accuracy, normalized for ARC-C, ARC-E, HellaSwag, OBQA and PIQA, with the seed for selecting prompt examples used in

the paper's language tables. The seeds change only the few-shot examples, so the intervals describe the sample of questions, not training variation. Measured on the same questions, MV's average stays near its 16K extended checkpoint on every route: the main route, Tulu3 SFT then OT3-30K, changes it by $+0.72$ ($+0.12$ to $+1.35$) points, the full pipeline by $+0.78$ ($+0.19$ to $+1.41$), and OT3-30K without Tulu3 by $+0.68$ ($+0.30$ to $+1.10$). ARC-C is the exception: it falls on every route through Phase 1, by -2.5 points through SFT and DPO, -2.3 through the main route and -2.0 through the full pipeline. Comma0.1 loses -0.80 and -1.89 points on its two routes, both with intervals below zero, and Qwen2.5 and Qwen3 end 2.6 to 2.7 and 6.9 to 7.0 points below their base models.

Table 35: Language suite results after Phase 2 OpenThoughts3 30K post-training for the models evaluated in the OT3 experiments. MV-16K's general language understanding stays near its base level, after OT3 from the SFT checkpoint, after the full pipeline and after direct OT3 post-training from the 16K extended checkpoint alike. Each value is one run, printed with one decimal, and Avg is the mean of the eleven tasks. Nemotron-16k is the YaRN-extended 16K model of Table 18, the one of the reasoning tables.

Model	ARC-C	ARC-E	BoolQ	CSQA	COPA	HellaS	LAMBADA	MMLU	OBQA	PIQA	WinoG	Avg
MV-16k												
→SFT →OT3-30K	39.3	70.3	73.5	49.6	71.0	55.2	49.0	41.3	34.4	70.1	58.6	55.7
→SFT →DPO →OT3-30K	39.7	70.0	73.3	50.4	71.0	55.5	48.9	41.3	33.8	70.4	58.9	55.7
→OT3-30K (**without Tulu3**)	42.5	70.9	74.5	49.1	69.0	54.1	48.7	41.3	33.2	70.1	58.6	55.6
Comma-16k												
→SFT →OT3-30K	38.0	66.5	66.1	21.6	71.0	53.8	53.7	27.6	33.8	71.4	60.9	51.3
→SFT →DPO →OT3-30K	36.0	65.8	65.1	22.0	73.0	51.1	49.3	27.5	33.4	70.1	58.9	50.2
FineWeb-Edu-16k												
→SFT →OT3-30K	47.4	78.5	69.5	21.4	83.0	66.7	54.0	26.8	40.8	77.1	63.2	57.1
Nemotron-16k												
→SFT →OT3-30K	51.6	80.7	77.9	52.2	84.0	72.0	57.2	45.5	43.0	78.6	65.4	64.4
→SFT →DPO →OT3-30K	51.8	80.7	77.8	51.9	84.0	72.0	57.2	45.5	43.0	78.8	65.4	64.4
Qwen2.5-1.5B												
→SFT →OT3-30K	50.8	78.1	76.7	72.8	84.0	67.3	46.3	58.8	41.2	76.9	63.1	65.1
→SFT →DPO →OT3-30K	50.9	78.0	76.9	72.6	84.0	67.5	46.2	58.8	40.6	77.0	63.3	65.1
Qwen3-1.7B												
→SFT →OT3-30K	49.1	76.1	74.5	65.0	81.0	59.3	42.8	53.0	38.6	74.0	60.5	61.3
→SFT →DPO →OT3-30K	49.0	75.8	74.8	64.5	81.0	59.5	43.0	53.1	38.6	73.6	60.5	61.2
SmolLM2-1.7B-16k												
→SFT →OT3-30K	52.6	80.8	70.4	58.6	88.0	70.9	64.1	44.4	43.0	77.5	66.1	65.1
SmolLM2-1.7B-Instruct-16k												
→OT3-30K	51.4	79.4	75.9	62.2	83.0	71.1	58.6	47.4	44.4	77.9	65.7	65.2

Table 36: MixtureVitae: change on each language task at each training step, in points, and the 11-task average with two decimals. Each task cell is the estimated change, and Table 37 gives the interval of the average. Bold marks a change with $q < 0.05$ after the Benjamini-Hochberg correction over the model's per-task changes. Same records and conventions as Table 37.

Change	ARC-C	ARC-E	BoolQ	CSQA	COPA	HellaS	LAMBADA	MMLU	OBQA	PIQA	WinoG	Avg
Base, 4K → base, 16K	+1.4	+0.5	−0.9	−2.5	−4.0	**−1.4**	**+1.2**	−0.9	−2.4	−0.7	+0.2	−0.87
Base, 16K → SFT	**−2.5**	−1.2	+0.3	+1.6	+2.0	**+1.8**	+0.6	+0.8	**+2.4**	+0.1	−0.5	+0.49
SFT → SFT, DPO	+0.0	**+0.3**	−0.1	−0.1	+0.0	**+0.1**	+0.2	+0.1	+0.2	−0.1	−0.1	+0.06
Base, 16K → OT3-30K	+0.9	−0.5	**+1.9**	+0.3	+2.0	**+1.0**	−0.7	+0.2	+1.4	+0.4	+0.7	+0.68
SFT → SFT, OT3-30K	+0.2	+0.1	+0.6	−0.7	+2.0	+0.4	**−1.0**	−0.6	+0.2	+0.3	+1.2	+0.23
SFT, DPO → SFT, DPO, OT3-30K	+0.5	−0.5	+0.4	+0.1	+2.0	**+0.4**	**−1.3**	**−0.7**	−0.6	+0.6	+1.5	+0.23
Base, 16K ⇒ SFT, DPO	**−2.5**	−0.9	+0.3	+1.5	+2.0	**+1.9**	+0.8	+0.9	**+2.6**	+0.1	−0.6	+0.56
Base, 16K ⇒ SFT, OT3-30K	**−2.3**	−1.1	+0.9	+0.8	+4.0	**+2.1**	−0.4	+0.2	**+2.6**	+0.4	+0.7	+0.72
Base, 16K ⇒ SFT, DPO, OT3-30K	−2.0	**−1.4**	+0.7	+1.6	+4.0	**+2.3**	−0.5	+0.3	+2.0	+0.7	+0.9	+0.78

E.7 How the answers end

A score can fall because answers are wrong or because they do not finish. Panel (a) of Table 38 gives the share of answers near the generation limit of Table 15, and panel (b) the share with no final answer or, for code, no code block. After OpenThoughts3, 53.2 to 100.0% of the matched-compute references' answers are near the limit on every task but GSM8K, against at most 12.2% at their Tulu3 SFT checkpoints, so their declines coincide with answers that do not finish. MV's answers are near the limit less often, on 15.9% of MATH500. Qwen3 shows the other pattern: at most 0.1% of its

Table 37: Changes in the 11-task language average at each training step, with 95% intervals. Each change is a checkpoint minus the checkpoint it was trained from, except those marked ⇒, which span several training steps. The intervals resample the questions within each task, with 2,000 replicates, and q is Benjamini–Hochberg corrected over the average changes of each model. The last column averages each question over the three seeds for selecting prompt examples. FineWeb-Edu's 16K extended checkpoint is the one its 16K SFT checkpoint was trained from. Nemotron-CC-HQ's checkpoints are those of Table 35, the YaRN-extended model of the reasoning tables.

Model	Change	From	To	Change [95% CI]	q	Three seeds
MixtureVitae	Base, 4K → base, 16K	55.82	54.96	−0.87 [−1.56, −0.15]	0.042	−0.96 [−1.61, −0.24]
	Base, 16K → SFT	54.96	55.45	+0.49 [−0.12, +1.13]	0.167	+0.50 [−0.08, +1.10]
	SFT → SFT, DPO	55.45	55.51	+0.06 [−0.01, +0.13]	0.152	+0.10 [+0.04, +0.16]
	Base, 16K → OT3-30K	54.96	55.64	+0.68 [+0.30, +1.10]	0.004	+0.68 [+0.33, +1.03]
	SFT → SFT, OT3-30K	55.45	55.68	+0.23 [−0.23, +0.66]	0.344	+0.34 [−0.09, +0.77]
	SFT, DPO → SFT, DPO, OT3-30K	55.51	55.74	+0.23 [−0.24, +0.67]	0.344	+0.27 [−0.17, +0.69]
	Base, 16K ⇒ SFT, DPO	54.96	55.51	+0.56 [−0.07, +1.18]	0.152	+0.60 [+0.01, +1.20]
	Base, 16K ⇒ SFT, OT3-30K	54.96	55.68	+0.72 [+0.12, +1.35]	0.042	+0.84 [+0.26, +1.42]
	Base, 16K ⇒ SFT, DPO, OT3-30K	54.96	55.74	+0.78 [+0.19, +1.41]	0.042	+0.86 [+0.28, +1.44]
Comma0.1	Base, 4K → base, 16K	50.05	52.10	+2.05 [+1.38, +2.76]	0.002	+2.00 [+1.37, +2.67]
	Base, 16K → SFT	52.10	51.15	−0.95 [−1.58, −0.29]	0.005	−0.82 [−1.41, −0.21]
	SFT → SFT, DPO	51.15	51.34	+0.20 [+0.04, +0.42]	0.007	+0.21 [+0.08, +0.42]
	SFT → SFT, OT3-30K	51.15	51.30	+0.16 [−0.42, +0.71]	0.613	+0.21 [−0.32, +0.70]
	SFT, DPO → SFT, DPO, OT3-30K	51.34	50.21	−1.14 [−1.79, −0.47]	0.004	−1.30 [−1.89, −0.69]
	Base, 16K ⇒ SFT, DPO	52.10	51.34	−0.75 [−1.37, −0.13]	0.018	−0.61 [−1.18, −0.04]
	Base, 16K ⇒ SFT, OT3-30K	52.10	51.30	−0.80 [−1.45, −0.18]	0.010	−0.61 [−1.26, −0.06]
	Base, 16K ⇒ SFT, DPO, OT3-30K	52.10	50.21	−1.89 [−2.55, −1.28]	0.002	−1.91 [−2.52, −1.33]
Nemotron-CC-HQ	Base, 4K → base, 16K	64.17	64.06	−0.11 [−0.56, +0.35]	0.751	−0.13 [−0.59, +0.30]
	Base, 16K → SFT	64.06	64.38	+0.31 [−0.17, +0.82]	0.479	+0.38 [−0.10, +0.88]
	SFT → SFT, DPO	64.38	64.54	+0.17 [+0.04, +0.30]	0.068	+0.11 [+0.00, +0.22]
	SFT → SFT, OT3-30K	64.38	64.37	+0.00 [−0.48, +0.48]	0.965	+0.00 [−0.48, +0.44]
	SFT, DPO → SFT, DPO, OT3-30K	64.54	64.38	−0.16 [−0.67, +0.32]	0.699	−0.08 [−0.56, +0.38]
	Base, 16K ⇒ SFT, DPO	64.06	64.54	+0.48 [−0.01, +0.97]	0.230	+0.49 [+0.01, +0.98]
	Base, 16K ⇒ SFT, OT3-30K	64.06	64.37	+0.31 [−0.25, +0.89]	0.479	+0.38 [−0.14, +0.90]
	Base, 16K ⇒ SFT, DPO, OT3-30K	64.06	64.38	+0.32 [−0.24, +0.90]	0.479	+0.42 [−0.10, +0.93]
FineWeb-Edu	Base, 4K → SFT, 4K	56.75	57.37	+0.62 [+0.07, +1.15]	0.081	+0.79 [+0.26, +1.27]
	SFT, 4K → SFT, DPO, 4K	57.37	57.46	+0.09 [+0.03, +0.15]	0.012	+0.08 [+0.04, +0.14]
	Base, 4K → base, 16K	56.75	57.30	+0.55 [+0.00, +1.10]	0.097	+0.63 [+0.12, +1.17]
	Base, 16K → SFT	57.30	56.78	−0.52 [−1.11, +0.04]	0.105	−0.47 [−1.02, +0.07]
	SFT → SFT, DPO	56.78	56.84	+0.06 [+0.01, +0.11]	0.070	+0.06 [+0.02, +0.10]
	SFT → SFT, OT3-30K	56.78	57.12	+0.34 [−0.13, +0.81]	0.164	+0.31 [−0.14, +0.77]
	Base, 16K ⇒ SFT, DPO	57.30	56.84	−0.46 [−1.05, +0.09]	0.139	−0.41 [−0.96, +0.13]
	Base, 16K ⇒ SFT, OT3-30K	57.30	57.12	−0.18 [−0.80, +0.46]	0.568	−0.15 [−0.77, +0.45]

answers are near the limit, while the grader finds no final answer in 76.2 to 95.6% of its answers on AIME24, AIME25, AMC23 and MATH500. On IFEval the share near the limit rises under the post-training, from 3.7% to 17.4% for MV and to 64.9 to 79.9% for the matched-compute references. Through Phase 1, MV's answers are near the limit less often than at its 16K extended checkpoint: under Protocol B, the Phase 1 decline is not accompanied by an increase in outputs near the generation limit. The share of answers without an extracted final answer does rise, from 6.9% to 15.3% on MATH500 and from 3.9% to 23.0% on JEEBench.

Table 38: **How the Phase 2 answers end.** Shares of the answers, in percent over the three runs, for the checkpoints of Tables 26 and 3, excluding the additional front-loading controls of Appendix E.5. (a) Answers whose length, counted with each checkpoint's own tokenizer, is at least 95% of the generation limit of Table 15, since re-encoding a sampled answer can give a few percent fewer tokens than were generated. This is a length criterion, not the reason a generation stopped. (b) For the mathematics and science tasks, answers from which the grader extracts no final answer, so that the answer is scored wrong. For the code tasks, answers with no code block, which the checker then runs whole, so that they are not scored wrong for that reason alone. GSM8K and GPQA-Diamond are left out of (b), because their graders take an answer from every response, and IFEval has no answer to extract. A dash marks a cell whose records do not hold the extracted answer or code.

(a) Answers near the generation limit

Model	A24	A25	AMC	M500	GSM8K	HE	LCB	MBPP	GPQA	JEE	IFE
MixtureVitae-16K											
16K extended checkpoint	21.1	13.3	14.2	6.1	0.0	0.0	9.8	0.3	0.0	0.1	1.1
→ OT3-30K (without Tulu3)	77.8	57.8	40.0	18.1	0.3	38.2	80.6	42.9	32.2	46.9	21.5
→ SFT	0.0	10.0	1.7	0.3	0.5	0.9	0.2	0.2	0.0	0.5	3.7
→ SFT → OT3-30K	70.0	65.6	36.7	15.9	3.8	22.8	80.6	42.2	31.3	47.0	17.4
→ SFT → DPO	0.0	7.8	0.0	0.3	0.5	0.2	0.4	0.1	0.0	0.5	3.1
→ SFT → DPO → OT3-30K	24.4	31.1	9.2	12.5	1.3	6.9	4.4	1.7	15.7	12.2	24.0
FineWeb-Edu-16K, Tulu3 SFT	3.3	12.2	0.0	2.2	1.4	2.4	5.3	1.8	0.2	3.2	1.9
→ OT3-30K	86.7	91.1	84.2	82.0	14.5	97.4	99.6	96.7	84.7	83.1	79.9
Nemotron-CC-HQ-16K, Tulu3 SFT	4.4	12.2	0.8	1.3	1.7	1.8	2.6	0.7	0.0	1.9	4.7
→ OT3-30K	86.7	100.0	73.3	70.3	19.4	87.6	98.1	92.8	53.2	62.1	64.9
Comma0.1-16K, Tulu3 SFT	3.3	11.1	3.3	1.8	2.0	0.0	0.9	0.5	0.0	1.7	3.2
→ OT3-30K	85.6	100.0	86.7	79.0	11.9	91.1	99.3	95.1	68.0	79.0	71.9
SmolLM2-16K, Tulu3 SFT	1.1	7.8	0.0	0.8	3.8	0.6	0.5	0.3	0.0	1.5	3.5
→ OT3-30K	77.8	91.1	72.5	61.1	25.2	27.8	61.7	47.5	3.4	74.2	6.9
Qwen2.5-1.5B, Tulu3 SFT	2.2	1.1	1.7	0.2	0.3	0.0	0.4	0.1	0.0	0.5	3.6
→ OT3-30K	64.4	73.3	48.3	57.1	8.6	90.2	93.6	84.0	63.3	42.6	91.3
Qwen3-1.7B, Tulu3 SFT	1.1	1.1	0.0	0.4	0.3	0.2	0.3	0.0	0.0	0.3	3.3
→ OT3-30K	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
OT3-excluded MixtureVitae-16K											
Base without OT3, before reasoning post-training	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.3
OT3 removed from pretraining, → OT3-30K	0.0	0.0	0.0	0.0	0.4	0.0	26.1	0.2	0.0	0.0	2.5

(b) No final answer, or no code block

Model	A24	A25	AMC	M500	HE	LCB	MBPP	JEE
MixtureVitae-16K								
16K extended checkpoint	18.9	16.7	15.0	6.9	0.0	5.8	0.0	3.9
→ OT3-30K (without Tulu3)	71.1	52.2	35.0	15.2	4.3	32.6	5.4	39.7
→ SFT	24.4	16.7	19.2	15.3	–	0.0	–	23.0
→ SFT → OT3-30K	65.6	58.9	30.8	13.0	5.3	33.6	7.0	38.9
→ SFT → DPO	23.3	21.1	11.7	13.8	0.0	0.0	0.0	20.6
→ SFT → DPO → OT3-30K	16.7	15.6	4.2	5.8	0.0	0.9	0.1	4.7
FineWeb-Edu-16K, Tulu3 SFT	58.9	63.3	55.0	58.2	0.4	0.7	0.9	43.9
→ OT3-30K	77.8	90.0	70.8	71.3	16.1	60.5	13.2	71.0
Nemotron-CC-HQ-16K, Tulu3 SFT	45.6	51.1	33.3	44.3	0.0	74.3	8.8	45.8
→ OT3-30K	75.6	–	66.7	60.4	0.0	98.2	88.7	49.1
Comma0.1-16K, Tulu3 SFT	47.8	23.3	50.0	39.6	1.2	0.1	0.0	15.8
→ OT3-30K	80.0	97.8	78.3	67.1	9.1	53.3	5.0	66.9
SmolLM2-16K, Tulu3 SFT	31.1	23.3	22.5	22.8	0.2	0.0	0.0	10.6
→ OT3-30K	81.1	94.4	65.8	54.1	0.8	6.4	3.4	59.8
Qwen2.5-1.5B, Tulu3 SFT	–	–	–	11.8	0.0	0.0	0.1	27.1
→ OT3-30K	42.2	55.6	20.8	21.3	3.0	23.6	2.3	19.3
Qwen3-1.7B, Tulu3 SFT	5.6	1.1	12.5	16.2	0.0	0.0	0.0	10.5
→ OT3-30K	95.6	90.0	85.0	76.2	0.8	36.7	7.1	79.4
OT3-excluded MixtureVitae-16K								
Base without OT3, before reasoning post-training	–	–	–	6.1	0.0	0.0	0.0	4.5
OT3 removed from pretraining, → OT3-30K	13.3	8.9	3.3	6.5	0.0	12.3	0.1	19.4

F Auxiliary experiments

Four smaller experiments, each with its reasoning and language results: the mixture composition ablation of Section 3.2, the historical filtering, a 0.4B comparison and prompt token loss masking. Every reasoning table of this appendix is under Protocol A (Table 15), as the mean and sample SD over decoding seeds 42, 1234 and 2025. Every language table gives one run, with one decimal, and Avg is the mean of the eleven tasks.

F.1 Mixture composition at a 100B budget

At a 100B-token budget, the full MixtureVitae mixture (MV-100) is compared with two variants that remove either the instruction and reasoning subset (MV-noinstruct) or the curated web subset (MV-noweb), each followed by the same 16K extension and Phase 1 pipeline (Tables 39 and 40). Removing the instruction and reasoning subset lowers the reasoning suite average at every stage and context length, by -3.7 to -12.9 points with every interval below zero (Table 41), and after SFT at 16K the largest losses are on GSM8K, HumanEval and MBPP. Removing the curated web subset is never significantly worse than the full mixture, and it is ahead at four of the six stages and context lengths. On the language suite, removing the subset costs 2.8 to 4.4 points that Phase 1 does not close, while removing the web subset stays within 1.6 points of the full mixture. These results agree with the composition dependence seen at 300B. They do not show how the variants would respond to other post-training recipes or budgets.

F.2 Effect of the historical filtering

This control compares the 300B MixtureVitae checkpoints of the main paper, trained on data after the historical 13-gram filtering, with checkpoints trained on the same mixture without it, through the same pipeline (Tables 42 and 43). It compares training with and without that filtering procedure, not with data disjoint from the benchmarks: the filtered data still hold the items of Appendix A.3. The largest change is at the 4K pretrained base, where the filtering lowers the reasoning suite average by 5.1 points. The 4K post-trained scores move by at most 1.2 points, and the 16K post-trained scores are higher after the filtering, by 2.7 points after SFT and 2.9 after DPO. On the language suite, the means of the two versions differ by less than one point (55.4 to 55.6 without the filtering, 55.0 to 55.8 with it). The filtering therefore does not account for the post-training differences this paper reports. The control covers MixtureVitae only, since the reference datasets are used as released (Appendix A.3).

F.3 Smaller scale check at 0.4B

This check compares a 0.4B MV model with Baguettotron, a 0.32B model trained on the PleIAs SYNTH mix [9], under the same Phase 1 pipeline (Tables 44 and 45). The two differ in size and in data, so it checks the direction of the effect against another model, not a second scale of the main comparison. The benchmark overlap of the SYNTH mix is not documented, and MixtureVitae's pretraining pool holds benchmark items (Appendix A.3), so part of either model's score may reflect overlap with the benchmarks. On the reasoning suite, Baguettotron stays at 5.3 to 7.7 across all stages and context lengths, and its HumanEval and MBPP scores are zero in all but one row each. MV (0.4B) reaches 14.1 after SFT and DPO at 4K, with 29.3 on HumanEval and 21.6 on MBPP. On the language suite MV is ahead as well at 4K, 44.1 to 44.3 after Phase 1 against 42.3 to 43.1, most on LAMBADA, HellaSwag and CommonsenseQA. Under this recipe, Phase 1 gives Baguettotron no code performance. Other recipes are not tested.

Table 39: Reasoning suite of the 100B-token mixture composition ablation: the full MixtureVitae mixture (MV) and the variants without the instruction and reasoning subset (MV-noinstruct) and without the curated web subset (MV-noweb), before and after Phase 1.

(a) Mathematics

Model	A24	A25	AMC	M500	GSM8K
MV-noinstruct	0.0±0.00	0.0±0.00	0.0±0.00	0.4±0.00	2.1±0.00
→ SFT	0.0±0.00	0.0±0.00	0.8±1.44	2.3±0.12	3.9±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	4.2±1.44	2.6±0.35	4.0±0.00
MV-noinstruct-16k	0.0±0.00	0.0±0.00	0.0±0.00	1.9±0.31	4.2±0.00
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	1.8±0.20	4.5±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	0.0±0.00	2.4±0.35	4.8±0.00
MV-noweb	1.1±1.92	0.0±0.00	14.2±3.82	27.3±0.64	42.7±0.00
→ SFT	0.0±0.00	0.0±0.00	7.5±0.00	13.0±0.20	35.8±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	12.5±0.00	16.9±0.46	36.6±0.00
MV-noweb-16k	0.0±0.00	3.3±0.00	10.0±0.00	34.0±0.69	44.6±0.00
→ SFT	3.3±0.00	0.0±0.00	7.5±0.00	22.8±0.53	33.4±0.00
→ SFT → DPO	5.6±1.92	0.0±0.00	10.0±0.00	22.9±1.17	34.3±0.00
MV	0.0±0.00	0.0±0.00	2.5±0.00	14.3±0.42	42.7±0.00
→ SFT	0.0±0.00	0.0±0.00	5.8±1.44	10.3±0.23	37.8±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	6.7±1.44	12.0±0.53	38.7±0.00
MV-16k	0.0±0.00	0.0±0.00	9.2±1.44	33.9±0.12	40.3±0.00
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	5.3±0.23	36.6±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	0.0±0.00	5.2±0.00	37.1±0.00

(b) Code, science and instruction following, and the average

Model	HE	LCB	MBPP	GPQA	JEE	IFE	Avg
MV-noinstruct	0.0±0.00	0.0±0.00	14.5±0.23	23.1±0.29	0.0±0.00	30.8±0.00	6.4±0.02
→ SFT	12.8±0.00	0.5±0.00	14.9±0.23	23.1±0.29	2.5±0.10	35.0±0.30	8.7±0.21
→ SFT → DPO	14.0±0.00	0.1±0.16	14.3±0.12	22.4±0.29	3.9±0.34	36.7±0.39	9.3±0.05
MV-noinstruct-16k	10.4±0.00	0.0±0.00	12.3±0.12	25.4±0.29	5.0±0.26	37.0±0.20	8.8±0.07
→ SFT	15.0±0.35	0.2±0.16	15.9±0.12	25.9±0.77	2.1±0.15	40.4±0.08	9.6±0.09
→ SFT → DPO	14.4±0.70	0.0±0.00	15.5±0.12	26.3±0.00	2.0±0.25	41.9±0.13	9.8±0.10
MV-noweb	43.7±0.35	4.8±0.41	33.0±0.00	19.9±0.77	3.1±0.25	31.3±0.26	20.1±0.43
→ SFT	44.1±0.35	0.9±0.16	37.3±0.31	26.3±0.00	3.2±0.34	40.5±0.20	19.0±0.04
→ SFT → DPO	44.1±0.35	0.6±0.16	37.1±0.23	25.6±0.29	4.3±0.28	40.8±0.23	19.9±0.11
MV-noweb-16k	39.0±0.00	2.1±0.41	36.4±0.00	21.7±0.00	6.0±0.46	41.4±0.27	21.7±0.09
→ SFT	41.3±0.35	1.9±0.27	34.7±0.12	23.1±0.29	3.6±0.32	43.5±0.49	19.6±0.09
→ SFT → DPO	39.8±0.35	1.6±0.27	35.7±0.12	23.2±0.00	3.7±0.07	43.8±0.91	20.1±0.18
MV	0.0±0.00	0.0±0.00	0.0±0.00	22.4±2.59	3.0±0.39	26.7±0.00	10.1±0.28
→ SFT	41.9±0.35	2.7±0.00	35.6±0.20	17.2±0.51	2.8±0.49	40.9±0.40	17.7±0.09
→ SFT → DPO	41.5±0.00	2.5±0.16	36.2±0.20	19.5±0.29	1.8±0.17	43.3±0.15	18.4±0.15
MV-16k	37.8±0.00	3.8±0.27	35.1±0.31	24.9±0.29	8.6±0.18	44.0±0.40	21.6±0.10
→ SFT	38.4±0.00	0.5±0.16	34.8±0.00	24.2±0.00	4.1±0.06	43.8±0.15	17.1±0.02
→ SFT → DPO	40.9±0.00	0.4±0.16	35.5±0.42	23.7±0.00	3.7±0.34	45.0±0.20	17.4±0.05

Table 40: Language suite of the 100B-token mixture composition ablation, before and after Phase 1.

Model	ARC-C	ARC-E	BoolQ	CSQA	COPA	HellaS	LAMBADA	MMLU	OBQA	PIQA	WinoG	Avg
MV-noinstruct	32.1	60.6	58.3	18.3	66.0	45.7	45.6	25.1	32.6	68.2	54.6	46.1
→SFT	32.1	60.7	57.2	19.3	66.0	46.5	46.6	25.1	32.8	67.9	55.1	46.3
→SFT →DPO	32.3	60.8	57.4	19.2	66.0	46.5	46.6	25.2	32.8	67.8	55.1	46.3
MV-noinstruct-16k	32.7	60.3	63.3	19.5	69.0	45.6	47.0	24.8	32.6	66.8	54.7	46.9
→SFT	31.3	59.8	60.7	18.3	67.0	46.3	46.7	24.9	32.8	66.9	55.7	46.4
→SFT →DPO	31.3	59.8	60.7	17.9	67.0	46.5	46.9	24.9	32.4	66.9	55.9	46.4
MV-noweb	34.5	64.7	66.6	34.6	67.0	46.2	38.7	31.4	33.0	67.5	56.0	49.1
→SFT	32.8	63.0	64.8	35.2	69.0	46.7	40.6	33.0	31.8	67.2	55.2	49.0
→SFT →DPO	32.8	63.0	64.8	35.5	69.0	46.8	40.7	33.0	32.0	67.2	55.0	49.1
MV-noweb-16k	35.6	65.2	69.8	36.6	70.0	45.4	40.5	32.1	32.4	67.1	55.8	50.0
→SFT	32.8	63.3	66.4	34.2	69.0	46.4	41.0	33.1	32.8	67.0	56.2	49.3
→SFT →DPO	32.9	63.5	66.6	34.2	69.0	46.5	41.2	33.1	33.2	67.0	56.4	49.4
MV	34.4	64.6	70.9	31.4	67.0	48.5	43.8	32.5	33.8	68.4	57.7	50.3
→SFT	33.9	63.8	70.5	31.9	69.0	49.3	45.0	33.0	33.6	68.7	57.7	50.6
→SFT →DPO	33.9	64.0	70.7	32.0	70.0	49.4	45.0	33.1	33.4	68.7	57.5	50.7
MV-16k	34.7	65.0	67.7	29.9	67.0	47.1	44.6	32.6	32.8	67.9	57.1	49.7
→SFT	33.4	64.2	69.2	29.2	68.0	48.8	45.0	32.4	32.6	68.3	58.7	50.0
→SFT →DPO	33.4	64.3	69.4	29.6	68.0	49.0	45.2	32.4	32.8	68.5	58.9	50.1

Table 41: Statistics for the 100B mixture composition ablation of Section 3.2, under Protocol A like Table 39. Differences are the ablation minus the full 100B mixture (MV-100), in points, so negative values mean the ablation is worse. The method is that of Table 19: each item is scored as its mean over the three decoding seeds, and the 95% intervals come from the paired item-level bootstrap of Appendix B.2, with items resampled within each benchmark, IFEval's prompts with all their instructions, and 20,000 replicates. q is the Benjamini–Hochberg adjusted q-value, computed within each column for the per-task rows and over the twelve suite average rows. n is the number of scored items: HumanEval counts its 164 Python problems and IFEval the 769 instructions of its 500 prompts. A dash marks a benchmark on which both models score zero on every item, which is not tested.

	MV-noweb − MV-100			MV-noinstruct − MV-100		
Setting or task	Δ pp	95% CI	q	Δ pp	95% CI	q
Eleven-task suite average						
Base, 4K	$+10.0$	$[+8.6, +11.4]$	< 0.001	-3.7	$[-4.7, -2.7]$	< 0.001
Tulu3 SFT, 4K	$+1.2$	$[+0.0, +2.5]$	0.060	-9.0	$[-10.3, -7.8]$	< 0.001
SFT→DPO, 4K	$+1.5$	$[+0.1, +3.0]$	0.047	-9.1	$[-10.5, -7.7]$	< 0.001
Base, 16K	$+0.1$	$[-1.4, +1.6]$	0.932	-12.9	$[-14.3, -11.5]$	< 0.001
Tulu3 SFT, 16K	$+2.5$	$[+1.1, +3.9]$	< 0.001	-7.4	$[-8.6, -6.3]$	< 0.001
SFT→DPO, 16K	$+2.7$	$[+1.2, +4.2]$	< 0.001	-7.7	$[-8.8, -6.5]$	< 0.001
Per task, after Tulu3 SFT at 16K						
AIME24 (n=30)	$+3.3$	$[+0.0, +10.0]$	0.916	—	—	—
AIME25 (n=30)	—	—	—	—	—	—
AMC23 (n=40)	$+7.5$	$[+0.0, +17.5]$	0.210	—	—	—
MATH500 (n=500)	$+17.5$	$[+13.8, +21.4]$	< 0.001	-3.5	$[-5.5, -1.5]$	< 0.001
GSM8K (n=1319)	-3.2	$[-6.0, -0.3]$	0.155	-32.1	$[-34.7, -29.4]$	< 0.001
HumanEval (n=164)	$+2.8$	$[-3.9, +9.6]$	0.838	-23.4	$[-30.5, -16.5]$	< 0.001
LiveCodeBench (n=369)	$+1.4$	$[+0.0, +3.0]$	0.170	-0.3	$[-1.1, +0.4]$	0.677
MBPP (n=500)	-0.1	$[-3.4, +3.1]$	0.942	-18.9	$[-23.1, -14.9]$	< 0.001
GPQA-Diamond (n=198)	-1.2	$[-7.1, +4.7]$	0.916	$+1.7$	$[-6.7, +9.9]$	0.713
JEEBench (n=515)	-0.5	$[-2.5, +1.4]$	0.916	-2.1	$[-3.9, -0.3]$	0.036
IFEval (n=769)	-0.3	$[-3.7, +3.2]$	0.942	-3.4	$[-6.9, +0.1]$	0.079

Table 42: Reasoning suite of the 300B MixtureVitae checkpoints trained without and with the historical 13-gram filtering of their data, before and after Phase 1: a check of that procedure, not of benchmark-disjoint data (Appendix A.3).

(a) Mathematics

	Model	A24	A25	AMC	M500	GSM8K
Without filtering	MV	0.0±0.00	0.0±0.00	13.3±1.44	13.7±0.23	53.6±0.00
	→ SFT	3.3±0.00	0.0±0.00	10.0±2.50	19.7±0.81	50.0±0.00
	→ SFT → DPO	1.1±1.92	0.0±0.00	10.0±0.00	20.3±0.42	51.1±0.00
	MV-16k	1.1±1.92	0.0±0.00	16.7±2.89	38.6±0.35	62.0±0.00
	→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	25.3±0.12	51.8±0.00
	→ SFT → DPO	0.0±0.00	0.0±0.00	0.0±0.00	27.3±0.42	52.2±0.00
With filtering	MV	0.0±0.00	7.8±5.09	5.0±0.00	17.8±0.72	54.1±0.00
	→ SFT	0.0±0.00	0.0±0.00	12.5±0.00	29.5±0.23	49.9±0.00
	→ SFT → DPO	0.0±0.00	3.3±0.00	10.8±1.44	30.5±0.12	50.2±0.00
	MV-16k	1.1±1.92	3.3±0.00	23.3±8.04	47.0±0.53	60.0±0.00
	→ SFT	3.3±0.00	3.3±0.00	15.0±0.00	35.0±0.53	51.6±0.00
	→ SFT → DPO	3.3±0.00	1.1±1.92	17.5±0.00	37.0±0.80	52.6±0.00

(b) Code, science and instruction following, and the average

	Model	HE	LCB	MBPP	GPQA	JEE	IFE	Avg
Without filtering	MV	46.5±0.35	6.3±0.41	43.5±0.42	25.1±1.27	5.6±0.29	36.4±0.08	22.2±0.14
	→ SFT	51.4±0.35	7.0±0.31	42.7±0.12	26.8±0.87	3.0±0.40	49.3±0.27	23.9±0.21
	→ SFT → DPO	56.3±0.70	7.5±0.16	43.7±0.12	27.8±0.00	3.3±0.23	51.8±0.64	24.8±0.22
	MV-16k	51.0±0.35	7.7±0.56	42.0±0.35	27.9±1.27	13.4±0.59	47.7±0.38	28.0±0.29
	→ SFT	52.2±0.35	5.7±0.00	43.5±0.12	24.6±0.29	4.1±0.47	56.0±0.38	23.9±0.08
	→ SFT → DPO	52.8±0.35	6.6±0.16	42.5±0.42	23.4±0.77	3.4±0.20	56.5±0.78	24.1±0.03
With filtering	MV	13.6±0.93	7.7±0.16	19.5±0.12	24.4±0.77	2.9±0.51	34.7±0.13	17.0±0.38
	→ SFT	44.3±0.70	1.2±0.16	39.9±0.23	20.7±0.87	3.7±0.55	48.0±0.08	22.7±0.14
	→ SFT → DPO	47.4±0.35	1.0±0.16	41.2±0.20	21.2±0.00	4.5±0.34	49.8±0.08	23.6±0.17
	MV-16k	54.1±0.35	11.5±0.16	40.1±0.23	23.4±0.29	8.0±0.24	47.7±0.34	29.1±0.69
	→ SFT	54.1±0.35	9.7±0.16	40.4±0.35	23.4±0.29	5.1±0.27	51.9±0.46	26.6±0.11
	→ SFT → DPO	52.6±0.70	9.4±0.31	40.3±0.12	23.4±0.29	6.2±0.20	52.6±0.60	26.9±0.04

Table 43: Language suite of the 300B MixtureVitae checkpoints trained without and with the historical 13-gram filtering of their data, before and after Phase 1.

Base	Model	ARC-C	ARC-E	BoolQ	CSQA	COPA	HellaS	LAMBADA	MMLU	OBQA	PIQA	WinoG	Avg
Without filtering	MV	39.6	71.3	75.0	48.7	73.0	53.9	48.2	38.0	35.2	69.6	58.3	55.5
	→SFT	39.0	70.4	74.6	46.8	71.0	54.9	49.8	39.8	35.0	70.1	58.3	55.4
	→SFT →DPO	39.1	70.5	74.7	47.0	71.0	55.1	49.9	39.9	35.0	70.1	58.3	55.5
	MV-16k	39.0	71.0	74.6	50.1	72.0	54.3	48.7	37.7	35.6	70.8	57.7	55.6
	→SFT	38.8	70.5	74.4	48.4	72.0	54.8	49.6	38.0	34.6	70.5	58.7	55.5
	→SFT →DPO	39.1	70.7	74.5	48.2	72.0	55.0	49.9	38.1	35.0	70.5	58.7	55.6
With filtering	MV	40.3	70.9	73.5	51.4	71.0	54.5	48.1	41.9	34.2	70.4	57.8	55.8
	→SFT	38.8	70.2	73.2	49.6	70.0	55.7	49.6	42.7	34.4	70.0	58.2	55.7
	→SFT →DPO	38.8	70.2	73.2	50.0	71.0	55.8	49.8	42.7	34.4	70.0	58.2	55.8
	MV-16k	41.6	71.4	72.6	48.8	67.0	53.1	49.4	41.1	31.8	69.7	57.9	55.0
	→SFT	39.2	70.2	72.9	50.4	69.0	54.9	50.0	41.9	34.2	69.9	57.5	55.5
	→SFT →DPO	39.2	70.5	72.9	50.3	69.0	55.0	50.2	42.0	34.4	69.8	57.4	55.5

Table 44: Reasoning suite of the 0.4B comparison between MV and Baguettotron, before and after Phase 1.

(a) Mathematics

Model	A24	A25	AMC	M500	GSM8K
Baguettotron	0.0±0.00	0.0±0.00	0.0±0.00	0.5±0.31	2.4±0.00
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	2.1±0.23	8.0±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	3.3±1.44	2.8±0.20	7.7±0.00
Baguettotron 16k	0.0±0.00	0.0±0.00	0.0±0.00	1.8±0.20	5.4±0.00
→ SFT	0.0±0.00	0.0±0.00	5.0±0.00	2.0±0.20	8.3±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	2.5±0.00	2.9±0.23	8.3±0.00
MV (0.4B)	0.0±0.00	0.0±0.00	2.5±0.00	1.0±0.00	29.2±0.00
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	9.1±0.12	23.4±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	2.5±0.00	11.7±0.23	23.4±0.00
MV-16k (0.4B)	0.0±0.00	0.0±0.00	8.3±1.44	19.6±0.00	29.4±0.00
→ SFT	0.0±0.00	0.0±0.00	2.5±0.00	9.4±0.20	23.6±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	5.0±0.00	10.1±0.42	25.0±0.00

(b) Code, science and instruction following, and the average

Model	HE	LCB	MBPP	GPQA	JEE	IFE	Avg
Baguettotron	0.0±0.00	0.0±0.00	0.0±0.00	21.9±2.10	0.2±0.00	33.7±0.60	5.3±0.12
→ SFT	0.0±0.00	0.0±0.00	0.0±0.00	26.3±0.00	1.7±0.06	36.2±0.27	6.8±0.02
→ SFT → DPO	0.0±0.00	0.0±0.00	0.0±0.00	25.3±0.00	1.6±0.15	36.8±0.30	7.0±0.12
Baguettotron 16k	1.2±0.00	0.0±0.00	0.0±0.00	21.0±0.29	1.8±0.22	40.4±0.53	6.5±0.06
→ SFT	0.0±0.00	0.0±0.00	0.2±0.00	30.3±0.00	1.4±0.41	37.3±0.33	7.7±0.06
→ SFT → DPO	0.0±0.00	0.0±0.00	0.0±0.00	29.8±0.00	1.9±0.25	37.5±0.40	7.5±0.04
MV (0.4B)	25.6±0.00	0.5±0.00	8.6±0.00	22.9±0.58	0.8±0.13	25.6±0.00	10.6±0.04
→ SFT	28.0±0.00	2.7±0.00	22.4±0.35	27.3±0.00	2.3±0.08	32.5±0.00	13.4±0.04
→ SFT → DPO	29.3±0.00	2.7±0.00	21.6±0.20	27.8±0.00	2.0±0.19	34.3±0.00	14.1±0.03
MV-16k (0.4B)	23.2±0.00	1.1±0.00	22.5±0.23	25.8±0.00	4.1±0.82	35.2±0.00	15.4±0.07
→ SFT	28.7±0.00	1.2±0.16	20.5±0.12	24.1±0.29	2.8±0.27	37.7±0.00	13.7±0.06
→ SFT → DPO	28.0±0.00	1.4±0.00	20.6±0.00	22.6±0.29	2.1±0.10	38.8±0.00	14.0±0.02

Table 45: Language suite of the 0.4B comparison between MV and Baguettotron, before and after Phase 1.

Model	ARC-C	ARC-E	BoolQ	CSQA	COPA	HellaS	LAMBADA	MMLU	OBQA	PIQA	WinoG	Avg
Baguettotron	32.8	58.5	60.7	19.6	56.0	36.3	29.8	26.5	30.8	60.9	53.3	42.3
→SFT	30.5	57.5	61.1	21.5	61.0	36.3	33.0	26.2	31.6	60.9	54.6	43.1
→SFT →DPO	30.7	57.5	61.6	21.0	61.0	36.4	33.2	26.2	31.6	60.8	54.4	43.1
Baguettotron 16k	31.6	57.8	57.4	20.6	60.0	35.8	33.7	26.5	31.0	61.5	55.2	42.8
→SFT	30.5	57.2	63.6	19.7	61.0	36.3	33.0	25.8	31.4	60.9	54.4	43.1
→SFT →DPO	30.5	57.2	63.7	19.5	61.0	36.3	33.4	25.8	31.6	61.0	54.3	43.1
MV (0.4B)	31.7	60.6	59.8	25.6	63.0	42.0	37.6	28.4	30.8	65.5	53.1	45.3
→SFT	29.0	58.7	50.1	24.9	63.0	42.3	39.7	28.5	30.2	65.4	53.4	44.1
→SFT →DPO	29.2	59.0	50.2	25.6	63.0	42.4	39.8	28.7	30.2	65.5	53.9	44.3
MV-16k (0.4B)	32.0	61.1	57.9	25.8	63.0	40.6	40.6	28.4	31.8	64.7	53.6	45.4
→SFT	28.2	58.7	53.4	25.1	64.0	41.7	40.4	28.6	30.6	64.9	55.1	44.6
→SFT →DPO	28.3	58.8	53.5	25.1	63.0	41.7	40.7	28.6	30.6	64.6	54.9	44.5

F.4 Effect of post-training masking

Prompt token loss masking computes the Phase 1 loss on the response tokens only. It is compared with the unmasked Phase 1 of the main MV models, the MV rows of Tables 17 and 18, which Tables 46 and 47 do not repeat. Masking raises the reasoning suite average slightly at every stage (MV 24.1 and 24.5 against 22.7 and 23.6 at 4K). GSM8K, HumanEval, MBPP and IFEval are higher with masking at every stage, while AMC23 and MATH500 are lower at three of the four stages. On the language suite the average barely moves, but HellaSwag is 0.8 to 1.3 points lower with masking in all four rows. Each arm is a single checkpoint per stage, so these small differences do not separate masking from the variation between training runs.

Table 46: Reasoning suite with prompt token loss masking in Phase 1 post-training (1.7B MixtureVitae, 300B tokens, historically filtered): separate Tulu3 SFT and DPO runs with the loss on the response tokens only, those of Table 47. The unmasked counterparts, separate training runs, are the main MV model's rows of Table 17. The change in parentheses is the suite average against them: masking raises it by 0.4 to 1.4 points at every stage, and these differences do not affect our comparison.

(a) Mathematics

Model	A24	A25	AMC	M500	GSM8K
MV					
→ SFT	0.0±0.00	0.0±0.00	10.8±3.82	27.0±0.35	52.2±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	11.7±1.44	28.1±0.42	52.7±0.00
MV-16k					
→ SFT	0.0±0.00	0.0±0.00	10.0±0.00	35.5±0.23	57.5±0.00
→ SFT → DPO	0.0±0.00	0.0±0.00	15.0±0.00	35.7±0.46	57.8±0.00

(b) Code, science and instruction following, and the average

Model	HE	LCB	MBPP	GPQA	JEE	IFE	Avg
MV							
→ SFT	53.7±0.00	2.3±0.41	42.6±0.00	19.5±0.58	5.0±0.46	51.7±0.08	24.1±0.36 (+1.4)
→ SFT → DPO	53.0±0.61	2.8±0.41	42.1±0.31	20.7±0.51	4.0±0.33	54.6±0.00	24.5±0.20 (+0.9)
MV-16k							
→ SFT	56.7±0.00	8.1±0.27	41.0±0.40	25.8±0.00	7.2±0.74	55.5±0.26	27.0±0.09 (+0.4)
→ SFT → DPO	54.7±0.70	8.1±0.27	41.3±0.31	26.9±0.29	6.8±0.34	56.4±1.03	27.5±0.15 (+0.6)

Table 47: Language understanding with prompt token loss masking in Phase 1 post-training (1.7B MixtureVitae, historically filtered): separate Tulu3 SFT and DPO runs with the loss on the response tokens only. The unmasked counterparts, separate training runs, are the main MV model's rows of Table 18.

Model	ARC-C	ARC-E	BoolQ	CSQA	COPA	HellaS	LAMBADA	MMLU	OBQA	PIQA	WinoG	Avg
MV												
→ SFT	38.7	70.1	74.6	51.8	70.0	54.9	50.1	42.5	34.6	70.2	58.5	56.0
→ SFT → DPO	38.5	70.2	74.6	51.7	71.0	55.0	50.3	42.6	34.8	70.2	57.8	56.1
MV-16k												
→ SFT	39.3	70.5	74.9	50.9	68.0	53.6	50.5	42.1	34.4	70.0	57.6	55.6
→ SFT → DPO	39.4	70.5	75.1	50.6	69.0	53.7	50.7	42.1	34.2	70.1	57.6	55.7

G Limitations and broader impact

Experimental coverage. Results come from two model scales, 1.7B parameters with a 0.4B cross-check, and two token budgets, 300B and 100B. These are reference points, not a scaling study, so we do not claim that the comparisons hold at substantially larger scale. Post-training covers Tulu3 SFT and DPO and reasoning SFT on 30K samples, without reinforcement learning with verifiable rewards or larger reasoning budgets. A deficit that these recipes do not recover may be recovered by others, so we claim only that it is not recovered under the recipes tested. The full four-dataset comparison is not carried through reasoning SFT: seven of its eight arms undergo reasoning SFT on

OpenThoughts3, all but MixtureVitae without the subset, and only MixtureVitae receives the three other reasoning corpora.

Attribution and measurement. The matched-compute references are the controlled comparisons. The external open-weights references differ in architecture, tokenizer and pretraining compute, so comparisons with them attribute outcomes to data and compute jointly. Every comparison is between single trained checkpoints, and our intervals describe evaluation on fixed checkpoints, so separating mixture effects from the variation between pretraining runs would require replicate runs.

Truncated reasoning targets. In the OpenThoughts3 sample, 62.4% of the training targets are cut at 16,384 tokens without a supervised end-of-turn token, and 62.7% of its answers as released never close their reasoning, before any cut (Appendix A.2). Every OpenThoughts3 run therefore trains on many answers that do not end, which bears on every Phase 2 result. Under AceReason and Nemotron-Cascade, whose targets end in 95.5% and 96.8% of samples, the original reference mixtures gain at most 1.6 MATH500 points (Table 4). Limited MATH500 improvement therefore also occurs with training data with substantially less truncation. The contribution of truncation to the OpenThoughts3 declines is not isolated.

Benchmark overlap. The audit of Appendix A.3 covers the post-training data and MixtureVitae's pretraining reasoning sources. The rest of MixtureVitae's pretraining data, the reference pretraining datasets, which are used as released, and the data of the external models were not audited, and the benchmark overlap of the SYNTH mix of the 0.4B comparison is not documented. Whether a 300B-token run drew a given document of the pool is unknown, and the Phase 2 suite averages retain the 199 LiveCodeBench and 18 MBPP problems found in the audited training data, because only aggregate coding scores are available. Scores on complete benchmarks can therefore include problems found in training data, and the comparisons without the identified items are a sensitivity analysis, not an evaluation on uncontaminated items. Moreover, 46.0 to 66.4% of the alternative corpora's sampled questions match non-OT3 sources in MixtureVitae's pretraining reasoning pool, so those comparisons test robustness across data releases without establishing that the fine-tuning questions were unseen during pretraining. The enriched reference mixtures' copy of the subset also holds OpenThoughts2 documents that the audit did not cover (Appendix E.5).

Generalization. The study makes no claim about permissive or non-permissive data as classes. The pre-training datasets tested are instances, not a sample of their populations, and in both classes the outcome depends on mixture composition, the factor the design isolates. What carries forward is the statement about reasoning enrichment. Testing it at larger scale, for example by deriving and checking scaling laws for the front-loading effect, is left to future work.

Broader impact. In the setting tested, a model pretrained on permissive data can remain competitive after post-training. Permissive data can be shared together with its recipes and checkpoints, which is meant to make such experiments easier to inspect, reuse and extend. The models are 1.7B research checkpoints with the capability and evaluation limits stated above, and the general dual-use considerations of open model releases apply to them.

H Author contributions

- **Ali Elganzory**: Phase 1 post-training. Trained every Tulu3 SFT and DPO model in this work and ran their reasoning and instruction-following evaluations, including the three-seed evaluation under one protocol that the main tables use. Co-designed the experiments and contributed to the analysis and interpretation of the results. Wrote the table generation code, prepared the final submission, co-wrote the rebuttal, and released checkpoints and evaluation records.
- **Harsh Raj**: Pretraining and data. Extended every base model to 16K context, trained the new base models (the retrained reference models, the enriched reference mixtures and MixtureVitae without OpenThoughts3 at 100B and 300B tokens), and prepared the decontaminated Dolci data. Ran the general language understanding evaluations. Found the code execution error that lowered the HumanEval scores, checked the post-training setup, and co-wrote the rebuttal.

- **Marianna Nezhurina**: Phase 2 post-training. Set up reasoning post-training on OpenThoughts3, trained the reasoning SFT runs other than those of the enriched reference mixtures and of Nemotron-CC-HQ, ran the Phase 2 evaluations with 32K-token generation, and released the Phase 2 responses and scores.
- **Victor May**: Proposed and specified the factorial design of reasoning enrichment and permissiveness, and carried it through reasoning post-training: trained the OpenThoughts3 runs of the three enriched reference mixtures and every reasoning SFT run of Nemotron-CC-HQ, and evaluated them and their Tulu3 SFT checkpoints under Protocol B, and Nemotron-CC-HQ's OpenThoughts3 models on the language suite. Found that the evaluation samples all MATH500 questions of a run with one seed, and evaluated MATH500 again with a seed for each question. Decontamination analysis of the post-training data and the benchmark-overlap audit. Statistical analysis. Paper writing for the submission, the rebuttal and the camera-ready revision, and figures. Built the reproduction repository: assembled and checked the evaluation records, generated every table and figure from them, and released the records and code. Coordinated the camera-ready revision.
- **Van Khue Nguyen**: Evaluation support. Analysis of MATH500 answer extraction, and GSM8K evaluation code and runs.
- **David Salinas**: Advice on post-training methods. Diagnosis of the MATH500 drop after Tulu3 SFT from sampled outputs. A writing pass on the submission, with the pipeline figure and the token count table.
- **Huu Nguyen**: Built the MixtureVitae pretraining corpus and contributed data for this work. Decontamination of MixtureVitae, input to the study design, the framing of permissiveness and its legal limits, paper co-writing and supervision.
- **Jenia Jitsev**: Project lead and supervision. Research question and scope, experiment design, compute acquisition, task assignment and deadlines. Provided the open-sci-ref reference models, diagnosed corrupted training data, and ran early reasoning post-training experiments. Drafted and revised the paper, made the first submissions, and edited and posted the rebuttal.

Table 48: Author contributions to this work. Large squares (■) indicate a major contribution and plus symbols (+) indicate a supporting contribution.

	Study design	Data and pretraining	Post-training	Evaluation	Decontamination	Statistical analysis	Writing	Reproducibility and release	Leadership and coordination	Advising and supervision
Ali Elganzory	+		■	■		+	+	+	+	
Harsh Raj	+	■	+	■	+		+	+		
Marianna Nezhurina	+	+	■	■			+	+		
Victor May	■		+	■	■	■	■	■	■	
Van Khue Nguyen				+						
David Salinas				+			+			+
Huu Nguyen	+	■			+		+			■
Jenia Jitsev	■	■	+	+			■		■	■