KNOWLEDGE DISTILLATION DURING MID-TRAINING FAVORS REASONING OVER FACTUAL RECALL

Anonymous authors
Paper under double-blind review

ABSTRACT

Logit-based knowledge distillation (KD) is used to train smaller language models (LMs) via supervision from stronger teachers, but whether its benefits are consistent across training stages remains unclear. Through controlled experiments, we find that forward Kullback-Leibler (KL) distillation—the standard KD formulation—with post-trained teachers behaves fundamentally differently during mid-training, an intermediate phase of self-supervised learning on curated corpora. Surprisingly, while forward KD simultaneously improves reasoning and factual recall during pre-training relative to standard next-token prediction (NTP), it instead slows factual recall acquisition during mid-training despite continued reasoning gains. We trace this stage dependence to an asymmetry in teacher confidence across data domains and the student's evolving knowledge state: teachers are more confident on procedural than unstructured text data, while students acquire low-entropy factual knowledge earlier in training. To mitigate this imbalance, we propose SWITCH DISTILLATION, a simple mid-training objective that distills on tokens where the teacher is confident, using teacher *predictive entropy* as a lightweight routing signal, and otherwise falls back to cross-entropy. SWITCH DISTILLATION consistently outperforms existing distillation objectives across teacher sizes. Relative to NTP, it achieves 1.61–1.71× the reasoning performance and 1.13–1.19× the knowledge and commonsense performance while preserving 96.7–96.8% of factual recall. Crucially, these benefits persist after post-training: SWITCH DISTILLATION closes the factual recall gap while maintaining 1.25–1.32× and 1.13–1.20× gains in reasoning and knowledge and commonsense, respectively. Our results demonstrate that stage-aware KD extracts substantially more useful supervision per token under a fixed data budget.

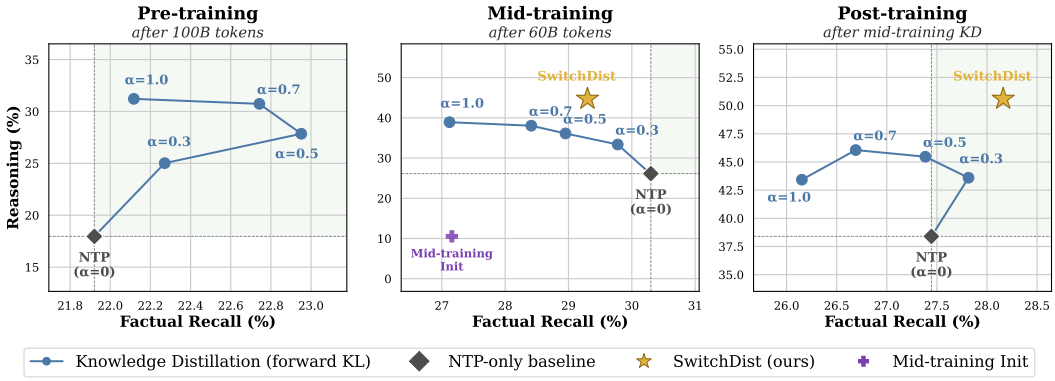


Figure 1: **Knowledge distillation (KD) generally improves reasoning, but its effect on factual recall changes across training stages.** Using an OLMo-2 1B student and 7B-Instruct teacher, we sweep the distillation strength α, which interpolates between NTP ($\alpha = 0$) and forward KL distillation. Distillation improves both reasoning and factual recall during pre-training (left), but favors reasoning at the expense of factual recall during mid-training (middle). SWITCH DISTILLATION (★) mitigates this imbalance and Pareto-dominates NTP after post-training (right).

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

1 INTRODUCTION

Modern language models (LMs) increasingly rely on a dedicated *mid-training* stage between pre-training and post-training. During mid-training, self-supervised next-token prediction continues on a smaller, high-quality corpus curated to improve capabilities such as factuality, reasoning, coding, and instruction following (Grattafiori et al., 2024; Allal et al., 2025; Walsh et al., 2025; Liu et al., 2026; Meta Superintelligence Labs, 2026). Because mid-training uses far fewer tokens than pre-training, extracting more learning signal from each token becomes especially important. Knowledge distillation (KD) (Buciluǎ et al., 2006; Hinton et al., 2015) offers a natural approach by augmenting ground-truth next-token supervision with the richer predictive distribution of a stronger teacher, typically by minimizing the forward Kullback-Leibler (KL) divergence between teacher and student predictive distributions (Gu et al., 2024; Zhong et al., 2024). Yet despite its growing adoption in frontier language modeling pipelines (Team Gemma et al., 2025; Meta Superintelligence Labs, 2026), KD has been studied almost exclusively in pre-training and post-training (Busbridge et al., 2025; Lu & Liu, 2026; Agarwal et al., 2024), leaving it unclear whether its benefits transfer to the relatively nascent mid-training stage.

Surprisingly, we find that knowledge distillation behaves qualitatively differently in mid-training than in pre-training. Using the OLMo-2 ecosystem (Walsh et al., 2025), one of the most recent fully open model families with intermediate checkpoints, training recipes, and multiple model scales, we conduct controlled pre-training and mid-training experiments with 1B students. As Fig. 1 shows, increasing distillation strength (α) generally improves reasoning across training stages, but its effect on factual recall differs sharply. While traditional forward KL distillation improves both reasoning and factual recall over vanilla next-token prediction (NTP) during pre-training, its reasoning gains are accompanied by comparatively lower factual recall during mid-training. We refer to this phenomenon as the *reasoning–recall tradeoff*.

This tradeoff is remarkably robust: across instruction-tuned teacher sizes, KL directions, and interpolation coefficients, no KD objective Pareto-dominates NTP during mid-training. We track this behavior to an interaction between teacher confidence, the student's evolving knowledge state, and the distillation objective. To begin, our teachers exhibit substantially lower predictive entropy on procedural data, such as math and instruction-following, than on unstructured data such as general web text; lower entropy is also strongly correlated with higher-quality supervision. Students acquire factual knowledge associated with lower teacher entropy earlier during pre-training, such that facts that have not been learned at the start of mid-training receive disproportionately weak teacher supervision. Finally, as teacher entropy rises, KD increasingly attenuates the ground-truth learning signal relative to NTP. Together, these effects provide evidence for why distillation preferentially accelerates reasoning while reducing open-ended factual recall relative to NTP during mid-training.

Motivated by this analysis, we propose SWITCH DISTILLATION, a drop-in mid-training objective that uses teacher predictive entropy to route each token between distillation and next-token prediction. Because this routing signal is computed from the teacher logits already required for KD, SWITCH DISTILLATION incurs minimal additional computation. Across 7B and 13B teachers, SWITCH DISTILLATION substantially improves the reasoning–recall tradeoff. With an OLMo-2 7B Instruct teacher, for example, SWITCH DISTILLATION improves average reasoning performance by 71% and knowledge and commonsense performance by 19% over NTP, while reducing factual recall by just 1 percentage point. Given that the purpose of mid-training is to provide a strong prior for alignment, we show that SWITCH DISTILLATION's gains persist through post-training: reasoning remains 32% higher, knowledge tasks improve by 20%, and the factual recall gap closes entirely. Our contributions are threefold:

1. **Empirical finding:** We uncover a robust reasoning–recall tradeoff during the mid-training regime: knowledge distillation of a substantially pre-trained student improves reasoning while slowing the acquisition of factual recall relative to NTP.
2. **Explanatory analysis:** We explain this tradeoff through the interaction between teacher confidence, student learning dynamics, and the distillation objective, showing that facts not yet learned by the pre-trained student disproportionately receive weak teacher supervision.
3. **Mid-training objective:** Our analysis naturally motivates SWITCH DISTILLATION, which routes tokens between KD and NTP using teacher predictive entropy. Our method substantially mitigates the tradeoff across teacher sizes and retains its gains after post-training.

2 BACKGROUND

2.1 PRELIMINARIES

Language modeling. Given a token sequence $\mathbf{x} = (x_1, \ldots, x_N)$ of length N and an autoregressive language model distribution p_θ over vocabulary $\mathcal{V}$, standard next-token prediction minimizes the expected cross-entropy loss

$$\mathcal{L}_{\mathrm{CE}} = \mathbb{E}_{(\mathbf{x},n)} \left[-\log p_\theta(x_n \mid x_{<n}) \right], \tag{1}$$

where x_n is the target next token and the expectation is over training sequences and token positions.

Knowledge distillation. Logit-based knowledge distillation (KD) trains a student to match the output distribution of a vocabulary-compatible teacher (Hinton et al., 2015; Busbridge et al., 2025):

$$\mathcal{L}_{\mathrm{KD}} = (1-\alpha)\mathcal{L}_{\mathrm{CE}} + \alpha\mathcal{L}_{\mathrm{KL}}, \tag{2}$$

where $\alpha \in [0,1]$ controls the distillation strength. Let $p_T^{(\tau)}(\cdot \mid x_{<n})$ and $p_S^{(\tau)}(\cdot \mid x_{<n})$ denote the next-token distributions scaled by temperature $\tau > 0$ for a teacher T and student S, respectively. Standard KD instantiates $\mathcal{L}_{\mathrm{KL}}$ using the forward KL (FKL) divergence (Kullback & Leibler, 1951):

$$\mathcal{L}_{\mathrm{FKL}} = \tau^2 \, \mathbb{E}_{(\mathbf{x},n)} \left[\sum_{v \in \mathcal{V}} p_T^{(\tau)}(v \mid x_{<n}) \log \frac{p_T^{(\tau)}(v \mid x_{<n})}{p_S^{(\tau)}(v \mid x_{<n})} \right], \tag{3}$$

Recent work has advocated for distillation using the reverse KL (RKL) divergence, which discourages student mass on the teacher's low-probability regions (Agarwal et al., 2024; Gu et al., 2024):

$$\mathcal{L}_{\mathrm{RKL}} = \tau^2 \, \mathbb{E}_{(\mathbf{x},n)} \left[\sum_{v \in \mathcal{V}} p_S^{(\tau)}(v \mid x_{<n}) \log \frac{p_S^{(\tau)}(v \mid x_{<n})}{p_T^{(\tau)}(v \mid x_{<n})} \right]. \tag{4}$$

Throughout this paper, we instantiate $\mathcal{L}_{\mathrm{KL}}$ as either $\mathcal{L}_{\mathrm{FKL}}$ or $\mathcal{L}_{\mathrm{RKL}}$, corresponding to forward-KL distillation (FKD) and reverse-KL distillation (RKD), respectively.

2.2 EXPERIMENTAL SETUP

Training regimes. We build on the open-source OLMo-2 ecosystem (Walsh et al., 2025). We pre-train and mid-train on Dolmino Mix 1124, a data mixture of filtered DCLM web text, FLAN instruction-following data, Dolmino Math, peS2o, Wikipedia (including Wikibooks), and Stack Exchange. For pre-training, we initialize 1B students from random weights and train beyond Chinchilla optimality for 100B tokens (Hoffmann et al., 2022). For mid-training, we initialize from the OLMo-2 1B Stage 1 checkpoint, already pre-trained on 4T tokens, and continue training for 60B tokens.

Teacher models. Post-trained models are increasingly used as teachers for reference-based language modeling in recent research (Goyal et al., 2026; Huang et al., 2026; Jin et al., 2026; Tan et al., 2026) and frontier LLM pipelines (Team Gemma et al., 2025; Meta Superintelligence Labs, 2026). Their stronger instruction-following and reasoning abilities make them natural choices for capability transfer. Accordingly, we employ OLMo-2 1B Instruct, 7B Instruct, and 13B Instruct as teachers.

Evaluation. We evaluate with the standardized OLMES (Gu et al., 2025) harness, grouping benchmark tasks into REASONING (generative problem solving), FACTUAL RECALL (open-ended generative retrieval of factual knowledge), KNOWLEDGE & COMMONSENSE (multiple choice world knowledge and commonsense reasoning), and INSTRUCTION FOLLOWING (post-training only). We report the macro-average performance of each task group and focus our stage-dependent analysis on REASONING and FACTUAL RECALL. See Table 7 for full task suite and evaluation settings.

3 CHARACTERIZING MID-TRAINING DISTILLATION

Fig. 2 compares reasoning and factual recall across distillation strength ($\alpha \in [0.0, 0.3, 0.5, 0.7, 1.0]$), KL direction (forward and reverse), training regime (pre-training and mid-training), and Instruct teacher size (1B, 7B, and 13B).[1]

[1]We observe the same qualitative trends for another model family, SmolLM2 (Allal et al., 2025), using a 1.7B Instruct teacher and 360M student, in App. D.1.

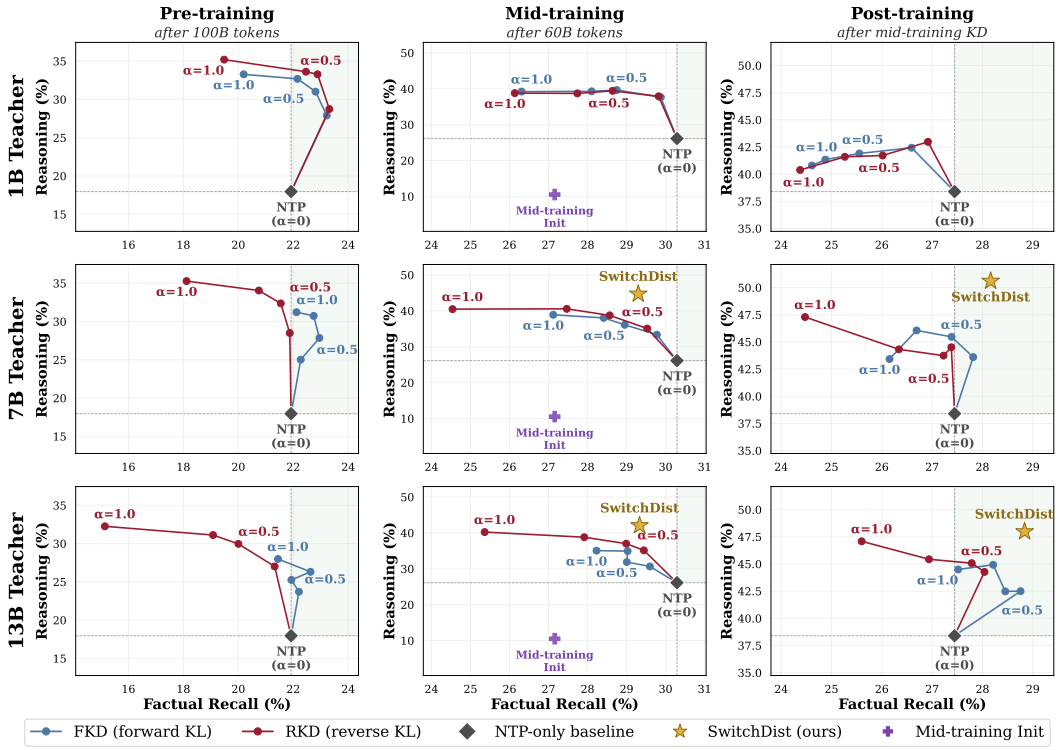


Figure 2: **Knowledge distillation exhibits a reasoning-recall tradeoff that falls below the NTP frontier during mid-training.** SWITCH DISTILLATION mitigates this tradeoff and yields Pareto improvements over NTP after post-training. Rows correspond to 1B, 7B, and 13B teachers. We sweep over the distillation weight α for forward KL and reverse KL distillation; ◆ denotes the NTP baseline ($\alpha = 0$), and mid-training panels show the pre-trained student at initialization.

Increasing distillation strength consistently shifts student performance toward reasoning. Across teacher sizes and KL directions, increasing the contribution of the teacher (via larger α) generally moves the operating point toward higher reasoning performance, with diminishing returns at stronger distillation. This trend holds across both pre-training and mid-training, suggesting that the teacher can reliably impart reasoning-relevant behavior even when the student has already acquired substantial knowledge. This is consistent with recent work using knowledge distillation to transfer reasoning capabilities from stronger teachers (Kim & Baek, 2026).

The effect of distillation on factual recall is stage-dependent. During pre-training (Fig. 2 left), moderate forward-KL distillation tends to improve factual recall alongside reasoning, yielding Pareto improvements over NTP across teacher sizes. At stronger distillation strengths, however, factual recall begins to decline especially under reverse KL. During mid-training (Fig. 2 middle), the pattern changes such that no KD operating point outperforms NTP on factual recall even as reasoning improves. Post-training changes this tradeoff asymmetrically: after applying the same standard post-training procedure to all mid-trained settings, without further distillation (Fig. 2 right), the factual-recall gap relative to NTP narrows or reverses for larger teachers, while the reasoning gains from distillation persist. This convergence partly reflects greater factual degradation of NTP during post-training, whereas distilled checkpoints retain more of their mid-training factual recall.

Thus, the reasoning–recall tradeoff changes qualitatively once distillation is applied to a substantially pre-trained student. Because our pre-training and mid-training experiments use the same data mixture, this shift cannot be attributed to differences in training data. Instead, it points to an interaction between teacher supervision and the student's prior training state: unlike a randomly initialized student, the mid-training student has already acquired substantial knowledge through pre-training. We investigate this interaction in the next section.

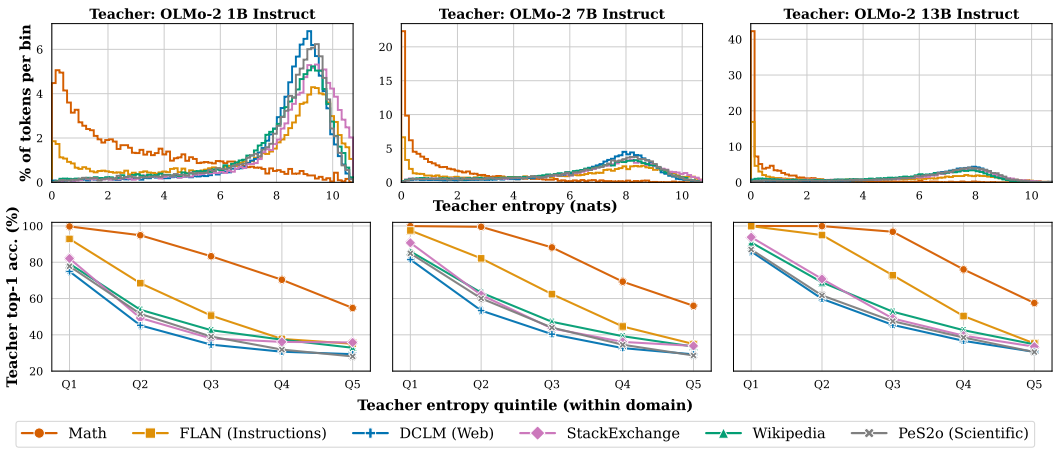


Figure 3: **Lower teacher entropy corresponds to more reliable teacher supervision.** Tokens are partitioned by teacher entropy into quantiles. Q1: lowest entropy; Q5: highest entropy. **Top:** Procedural domains (e.g., math, instruction-following) concentrate at lower teacher predictive entropy than unstructured domains (e.g., web text, Wikipedia). **Bottom:** For each domain, lower entropy tokens exhibit substantially higher teacher top-1 agreement with the ground-truth token.

4 Why Does the Reasoning–Recall Tradeoff Occur?

We attribute KD's stage-dependent behavior to the interaction of three factors: (i) the teacher's predictive confidence, (ii) the student's existing knowledge state, and (iii) the optimization dynamics induced by distillation. We study each factor in turn below.

4.1 Teacher Supervision is Asymmetric Across Data Domains

We begin by asking whether teacher supervision is uniformly reliable across the training corpus. Using OLMo-2 Instruct 1B, 7B, and 13B as teachers, we compute the *teacher predictive entropy* H_n for every token at position n from randomly sampled Dolmino documents. Formally, let

$$H_n = -\sum_{v \in \mathcal{V}} p_{\mathrm{T}}^{(\tau)}(v \mid x_{<n}) \log p_{\mathrm{T}}^{(\tau)}(v \mid x_{<n}) \tag{5}$$

denote the entropy of the teacher distribution T (scaled with temperature τ, and with vocabulary $\mathcal{V}$). Teacher entropy varies systematically across domains (Fig. 3, top): procedural domains (e.g., math and instruction-following) exhibit substantially lower entropy than unstructured text domains.[2] This trend is consistent across teacher sizes. Domain-level differences alone, however, do not establish that predictive entropy reflects supervision quality. Teacher predictive entropy is also indicative of correctness *within* each domain, where correctness is defined as agreement with the ground-truth corpus token. In Fig. 3 (bottom), the probability that the teacher's top-1 prediction equals the ground-truth token decreases monotonically across entropy quintiles for every data domain and teacher size, showing that teacher entropy is predictive of supervision quality.

Together, these results indicate that teacher supervision is asymmetric across domains: in distillation, the teacher provides confident, lower-entropy supervision on tokens from procedural domains, but substantially more diffuse, higher-entropy supervision on tokens from unstructured ones. This asymmetry may bias distillation toward learning reasoning better over factual recall.

4.2 Teacher Entropy Predicts Student Factual Acquisition

Given qualitatively consistent trends across teacher sizes, we use OLMo-2 7B Instruct as a representative teacher for the remaining analyses; corresponding 1B and 13B results are shown in App. D.3.

We track the acquisition of open-ended factual recall across NTP checkpoints on aggregated QA samples from our FACTUAL RECALL tasks (TriviaQA, Natural Questions, and SimpleQA). For each

[2]This pattern generalizes across model families and training stages (App. D).

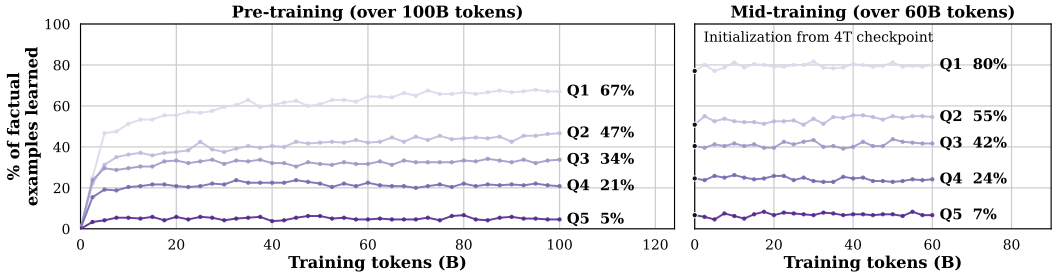


Figure 4: **Unresolved facts become concentrated at higher teacher entropy (higher quintiles).** Lower-entropy facts are acquired earlier during pre-training; by mid-training initialization (after 4T tokens), unresolved facts are concentrated in the highest-entropy quintiles. Q1: lowest entropy; Q5: highest entropy. Results for other teacher sizes are in Fig. 10.

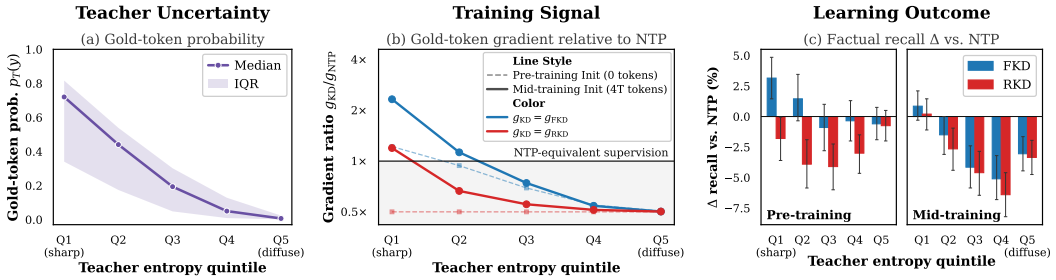


Figure 5: **High teacher entropy provides weaker supervision for factual acquisition, translating to worse factual recall.** Upon stratifying factual recall examples into teacher-entropy quintiles using an OLMo-2 7B Instruct teacher, the highest teacher entropy tokens (a) have lower teacher-assigned ground-truth probability, which (b) produces weaker optimization signal via gradient updates during KD, and (c) results in worse factual recall. Q1: lowest entropy; Q5: highest entropy.

sample, we teacher-force the prompt to the first answer token and measure (i) the teacher entropy at that position and (ii) whether the student's top-1 prediction matches the first token of any gold answer alias. We use this token-level criterion to operationalize whether a fact has been acquired. Teacher entropy is used only to stratify factual examples; students are trained purely with NTP.

As shown in Fig. 4, teacher entropy strongly predicts factual acquisition: by the end of pre-training, the student has learned 67% of facts in the lowest-entropy quintile (Q1), and only 5% in the highest (Q5), with intermediate quintiles progressing monotonically. By the start of mid-training, this stratification has largely saturated, leaving unresolved factual-recall examples concentrated in the highest-entropy quintiles, which is precisely where teacher supervision is least confident.

4.3 Knowledge Distillation Attenuates Factual Supervision

Having characterized the teacher's predictive confidence and the student's knowledge state, we next examine their interaction through the distillation objective. Using the entropy-stratified factual recall examples from §4.2, we trace three cascading effects: (i) the teacher's probability assigned to that token, (ii) the supervision placed on the token under forward and reverse KD, and (iii) the resulting factual recall difference between KD and NTP. Across all three analyses, higher teacher entropy consistently corresponds to progressively weaker factual supervision and worse performance.

First, teacher probability on the ground-truth token strictly decreases with predictive entropy (Fig. 5a), meaning that the teacher places less weight on the correct answer as entropy increases.

We next measure the gold-token directional gradient g, which captures how strongly the objective reinforces the correct next token. Because g depends on both teacher and student distributions, we consider two student initializations representing the start of pre- and mid-training: random initialization and the 4T-token checkpoint, respectively. We compute the ratio of the gold-token gradient

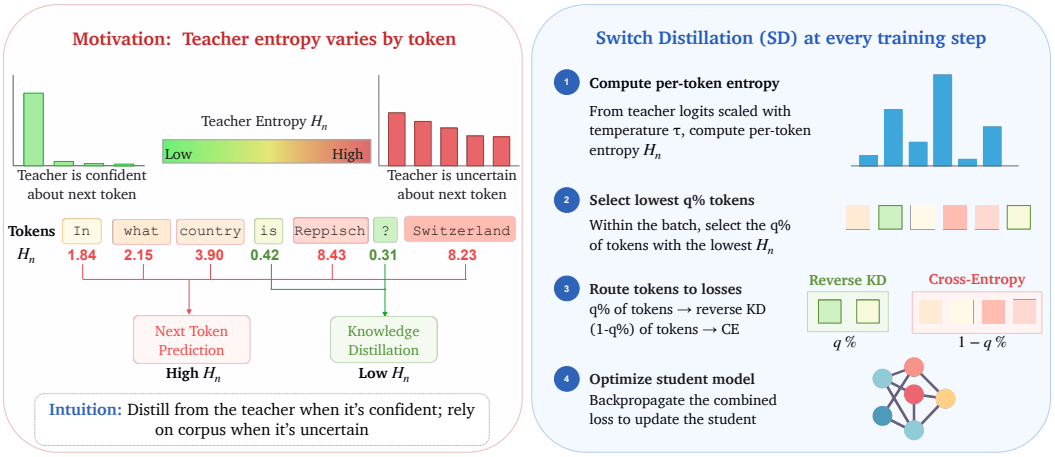


Figure 6: **SWITCH DISTILLATION overview.**

under FKD (g_{FKD}) or RKD (g_{RKD}) to NTP (g_{NTP}).[3] Ratios below 1 indicate weaker ground-truth supervision than NTP. Both FKD and RKD increasingly attenuate ground-truth supervision as teacher entropy rises, reaching approximately $0.5\times$ NTP for the highest-entropy facts (Fig. 5b).

To connect this weakened gradient signal to downstream acquisition, we compare KD and NTP factual recall within each entropy quintile (Fig. 5c). After pre-training, FKD outperforms NTP on low-entropy facts (Q1–Q2), but this advantage disappears as entropy rises; RKD underperforms NTP across all quintiles at this stage. During mid-training, this effect is substantially stronger: low-entropy facts have largely already been acquired, leaving unresolved facts concentrated in high-entropy regions where KD provides the weakest supervision. Consequently, both FKD and RKD incur their largest factual recall deficits on high-entropy facts.

5 SWITCH DISTILLATION IMPROVES THE REASONING–RECALL TRADEOFF

Evidence thus far suggests that teacher supervision is not equally beneficial across all tokens: teacher predictions tend to be concentrated on procedural reasoning trajectories but more diffuse on factual payload tokens. This suggests that a uniform distillation objective may be suboptimal during mid-training, and that teacher supervision should be applied only at token positions where it is most reliable. To this end, we introduce SWITCH DISTILLATION (Fig. 6).

Using teacher entropy H_n (Eq. (5)), we route the lowest-$q\%$ of in-batch tokens to the reverse KL loss,[4] defining $\mathcal{S}_q = \{n : H_n \le \text{Quantile}_q(\{H_n\})\}$, the set of tokens assigned to distillation. Reverse KL is particularly well-suited to low-entropy teacher predictions as its mode-seeking behavior reinforces the teacher's preferred continuation. We optimize the following objective

$$\mathcal{L}^{\text{SwitchDist}} = \tau^2 \frac{1}{|\mathcal{S}_q|} \sum_{n \in \mathcal{S}_q} \mathcal{L}_{\text{RKL}}\left(p^{(\tau)}_{\text{S},n} \,\|\, p^{(\tau)}_{\text{T},n}\right) + \frac{1}{|\bar{\mathcal{S}}_q|} \sum_{n \in \bar{\mathcal{S}}_q} \mathcal{L}_{\text{CE},n}, \tag{6}$$

where $\bar{\mathcal{S}}_q$ denotes the complement of $\mathcal{S}_q$ over supervised tokens. Note that SWITCH DISTILLATION adds only negligible entropy and quantile computations beyond standard online KD, requiring no additional parameters or model forward passes.

6 SWITCH DISTILLATION EXPERIMENTS

6.1 EXPERIMENTAL SETUP

We evaluate SWITCH DISTILLATION under the same mid-training setup as in §2.2, and consider OLMo-2 7B Instruct and 13B Instruct as teacher models. Our baselines include standard next-

[3] We evaluate at $\alpha = 0.5$. Closed-form derivations are provided in App. C.3.

[4] We normalize the terms separately so that each partition's aggregate contribution is independent of its size; consequently, relative per-token weights vary with q. We select $q = 20\%$; full sweep results are in App. B.2.

Table 1: **Full downstream results after mid-training.** The NTP baseline is duplicated because it has no teacher ($T = \mathrm{N/A}$) and therefore serves as the shared reference for both teacher size blocks. **Bold** denotes the best result per teacher block, and * indicates a statistically significant improvement over the strongest competing baseline ($p < 0.05$, paired bootstrap). Benchmark names are abbreviated; see Table 7 for full task names.

		Reasoning						Factual Recall			Knowledge & Commonsense					
T	Method	GSM8K	GSM-S	GSM+	BBH	DROP	MATH	TQA	NQ	SQA	MMLU	MMLU-P	ARC-C	OBQA	Wino	AGI
N/A	NTP	40.4	29.8	23.1	29.9	29.8	3.8	**56.7***	**25.5**	**8.7**	43.6	15.5	51.1	51.8	51.4	34.1
7B	FKD	57.8	42.6	36.2	30.6	41.9	7.6	53.9	24.7	8.4	49.5	18.7	61.9	61.8	51.5	38.6
7B	RKD	62.1	46.2	39.2	31.6	43.1	10.4	53.1	24.1	8.4	50.1	18.9	62.5	62.8	52.5	39.3
7B	TRKD	52.5	36.9	31.0	30.3	36.8	5.6	54.0	24.1	7.8	47.9	17.4	58.9	60.0	51.2	37.5
7B	**SD**	**69.7***	**55.3***	**46.1***	**32.8***	**49.6***	**14.8***	54.9	24.6	8.4	**51.6***	**19.8***	**64.7***	**64.2**	**53.8**	**41.5***
N/A	NTP	40.4	29.8	23.1	29.9	29.8	3.8	**56.7***	**25.5**	8.7	43.6	15.5	51.1	51.8	51.4	34.1
13B	FKD	52.7	37.9	31.8	29.4	33.1	6.4	54.2	24.7	8.1	47.5	16.3	55.3	56.2	51.2	36.2
13B	RKD	59.2	48.7	37.4	31.4	37.3	8.0	53.4	24.0	8.5	48.4	17.3	57.8	60.0	51.0	38.3
13B	TRKD	47.8	32.7	28.4	29.3	34.0	5.6	53.8	24.2	8.3	45.7	15.4	53.0	56.6	50.7	35.2
13B	**SD**	**66.0***	**54.8***	**42.2***	**31.6**	**45.8***	**12.0***	54.1	24.9	**9.0**	**48.5**	**17.7**	**58.5**	**63.8***	**51.5**	**39.2**

token prediction (**NTP**), as well as forward and reverse knowledge distillation (**FKD** and **RKD**, respectively) at $\alpha = 0.5$, which empirically provides the best balance between factual recall and reasoning. We also compare against token-routing KD (**TRKD**) (Goyal et al., 2026). TRKD applies forward-KL distillation to high-entropy tokens while retaining CE on all tokens, whereas SWITCH DISTILLATION hard-switches between reverse-KL on low-entropy tokens and CE otherwise.[5]

6.2 SWITCH DISTILLATION IMPROVES REASONING WHILE PRESERVING FACTUAL RECALL

Table 1 compares SWITCH DISTILLATION against mid-training baselines. Across both teacher sizes, SWITCH DISTILLATION achieves the strongest REASONING, improving the macro-average from 26.1% under NTP to 44.7%/42.1% with the 7B and 13B teachers, while remaining the KD baseline closest to NTP on FACTUAL RECALL (29.3%/29.3% vs. 30.3%). SWITCH DISTILLATION also achieves the strongest KNOWLEDGE & COMMONSENSE performance (49.3%/46.5%). Distillation with the 7B teacher generally outperforms the 13B teacher, corroborating work showing that a large size difference between teacher and student—defined as the *capacity gap*—may reduce distillation effectiveness (Mirzadeh et al., 2019; Panigrahi et al., 2024; Busbridge et al., 2025).

6.3 SWITCH DISTILLATION'S BENEFITS PERSIST THROUGH POST-TRAINING

We next ask whether SWITCH DISTILLATION's mid-training gains are sustained after post-training, as stronger base models do not necessarily translate to stronger final models after alignment (Springer et al., 2025; Lu & Liu, 2026; Watts et al., 2026). We apply OLMo-2 1B's four-stage post-training pipeline—supervised fine-tuning (SFT), direct preference optimization (DPO), and two rounds of reinforcement learning with verifiable rewards (RLVR1, RLVR2)—to each mid-trained model and report final performance in Table 2.[6] Post-training improves reasoning across all methods, and SWITCH DISTILLATION remains the strongest method on the REASONING task group, with significant gains on all six tasks with the 7B teacher and five of six with the 13B teacher. With the 7B teacher, SWITCH DISTILLATION's macro-average increases from 44.7% to 50.6%; with the 13B teacher, it increases from 42.1% to 48.0%. Consistent with work showing that post-training can degrade factual recall and broader knowledge (Gekhman et al., 2024; Ghosal et al., 2024; Yuan et al., 2024; Kaplan et al., 2026), we observe modest degradation in KNOWLEDGE & COMMONSENSE and FACTUAL RECALL across methods. However, SWITCH DISTILLATION is the most robust: despite entering post-training with a small factual-recall deficit relative to NTP, it experiences the least forgetting and finishes with the highest FACTUAL RECALL macro-average.

[5]TRKD was originally proposed to improve in-context learning during distillation pre-training; we provide a more detailed description of the method and its differences from SWITCH DISTILLATION in App. A.1.

[6]We report full evaluation results after each intermediate post-training stage in App. E.

Table 2: **Downstream results after post-training.** Notation and significance testing follow Table 1.

		Reasoning						Factual Recall			Knowledge & Commonsense						Inst.
T	Method	GSM8K	GSM-S	GSM+	BBH	DROP	MATH	TQA	NQ	SQA	MMLU	MMLU-P	ARC-C	OBQA	Wino	AGI	IFE
N/A	NTP	67.0	46.8	40.6	31.1	32.7	12.6	53.3	21.6	7.5	32.3	16.6	52.4	52.4	51.4	32.3	62.1
7B	FKD	76.2	56.4	51.6	33.8	38.4	18.0	51.9	22.2	8.1	42.2	18.6	60.2	57.4	51.5	38.0	64.7
7B	RKD	73.8	53.4	49.5	31.6	40.5	17.6	51.8	21.9	7.9	46.2	18.0	61.1	53.0	51.2	37.1	61.6
7B	TRKD	70.7	52.9	46.3	31.8	36.8	14.4	51.7	22.3	8.0	40.9	17.4	59.0	53.6	51.5	35.6	65.2
7B	**SD**	**79.8***	**65.0***	**52.7***	**35.6***	**48.2***	**22.4***	**53.6**	**22.7**	**8.2**	**48.1***	**19.8***	**62.6**	**59.4**	**54.6***	**39.6***	**69.5***
N/A	NTP	67.0	46.8	40.6	31.1	32.7	12.6	53.3	21.6	7.5	32.3	16.6	52.4	52.4	51.4	32.3	62.1
13B	FKD	72.1	54.3	46.5	31.4	35.7	14.6	54.1	23.2	8.0	36.1	**18.1**	55.0	56.4	51.2	35.9	64.9
13B	RKD	75.1	58.1	49.3	**33.3**	38.7	16.0	53.9	22.3	7.7	38.2	17.0	56.0	55.2	51.1	36.6	64.1
13B	TRKD	69.6	48.5	42.6	31.5	34.6	12.2	53.9	22.2	8.2	37.0	17.3	55.4	54.0	51.3	33.6	64.1
13B	**SD**	**77.8***	**62.8***	**51.8***	**33.3**	**42.8***	**19.6***	**54.2**	**24.1**	**8.4**	**43.1***	**18.1**	**56.5**	**59.0**	**52.5**	**38.1**	**67.1**

6.4 Ablations

We further ablate the design choices to SWITCH DISTILLATION using the OLMo-2 7B Instruct teacher and report mid-training macro-averages in Table 3 (per-task results in App. E.2).

Table 3: **Ablation results.** The first row reports SWITCH DISTILLATION's absolute performance; subsequent rows report relative changes from other design choices.

Method / Ablation	**Reasoning**	**Factual Recall**	**Knowledge**
SWITCH DISTILLATION	44.7	29.3	49.3
Distillation Objective			
SWITCH DISTILLATION$_{\text{FKL}}$	-2.9	-0.2	-1.3
Routing Policy			
Teacher-Correct Routing	-4.4	-5.1	-1.0
Random Routing	-6.5	-0.8	-2.0
Oracle Domain Routing	-7.2	-1.3	-2.3
Supervision Objective			
Always CE	-0.3	+0.1	-0.6
Teacher Top-1 Labels	-6.4	+1.3	-2.8

We examine three design choices using the OLMo-2 7B Instruct teacher: *KL direction*, replacing RKL with FKL (**SWITCH DISTILLATION$_{\text{FKL}}$**); *routing signal*, replacing teacher entropy with whether the teacher's top-1 prediction matches the target (**Teacher-Correct Routing**) or with a random mask (**Random Routing**) (each with a fixed routing budget of $q = 20\%$), or whether the token comes from MATH or FLAN (**Oracle Domain Routing**); and *supervision objective*, retaining CE on all tokens (**Always CE**) or replacing soft distillation with the teacher's top-1 predictions (**Teacher Top-1 Labels**).

Replacing RKL with FKL modestly reduces REASONING and KNOWLEDGE & COMMONSENSE, suggesting that RKL may better exploit the sharp, low-entropy teacher distributions selected by our routing strategy. Alternative routing signals consistently underperform teacher entropy, while **Always CE** has little effect. Finally, **Teacher Top-1 Labels** improves FACTUAL RECALL at the expense of the other categories. These results identify entropy-based routing as the driver of our method's gains, while soft teacher distributions provide useful supervision beyond top-1 predictions.

7 Discussion

Enabling language models to learn more from a fixed data pool remains a longstanding challenge, typically addressed by improving aggregate training data quality. We identify a complementary direction: improving *how* existing tokens are learned during mid-training, where high-quality tokens are scarce and expensive. We argue that knowledge distillation should not be stage-agnostic: while the standard formulation is effective during pre-training, it exhibits fundamentally different behavior during mid-training, where teacher uncertainty and student knowledge interact to produce a reasoning–recall tradeoff. By explicitly accounting for this interaction, SWITCH DISTILLATION consistently improves reasoning and largely preserves factual recall in a token-matched setting. More broadly, our findings suggest that objectives themselves ought to be stage-aware. While we study mid-training, this principle may extend to other phases where the student's knowledge has substantially evolved, such as late-stage or continual pre-training. We hope our findings motivate further investigation into stage-aware optimization methods for data-efficient language modeling.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan Son NGUYEN, Ben Burtenshaw, Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, Colin Raffel, Leandro Von Werra, and Thomas Wolf. SmolLM2: When smol goes big — data-centric training of a fully open small language model. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=3JiCl2A14H.

Allen Institute for AI. Open instruct: Allenai's post-training codebase. https://github.com/allenai/open-instruct, 2025. Accessed: 2026-07-28.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pp. 535–541, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933395. doi: 10.1145/1150402.1150464. URL https://doi.org/10.1145/1150402.1150464.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russell Webb. Distillation scaling laws. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=1nEBAkpfb9.

Sungmin Cha and Kyunghyun Cho. Why knowledge distillation works in generative models: A minimal working explanation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 30017–30037. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2b13864517555dd14f492abdce0469f3-Paper-Conference.pdf.

Mayee F Chen, Tyler Murray, David Heineman, Matt Jordan, Hannaneh Hajishirzi, Christopher Re, Luca Soldaini, and Kyle Lo. Olmix: A framework for data mixing throughout LM development. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=8pOl3azhbL.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

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

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs, 2019. URL https://arxiv.org/abs/1903.00161.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.444. URL https://aclanthology.org/2024.emnlp-main.444/.

Gaurav Rohit Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. Understanding finetuning for factual knowledge extraction. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=cPsn9AcOYh.

Sachin Goyal, David Lopez-Paz, and Kartik Ahuja. Distilled pretraining: A modern lens of data, in-context learning and test-time scaling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PNm2dl7HcY.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le,

Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A standard for language model evaluations. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5020–5048, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.282. URL https://aclanthology.org/2025.findings-naacl.282/.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023. URL https://arxiv.org/abs/2306.11644.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In

Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Junjie Huang, Jiarui Qin, Di Yin, Weiwen Liu, Yong Yu, Xing Sun, and Weinan Zhang. Remit: Rl-guided mid-training for iterative llm evolution, 2026. URL https://arxiv.org/abs/2602.03075.

Hugging Face Team. Smollm3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

IBM Research. Granite 3.3 language models. https://huggingface.co/ibm-granite/granite-3.3-8b-instruct, 2025. Accessed: 2026-08-18.

Woogyeol Jin, Taywon Min, Yongjin Yang, Dennis Wei, Yi Zhou, Swanand Ravindra Kadhe, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models, 2026. URL https://arxiv.org/abs/2603.07079.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Guy Kaplan, Zorik Gekhman, Zhen Zhu, Lotem Rozner, Yuval Reif, Swabha Swayamdipta, Derek Hoiem, and Roy Schwartz. Why fine-tuning encourages hallucinations and how to fix it, 2026. URL https://arxiv.org/abs/2604.15574.

Konwoo Kim, Suhas Kotha, Yejin Choi, Tatsunori Hashimoto, Nick Haber, and Percy Liang. Data-efficient pre-training by scaling synthetic megadocs, 2026a. URL https://arxiv.org/abs/2603.18534.

Konwoo Kim, Suhas Kotha, Percy Liang, and Tatsunori Hashimoto. Pre-training under infinite compute. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=ck0aZTAnwK.

Minsang Kim and Seung Jun Baek. Explain in your own words: Improving reasoning via token-selective dual knowledge distillation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zph7e5JaXc.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation, 2016. URL https://arxiv.org/abs/1606.07947.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Margaret Li, Sneha Kudugunta, Danielle Rothermel, and Luke Zettlemoyer. Slicing and dicing: Configuring optimal mixtures of experts, 2026. URL https://arxiv.org/abs/2605.11689.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2961–2984, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.163. URL https://aclanthology.org/2024.acl-long.163/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, yelong shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Not all tokens are what you need for pretraining. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0NMzBwqaAJ.

Emmy Liu, Graham Neubig, and Chenyan Xiong. Midtraining bridges pretraining and posttraining distributions, 2026. URL https://arxiv.org/abs/2510.14865.

Taiming Lu and Zhuang Liu. Strong teacher not needed? on distillation in llm pretraining, 2026. URL https://arxiv.org/abs/2605.23857.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14044–14072, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.757. URL https://aclanthology.org/2024.acl-long.757/.

Meta Superintelligence Labs. Introducing muse glimmer: An open agentic model that runs on your device. https://research.meta.ai/blog/introducing-muse-glimmer-open-agentic-model, August 2026. Accessed: 2026-08-12.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018. URL https://arxiv.org/abs/1809.02789.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2025. URL https://arxiv.org/abs/2410.05229.

Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019. URL https://arxiv.org/abs/1902.03393.

Thao Nguyen, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, and Xian Li. Recycling the web: A method to enhance pre-training data quality and quantity for language models. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=lkjhBdz3rn.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Andrej Risteski, and Surbhi Goel. Progressive distillation induces an implicit curriculum, 2024. URL https://arxiv.org/abs/2410.05464.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=YW6edSufht.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Ellen Xiaoqing Tan, Jack Lanchantin, Shehzaad Dhuliawala, Danwei Li, Thao Nguyen, Jing Xu, Ping Yu, Ilia Kulikov, Sainbayar Sukhbaatar, Jason Weston, Xian Li, and Olga Golovneva. Self-improving pretraining: using post-trained models to pretrain better models, 2026. URL https://arxiv.org/abs/2601.21343.

Team Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya,

Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL https://arxiv.org/abs/2512.13961.

Mathurin Videau, Badr Youbi Idrissi, Daniel Haziza, Luca Wehrstedt, Jade Copet, Olivier Teytaud, and David Lopez-Paz. Meta Lingua: A minimal PyTorch LLM training library, 2024. URL https://github.com/facebookresearch/lingua.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=2ezugTT9kU.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: a more robust and challenging multi-task language understanding benchmark. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling, 2025. URL https://arxiv.org/abs/2506.20512.

Ishaan Watts, Catherine Li, Sachin Goyal, Jacob Mitchell Springer, and Aditi Raghunathan. Sharpness-aware pretraining mitigates catastrophic forgetting, 2026. URL https://arxiv.org/abs/2605.02105.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking Kullback-Leibler divergence in knowledge distillation for large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5737–5755, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.383/.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*,

volume 36, 2023a. URL https://papers.nips.cc/paper_files/paper/2023/hash/dcba6be91359358c2355cd920da3fcbd-Abstract-Conference.html.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=uPSQv0leAu.

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

Jiaqing Yuan, Lin Pan, Chung-Wei Hang, Jiang Guo, Jiarong Jiang, Bonan Min, Patrick Ng, and Zhiguo Wang. Towards a holistic evaluation of llms on factual knowledge recall, 2024. URL https://arxiv.org/abs/2404.16164.

Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and rl on reasoning language models, 2025. URL https://arxiv.org/abs/2512.07783.

Qihuang Zhong, Liang Ding, Li Shen, Juhua Liu, Bo Du, and Dacheng Tao. Revisiting knowledge distillation for autoregressive language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10900–10913, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.587. URL https://aclanthology.org/2024.acl-long.587/.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023. URL https://arxiv.org/abs/2304.06364.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

APPENDIX

A General **19**

A.1 Related Work 19

A.2 Limitations 20

A.3 AI Usage Statement 20

A.4 Reproducibility Statement 21

B SWITCH DISTILLATION Details **21**

B.1 Pseudocode 21

B.2 Ablating q 21

B.3 Partition normalization. 21

B.4 Distributed quantile estimation. 22

C Experimental Details **22**

C.1 Training setup 22

C.2 Evaluation 23

C.3 Analysis methodology 23

C.4 Standalone student and teacher results 26

D Supplementary Analyses **26**

D.1 How generalizable is the mid-training tradeoff across model families? 26

D.2 How generalizable is teacher supervision asymmetry? 26

D.3 Robustness across teacher sizes 27

D.4 Forward and reverse KL exhibit different optimization geometry. 29

D.5 SWITCH DISTILLATION accelerates mid-training reasoning acquisition 29

D.6 SWITCH DISTILLATION improves pass@k at low sampling budgets 30

D.7 SWITCH DISTILLATION improves code generation performance 31

E Full Results **32**

E.1 Intermediate post-training results 32

E.2 Ablation mid-training results 34

A General

A.1 Related Work

Mid-training and its origins. Foundation model development has converged on mid-training as a distinct stage between pre-training and post-training, during which models are further optimized in a self-supervised fashion on curated data mixtures (Zhang et al., 2025; Liu et al., 2026). While mid-training has its roots in continued pre-training, its modern formulation emphasizes capability-focused data mixtures designed to better prime models for subsequent post-training (Gururangan et al., 2020).

Recent work has indicated that surfacing post-training capabilities earlier during mid-training can further strengthen desirable downstream attributes; carefully designed mid-training recipes have been found to particularly benefit subsequent reinforcement learning (Wang et al., 2025; Huang et al., 2026; Liu et al., 2026; Tan et al., 2026). To date, most of the mid-training literature employs the standard next-token prediction objective. We revisit knowledge distillation in this setting and uncover a previously uncharacterized reasoning–recall tradeoff.

Data-efficient language modeling. As the growth of compute is on track to outpace the supply of organic web text, high-quality human-written data is becoming scarce for language model training (Kim et al., 2026b). This impending constraint has motivated a body of work on *data-efficient language modeling*. Prior approaches have largely pursued data efficiency by improving the training corpus itself through curation, augmentation, selection, or mixture optimization (Gunasekar et al., 2023; Xie et al., 2023b;a; Lin et al., 2024; Maini et al., 2024; Nguyen et al., 2025; Chen et al., 2026; Kim et al., 2026a), often with guidance from stronger reference models. More broadly, Kim et al. (2026b) argue that algorithmic interventions such as ensembling or self-distillation may soon serve as important avenues for tackling the data wall. We study this algorithmic perspective at the mid-training stage, which necessitates substantially higher-quality data than large-scale pre-training while consuming orders of magnitude more tokens than post-training. Our approach is complementary to these data-centric methods: rather than modifying the training corpus, we improve how supervision is extracted from each observed token.

Knowledge distillation across training stages. Knowledge distillation broadly encompasses both hard (sequence-level) distillation, in which a teacher generates synthetic training sequences for subsequent next-token prediction (Kim & Rush, 2016), and soft (logit-based) distillation, in which the student is trained to match the teacher's predictions by minimizing its distributional divergence (Buciluǎ et al., 2006; Hinton et al., 2015). We focus on the latter, whose role has been studied extensively during language model pre-training and post-training.

In the pre-training setting, Busbridge et al. (2025) derive scaling laws for language model distillation and characterize compute-optimal teacher-student configurations under fixed compute budgets. Goyal et al. (2026) show that pre-training distillation improves test-time scaling at the cost of in-context learning and propose token-routing KD (TRKD) to mitigate this degradation. While TRKD targets in-context learning during pre-training, Switch Distillation addresses the reasoning–recall tradeoff that emerges during mid-training. Finally, Cha & Cho (2025) show that generative distillation induces a precision–recall tradeoff, whereby lower-entropy teachers produce sharper but lower-coverage students.

Recent work has revisited the standard KD formula primarily in the post-training setting. MiniLLM, for one, advocates reverse-KL distillation to improve generative capabilities of LMs, and subsequent work has proposed adaptive combinations of forward and reverse KL to exploit their distinct optimization behaviors (Gu et al., 2024; Zhong et al., 2024; Wu et al., 2025). Reverse-KL is the primary KD direction for *on-policy knowledge distillation*, a relatively new post-training paradigm in which the student is trained on its own sampled trajectories under teacher guidance (Agarwal et al., 2024). Token-Selective Dual Knowledge Distillation (TSD-KD) (Kim & Baek, 2026) is an on-policy KD method that selectively applies teacher supervision based on teacher–student confidence discrepancies and teacher-ranked student-generated reasoning trajectories. Entropy-Aware On-Policy Distillation (EOPD) proposes to use teacher predictive entropy as a signal to interpolate between reverse and forward KL for high-entropy teacher distributions (Jin et al., 2026). Among prior work, EOPD is conceptually closest to Switch Distillation, but differs in both training

regime and use of entropy: EOPD operates during post-training on student-generated trajectories and always applies teacher-based distillation, adapting the KL objective as teacher uncertainty varies. In contrast, SWITCH DISTILLATION operates at the self-supervised mid-training stage on fixed tokens and uses entropy to determine whether to obtain supervision from the corpus or the teacher model.

Collectively, these works establish knowledge distillation as an effective supervision strategy during pre-training and post-training. We extend this line of work to the emerging mid-training regime and show that KD exhibits a fundamentally different tradeoff. This behavior consequently motivates a stage-specific adaptation of the standard distillation objective.

A.2 LIMITATIONS

Our main experiments are conducted with the OLMo-2 recipe, which enables controlled and replicable experimentation across pre-training, mid-training, and post-training. We provide supplementary evidence that our findings extend beyond this setting: the stage-dependent tradeoff exhibited by KD also appears in the SmolLM2 family (App. D.1), while similar asymmetries in teacher supervision emerge across teachers from different training stages and model families (App. D.2). Broader controlled validation is difficult because isolating stage-dependent distillation effects requires access to more than model weights: it requires intermediate pre- and mid-training checkpoints, the corresponding data mixtures, and sufficiently complete training recipes to reproduce transitions between stages. Few model families currently release all of these artifacts. Moreover, logit-based distillation mathematically requires compatible output vocabulary between student and teacher, further restricting the set of viable model pairs. We therefore center our controlled experiments on OLMo-2, with SmolLM2 and cross-family teacher analyses as complementary tests for generality.

Similarly, we adopt competitive defaults wherever possible: while our student model sizes are relatively small (1B parameters), we train substantially beyond Chinchilla-optimal token budgets and distill from Instruct models that outperform their base counterparts both as standalone models and as teachers. We do not exhaustively ablate these experimental choices, such as the effect of teacher post-training or broader teacher scales, as doing so would require substantial additional compute and prior work has already characterized several of these dimensions; for example, increasing the teacher-student capacity gap can impair distillation effectiveness (Mirzadeh et al., 2019). Our experiments instead focus compute on isolating how distillation behavior changes across training stages and objectives.

Exploring better strategies for selective distillation is an interesting future extension. SWITCH DISTILLATION routes supervision using teacher predictive entropy, a simple primitive that requires no additional supervision or parameters. While our proposed algorithm is cheap and effective, richer and more expressive strategies (i.e., a learned router network, analogous to those employed by Mixture-of-Experts architectures (Shazeer et al., 2017; Li et al., 2026)) may better capture when and where teacher supervision is beneficial. As SWITCH DISTILLATION arises from our study on how best to leverage teacher supervision conditional on fixed data, it may be broadly compatible with approaches that instead optimize the data pool itself.

Finally, while we develop and evaluate SWITCH DISTILLATION as a *mid-training* strategy, our tradeoff analysis suggests that it may be beneficial more broadly whenever factual acquisition slows under teacher supervision. In particular, we hypothesize that SWITCH DISTILLATION may also improve upon standard KD during late-stage pre-training, when the student has already acquired much of the easily transferred knowledge from the teacher. Characterizing when standard KD ceases to yield Pareto improvements and begins to induce such a tradeoff is an important future direction.

A.3 AI USAGE STATEMENT

We used generative AI tools in this work for lightweight assistance with copy-editing the manuscript, creating and improving the presentation of scientific figures, debugging implementations, and automating our training and evaluation scripts. We ran our submission through Google's Paper Assistant Tool (PAT) as made available by the ICLR 2027 committee. We did not use any generative AI for methodological or experimental design, the analysis and interpretation of results, or the identification of relevant prior work. All AI-assisted code, figures, and text were manually reviewed before

Algorithm 1 SWITCH DISTILLATION

Input: Student model p_S, teacher model p_T, local token batch $x_{1:B}$, routing quantile $q \in (0, 1)$, temperature τ

Output: Training loss $\mathcal{L}_{\text{SWITCH DISTILLATION}}$

$\mathcal{B}_{\text{tok}} \leftarrow$ all valid next-token positions in $x_{1:B}$, with target y_n at each position n

$z_T, z_S \leftarrow \text{Forward}(p_T, x_{1:B}), \text{Forward}(p_S, x_{1:B})$ ▷ Teacher and student logits

▷ Compute softened distributions used for logit-based distillation.

$p_{T,n}^{(\tau)} \leftarrow \text{softmax}(z_T^n/\tau)$ for all $n \in \mathcal{B}_{\text{tok}}$

$p_{S,n}^{(\tau)} \leftarrow \text{softmax}(z_S^n/\tau)$ for all $n \in \mathcal{B}_{\text{tok}}$

▷ Score tokens by teacher predictive entropy and route by quantile.

$H_n \leftarrow -\sum_{v \in \mathcal{V}} p_{T,n}^{(\tau)}(v) \log p_{T,n}^{(\tau)}(v)$ for all $n \in \mathcal{B}_{\text{tok}}$

$\mathcal{S}_q \leftarrow \left\{ n \in \mathcal{B}_{\text{tok}} : H_n \leq \text{Quantile}_q(\{H_{n'} : n' \in \mathcal{B}_{\text{tok}}\}) \right\}$ ▷ Route low-entropy tokens to KD

$\bar{\mathcal{S}}_q \leftarrow \mathcal{B}_{\text{tok}} \setminus \mathcal{S}_q$

▷ Compute separately normalized RKL and CE objectives.

$$\mathcal{L}_{\text{RKL}} \leftarrow \frac{\tau^2}{|\mathcal{S}_q|} \sum_{n \in \mathcal{S}_q} \text{KL}\left(p_{S,n}^{(\tau)} \,\middle\|\, p_{T,n}^{(\tau)} \right)$$

$$\mathcal{L}_{\text{CE}} \leftarrow \frac{1}{|\bar{\mathcal{S}}_q|} \sum_{n \in \bar{\mathcal{S}}_q} [-\log p_S(y_n \mid x_{<n})]$$

$\mathcal{L}_{\text{SWITCH DISTILLATION}} \leftarrow \mathcal{L}_{\text{RKL}} + \mathcal{L}_{\text{CE}}$

return $\mathcal{L}_{\text{SWITCH DISTILLATION}}$

use. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the assistance of generative AI.

A.4 REPRODUCIBILITY STATEMENT

All experiments are conducted using open-source training and evaluation stacks, with publicly available models and datasets. We provide complete training configuration, hyperparameter, and evaluation details in App. C. Moreover, SWITCH DISTILLATION requires only a minimal modification to standard knowledge distillation; we describe the mechanics of it in considerable detail in §5 and provide the pseudocode in App. B.1.

B SWITCH DISTILLATION DETAILS

B.1 PSEUDOCODE

We provide the pseudocode for SWITCH DISTILLATION in Alg. 1.

B.2 ABLATING q

We sweep the routing threshold $q \in \{10\%, 20\%, 30\%\}$ for both the 7B and 13B teacher settings. Table 4 shows per-task results; overall, we choose $q = 20\%$ as the default routing threshold across teacher sizes, as it provides the highest reasoning performance, while remaining competitive or best on FACTUAL RECALL and KNOWLEDGE & COMMONSENSE. Performance is relatively stable between $q = 20\%$ and $q = 30\%$, suggesting that the method is not highly sensitive to the precise routing threshold.

B.3 PARTITION NORMALIZATION.

We additionally evaluate uniform token normalization across the batch by removing the relative per-token upweighting induced by separate partition normalization. For the 7B teacher case, this re-

Table 4: **Downstream results for SWITCH DISTILLATION at routing thresholds $q \in \{10\%, 20\%, 30\%\}$. Bold** denotes the best result per teacher block.

		Reasoning							Factual Recall				Knowledge & Commonsense						
T	Method	GSM8K	GSM-S	GSM+	BBH	DROP	MATH	**Avg.**	TQA	NQ	SQA	**Avg.**	MMLU	MMLU-P	ARC-C	OBQA	Wino	AGI	**Avg.**
7B	$q = 10\%$	61.6	48.8	39.4	32.5	**50.2**	11.4	40.6	53.7	23.9	8.6	28.7	50.6	19.0	62.0	63.8	**55.5**	40.4	48.5
7B	$q = 20\%$	69.7	**55.3**	**46.1**	32.8	49.6	**14.8**	**44.7**	**54.9**	24.6	8.4	29.3	**51.6**	**19.8**	**64.7**	**64.2**	53.8	**41.5**	**49.3**
7B	$q = 30\%$	**69.8**	54.1	45.6	**33.8**	47.2	13.0	43.9	54.7	**25.3**	**8.8**	**29.6**	51.0	19.5	63.5	62.4	53.0	40.7	48.4
13B	$q = 10\%$	62.1	49.6	38.3	28.6	44.2	7.8	38.5	52.2	23.1	**9.2**	28.1	46.8	17.1	55.5	59.6	52.7	37.9	45.0
13B	$q = 20\%$	**66.0**	**54.8**	42.2	31.6	**45.8**	**12.0**	**42.1**	54.1	**24.9**	9.0	**29.3**	48.5	17.7	58.5	**63.8**	51.2	**39.2**	46.5
13B	$q = 30\%$	**66.0**	53.1	**42.9**	**32.5**	44.4	11.0	41.6	**54.2**	24.7	8.6	29.2	**49.0**	**18.4**	**60.9**	63.6	**52.9**	38.7	**47.2**

duces average reasoning by 5.4 points while improving factual recall modestly by 1 point. This shift is consistent with the reasoning–recall tradeoff observed under increasing distillation strength (§3). Notably, reasoning still remains substantially above NTP under uniform normalization, while random routing under our default normalization also substantially underperforms entropy-based routing (Table 3), indicating that both entropy-based routing and separate partition normalization contribute to the reasoning gains.

B.4 DISTRIBUTED QUANTILE ESTIMATION.

Our default implementation is communication-free: it computes the routing quantile over the local micro-batch and avoids an additional synchronization point in the training loop. While this overhead is modest relative to model computation at small scale, it becomes less desirable as teacher and student models grow and available memory and communication headroom decrease. In practice, our local micro-batches are sufficiently large and representative of the data mixture such that we observe negligible differences in downstream performance compared with global quantile routing.

However, this assumption may break down when the micro-batch size is constrained to be very small, which may occur with larger teacher and student models. In this setting, one potential alternative is to employ a lagged global quantile. At each optimizer step, we can aggregate teacher entropies over the full global batch and use its q-quantile to route the subsequent batch. This decouples quantile estimation from micro-batch size while avoiding a blocking global-quantile computation before routing the current batch. We leave this exploration to future work.

C EXPERIMENTAL DETAILS

C.1 TRAINING SETUP

All our experiments make use of open-source code, checkpoints, and data. We use the `lingua` (Videau et al., 2024) framework for pre-training and mid-training, and the `open-instruct` (Allen Institute for AI, 2025) repository for post-training.

We run pre-training, mid-training, and all stages of post-training except DPO on 32 NVIDIA H200 Tensor Core GPUs across 4 nodes; DPO is conducted on a single node. For efficient distillation, our 7B and 13B teachers are loaded in FP8 quantization, while 1B teachers are kept in BF16; we find that teacher quantization does not substantially affect the predictive entropy ranking of tokens. We set our distillation temperature to $\tau = 2$ for all settings.

Student model architecture. We provide architecture details for our 1B student model in Table 5.

Training hyperparameters. We provide pre-training and mid-training hyperparameters in Table 6. For post-training, we defer hyperparameter choices to the official setup from Walsh et al. (2025), with the exception of a lower learning rate ($5e^{-6}$) during supervised fine-tuning. We find that the catastrophic forgetting of factual knowledge is most pronounced during SFT. Consistent with recommendations from prior work, a smaller learning rate mitigates this loss (Springer et al., 2025), although it comes at the cost of slightly weaker reasoning performance overall.

Table 5: **1B student architecture.** We follow the OLMo-2 1B model configuration from Walsh et al. (2025).

SETTING	VALUE
Architecture	OLMo-2 1B (`Olmo2ForCausalLM`)
Hidden dimension	2048
Transformer layers	16
Attention heads	16
KV heads	16 (full multi-head attention)
FFN intermediate dimension	8192
Normalization	RMSNorm ($\epsilon = 10^{-6}$)
QK normalization	Enabled
Positional encoding	RoPE ($\theta = 500{,}000$)
Maximum sequence length	4096
Weight tying	Disabled
Parameters	1,484,613,632
Vocab Size	100352

Table 6: **Hyperparameters for pre-training and mid-training.**

SETTING	PRE-TRAINING	MID-TRAINING
Training steps	48,000	28,800
Training tokens	100.7B	60.4B
Peak learning rate	4×10^{-4}	7.45×10^{-5}
Learning rate schedule	Cosine	Linear decay
Minimum LR ratio	0.1	0.0
Warmup	4,000 steps	None
Initialization	Random	OLMo-2-0425-1B-stage1-4001B
	Shared	
Global batch size	2,097,152 tokens/step	
Optimizer	AdamW	
β_1, β_2	(0.9, 0.95)	
Weight decay	0.1	
Gradient clipping	1.0	

C.2 EVALUATION

Evaluation tasks. Table 7 shows the evaluation tasks used in this paper. We adopt the same defaults (e.g., few-shot exemplars, sampling strategy) as Gu et al. (2025).

Baselines. In our implementation of TRKD (Goyal et al., 2026), we disable the distillation loss on the 15% lowest-teacher-entropy tokens, retaining only ground-truth supervision (CE loss) on these tokens (following the original paper's suggestions); the remaining 85% receive a convex blend $(1-\lambda), \mathrm{CE} + \lambda, T^2, \mathrm{KL}(p_T|p_S)$. We set the forward KL mixing coefficient λ to 0.5.

C.3 ANALYSIS METHODOLOGY

Teacher supervision analysis. For each teacher, we score the same 240 Dolmino documents (40 per domain, sampled with a fixed seed), yielding 107,555 token-level next-token predictions in total, decomposed by domain as 25,981 from DCLM, 22,514 from Wikipedia, 16,692 from StackExchange, 16,599 from PeS2o, 12,886 from Math, and 12,883 from FLAN. All three teachers therefore probe identical positions; between- and within-teacher comparisons share the same support.

Gold-answer gradient analysis. In §4.3, we analyze the gradients induced by the NTP, FKD, and RKD training objectives with respect to the ground-truth token. For FKD and RKD, we evaluate at $\alpha = 0.5$; thus, the dashed $0.5\times$ NTP level in Fig. 5 corresponds to the $1-\alpha$ floor as the distillation-gradient contribution vanishes.

Table 7: Evaluation tasks used for mid-training (**Mid**) and post-training (**Post**). We follow the standard evaluation settings used in the OLMES evaluation harness (Gu et al., 2025).

Category	Format	Tasks	Mid	Post
REASONING	Generation	**GSM8K** (Cobbe et al., 2021), **GSM-Symbolic** (Mirzadeh et al., 2025), **GSM-Plus** (Li et al., 2024), **BBH** (Suzgun et al., 2022), **DROP** (Dua et al., 2019), **MATH** (Lightman et al., 2023)	✓	✓
FACTUAL RECALL	Generation	**TriviaQA** (Joshi et al., 2017), **Natural Questions** (Kwiatkowski et al., 2019), **SimpleQA** (Wei et al., 2024)	✓	✓
KNOWLEDGE & COMMONSENSE	Multiple choice	**MMLU** (Hendrycks et al., 2021), **MMLU-Pro** (Wang et al., 2024), **ARC-Challenge** (Clark et al., 2018), **OpenBookQA** (Mihaylov et al., 2018), **WinoGrande** (Sakaguchi et al., 2019), **AGIEval** (Zhong et al., 2023)	✓	✓
INSTRUCTION FOLLOWING	Generation	**IFEval** (Zhou et al., 2023)		✓

Let A denote the set of accepted first-token IDs obtained from the gold answer aliases (e.g., if a prompt is "Name a European capital: ", then A might include the token IDs corresponding to {"Paris", "London", "Berlin"}). We define the *gold-answer gradient* g as the negative gradient of some loss $\mathcal{L}$ with respect to the student logits corresponding to the set A:

$$g = -\frac{\partial \mathcal{L}}{\partial z_{S,A}}, \quad \text{where} \quad \frac{\partial}{\partial z_{S,A}} = \sum_{y \in A} \frac{\partial}{\partial z_{S,y}}. \tag{7}$$

Intuitively, g is a scalar that measures the strength of the learning signal induced by $\mathcal{L}$ on all accepted answer tokens in A; larger values of g correspond to stronger pressure to increase the probability of the correct answer. We next derive the closed forms of g for the training objectives in our study: g_{NTP}, g_{FKD}, and g_{RKD}.

Deriving g_{NTP}. Recall that the NTP objective is the standard cross-entropy loss ($\mathcal{L}_{\mathrm{CE}}$). Let $p_S^\tau = \mathrm{softmax}(z_S/\tau)$ denote the student distribution at temperature τ. The cross-entropy term is computed at τ=1 against the observed next token $y^\star$. By the construction of our probe, we guarantee that $y^\star \in A$. Since $\mathcal{L}_{\mathrm{CE}} = -\log p_S^1(y^\star)$ and using the standard Jacobian of the log-softmax, $\frac{\partial \log p_S^1(y^\star)}{\partial z_{S,j}} = \delta_{jy^\star} - p_S^1(j)$ for any arbitrary student logit $z_{S,j}$,

$$\begin{aligned} g_{\mathrm{NTP}} &= -\sum_{y \in A} \frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial z_{S,y}} && (8)\\ &= -\sum_{y \in A} \left(p_S^1(y) - \delta_{yy^\star}\right) && (9)\\ &= -\sum_{y \in A} p_S^1(y) + \sum_{y \in A} \delta_{yy^\star} && (10)\\ &= 1 - p_S^1(A). && (11) \end{aligned}$$

Intuitively, this gradient represents the probability mass still missing from the accepted answer set. The cross-entropy loss pulls these logits upward until $p_S^1(A) = 1$. Note that this target is absolute and does not depend on any reference teacher.

Deriving g_{FKD}. We next examine the FKD objective, which consists of a convex combination of the forward KL (FKL) loss and the cross-entropy loss as weighted by $\alpha \in [0, 1]$.

Let us tackle the FKL loss first, which depends on a reference teacher T. Recall that the FKL equation is, by definition, $\mathrm{KL}(p_T^\tau \| p_S^\tau) = \sum_y p_T^\tau(y) \log p_T^\tau(y) - \sum_y p_T^\tau(y) \log p_S^\tau(y)$, with p_T^τ as the teacher distribution and p_S^τ as the student distribution, both scaled by temperature τ. Since the teacher distribution p_T^τ is fixed with respect to the student logits z_S, the first term is constant and has zero derivative, and only the second term $-\sum_y p_T^\tau(y) \log p_S^\tau(y)$ contributes to the gradient.

Using the standard Jacobian of the log-softmax, $\frac{\partial \log p_S^\tau(y)}{\partial z_{S,j}} = \frac{1}{\tau}\big(\delta_{yj} - p_S^\tau(j)\big)$, and the observation that $\sum_y p_T^\tau(y) = 1$,

$$\frac{\partial \,\mathrm{KL}(p_T^\tau \| p_S^\tau)}{\partial z_{S,j}} = -\sum_y p_T^\tau(y)\,\frac{1}{\tau}\big(\delta_{yj} - p_S^\tau(j)\big) \tag{12}$$

$$= \frac{1}{\tau}\big(p_S^\tau(j) - p_T^\tau(j)\big). \tag{13}$$

Summing the negative gradients over $y \in A$ and combining with the cross-entropy term in the full FKD objective $\mathcal{L}_{\mathrm{FKD}} = (1-\alpha)\mathcal{L}_{\mathrm{CE}} + \alpha\tau^2\,\mathrm{KL}(p_T^\tau \| p_S^\tau)$, where the τ^2 cancels one factor of $1/\tau$, leads to

$$g_{\mathrm{FKD}} = -\sum_{y\in A} \frac{\partial \mathcal{L}_{\mathrm{FKD}}}{\partial z_{S,y}} \tag{14}$$

$$= \underbrace{(1-\alpha)\big(1 - p_S^1(A)\big)}_{\text{CE term}} + \underbrace{\alpha\tau\big(p_T^\tau(A) - p_S^\tau(A)\big)}_{\text{FKL term}}. \tag{15}$$

Comparing against g_{NTP}, the FKL term replaces the cross-entropy target of 1 with $p_T^\tau(A)$, the teacher's probability mass on the accepted answer set. Thus, unlike cross-entropy, the distillation component does not continue increasing the student's accepted-answer mass once it reaches the teacher's: when $p_S^\tau(A) = p_T^\tau(A)$, the FKL contribution vanishes.

When $p_S^\tau(A) > p_T^\tau(A)$, it becomes negative and actively pushes the student back toward the teacher distribution. In the full FKD objective (Eq. (15)), this negative distillation gradient competes with the remaining positive cross-entropy gradient and can dominate when the teacher assigns substantially less mass to A than the student. This regime may be especially relevant during mid-training, where the student may already assign high probability to factual continuations for which the teacher remains uncertain.

Deriving g_{RKD}. We now examine the RKD objective, which substitutes the Forward KL divergence with the Reverse KL divergence, $\mathrm{RKL}(p_S^\tau \| p_T^\tau) = \sum_y p_S^\tau(y) \log \frac{p_S^\tau(y)}{p_T^\tau(y)}$. To find the gradient with respect to any arbitrary student logit $z_{S,j}$, we apply the product rule and the softmax Jacobian $\frac{\partial p_S^\tau(y)}{\partial z_{S,j}} = \frac{1}{\tau} p_S^\tau(y)\big(\delta_{yj} - p_S^\tau(j)\big)$:

$$\frac{\partial \,\mathrm{RKL}(p_S^\tau \| p_T^\tau)}{\partial z_{S,j}} = \sum_y \frac{\partial p_S^\tau(y)}{\partial z_{S,j}} \left(\log \frac{p_S^\tau(y)}{p_T^\tau(y)} + 1\right) \tag{16}$$

$$= \frac{1}{\tau}\sum_y p_S^\tau(y)\big(\delta_{yj} - p_S^\tau(j)\big) \left(\log \frac{p_S^\tau(y)}{p_T^\tau(y)} + 1\right). \tag{17}$$

Because the sum of probabilities is 1, the gradient of that sum is zero ($\sum_y \frac{\partial p_S^\tau(y)}{\partial z_{S,j}} \cdot 1 = 0$), causing the $+1$ term to vanish. Distributing the remaining terms yields

$$\frac{\partial \,\mathrm{RKL}(p_S^\tau \| p_T^\tau)}{\partial z_{S,j}} = \frac{1}{\tau}\left[p_S^\tau(j) \log \frac{p_S^\tau(j)}{p_T^\tau(j)} - p_S^\tau(j) \sum_y p_S^\tau(y) \log \frac{p_S^\tau(y)}{p_T^\tau(y)}\right] \tag{18}$$

$$= \frac{1}{\tau} p_S^\tau(j)\Big(\log \frac{p_S^\tau(j)}{p_T^\tau(j)} - \mathrm{RKL}(p_S^\tau \| p_T^\tau)\Big). \tag{19}$$

Summing the negative gradients over $y \in A$ and combining with the cross-entropy term under the full RKD objective $\mathcal{L}_{\mathrm{RKD}} = (1-\alpha)\mathcal{L}_{\mathrm{CE}} + \alpha\tau^2\,\mathrm{RKL}(p_S^\tau \| p_T^\tau)$, we obtain:

$$g_{\mathrm{RKD}} = -\sum_{y\in A} \frac{\partial \mathcal{L}_{\mathrm{RKD}}}{\partial z_{S,y}} \tag{20}$$

$$= \underbrace{(1-\alpha)\big(1 - p_S^1(A)\big)}_{\text{CE term}} \underbrace{- \alpha\tau \sum_{y\in A} p_S^\tau(y)\Big(\log \frac{p_S^\tau(y)}{p_T^\tau(y)} - \mathrm{RKL}(p_S^\tau \| p_T^\tau)\Big)}_{\text{RKL term}}. \tag{21}$$

Table 8: **Downstream performance of the student initialization and teacher models.** We report performance of the OLMo-2 1B Stage 1 student initialization and the OLMo-2 1B, 7B, and 13B Instruct teachers on the downstream evaluation suite. Benchmark names are abbreviated; see Table 7 for full task names.

	Reasoning						Factual Recall			Knowledge & Commonsense					
Model	**GSM8K**	**GSM-S**	**GSM+**	**BBH**	**DROP**	**MATH**	**TQA**	**NQ**	**SQA**	**MMLU**	**MMLU-P**	**ARC-C**	**OBQA**	**Wino**	**AGI**
OLMo-2 1B STAGE 1	3.2	1.9	2.1	29.9	24.8	1.2	52.5	21.0	8.0	26.9	11.3	26.2	26.2	50.6	23.9
OLMo-2 1B-INST	67.3	49.1	42.7	35.2	32.3	9.0	50.8	21.1	7.5	41.9	17.8	47.2	48.8	50.4	35.0
OLMo-2 7B-INST	79.2	69.6	57.0	50.4	59.2	17.4	75.8	38.1	10.6	61.8	32.4	80.1	78.4	66.9	50.1
OLMo-2 13B-INST	82.9	72.6	59.9	58.2	70.4	22.0	83.2	49.4	10.9	66.7	37.2	83.4	85.0	71.3	54.0

Note that in Eq. (21), RKL can likewise exert active negative pressure on ground-truth answer logits rather than merely reducing their positive supervision. The contribution for an accepted answer token y is negative to g_{RKD} whenever its log-probability ratio $\log \frac{p_S^\tau(y)}{p_T^\tau(y)}$ exceeds the student-averaged log ratio $\mathrm{RKL}(p_S^\tau \| p_T^\tau)$.

Thus, although their respective gradient structure may differ, both KL directions can penalize student predictions on the ground truth answers if they exceed the teacher's relative preference.

C.4 STANDALONE STUDENT AND TEACHER RESULTS

For reference, we report student (OLMo-2 1B, after pre-training but before mid-training) and teacher (OLMo-2 1B, 7B, and 13B Instruct) performance in Table 8. We remark that the 1B student already matches the 1B Instruct teacher on factual recall, despite substantially weaker reasoning performance, suggesting that much of the teacher's advantage at the start of mid-training lies in procedural capabilities rather than factual knowledge. The larger teachers, however, consistently outperform the student across all evaluated capabilities, indicating that the observed reasoning–recall tradeoff cannot simply be attributed to teachers lacking the factual knowledge being learned.

D SUPPLEMENTARY ANALYSES

D.1 HOW GENERALIZABLE IS THE MID-TRAINING TRADEOFF ACROSS MODEL FAMILIES?

Our main experiments use the OLMo-2 model family and training pipeline. To test whether this observed mid-training tradeoff is generalizable, we move off OLMo-2 entirely and repeat our tradeoff analysis with SmolLM: namely, a substantially smaller SmolLM2 360M student, a SmolLM2 1.7B Instruct teacher (Allal et al., 2025), and SmolLM3 Stage-3 training data (Hugging Face Team, 2025). In line with our main experiments, we train for 100B tokens during pre-training and 60B tokens during mid-training, using the same SmolLM3 Stage-3 data in both stages, and retain the same evaluation setup. We also include SWITCH DISTILLATION at mid-training (keeping the same routing threshold $q = 20\%$).

Despite noisier and less monotonic tradeoff curves at this substantially smaller scale, we recover the same qualitative behavior (Fig. 7): during pre-training, moderate forward KL distillation admits Pareto improvements in both reasoning and factual recall over NTP. During mid-training, however, standard forward and reverse KL improve reasoning only at the expense of factual recall. SWITCH DISTILLATION is able to mitigate this tradeoff *at mid-training*, jointly improving reasoning and factual recall over NTP.

These results suggest that the stage-dependent behavior of distillation, as well as the benefit of SWITCH DISTILLATION, extends beyond the OLMo-2 model and data ecosystem.

D.2 HOW GENERALIZABLE IS TEACHER SUPERVISION ASYMMETRY?

One of our main findings is that teacher supervision is asymmetric across domains using the OLMo-2 Instruct models as teachers; specifically, tokens from procedural domains (math, instruction-

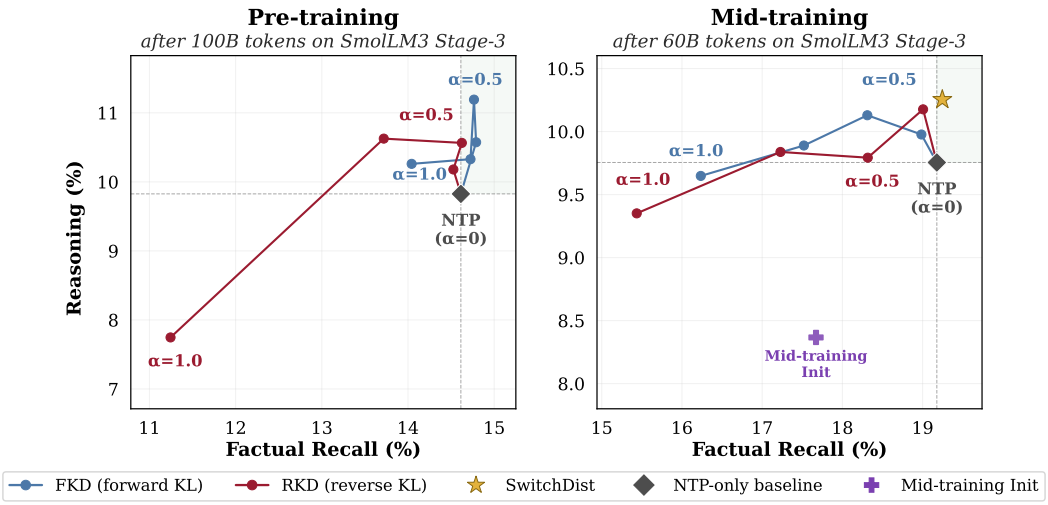


Figure 7: **Reasoning-recall tradeoff using the SmolLM ecosystem.**

following) tend to have lower teacher entropy than those from knowledge-intensive domains. Here, we ask whether this phenomenon generalizes across training stages and model families.

Across training stages. Figure 8 repeats the analysis using each teacher model at different stages of training and parameter size. To quantify the separation, we report the receiver operating characteristic area under the curve (ROC AUC) with 95% CI. The ROC AUC quantifies how well teacher entropy distinguishes procedural (Math and FLAN) from knowledge-intensive (DCLM, Wikipedia, StackExchange, and PeS2o) domains using the teacher predictive entropy of tokens. More specifically, an ROC AUC of x indicates that a randomly selected procedural token has lower teacher entropy than a randomly selected knowledge-intensive token with probability x; 0.5 is random chance.

Across all 12 stage-size cells, the ROC AUC stays within $[0.744, 0.826]$ (with 95% CI width ≤ 0.005), i.e., the entropy gap between procedural and knowledge-intensive tokens is a pretraining-era property that neither SFT, DPO, nor the final Instruct stage removes; the small monotone erosion visible from Base to Instruct at every size (e.g. $0.816 \rightarrow 0.761$ at 7B) is the only stage-level effect and never approaches chance.

Across model families. To test whether this finding generalizes across other model families, we additionally repeat our analysis on OLMo-3 7B Instruct (Team Olmo et al., 2026), Qwen 3 8B (Yang et al., 2025), Gemma 3 12B Instruct (Team Gemma et al., 2025), and Granite 3.3 8B Instruct (IBM Research, 2025), all of which are recent open-weight instruction-tuned models. As these models have different vocabularies, we report the entropy normalized by the vocabulary size.

In Fig. 9, all four exhibit the same qualitative shape as OLMo-2: procedural mass concentrates at the low end of normalized entropy, and knowledge-intensive mass at the high end, and every AUC sits well above chance (0.696 to 0.771; 95% CI widths ≤ 0.008). The separation is thus not an OLMo-2-specific artifact but a general property of modern instruction-tuned models.

D.3 Robustness across teacher sizes

Factual acquisition over the course of training. In §4.2, we showed that a correlation between predictive entropy and factual acquisition exists using OLMo-2 7B Instruct as a teacher. In Fig. 10, we find that the same empirical trend persists across teacher sizes, using OLMo-2 Instruct 1B and 13B as teachers.

Factual recall analysis of KD. In a similar vein, we show that the KD analysis on factual recall examples is also largely consistent for OLMo-2 1B and 13B Instruct teachers in Fig. 11.

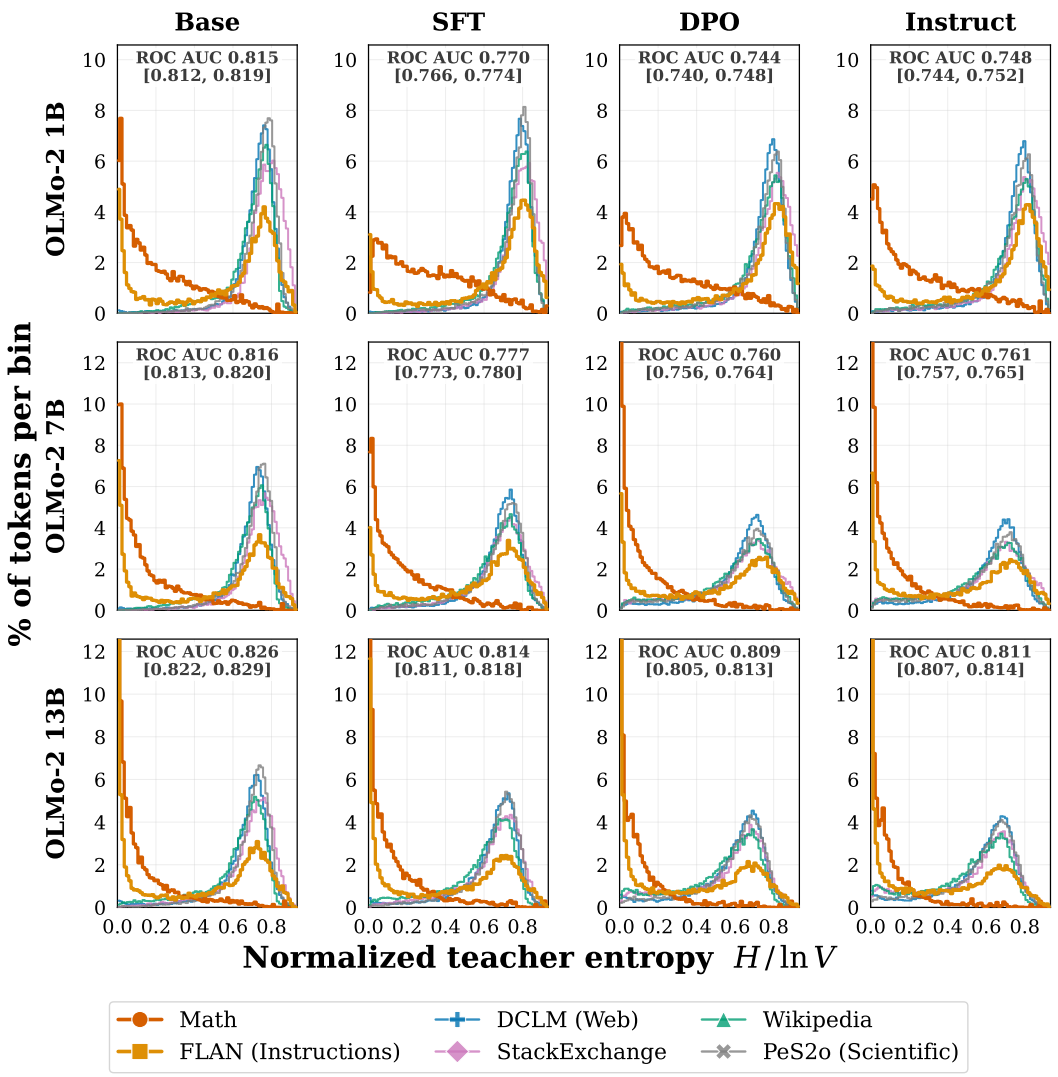


Figure 8: **Asymmetric teacher supervision holds across training stages.**

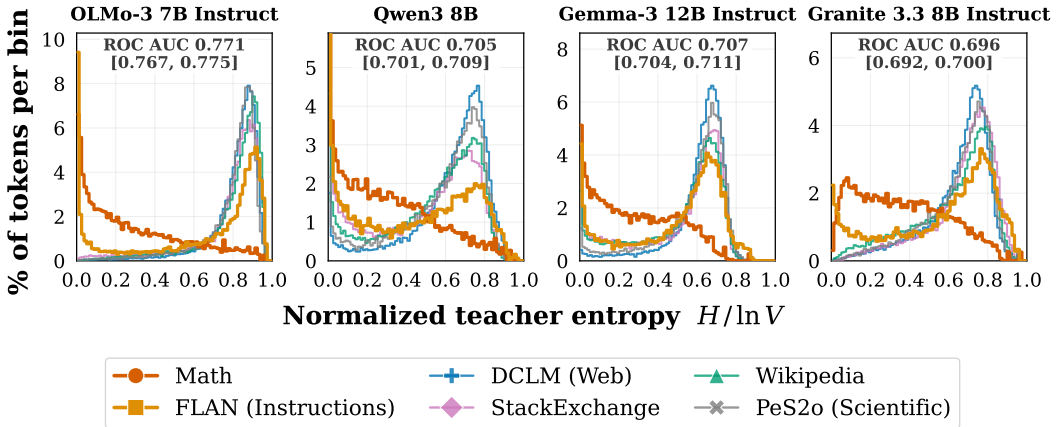


Figure 9: **Asymmetric teacher supervision holds across model families.**

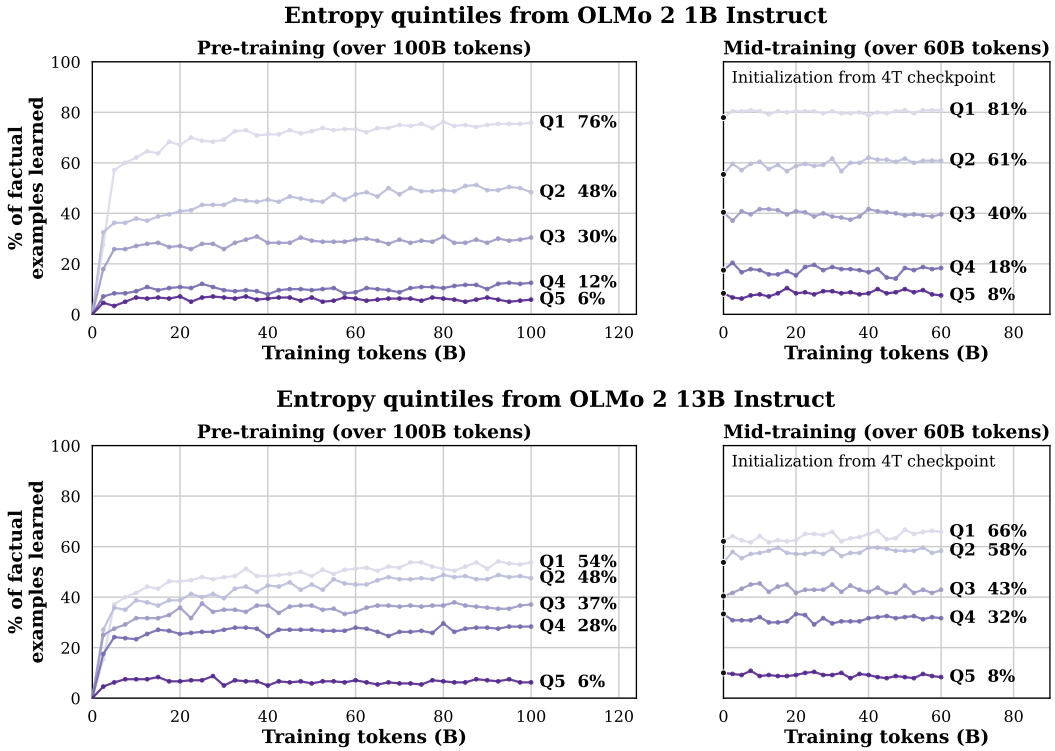


Figure 10: **Teacher entropy predicts factual acquisition under standard NTP.** The same qualitative trend holds when using OLMo-2 1B and 13B Instruct as teachers.

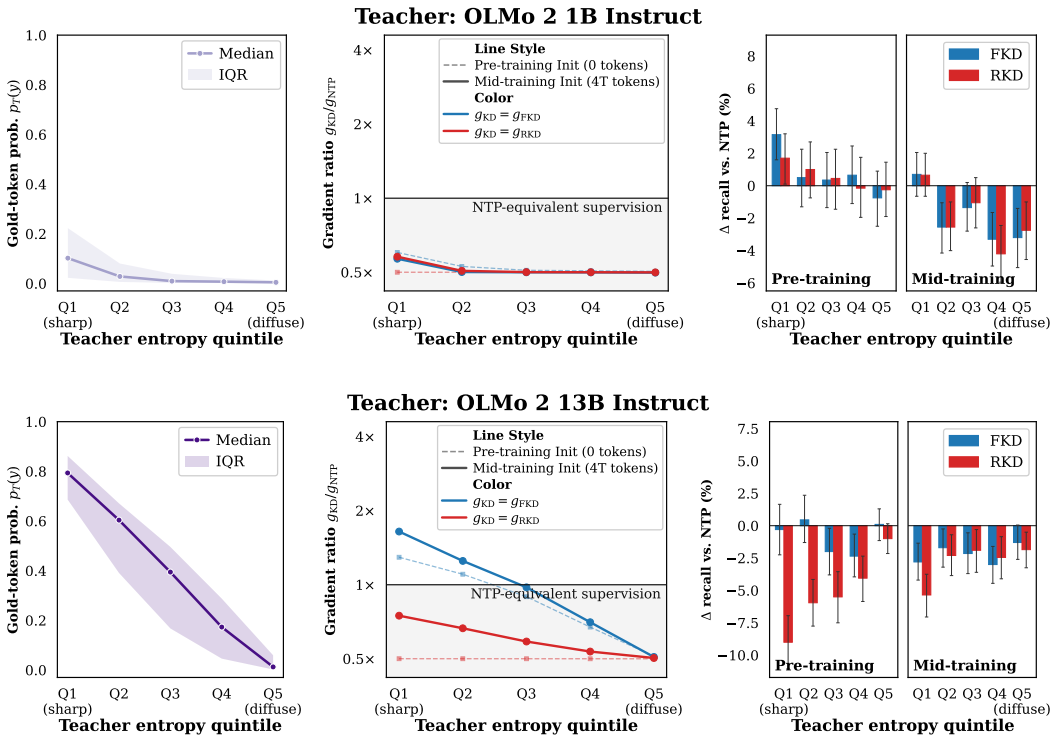
D.4 Forward and reverse KL exhibit different optimization geometry.

Forward and reverse KD differ only in the choice of divergence, yet consistently produce different reasoning–recall tradeoff patterns. To better understand this difference, we analyze how the KL divergence loss in the objective interacts with the CE component. Specifically, at fixed model parameters, we compute the cosine similarity between the closed-form logit gradients of the CE and KL losses (e.g., FKL and RKL) on held-out pretraining data. Higher cosine values indicate stronger alignment between the two objectives.

Fig. 12 reports the resulting gradient cosine throughout training across KD mixture weights. At larger teacher sizes (7B and 13B) especially, FKL consistently exhibits higher CE–KL gradient alignment than RKL during both pre-training and mid-training, with the difference persisting across training checkpoints and mixture weights. This indicates that, locally, the FKL objective favors update directions more similar to NTP than does RKL. Consequently, FKL perturbs the next-token prediction optimization direction less than RKL, providing one explanation for why the two KD objectives occupy systematically different positions on the reasoning–factual recall frontier. Notably, this distinction depends on teacher size: with the size-matched 1B teacher, FKL shows greater alignment than RKL primarily during early pre-training, while the difference largely disappears later in pre-training and throughout mid-training.

D.5 Switch Distillation accelerates mid-training reasoning acquisition

Our main experiments compare methods at the end of mid-training. Here, we instead examine their learning trajectories to understand *when* reasoning gains emerge. We evaluate intermediate checkpoints throughout the 60B-token mid-training run, comparing standard NTP, forward KD (at $\alpha = 0.5$), and Switch Distillation, using the OLMo-2 7B Instruct teacher. Fig. 13 reports macro-averaged reasoning performance as a function of the number of mid-training tokens consumed.



Teacher Uncertainty | **Training Signal** | **Learning Outcome**

(a) Gold-token probability (b) Gold-token gradient relative to NTP (c) Factual recall Δ vs. NTP

Figure 11: **KD analysis on factual recall examples.** The same qualitative trends hold when using OLMo-2 1B and 13B Instruct as teachers, in accord with using the 7B teacher in Fig. 5.

We observe that the reasoning advantage from distillation emerges remarkably early: the NTP baseline reaches a final reasoning macro-average of 26.1% after 60B mid-training tokens. In contrast, SWITCH DISTILLATION exceeds this level by the first evaluated checkpoint at 2.5B tokens, corresponding to $1/24$ of the NTP training token budget. FKD reaches the same threshold with twice that budget ($1/12$). Notably, these early gains do not simply reflect faster convergence to the same solution: both FKD and SWITCH DISTILLATION continue to improve throughout training and finish substantially above the NTP baseline, with SWITCH DISTILLATION maintaining the strongest reasoning performance across the trajectory.

D.6 SWITCH DISTILLATION IMPROVES PASS@k AT LOW SAMPLING BUDGETS

Fig. 14 examines whether the reasoning gains from SWITCH DISTILLATION persist beyond pass@1 as additional inference-time samples become available. After mid-training, SWITCH DISTILLATION and the other KD methods consistently outperform NTP across sampling budgets. On the GSM tasks, SWITCH DISTILLATION performs best among all methods, with its largest advantages in the low-budget regime ($k \leq 16$). On MATH-500, SWITCH DISTILLATION slightly trails FKD and RKD. The differences among KD methods progressively shrink as k increases, suggesting that additional test-time compute can partially compensate for weaker per-sample reasoning performance.

After post-training, the KD methods become substantially closer on GSM8K, GSM-Symbolic, and GSM-Plus, consistent with post-training narrowing the reasoning gaps observed after mid-training. Nevertheless, SWITCH DISTILLATION maintains a slight lead across the GSM tasks. Notably, the trend reverses on MATH-500: SWITCH DISTILLATION overtakes FKD and achieves the strongest pass@k performance across all sampling budgets, with its advantage persisting through $k = 64$. Overall, these results suggest that SWITCH DISTILLATION is particularly effective under constrained

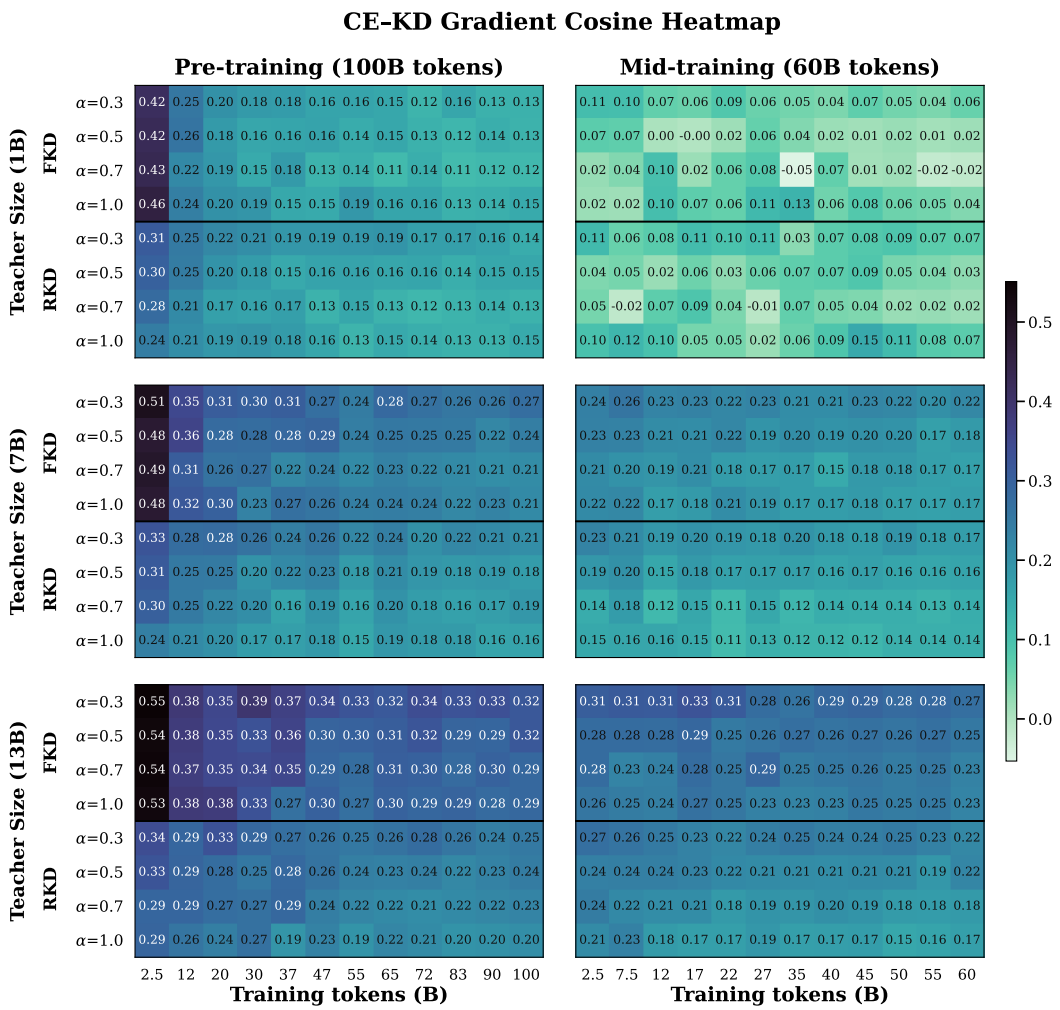


Figure 12: **FKL exhibits higher gradient alignment with the CE objective than RKL throughout training.**

inference-time sampling budgets, while its gains can persist even at larger budgets on more challenging reasoning tasks.

D.7 Switch Distillation improves code generation performance

We additionally provide code generation results for our mid-training settings on Mostly Basic Python Problems (Austin et al., 2021). We do not include code generation in our main evaluation because explicitly code-related data constitutes a relatively small portion of the OLMo-2 mid-training mixture; for example, StackExchange's sampling ratio is only $\sim$2.5% at mid-training. Nevertheless, for comprehensiveness, we evaluate on MBPP to test whether the observed mid-training gains in procedural reasoning extend to code generation.

As Table 9 shows, all distillation settings improve over NTP, with Switch Distillation achieving the highest pass@1 for both teacher sizes. The improvements from Switch Distillation over NTP are statistically significant, while the smaller differences among distillation methods are within the uncertainty of this evaluation.

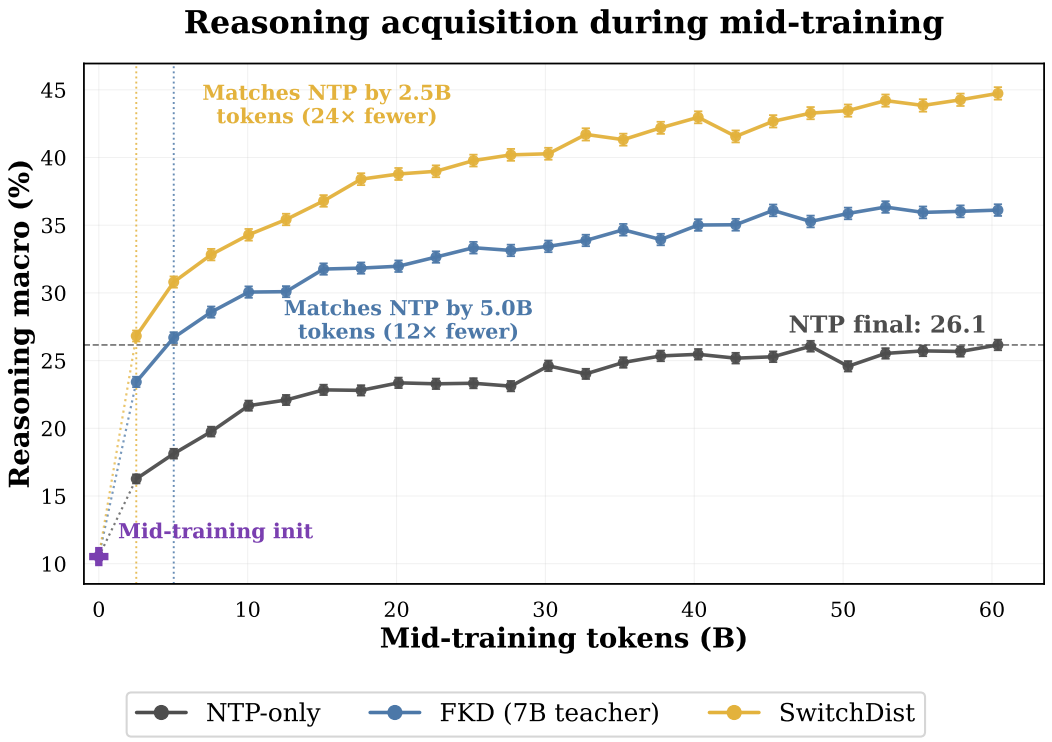


Figure 13: **SWITCH DISTILLATION substantially accelerates reasoning acquisition during mid-training.** SWITCH DISTILLATION surpasses the final reasoning performance of the 60B-token NTP baseline within the first 2.5B mid-training tokens evaluated, while standard forward KD does so with double the amount (5B tokens). Both distillation methods continue to improve with additional training, maintaining a substantial advantage over NTP throughout mid-training.

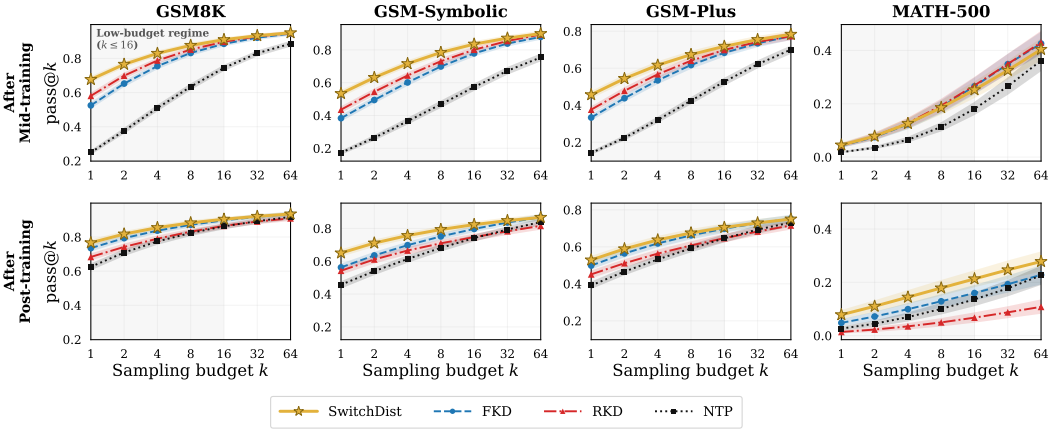


Figure 14: **SWITCH DISTILLATION improves pass@k performance for reasoning tasks (GSM8K, GSM-Symbolic, GSM-Plus, Math-500), particularly at low sampling budgets.** We report pass@k performance for $k \in \{1, 2, 4, 8, 16, 32, 64\}$ after mid-training (top) and post-training (bottom).

E FULL RESULTS

E.1 INTERMEDIATE POST-TRAINING RESULTS

The OLMo-2 1B post-training pipeline consists of SFT, DPO, and two stages of RLVR. We reported final post-training results in Table 2; intermediate results for SFT, DPO, and RLVR1 are in Table 10.

Table 9: **Code generation results on MBPP (Austin et al., 2021) after mid-training.** Following the OLMES evaluation setup, we report pass@1 (%) on MBPP (500 samples). $\pm$ denotes the 95% confidence interval, and stars indicate statistically significant differences from the shared NTP baseline under paired per-problem tests ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

T	Method	MBPP ($n = 500$)
N/A	NTP	4.80 ± 1.87
7B	FKD	$7.60 \pm 2.32^{*}$
7B	RKD	$7.80 \pm 2.35^{*}$
7B	TRKD	5.60 ± 2.02
7B	**SD**	$\mathbf{8.00 \pm 2.38}^{*}$
N/A	NTP	4.80 ± 1.87
13B	FKD	$8.60 \pm 2.46^{**}$
13B	RKD	$9.60 \pm 2.58^{***}$
13B	TRKD	6.80 ± 2.21
13B	**SD**	$\mathbf{10.00 \pm 2.63}^{***}$

Table 10: **Per-task results after intermediate post-training stages.** The NTP baseline is duplicated because it serves as the common reference for both teacher settings. **Bold** denotes the best result within each teacher block. * indicates a statistically significant improvement over the strongest competing baseline ($p < 0.05$, paired bootstrap). Benchmark names are abbreviated for space; see Table 7 for full task names.

		Reasoning						Factual Recall			Knowledge & Commonsense						Inst.
T	Method	GSM8K	GSM-S	GSM+	BBH	DROP	MATH	TQA	NQ	SQA	MMLU	MMLU-P	ARC-C	OBQA	Wino	AGI	IFE
After SFT																	
N/A	NTP	45.2	28.2	23.3	30.2	33.8	6.8	**56.3**	21.7	8.2	39.2	16.3	51.1	51.4	51.5	33.2	45.8
7B	FKD	54.7	34.2	30.7	31.3	39.1	9.4	54.1	23.3	8.2	49.1	18.5	**61.7**	59.4	51.9	39.7	**46.0**
7B	RKD	57.7	35.5	32.0	32.1	39.9	11.2	53.2	22.4	8.4	48.9	18.2	60.9	60.2	52.2	39.2	45.8
7B	TRKD	49.9	31.1	26.9	31.5	37.1	7.6	53.7	23.1	8.4	46.5	17.4	57.8	56.0	51.4	37.5	45.3
7B	**SD**	**63.7***	**42.7***	**36.7***	**33.5***	**48.9***	**12.8**	55.1	**24.5***	**8.5**	**50.3***	**19.4***	61.3	**61.4**	**56.1***	**40.6**	45.5
N/A	NTP	45.2	28.2	23.3	30.2	33.8	6.8	**56.3**	21.7	8.2	39.2	16.3	51.1	51.4	51.5	33.2	**45.8**
13B	FKD	54.1	32.5	28.0	31.8	36.3	7.6	54.9	23.5	7.9	45.9	17.1	55.6	57.2	51.2	37.1	44.2
13B	RKD	56.8	36.7	31.3	**32.1**	38.9	**11.6**	55.4	23.8	8.0	47.3	17.3	**57.8***	57.4	**52.2**	**38.1**	43.6
13B	TRKD	49.0	29.8	25.4	31.6	35.2	7.6	55.4	23.6	**8.8**	43.8	16.7	53.8	55.0	51.3	36.2	43.4
13B	**SD**	**62.7***	**42.5***	**34.8***	31.6	**43.5***	10.2	54.9	**24.1**	8.5	**48.0***	**17.7**	55.1	**59.4**	**52.2**	38.0	43.1
After DPO																	
N/A	NTP	52.4	33.4	28.2	32.3	33.5	6.6	**55.4**	21.5	7.7	41.1	16.4	50.3	50.6	51.5	34.1	59.9
7B	FKD	60.4	35.8	33.1	33.1	39.8	7.8	53.6	23.1	**8.3**	49.2	18.5	59.2	58.0	51.6	39.7	**64.1**
7B	RKD	64.5	40.0	36.4	34.2	40.5	10.0	52.8	22.3	7.8	48.8	17.3	57.4	55.0	52.5	38.2	**64.1**
7B	TRKD	57.2	34.9	32.3	32.4	36.6	7.2	53.1	23.2	7.9	47.0	17.9	56.7	57.2	51.9	37.5	62.5
7B	**SD**	**70.9***	**50.8***	**44.2***	**35.5**	**49.0***	**15.0***	54.7	**23.6**	8.1	**50.0**	**20.1***	**61.2**	**61.0**	**55.5***	**40.9**	62.1
N/A	NTP	52.4	33.4	28.2	32.3	33.5	6.6	55.4	21.5	7.7	41.1	16.4	50.3	50.6	51.5	34.1	59.9
13B	FKD	61.0	34.2	31.7	**33.4**	36.9	8.2	54.6	23.2	7.7	45.9	17.4	53.2	57.2	52.3	37.4	**64.7**
13B	RKD	65.6	41.2	35.0	33.2	39.9	7.4	53.5	23.7	7.9	**47.8**	**17.9**	**57.1**	56.2	51.9	38.5	64.5
13B	TRKD	53.1	34.9	30.7	32.9	35.3	6.8	**54.9**	22.4	**8.5**	43.4	16.5	53.2	56.0	52.0	35.3	61.7
13B	**SD**	**69.9***	**45.7***	**39.5***	32.2	**44.2***	**10.0**	54.5	**23.9**	8.2	47.3	**17.9**	56.1	**59.2**	**52.5**	**38.7**	61.0
After RLVR1																	
N/A	NTP	69.1	45.6	40.5	32.4	33.4	11.2	**54.5**	21.7	8.0	31.1	16.9	52.3	54.2	51.1	33.2	63.6
7B	FKD	76.7	59.4	51.9	34.4	39.2	16.8	52.8	**23.1**	**8.3**	41.4	18.3	59.4	56.2	51.2	38.0	67.1
7B	RKD	77.9	54.3	52.5	34.2	40.1	15.6	52.1	21.9	7.8	45.3	16.8	57.6	50.4	51.5	36.1	59.3
7B	TRKD	71.0	54.6	46.8	33.0	37.4	11.4	52.8	22.7	8.0	40.2	17.8	58.9	55.6	52.0	35.9	65.2
7B	**SD**	**79.3**	**63.4***	**52.8**	**36.1***	**48.8***	**17.4**	54.1	23.0	8.0	**47.2***	**19.7***	**61.7**	**59.8**	**53.0**	**39.7***	**68.8**
N/A	NTP	69.1	45.6	40.5	32.4	33.4	11.2	54.5	21.7	8.0	31.1	16.9	52.3	54.2	51.1	33.2	63.6
13B	FKD	73.0	55.5	47.3	**33.7**	36.1	14.0	**54.7**	23.3	8.3	35.0	17.2	54.0	56.6	**51.7**	35.4	63.2
13B	RKD	75.6	60.1	50.3	33.3	38.1	15.8	53.3	22.3	8.0	36.2	16.9	55.9	55.2	51.3	36.4	63.6
13B	TRKD	72.4	53.3	47.6	32.7	35.0	10.4	**54.7**	22.2	**8.6**	36.6	16.6	54.1	54.6	**51.7**	34.5	67.1
13B	**SD**	**78.8***	**65.6***	**52.5***	33.5	**43.3***	**17.8**	54.3	**24.0**	8.2	**38.7***	**17.7**	**56.2**	**57.8**	**51.7**	**37.7**	**68.9**

Table 11: **Per-task mid-training ablations to SWITCH DISTILLATION (SD), using OLMo-2 7B Instruct as the teacher.** Macro-averaged results appear in Table 3.

	Reasoning						Factual Recall			Knowledge & Commonsense					
Method	GSM8K	GSM-S	GSM+	BBH	DROP	MATH	TQA	NQ	SQA	MMLU	MMLU-P	ARC-C	OBQA	Wino	AGI
Distillation Objective															
SWITCH DISTILLATION$_{\text{FKL}}$	66.7	52.0	43.7	31.2	46.4	11.0	54.8	24.3	8.2	50.5	19.1	63.7	62.6	52.2	39.6
Routing Policy															
TEACHER-CORRECT ROUTING	63.8	49.5	41.7	31.8	45.2	9.8	44.5	19.4	8.7	50.6	19.3	63.2	64.6	52.0	39.9
RANDOM ROUTING	61.5	45.5	38.6	32.0	40.6	10.8	53.2	23.9	8.5	49.6	17.8	61.6	62.6	52.6	39.5
ORACLE DOMAIN ROUTING	61.0	45.6	38.5	31.6	40.1	8.0	53.0	22.6	8.3	49.4	18.8	61.1	61.4	52.6	38.9
Supervision Objective															
ALWAYS CE	69.1	55.2	46.5	33.3	50.1	12.0	54.3	24.5	9.4	51.1	19.9	63.7	63.6	52.7	41.1
TEACHER TOP-1	62.2	46.0	40.1	30.7	42.1	8.8	57.9	24.7	9.3	48.8	18.0	59.8	62.0	51.0	39.2

E.2 ABLATION MID-TRAINING RESULTS

We provide per-task results for the mid-training ablations to SWITCH DISTILLATION in Table 11 (macro-averages are in Table 3).