OLMo1 and OLMo3

Bae Sun Woo

January 26, 2026

Quote

“There are certain things that you need to get right at pre-training, especially as your post-training is shifting towards more reinforcement learning heavy stack and something with much more inference”

“All the data on pre-training … readjustment of the prioritization of pre-training much more heavily on reasoning”

- Nathan Lambert –

Table of Contents

1. Preliminary: SwiGLU, RoPE => YaRN, Adam => AdamW, DPO
2. OLMO1
3. OLMO3
4. Experiments: Adaptive Multi-Dataset Policy Optimization

SwiGLU : Swish + GLU

Swish

$Swish(x) = x \cdot \sigma(\beta x)$
σ: sigmoid function
β: trainable parameter or constant
$SiLU(x) = x \cdot \sigma(x)$

Smooth :
Differentiable at $x = 0$

Non-Monotonic:
Possess Small Negative Values.
While $ReLU(x) = 0$ for $x < 0$

GLU(Gated Linear Unit)

$GLU(x) = (xW + b) \otimes \sigma(xV + C)$
$(xW + b)$: Information Path
$\sigma(xV + C)$: Gate Path

Element-wise Product:
$\sigma(xV + C) \sim 1$ => $GLU(x) = xW + b$
$\sigma(xV + C) \sim 0$ => $GLU(x) = 0$
=> Information Filtering

SwiGLU

$SwiGLU(x) = SiLU(xW) \otimes (xV)$

RoPE: Rotary Positional Encoding

Absolute PE : limitation of addition

$$q_m = x_m + p_m \,, k_n = x_n + p_n$$

$$\begin{aligned} &Attention(q,k) \\ &= (x_m + p_m)^T (x_n + p_n) \\ &= x_m^T x_n + x_m^T p_n + p_m^T x_n + p_m^T p_n \end{aligned}$$

The absolute position matters

"I like dog more than cat" (dog:3, cat:6)

≠

"Well, I like dog more than cat" (dog:4, cat:7)

RoPE :

$$q_m = R_m x_m \,, k_n = R_n x_n$$

$$R_m = \begin{pmatrix} \cos(m\theta) & -\sin(m\theta) \\ \sin(m\theta) & \cos(m\theta) \end{pmatrix}$$

$$Attention(q,k) = (R_m x_m)^T (R_n x_n) = x_m^T R_m^T R_n x_n$$

$R_m^T R_n = R_{-m} R_n$

As, $e^{-im\theta} \cdot e^{n\theta} = e^{i(n-m)\theta}$

$$x_m^T R_m^T R_n x_n = x_m^T \boldsymbol{R_{n-m}} x_n$$

Only Relative Position(n-m) matters and remains

"I like dog more than cat" (cat-dog = 3)

=

"Well, I like dog more than cat" (cat-dog = 3)

RoPE: Rotary Positional Encoding

3.2.2 General form

In order to generalize our results in 2D to any $\boldsymbol{x}_i \in \mathbb{R}^d$ where d is even, we divide the d-dimension space into $d/2$ sub-spaces and combine them in the merit of the linearity of the inner product, turning $f_{\{q,k\}}$ into:

$$f_{\{q,k\}}(\boldsymbol{x}_m, m) = \boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_{\{q,k\}} \boldsymbol{x}_m \tag{14}$$

where

$$\boldsymbol{R}^d_{\Theta,m} = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix} \tag{15}$$

is the rotary matrix with pre-defined parameters $\Theta = \{\theta_i = 10000^{-2(i-1)/d}, i \in [1, 2, ..., d/2]\}$. A graphic illustration of RoPE is shown in Figure **(1)**. Applying our RoPE to self-attention in Equation (2), we obtain:

$$\boldsymbol{q}_m^\intercal \boldsymbol{k}_n = (\boldsymbol{R}^d_{\Theta,m} \boldsymbol{W}_q \boldsymbol{x}_m)^\intercal (\boldsymbol{R}^d_{\Theta,n} \boldsymbol{W}_k \boldsymbol{x}_n) = \boldsymbol{x}^\intercal \boldsymbol{W}_q R^d_{\Theta,n-m} \boldsymbol{W}_k \boldsymbol{x}_n \tag{16}$$

=> Smaller dimensions rotate fast, Larger dimensions rotate slowly

YaRN: Yet another RoPE extensioN method

When long context is given

Position Interpolation

$$\theta \rightarrow \frac{\theta}{2}$$

However,

Loss of important high frequency details

which the network needs in order to resolve tokens that are both very similar and very close together

YaRN :

$$s = \frac{L'}{L} \quad r(d) = \frac{L}{\lambda_d} = \frac{L}{2\pi b'^{\frac{2d}{|D|}}} \quad \gamma(r) = \begin{cases} 0, & \text{if } r < \alpha \\ 1, & \text{if } r > \beta \\ \dfrac{r-\alpha}{\beta-\alpha}, & \text{otherwise.} \end{cases}$$

γ: Retention rate / Degree of keeping the original features
$\gamma = 1$: High Freq (No Interpolation, Extrapolation, Keep original rotation)
$\gamma = 0$: Low Freq (Linear Interpolation, Slow down rotation)

$$h(\theta_d) = \Big(1 - \gamma\big(r(d)\big)\Big)\frac{\theta_d}{s} + \gamma\big(r(d)\big)\theta_d.$$

Additionally, introduce t, temperature

$$\text{softmax}\left(\frac{\mathbf{q}_m^T \mathbf{k}_n}{t\sqrt{|D|}}\right)$$

=> Sharpen Attention Logits(Counteract Entropy Increase)

Adam, AdamW = Adam + Corrected Weight Decay

Adam

Momentum(Overcome local minima)

$$m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$$

$$W := W - \alpha \cdot m_t$$

RMSProp(Adaptive Step Size)

$$v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$$

$$W := W - \alpha \cdot \frac{g_t}{\sqrt{v_t} + \epsilon}$$

(g_t^2 is used to take magnitude)
($\sqrt{v_t}$, square root is used to Dimension homogeneity)

Adam

$$W := W - \alpha \cdot \frac{m_t}{\sqrt{v_t} + \epsilon}$$

AdamW

$$W_{t+1} = W_t - \alpha \cdot \left(\frac{m_t}{\sqrt{v_t}+\epsilon} + \lambda \cdot W_t\right)$$

DPO: Direct Preference Optimization

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} \right) \right]$$

What does the DPO update do? For a mechanistic understanding of DPO, it is useful to analyze the gradient of the loss function $\mathcal{L}_{\text{DPO}}$. The gradient with respect to the parameters θ can be written as:

$$\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$

$$-\beta \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \Bigg[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \Bigg[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l} \Bigg] \Bigg],$$

Open **L**anguage **M**odel

OLMo Architecture

Size	L	D	H	Tokens	Peak LR	Warmup	Weight Tying	Batch size
1B	16	2048	16	2T	4.0E-4	2000 steps	yes	~4M
7B	32	4086	32	2.46T	3.0E-4	5000 steps	no	~4M

Table 1: OLMo model sizes, number of training tokens, and optimizer settings. In all runs, the optimizer was AdamW, with betas of 0.9 and 0.95, and an epsilon of 1.0E-5. **L** is number of layers, **D** is hidden dimension, **H** is number of attention heads, **WD** is weight decay.

- Decoder-only Transformer
- No Biases
- Non-parametric layer norm
- SwiGLU activation function
- Rotary positional embeddings (RoPE)
- BPE-based tokenizer from GPT-NeoX-20B
- AdamW Optimizer

OLMo Pretraining Data : Dolma

Source	Type	UTF-8 bytes *(GB)*	Docs *(millions)*	Tokens *(billions)*
Common Crawl	web pages	9,812	3,734	2,180
GitHub	code	1,043	210	342
Reddit	social media	339	377	80
Semantic Scholar	papers	268	38.8	57
Project Gutenberg	books	20.4	0.056	5.2
Wikipedia	encyclopedic	16.2	6.2	3.7
Total		**11,519**	**4,367**	**2,668**

Table 2: Composition of Dolma. Tokens counts are based on the GPT-NeoX tokenizer.

Pipeline of (1) language filtering, (2) quality filtering, (3) content filtering, (4) deduplication, (5) multi-source mixing, and (6)tokenization

Trained by ZeRO optimizer strategy via PyTorch's FSDP framework

OLMo Zero-shot Evaluation

Models	arc challenge	arc easy	boolq	hella-swag	open bookqa	piqa	sciq	wino-grande	avg.
StableLM 1.6B	43.8	63.7	76.6	68.2	45.8	74.0	94.7	64.9	66.5
Pythia 1B	33.1	50.2	61.8	44.7	37.8	69.1	86.0	53.3	54.5
TinyLlama 1.1B	34.8	53.2	64.6	58.7	43.6	71.1	90.5	58.9	59.4
OLMo-1B	34.5	58.1	60.7	62.5	46.4	73.7	88.1	58.9	60.4
Falcon-7B	47.5	70.4	74.6	75.9	53.0	78.5	93.9	68.9	70.3
LLaMA 7B	44.5	67.9	75.4	76.2	51.2	77.2	93.9	70.5	69.6
Llama 2 7B	48.5	69.5	80.2	76.8	48.4	76.7	94.5	69.4	70.5
MPT-7B	46.5	70.5	74.2	77.6	48.6	77.3	93.7	69.9	69.8
Pythia 6.9B	44.1	61.9	61.1	63.8	45.0	75.1	91.1	62.0	63.0
RPJ-INCITE-7B	42.8	68.4	68.6	70.3	49.4	76.0	92.9	64.7	66.6
OLMo-7B	48.5	65.4	73.4	76.4	50.4	78.4	93.8	67.9	69.3

Table 3: Zero-shot evaluation of OLMo-1B and OLMo-7B, with other publicly available comparable model checkpoints on 8 core tasks from the downstream evaluation suite described in Section 2.4. For OLMo-7B, we report results for the 2.46T token checkpoint.

OLMo Intrinsic Evaluation

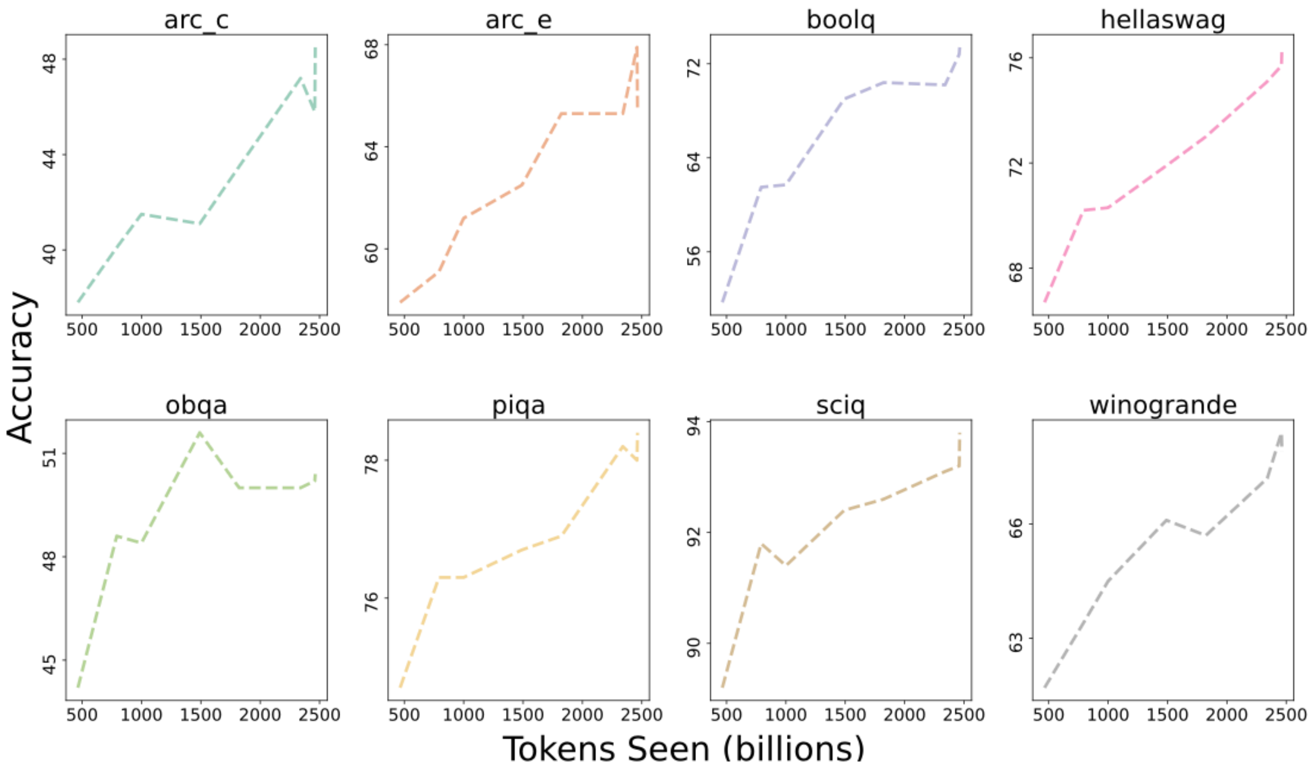


Figure 1: Accuracy score progression of OLMo-7B on 8 core end-tasks score from Catwalk evaluation suite described in Section 2.4. We can see the benefit of decaying LR to 0 in the final 1000 steps of training on most tasks.

OLMo Adaptation Evaluation

Model	MMLU 0-shot ↑	AlpacaEval %win ↑	ToxiGen % Toxic ↓	TruthfulQA %Info+True ↑
OLMo (base)	28.3	-	81.4	31.6
MPT Chat	33.8	46.8	0.1	42.7
Falcon Instruct	25.2	14.0	70.7	27.2
RPJ-INCITE Chat	27.0	38.0	46.4	53.0
Llama-2-Chat	46.8	87.3	0.0	26.3
TÜLU 2	50.4	73.9	7.0	51.7
TÜLU 2+DPO	50.7	85.1	0.5	-[7]
OLMo+SFT	47.3	57.0	14.4	41.2
OLMo+SFT+DPO	46.2	69.3	1.7	52.0

Table 4: Evaluation of various instruction-tuned 7B models, including OLMo-7B and before and after adaptation training. Lower is better for ToxiGen and higher is better for other metrics. We provide a detailed description of models and metrics in Appendix. E.

Evaluation after undergoing Instruction fine-tuning and DPO on OLMo as base model

OLMo 3

“All the data on pre-training … readjustment of the prioritization of pre-training much more heavily on reasoning”

- Nathan Lambert –

Model Flow for OLMo 3

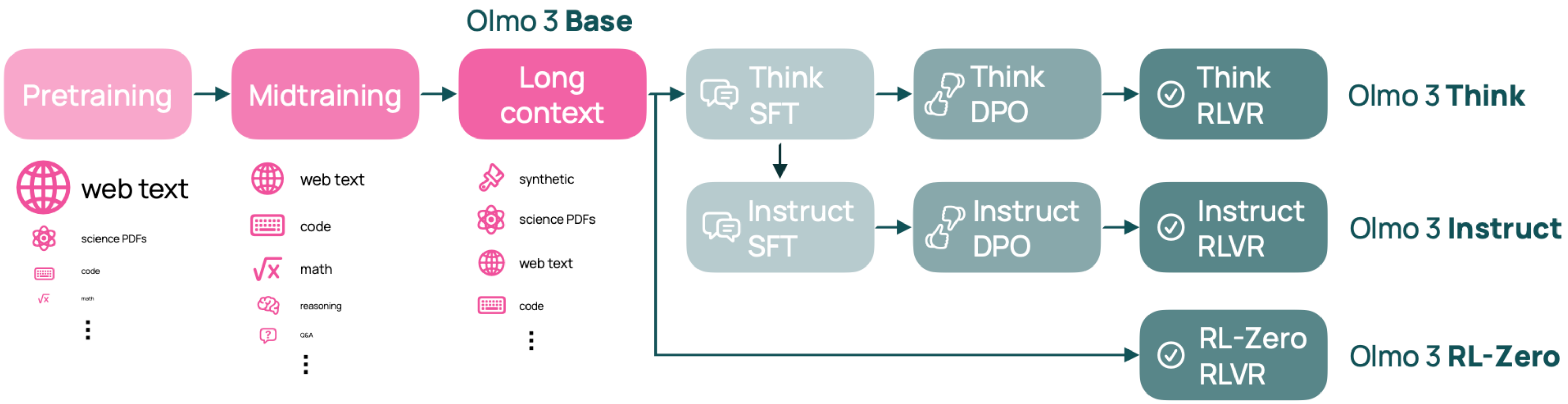


Figure 2 Depiction of model flow for Olmo 3. Development is divided into major **base model training (left)** and **post-training (right)** stages, each further divided into sub-stages with their own recipes (i.e., training data and method).

Olmo 3 Base Training, Data Curriculum

	Pretraining	Midtraining	Long-Context
Purpose	Build general language capacity and foundation	Shape reasoning-relevant capabilities and improve post-trainability	Enable long-context generalization and stable long-sequence processing
Dataset	Dolma 3 Mix	Dolma 3 Dolmino Mix	Dolma 3 Longmino Mix
Data Source	Web, Code, Academic PDFs(olmOCR), Docs	Math, Code, QA, Instruction, Thinking (synth + curated)	Long scientific PDFs (olmOCR)
Token	5.9T	100B	7B: 50B 32B: 100B

Olmo 3 Base Result – 7B

Model	# Toks	Math	Code	MC STEM	MC Non-STEM	GenQA	Minerva	GenXL	MMLU	BCB
		Base Aggregate Scores					Select Base Benchmarks			
7B scale										
OLMo 2 7B Stage 1	4T	12.7	7.1	61.0	70.6	68.6	5.6	15.8	59.8	81.6
OLMo 2 7B Stage 2 Ingredient 1	4.05T	40.4	10.4	64.1	74.6	72.1	18.9	21.3	63.1	85.1
OLMo 2 7B Stage 2 Ingredient 2	4.05T	41.4	10.4	64.3	74.9	71.8	18.7	21.0	63.8	85.8
OLMo 2 7B Stage 2 Ingredient 3	4.05T	40.8	10.1	64.0	74.9	72.1	19.1	21.9	63.8	85.6
OLMo 2 7B Stage 2 Soup	4.15T	41.7	10.4	64.6	75.2	72.4	19.1	21.2	63.7	85.7
Apertus 8B Phase 3	12T	19.2	9.9	61.1	68.4	68.3	7.3	19.0	58.3	81.4
Apertus 8B Phase 4	13.5T	26.0	16.2	65.1	73.8	69.7	10.8	30.5	63.3	86.8
Apertus 8B Phase 5	15T	29.3	19.0	66.7	75.0	70.1	12.9	31.0	65.0	88.6
Marin 8B Phoenix	11.1T	11.2	8.0	60.9	71.1	68.7	4.7	15.0	58.5	83.1
Marin 8B Starling	12.4T	40.5	20.8	**68.3**	78.7	75.7	23.2	36.2	**67.8**	89.1
Marin 8B Deeper Starling	12.7T	39.4	21.3	68.1	**78.8**	**75.9**	23.9	37.0	67.7	89.2
OLMO 3 7B Stage 1	5.9T	23.5	19.8	64.0	71.9	68.5	12.2	34.7	62.3	84.8
OLMO 3 7B Stage 2	6T	**59.8**	**31.9**	67.2	78.2	71.3	**41.4**	**49.1**	66.9	**89.7**
OLMO 3 7B Stage 3	6.05T	54.4	30.6	66.4	78.2	72.5	39.8	43.6	66.9	89.2

Olmo 3 Base Result – 32B

				32B scale						
OLMo 2 32B Stage 1	6.5T	33.2	16.0	73.0	81.7	75.8	13.6	29.2	72.3	93.5
OLMo 2 32B Stage 2 Ingredient 1	6.6T	51.6	19.9	75.1	84.5	78.5	30.3	36.8	75.5	94.8
OLMo 2 32B Stage 2 Ingredient 2	6.6T	51.9	20.0	74.1	83.8	79.1	30.7	35.2	74.0	94.1
OLMo 2 32B Stage 2 Ingredient 3	6.6T	51.5	19.6	74.4	83.6	79.0	29.2	35.7	74.3	93.8
OLMo 2 32B Stage 2 Ingredient 4	6.8T	51.9	19.2	74.6	83.3	78.3	31.0	37.1	74.3	94.0
OLMo 2 32B Stage 2 Soup	7.1T	53.9	20.5	75.3	84.2	79.1	31.0	37.1	75.0	94.4
Apertus 70B Phase 3	12T	34.2	17.8	68.6	78.2	74.6	13.4	31.9	67.3	88.8
Apertus 70B Phase 4	13.5T	39.8	21.5	70.5	79.5	75.8	16.3	34.8	69.5	91.0
Apertus 70B Phase 5	15T	40.6	23.0	70.5	79.4	75.5	17.5	37.7	69.3	91.4
K2 70B Stage 1	1.2T	34.0	27.5	69.5	78.2	73.8	16.2	42.6	67.9	89.2
K2 70B Stage 2	1.4T	43.3	29.6	68.2	78.0	73.5	25.7	46.6	67.8	88.3
Marin 32B Phase 3	5.4T	25.8	13.9	70.4	80.2	75.1	9.7	19.6	69.5	90.8
Marin 32B Mantis	6.5T	49.3	30.8	**75.9**	84.5	**80.3**	36.8	52.1	75.7	93.4
Olmo 3 32B Stage 1	5.5T	48.4	29.8	72.3	80.6	76.1	26.7	47.8	71.7	92.6
Olmo 3 32B Stage 2 Ingredient 1	5.6T	66.8	38.4	74.6	85.6	78.9	46.5	59.6	75.9	94.7
Olmo 3 32B Stage 2 Ingredient 2	5.6T	65.4	39.3	74.8	85.0	78.9	44.1	**60.0**	76.3	94.3
Olmo 3 32B Stage 2 Soup	5.7T	**69.7**	**39.7**	75.6	**85.7**	79.4	**46.9**	59.7	**76.9**	**95.0**
Olmo 3 32B Stage 3	6.2T	61.4	**39.7**	74.3	85.6	79.7	42.9	59.4	76.2	94.8

Table 13 Results comparing Olmo 3 to open base models across stages of pretraining, midtraining and long context. As of writing, Marin has undergone learning rate cooldown (Mantis), but not long-context (LC) extension stage. Apertus also has a two-stage cooldown (Phase 4 and 5) and performed long-context extension by mixing-in data to their Phase 5 training. Token counts are presented in "Cumulative training tokens", so each row denotes the number of tokens that model has seen up to that point in training. For OLMo 2 and Olmo 3 models, Stage 1 is the standard pretraining phase, Stage 2 is midtraining, and Stage 3 is LC extension.

Olmo 3 Think

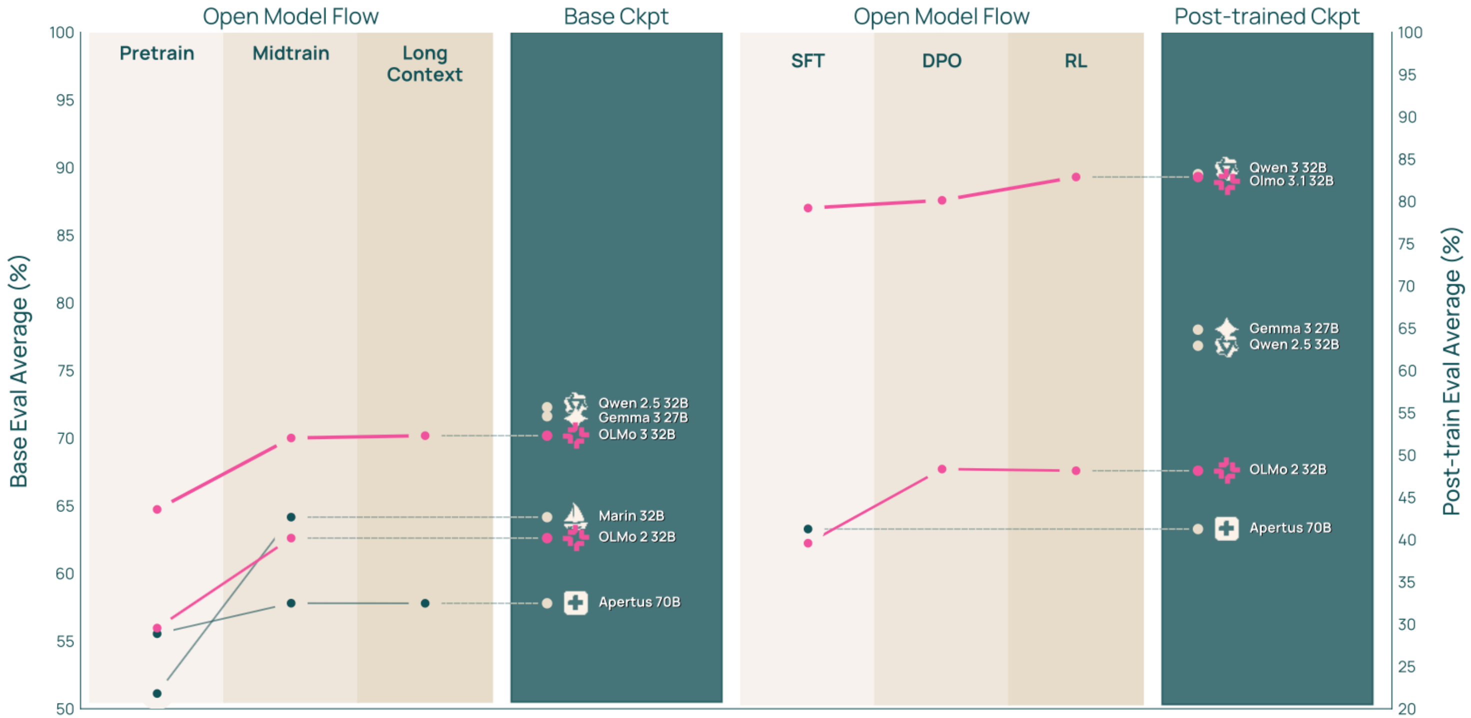


Figure 1 The model flow encompasses training data, code and intermediate checkpoints for all stages of development. While both fully-open and open-weights models release their final checkpoints **(dark teal)**, fully-open releases like Marin, Apertus, and Olmo provide data along their model flow, enabling the careful study of intermediate development stages **(beige)**. OLMO 3 THINK 32B is shown here along with other open models of comparable size and architecture. OLMO 3 THINK is competitive with Qwen 3 32B, which does not have a released base model. Its underlying OLMO 3 BASE 32B surpasses all other fully-open base models.

Olmo 3 Think: SFT,DPO,RL

	SFT	DPO	RL
Purpose	**Enable thinking traces as default behavior(activate reasoning mode)**	**Improve reasoning quality via contrastive delta learning**	**Lock in reasoning performance with verifiable rewards**
Dataset	Dolci Think SFT	Dolci Think DPO	Dolci Think RL
Data Type	Prompt → **Thinking trace + answer**	**Pairwise preference (chosen vs rejected)**with explicit capability gap	Prompt → rollout → **verifier reward**
Data Source	Curated + synthetic prompt thinking traces generated & filtered	chosen = Qwen3-32B (thinking) rejected = Qwen3-0.6B (thinking) (force quality delta)	Model rollouts evaluated by **verifiers** (math, code, IF, chat)

OLMo 3 Think - DPO

In particular, we find that further supervised finetuning on thinking traces generated by Qwen3 32B (one of the few open-thought models) outright hurts the performance of OLMo 3 Think SFT, indicating that we are approaching saturation on learning from imitation. To extract a useful training signal out of these

	Subset of Olmo 3 Think Benchmarks										
Name	**Avg.**	**MMLU**	**BBH**	**GPQA**	**Zebra**	**AGI**	**AIME25**	**AIME24**	**CHE**	**LCB**	**IFEval**
Qwen3 32B (chosen)	83.2	88.8	90.6	64.7	78.2	90.2	71.0	80.3	90.9	89.6	87.4
Qwen3 0.6B (rejected)	35.1	55.8	41.5	27.2	29.8	59.2	15.2	11.2	14.8	34.4	62.3
Dev. 7B SFT ckpt	70.3	**76.1**	**83.9**	45.1	56.5	76.4	58.8	71.0	88.1	67.0	**79.7**
Cont. SFT on chosen	64.5	72.6	80.2	40.2	49.8	73.9	52.8	61.0	83.4	55.1	76.0
Delta learning	**72.9**	75.5	82.8	**48.4**	**60.9**	**79.7**	**66.3**	**75.7**	**91.5**	**72.6**	75.2

Table 21 The delta between chosen and rejected responses is critical. Supervised finetuning directly on the chosen responses generated by Qwen3-32B Thinking hurts the Initial SFT model. In contrast, DPO tuning to prefer the 32B responses over weaker Qwen3-0.6B Thinking responses yields strong gains across math and code reasoning.

OLMo 3 Think – DPO, Delta Learning

	Subset of Olmo 3 Think Benchmarks										
Name	Avg.	MMLU	BBH	GPQA	Zebra	AGI	AIME25	AIME24	CHE	LCB	IFEval
SFT	70.1	74.9	84.1	45.8	57.9	77.2	57.6	69.6	88.2	67.8	77.9
SFT + DPO	72.7	74.8	83.7	48.6	60.6	79.1	62.7	**74.6**	**91.4**	**75.1**	75.9
SFT + RLVR	71.9	77.4	83.2	42.7	63.1	78.5	62.4	70.0	87.9	70.7	**82.8**
SFT + DPO + RLVR	**74.1**	**77.9**	**86.8**	**50.2**	**62.9**	**80.1**	**64.2**	73.2	89.9	73.4	82.3

Table 22 Delta learning provides a stronger initialization for subsequent RLVR than SFT alone. We show the effect of conducting RLVR for 1000 steps after DPO and SFT on our 7B model on a subset of our evaluation suite. Note that here evaluations are from one run only. Preference tuning with delta learning first followed by RLVR, yields the best overall performance. For RLVR, we use data offline-filtered by the corresponding starting point (SFT only or SFT + DPO).

Qwen3-32B Thinking Model ⟷ Delta ⟷ Qwen3-0.6B Thinking Model

OLMo 3 Think - RL

OlmoRL formulation Our final objective function includes a token-level loss, truncated importance sampling, clip-higher, and no standard deviation in the advantage calculation:

$$\mathcal{J}(\theta) = \frac{1}{\sum_{i=1}^{G} |y_i|} \sum_{i=1}^{G} \sum_{t=1}^{|y_i|} \min\left(\frac{\pi(y_{i,t} \mid x, y_{i,<t}; \theta_{\text{old}})}{\pi_{\text{vllm}}(y_{i,t} \mid x, y_{i,<t}; \theta_{\text{old}})}, \rho \right) \min\left(r_{i,t} A_{i,t},\ \text{clip}(r_{i,t},\ 1 - \varepsilon_{\text{low}},\ 1 + \varepsilon_{\text{high}}) A_{i,t} \right), \quad (1)$$

where $r_{i,t} = \frac{\pi(y_{i,t}|x, y_{i,<t}; \theta)}{\pi(y_{i,t}|x, y_{i,<t}; \theta_{\text{old}})}$, ε_{low} and $\varepsilon_{\text{high}}$ are the clipping hyperparameters. Here, $y_i \sim \pi_{\text{vllm}}(\cdot \mid x; \theta_{\text{old}})$ and $\pi_{\text{vllm}}(\cdot \mid x; \theta_{\text{old}})$ are the token probabilities returned from vLLM, ρ is the truncated importance sampling cap value (Yao et al., 2025), and the advantage $A_{i,t}$ for the t-th token t in the response y_i is calculated within the group G based on the relative reward of the outputs inside each group:

$$A_{i,t} = \left(r(x, y_i) - \text{mean}\left(\{ r(x, y_i) \}_{i=1}^{G} \right) \right). \quad (2)$$

$r(x, y_i)$ is the reward score returned by the corresponding verifier. Our hyperparameters for various runs are in Appendix Table 49.

GRPO + DAPO + Dr GRPO

OLMo 3 Think Result

	Olmo 3 32B Think				Baselines			
	SFT	DPO	Final Think 3.0	Final Think 3.1	Qwen 3 32B	Qwen 3 VL 32B Think	DS-R1 32B	K2-V2 70B Instruct
Math								
MATH	95.6	95.9	96.1	96.2	95.4	96.7	92.6	94.5
AIME 2024	73.5	76.0	76.8	80.6	80.8	86.3	70.3	78.4
AIME 2025	66.2	70.7	72.5	78.1	70.9	78.8	56.3	70.3
OMEGA	43.1	45.2	50.6	53.4	47.7	50.8	38.9	46.1
Reasoning								
BigBenchHard	88.8	89.1	89.8	88.6	90.6	91.1	89.7	87.6
ZebraLogic	70.5	74.5	76.0	80.1	88.3	96.1	69.4	79.2
AGI Eval English	85.9	87.8	88.2	88.8	90.0	92.2	88.1	89.6
Coding								
HumanEvalPlus	90.0	91.6	91.4	91.5	91.2	90.6	92.3	88.0
MBPP+	66.7	67.2	68.0	68.3	70.6	66.2	70.1	66.0
LiveCodeBench v3	75.8	81.9	83.5	83.3	90.2	84.8	79.5	78.4
IF								
IFEval	83.9	80.6	89.0	93.8	86.5	85.5	78.7	68.7
IFBench	37.0	34.4	47.6	68.1	37.3	55.1	23.8	46.3
Knowledge & QA								
MMLU	85.3	85.2	85.4	86.4	88.8	90.1	88.0	88.4
PopQA	33.1	37.0	31.9	30.9	30.7	32.2	26.7	32.2
GPQA	55.7	57.6	58.1	56.7	67.3	67.4	61.8	64.0
Chat								
AlpacaEval 2 LC	69.1	78.6	74.2	69.1	75.6	80.9	26.2	-
Safety	64.8	65.3	68.8	83.6	69.0	82.7	63.6	88.5

Table 14 Results on our flagship model Olmo 3 Think 32B on our post-training evaluation suite. OLMO 3.1 THINK 32B is the best fully-open model at 32B.

	Olmo 3 7B Think			Baselines					
	SFT	DPO	Final Think	Open-Thinker3 7B	Nemotron Nano 9B v2	DS-R1 Qwen 7B	Qwen 3 8B	Qwen 3 VL 8B Think	OR Nemotron 7B
Math									
MATH	94.4	92.4	95.1	94.5	94.4	87.9	95.1	95.2	94.6
AIME 2024	69.6	74.6	71.6	67.7	72.1	54.9	74.0	70.9	77.0
AIME 2025	57.6	62.7	64.6	57.2	58.9	40.2	67.8	61.5	73.1
OMEGA	37.8	40.5	45.0	38.4	42.4	28.5	43.4	38.1	43.2
Reasoning									
BigBenchHard	84.1	83.7	86.6	77.1	86.2	73.5	84.4	86.8	81.3
ZebraLogic	57.9	60.6	66.5	34.9	60.8	26.1	85.2	91.2	22.4
AGI Eval English	77.2	79.1	81.5	78.6	83.1	69.5	87.0	90.1	81.4
Coding									
HumanEvalPlus	88.2	91.4	89.9	87.4	89.7	83.0	80.2	83.7	89.7
MBPP+	63.2	63.0	64.7	61.4	66.1	63.5	69.1	63.0	61.2
LiveCodeBench v3	67.8	75.1	75.2	68.0	83.4	58.8	86.2	85.5	82.3
IF									
IFEval	77.9	75.9	88.2	51.7	86.0	59.6	87.4	85.5	42.5
IFBench	30.0	28.3	41.6	23.0	34.6	16.7	37.1	40.4	23.4
Knowledge & QA									
MMLU	74.9	74.8	77.8	77.4	84.3	67.9	85.4	86.5	80.7
PopQA	20.8	24.7	23.7	18.0	17.9	12.8	24.3	29.3	14.5
GPQA	45.8	48.6	46.2	47.6	56.2	54.4	57.7	61.5	56.6
Chat									
AlpacaEval 2 LC	43.9	50.6	52.1	24.0	58.0	7.7	60.5	73.5	8.6
Safety	65.8	67.7	70.7	31.6	72.1	54.0	68.3	82.9	30.3

Table 15 Overview of results of Olmo 3 Think 7B on our post-training evaluation suite. All numbers are the mean of three runs. We evaluate all models using our evaluation framework, generating up to a maximum of 32768 tokens.

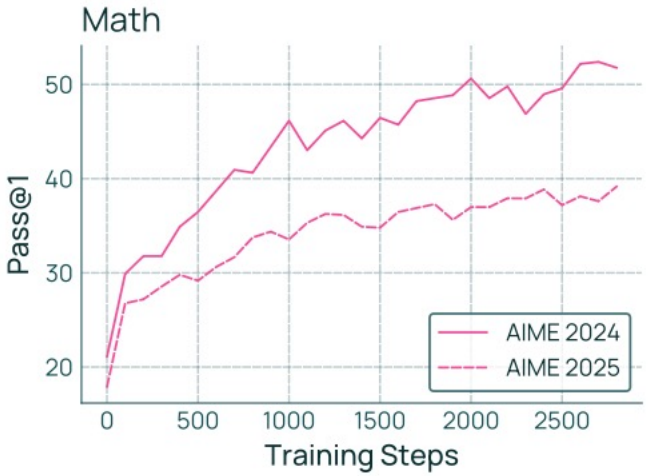
Olmo 3 RL-Zero

Olmo 3 Base ➕ OlmoRL ➡ Olmo 3 RL-Zero

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	**72.6**	**83.3**	**94.3**	**62.1**	**57.2**

Table 6 | Comparison of distilled and RL Models on Reasoning-Related Benchmarks.

DeepSeek R1-Zero-Qwen-32B



OLMo 3 RL-Zero