

TTT-NTP: Test-Time Training with Next-Token Prediction



Xuan Ouyang* Zefan Cai* Junjie Hu†

University of Wisconsin–Madison

* Equal contribution † Corresponding author

COLM 2026
2nd Workshop on Efficient Reasoning



Paper



Code

Motivation and contribution

1 Context capacity does not ensure context use

Modern LLMs accept long prompts, yet evidence inside the context window can still fail to affect later predictions. Extending the window or continuing pretraining does not directly solve this problem. Each observed prompt token already supplies a prefix-next-token pair, offering self-supervision for adapting to the current document, topic, style or retrieval problem.

2 What should a fast-weight write store?

Fast weights let an in-place TTT model retain prompt information inside selected MLP down-projections, making it available to later tokens. Existing recipes learn local activation proxies, often through a convolution over neighboring states. Their outer loss is language modeling, but the inner write target is not directly tied to next-token prediction. The target remains a design choice even when the backbone and update mechanism are fixed.

3 Predict the model's next contextual state

Computing full next-token gradients requires costly backpropagation through the decoder and disrupts chunk parallelism. TTT-NTP instead pairs each current gated MLP activation with the same layer's next contextual hidden state. A learned pointwise projection maps that dense state to the write target. Context enters through the causal forward pass itself, without aggregating a neighborhood of target states.

Can the prompt's next-token signal teach fast weights what to remember for later predictions?

One target connects causal training and prompt-specific inference. We test it on four backbones, three model families and two long-context benchmarks, with controlled target and solver ablations.

Causal writes during training

Reuse the existing MLP down-projection as fast memory. The current gated activation serves directly as its key, without an extra key projection.

$$\mathcal{L}_{\text{inner}} = -\langle Wz, v \rangle, \quad \Delta W = vz^\top$$
$$W_\ell^{(c)} = W_\ell^{\text{down}} + \eta \sum_{c' < c} \Delta W_\ell^{(c')}$$

Chunks contain 1,024 tokens. Exclude each chunk's last token from write pairs, and let each chunk read only earlier chunks' writes to prevent future-token leakage. Train the backbone and projection jointly under outer next-token cross-entropy, then freeze the projection at inference. For Llama-8B and Qwen3-4B, normalize each chunk write by its spectral norm to bound cumulative weight drift. Other backbones use raw writes.

Long-context retrieval: RULER

RULER Full-13 accuracy (%) at each context length

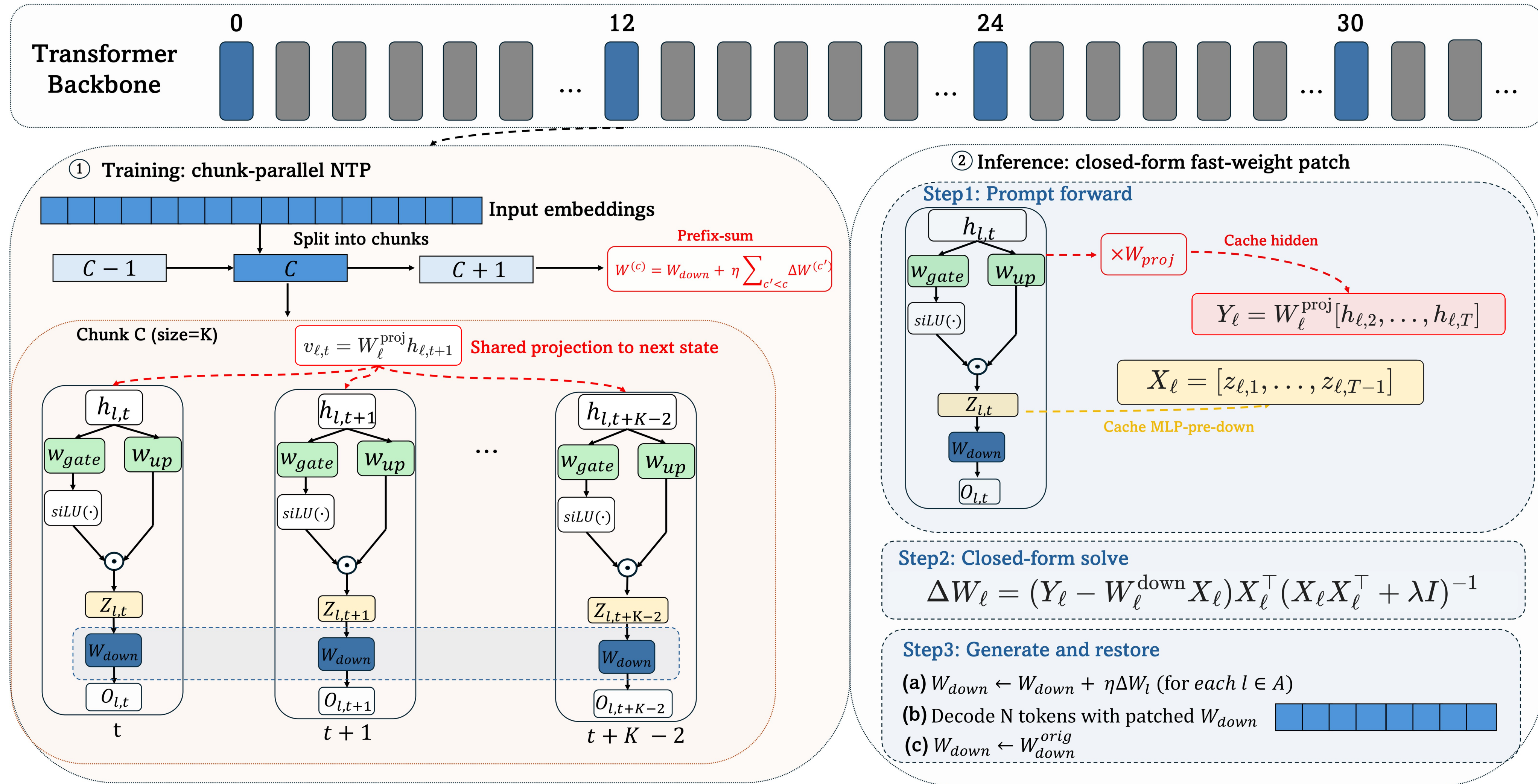
Backbone	Method	4k	8k	16k	32k	Avg
Llama-8B	Base	65.05	56.98	46.94	54.22	55.80
	qTTT	64.72	56.78	47.55	55.10	56.04
	CPT	55.09	51.14	46.61	46.93	49.94
	In-Place	57.82	52.65	47.23	49.34	51.76
TTT-NTP		63.39	59.77	53.90	61.73	59.70
Mistral-7B	Base	73.26	76.76	72.24	60.13	70.60
	qTTT	68.59	70.63	67.09	57.23	65.89
	CPT	69.25	68.44	61.88	52.15	62.93
	In-Place	68.41	66.61	62.27	51.31	62.15
TTT-NTP		81.51	79.13	72.64	61.23	73.63
Qwen3-4B	Base	93.23	84.03	74.96	75.47	81.92
	qTTT	93.75	84.88	75.34	76.18	82.53
	CPT	88.27	82.93	71.32	68.31	77.71
	In-Place	87.98	84.71	70.04	66.59	77.33
TTT-NTP		89.69	87.43	86.70	80.08	85.98
Qwen3-0.6B	Base	81.83	70.95	65.24	56.20	68.55
	qTTT	81.99	71.57	65.43	56.74	68.93
	CPT	82.08	70.33	65.63	57.38	68.86
	In-Place	72.41	68.40	67.57	59.18	66.89
TTT-NTP		82.71	73.31	71.39	58.30	71.43

Mean gains are +3.90 / +3.03 / +4.06 / +2.88 pp in the backbone order above. Qwen3-4B gains 11.74 pp at 16k (74.96 → 86.70), while Llama-8B and Qwen3-4B both decline at 4k.

The 13 tasks cover needle retrieval, variable tracking, word extraction and question answering. Shaded rows show TTT-NTP.

Scope: four full-attention backbones up to 8B, one corpus; Qwen3-0.6B uses a step-100 snapshot. Fig. 2 is a separate ablation setting. Future: larger models, diverse data, alternative attention, factuality and robustness.

TTT-NTP pipeline: the same next-position target in training and inference



Closed-form prompt adaptation

One prompt prefill caches gated activations X and projected next-position targets Y . Ridge regression fits a perturbation of the existing weight:

$$\min_{\Delta W} \|Y - (W^{\text{down}} + \Delta W)X\|_F^2 + \lambda \|\Delta W\|_F^2$$
$$\Delta W_\ell = (Y_\ell - W_\ell^{\text{down}} X_\ell) X_\ell^\top (X_\ell X_\ell^\top + \lambda I)^{-1}$$

Training omits the weight-dependent term for parallel prefix accumulation. Inference restores it in the complete regression. Fit the last $\leq 8,192$ tokens with a primal or dual solve, $\lambda = 1$ and $\eta = 0.1$. Cap each layer's relative Frobenius update at 0.1 and restore pretrained row norms. Reuse the original prompt KV cache so the patch acts only during decoding, then restore the weights after each sample.

Same target, different training and inference objectives.

Real-world QA and general ability

LongBench-v2 medium: 215 multiple-choice questions on 33k-128k-word documents. All models use a 32k-token budget with head-and-tail truncation.

Backbone	Method	SD	MD	ICL	Dlg	Code	Str	All
Llama-8B	Base	26.0	25.0	27.9	15.8	33.3	26.1	25.6
	qTTT	28.6	25.0	27.9	15.8	33.3	26.1	26.5
	CPT	31.2	27.3	25.6	31.6	44.4	26.1	29.3
	In-Place	35.1	20.5	27.9	21.1	44.4	26.1	28.8
TTT-NTP		33.8	29.5	30.2	21.1	33.3	34.8	31.2
Mistral-7B	Base	29.9	25.0	25.6	31.6	33.3	13.0	26.5
	qTTT	31.2	25.0	18.6	31.6	22.2	13.0	25.1
	CPT	35.1	25.0	25.6	10.5	33.3	21.7	27.4
	In-Place	32.5	29.5	25.6	15.8	33.3	17.4	27.4
TTT-NTP		32.5	31.8	23.3	31.6	33.3	30.4	30.2

SD/MD/ICL/Dialogue/Code/Structured n = 77/44/43/19/9/23.

Overall gains are +5.6 pp for Llama and +3.7 pp for Mistral. Retrieval-oriented domains drive the gain, with Mistral structured-data QA rising from 13.0 to 30.4 (+17.4 pp).

General capability of the trained backbone

Backbone	Method	HSwag	ARC-e	ARC-c	PIQA	MMLU	Avg
Llama-8B	Base	71.15	80.40	58.02	83.40	63.45	71.28
	TTT-NTP	71.50	80.80	58.70	83.40	63.37	71.56
Mistral-7B	Base	72.85	78.20	60.24	84.20	59.64	71.03
	TTT-NTP	69.05	78.40	59.64	85.00	58.27	70.07
Qwen3-4B	Base	65.20	74.20	50.20	79.60	73.80	68.60
	TTT-NTP	64.40	77.20	49.20	79.20	73.00	68.60
Qwen3-0.6B	Base	53.53	57.87	38.82	70.13	50.24	54.12
	TTT-NTP	54.03	59.72	38.48	70.13	50.56	54.59

Aggregate scores remain close to Base, but individual tasks can regress. Mistral loses 0.96 pp overall, including a 3.80 pp HellaSwag drop, despite its long-context gains.

Training recipe and controls

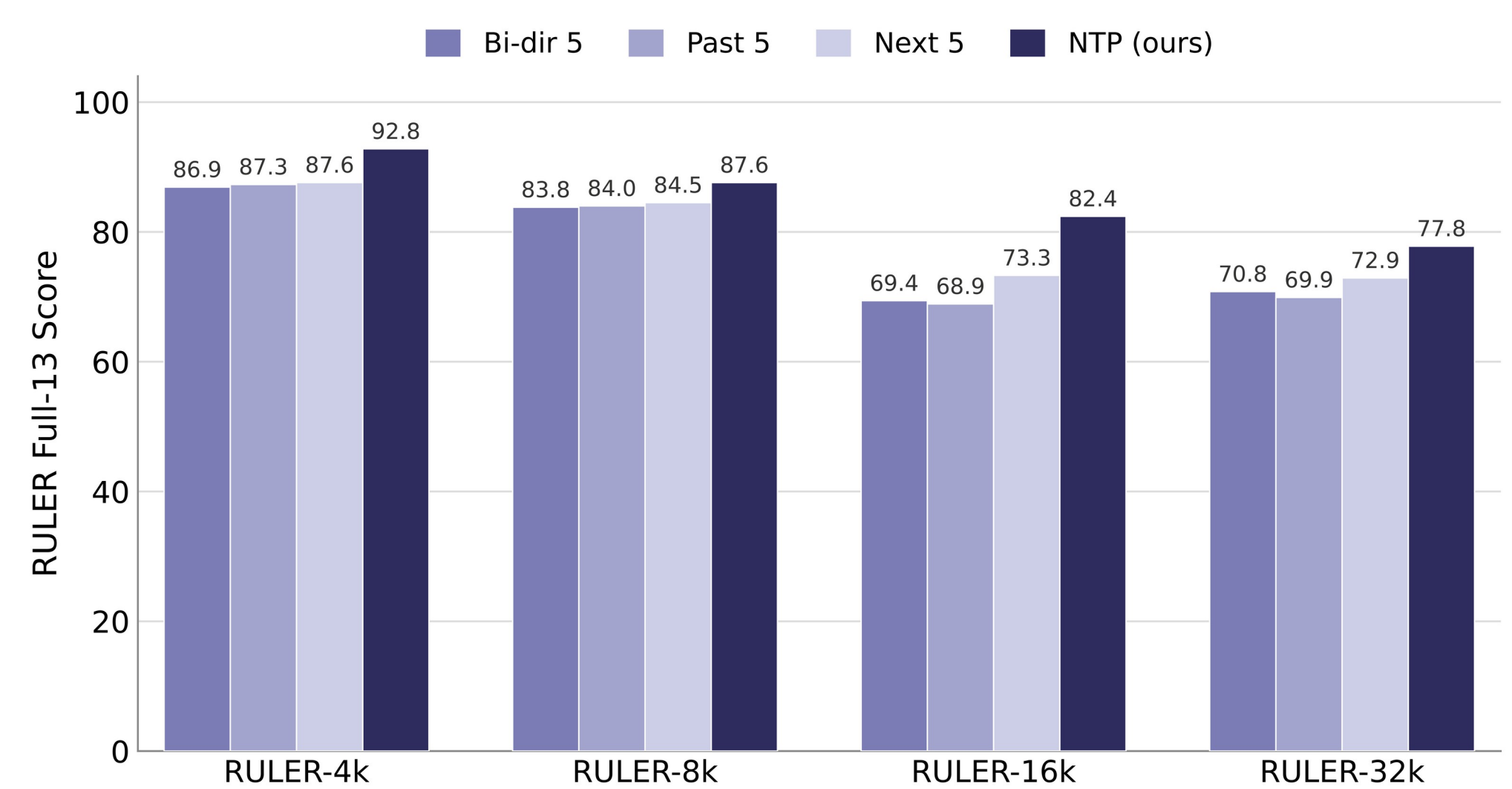
Train on Long-Data-Collections without Books, using 32,768-token single documents with no packing. All trained variants share AdamW, global batch 64, peak LR 5×10^{-6} , weight decay 0.1, 5% linear warmup and cosine decay, bf16 and FSDP.

Backbone	Tokens	Adapted layers	Train η
Llama-8B	0.4B	0, 6, 12, 18, 24, 30	0.3
Mistral-7B	0.1B	0, 6, 12, 18, 24, 30	0.15
Qwen3-4B	2B	0, 6, 12, 18, 24, 30	3.0
Qwen3-0.6B	0.2B	0, 4, 8, 12, 16, 20, 24	2.6

CPT uses no fast writes. In-Place TTT matches data, budget, placement, chunking and update scale, while learning convolutional targets. qTTT uses 32 steps on 128-token spans (LR 10^{-5} , weight decay 0.01, clipping 1.0).

Ablations: target and solver

Qwen3-4B with matched data, budget, placement and updates. Only targets differ: Past/Next use five states, and Bi-dir uses a symmetric 11-state window.



Next-position targets win at all four lengths. At 16k, NTP reaches 82.4 against 73.3 for the best convolutional target, a 9.1 pp gap.

Inference solver: whitening is decisive

Inference write	Llama-8B	Qwen3-4B
Inner product, no whitening	12.61	71.22
Ridge, no regularization	56.11	85.92
Ridge (TTT-NTP)	59.70	85.98

RULER averages (%). The Gram inverse corrects for correlated key directions. Unwhitened writes can over-amplify dominant directions. On Llama, regularization raises accuracy from 56.11 to 59.70 (+3.59 pp).