Puro-2B: Affordable From-Scratch Pretraining

Poor Lab's Qwen2-1.5B Trained on RTX 5090 within $5090

Paper by Kairong Luo, Jiarui Cui, Yaorui Yin,
Shengqi Chen et al.
Tsinghua University and Pengcheng Laboratory
Research paper discussion

arXiv:2608.27370v2 Reading: 14 October 2026

A complete training recipe makes pretraining experiments affordable.

Two budgets support two different claims

Checkpoint	Total tokens	GPU cost	Avg15 ↑
Puro UD, half Phase 2	918.83B	$4.37K	55.54
Puro CMA, canonical	≈1.40T	$6.89K	57.81
Qwen2-1.5B			55.14
Qwen2.5-1.5B			60.73

- The $4.4K result is an observed checkpoint, not only a fitted crossing.
- The best model uses $6.9K and remains 2.91 points below Qwen2.5.
- Qwen2 is the comparison target; Puro starts from random initialization.

Paper, Table 1, p.6; Tables 4–5, pp.21–22; Tables 18–19, p.53. Avg15 recomputed from the displayed task scores.

The headline is a marginal accelerator cost

$$\text{GPU cost} = \text{active GPU-hours} \times \text{normalized hourly rate}$$

Canonical accounting

$(6{,}009.46 + 16{,}504.95)$ GPU-hours $\times$ \$0.3061/hour $\approx$ **\$6,891**.

- Starts after training shards have been materialized.
- Excludes data processing, proxy runs, ablations, failed runs, SFT, evaluation, averaging operations, and research labor.
- The 5090 rate is a local amortization-based estimate, not a broadly available public cloud offer.

Paper, Sec.2.2, p.8; Table 2, p.9; Appendix B, pp.38–40. Fixed paper exchange rate: 6.8067 CNY/USD.

A 2B Transformer is trained in two phases

- Qwen3-1.7B architecture; untied embedding and LM head give about 2B parameters.
- Sequence length 4,096; global batch 1,536; micro-batch 2.

	Phase 1	Full Phase 2
Scheduled tokens	438.84B	959.99B
RTX 5090 GPUs	24	96
TP / PP / DP	1 / 2 / 12	1 / 4 / 24
Base LR trajectory	Power decay	Linear decay + CMA tail
Active time equivalent	10.43 days	7.16 days

The full run uses about 700 tokens per parameter: a data-rich regime for a compact model.

Paper, Sec.2.1, pp.6–8; Table 8, p.40; Appendix A, p.38.

Consumer hardware requires communication work

Measured communication	Before	After
One-way P2P bandwidth	31.5 GB/s	56 GB/s
Point-to-point latency	14.3 μs	0.4 μs
8-GPU AllReduce bus bandwidth	14.75 GB/s	27.34 GB/s
24-GPU AllReduce, GDR enabled	8.87 GB/s	19.93 GB/s

- 32 GB per GPU, no NVLink; 400 Gbps InfiniBand across nodes.
- Reported performance uses topology-specific driver modifications.
- GDR modification details are undisclosed; stock drivers give a slower communication path.

Paper, Sec.3.1.2, pp.9–10. AllReduce numbers are nccl-tests busbw.

Parallelism follows the communication bottleneck

- Use DP and PP; avoid frequent per-layer TP collectives.
- Give the expensive LM-head stage fewer Transformer layers.
- Distribute intact Muon matrices by memory footprint; fill remaining space with Adam states.

	24 GPUs	96 GPUs
Pipeline layers	18 | 10	9 | 9 | 9 | 1
Median TFLOP/s/GPU	238	192

$$P_{\text{eff}} = \left(\frac{0.72}{419} + \frac{0.28}{209.5}\right)^{-1} \approx 327, \quad \text{MFU}_{\text{P1}} \approx \frac{238}{327} = 73\%.$$

Paper, Sec.3.1.3, pp.10–11. MFU uses a mixed-precision peak, not a BF16-only peak.

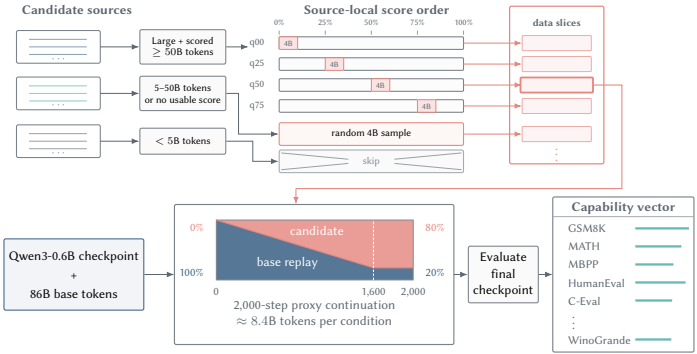
FP8 accelerates linear layers from initialization

Part of training	Representation
Linear-layer Fprop / Dgrad / Wgrad	E4M3 operands
Activation and activation-gradient scales	Groups of 128 values
Weight scales	128×128 blocks
Residual flow, outputs, checkpoint weights	BF16
Master state and gradient accumulation	FP32

- Local scales reduce the effect of outliers on quantization resolution.
- Power-of-two scales execute through the Blackwell MXFP8 path.
- About 72% of theoretical work uses FP8; saved inputs also use less memory.

Paper, Sec.3.2, p.11; Fig.14, p.43. Core attention retains BF16 arithmetic with FP32 softmax.

Proxy continuations guide data selection



- Common 0.6B checkpoint after 86B base tokens.
- 39 candidate slices; about 8.4B continuation tokens each.
- Candidate fraction rises to 80%.
- Capability profiles inform manually chosen mixture weights.

Paper, Fig.9, p.18; Appendix I, pp.60–63. Proxy tasks include C-Eval/CMMLU and differ from the final 15-task suite.

Phase 2 emphasizes mathematics and specialist data

Domain share	Phase 1	Phase 2 pool
English	73.2%	59.4%
Mathematics	7.2%	18.3%
Chinese	11.7%	9.4%
Code	7.9%	11.5%
Instruction-formatted	0.0%	1.3%

$$\underbrace{938.06\text{B}}_{\text{Phase 2 pool}} + \underbrace{21.93\text{B}}_{\text{Phase 1 replay}} = \underbrace{959.99\text{B}}_{\text{Phase 2 schedule}} .$$

The 43.87B transition includes replay and an early range already inside the Phase 2 pool.

Paper, Fig.8 / Table 3, p.18; Table 20, p.55. Semantic Code includes MegaMath-Code. Rounded shares may not sum to 100%.

MuonH directly schedules the relative update size

For a selected matrix W_t and a Muon update u_t:

$$\rho_t(W) = \frac{\eta_t \|u_t\|_F}{\|W_t\|_F} \qquad \text{(ordinary Muon; radial decay excluded).}$$

Fix $R = \|W_0\|_F$ and normalize $\widehat{u}_t = u_t / \|u_t\|_F$:

$$\widetilde{W}_{t+1} = W_t - \eta_t^H R \widehat{u}_t, \qquad W_{t+1} = R \frac{\widetilde{W}_{t+1}}{\|\widetilde{W}_{t+1}\|_F}.$$

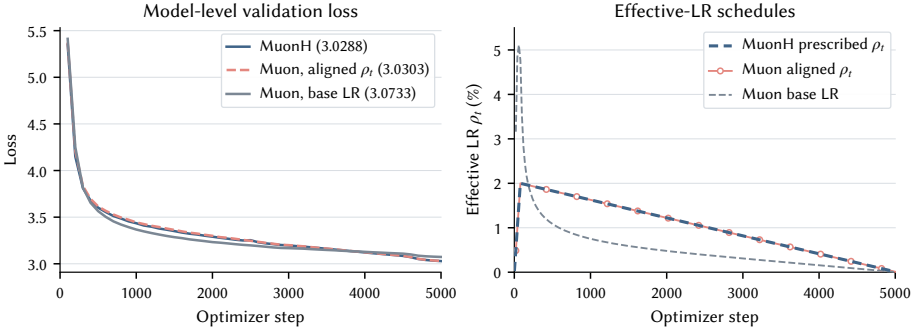
Effective LR becomes explicit

The pre-projection displacement relative to the fixed radius is η_t^H. Production MuonH groups use $\eta_t^H = 10\eta_t^{\text{base}}$.

Other parameters use AdamW. The motivation is approximate scale invariance of selected attention/MLP matrices.

Paper, Eqs.(1)–(4), p.12; Sec.3.3.1, p.13; Table 9, p.40.

Matching effective LR nearly matches final loss

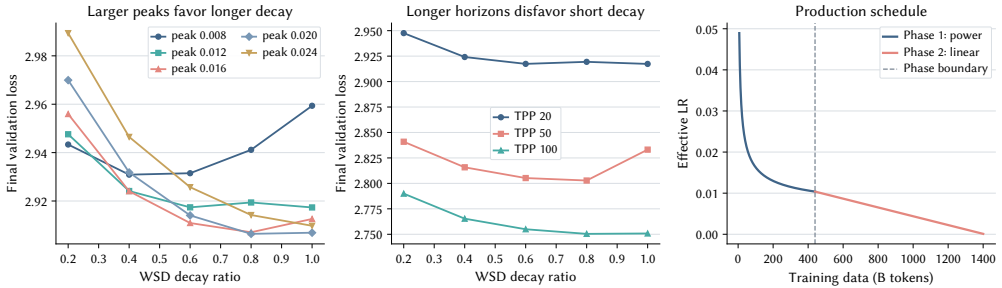


$$\eta_t^{\text{aligned}} = \rho_t^H \frac{\|W_t\|_F}{\|u_t\|_F} \quad \Rightarrow \quad L_{\text{aligned}} \approx 3.030, \quad L_{\text{MuonH}} \approx 3.029.$$

Ordinary-LR Muon ends near 3.073. The aligned control uses no fixed-radius projection.

Paper, Fig.5, p.14; Eq.(5), p.13. Matched 170M BF16 diagnostic; effective LR traces use selected MLP down-projections.

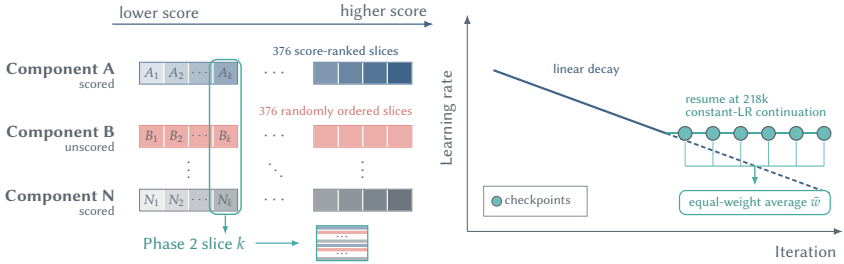
Higher peaks favor longer learning-rate decay



- In 0.6B WSD sweeps, the competitive region shifts toward longer decay as effective peak LR rises.
- Production combines an open-ended Phase 1 power schedule with a long Phase 2 linear decay.

Paper, Fig.6, pp.14–15; Appendix F.2, pp.51–52. One run per sweep setting; broad minima, not a precise monotone optimum.

CMA coordinates ordering, learning rate, and averaging



(a) Rank within each component and merged by progress

(b) Constant-LR continuation & Model Average

- Align 376 source-local token-percentile buckets to preserve the mixture.
- Put more preferred scored data later; unscored sources remain randomly ordered.
- At step 218k, hold base LR at 4.08×10^{-5}; average six late checkpoints.

Paper, Fig.7, p.16 (vector crop); Appendix H, pp.54–56. Scores are not compared across sources.

Averaging helps the selected constant-LR branch

Phase 2 recipe	Final checkpoint	Model average
Uniform + decay (UD)	56.00	55.57
Curriculum + decay (CD)	57.17	57.18
Curriculum + constant tail at 215k	55.64	56.80
Curriculum + constant tail at 218k	57.12	**57.81**

- Curriculum improves the decay endpoint by about 1.17 points.
- Averaging alone does not improve the uniform trajectory.
- Canonical CMA is the selected 218k branch with six-checkpoint averaging.

Paper, Table 19, p.53; Sec.3.4, p.16; Appendix H.3, p.56. Avg15 in percent. Fig.13b uses 55.99 for UD; the ledger uses 56.00.

The headline score gives each benchmark equal weight

Panel	Tasks	Evaluation
Math / code	GSM8K, MATH, MBPP, HumanEval	Generation
Complex reasoning	MMLU-Pro, BBH	Generation
Knowledge / commonsense	Nine remaining tasks	Answer likelihood

$$P = \frac{1}{15}\sum_{b} s_b = \frac{4\,\mathrm{Avg4} + 11\,\mathrm{Avg11}}{15}.$$

- Fixed base checkpoints, prompts, snapshots, greedy decoding, and parsers.
- Where applicable, select the better CF/MCF formulation separately for each model.
- Chinese proxy tasks are absent from this headline suite.

Paper, Sec.4.1, pp.20–21; Appendix A, p.38. MBPP means sanitized-MBPP.

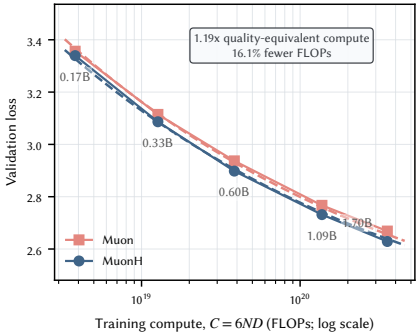
Broad capability improves, with uneven task trade-offs

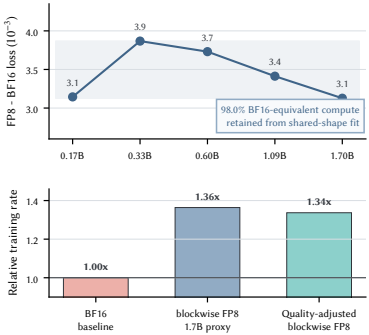
Base model	Math/code	Reasoning/knowledge	Avg15
Qwen2-1.5B	40.29	60.54	55.14
Qwen2.5-1.5B	47.52	65.53	60.73
Instella-3B	33.41	63.13	55.20
Yulan-Mini-2.4B	52.11	61.24	58.80
SmolLM3-3B	58.99	68.35	65.85
Puro UD, $4.4K	38.61	61.70	55.54
Puro CMA, $6.9K	**43.50**	**63.02**	**57.81**

The $4.4K model exceeds Qwen2 in Avg15 while trailing it on the four math/code tasks.

Paper, Tables 4–5, pp.21–22. All scores in percent, higher is better. Avg15 recomputed from individual displayed scores.

Ladders separate optimization and precision benefits

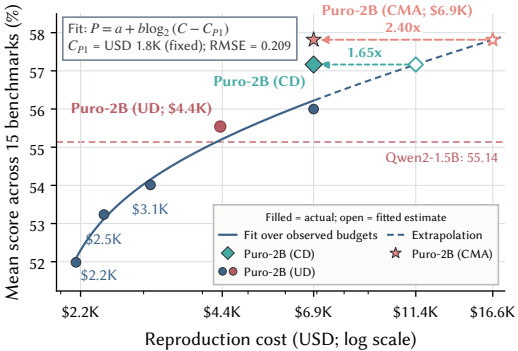




- MuonH: 1.19× fitted compute-equivalent gain with FP8 fixed.
- FP8: about 1.34× quality-adjusted speedup with MuonH fixed.

Paper, Fig.10, p.23; Appendix D, pp.42–44. 0.17B–1.7B at TPP=20; throughput proxy from 1.7B. Transfer to production TPP ≈700 is an estimate.

The cost law is a fixed-model scale-down fit



$$P = a + b\log_2(C - C_{\mathrm{P1}})$$

- $C_{\mathrm{P1}} = \$1.84\mathrm{K}$ fixed.
- Five UD budgets fit the incremental-cost trend.
- CD: fitted UD equivalent \$11.36K (1.65×).
- CMA: \$16.55K (2.40×).

Both equivalent budgets extrapolate beyond the largest observed UD budget of \$6.89K.

Paper, Fig.2b, p.5; Sec.4.3.1, pp.23–24; Tables 18–19, p.53. Filled: actual cost; hollow: fitted equivalent.

The CMA advantage survives matched SFT

SFT setting and metric	UD-based	CMA-based	Gain (pp)
GSM8K-based: GSM8K	66.89	68.66	+1.77
Math/code + replay: GSM8K	74.10	76.12	+2.02
Tulu-3: Avg15	53.96	55.13	+1.17
Tulu-3: IFEval	41.71	43.07	+1.36

- Same data, order, optimization, and budget within each paired comparison.
- Three matched SFT data orders; pretraining checkpoints are fixed.
- GSM8K checks use zero-shot ChatML and flexible numeric extraction.

Paper, Fig.11, p.25; Tables 13, 15, 17, pp.47–49. Scores in percent. IFEval is separate from Avg15.

Post-training gains are not uniform across tasks

Tulu-3 task	CMA − UD (pp)
HellaSwag	+7.25
ARC-Challenge	+3.95
MATH	+3.79
BoolQ	−4.61
HumanEval	−1.63
BBH	−1.06

Paper, Table 17, p.49; Appendix A, p.38; Sec.3.5, p.17.

- Ten of fifteen tasks favor CMA after Tulu-3 SFT.
- The paired result concerns the complete CMA recipe.
- SFT decontamination does not audit overlap already present in pretraining.

Three boundaries shape how the result transfers

Data evidence

No strict corpus-wide exact/near-duplicate audit against all proxy and final benchmarks. Late curriculum data may concentrate overlap.

Scale and attribution

TPP=20 ladder factors are transferred to TPP $\approx$700. Hardware, FP8, MuonH, and CMA factors use different baselines and should not be multiplied as a measured speedup.

Reproduction conditions

The accelerator-only boundary, local 5090 pricing, and communication adaptations determine the reported cost. Upstream data terms remain component-specific.

Paper, Sec.2.2, p.8; Sec.4.3, pp.22–24; Appendices A/B/D/I.

What this recipe makes possible

1. **Repeatable pretraining research:** useful 2B base models at a measured, explicitly bounded GPU budget.
2. **Observable optimization:** effective LR helps explain training dynamics beyond the scalar schedule.
3. **Joint data and schedule design:** ordering, late updates, and averaging should be studied together.

Next experiment (reading analysis): a decontaminated pool with matched ordering $\times$ tail LR $\times$ averaging controls.

Luo et al., Puro-2B: Poor Lab's Qwen2-1.5B Trained on RTX 5090 within $5090.
arXiv:2608.27370v2, 3 September 2026.
Paper-listed artifacts: Puro-2B collection; Puro-Megatron; Kaiyuan-Spark.

Backup: the two learning-rate coordinates

The production *base* schedule in Phase 1 is

$$\eta_{\text{base}}(k) = \begin{cases} 5 \times 10^{-3} k/1000, & 0 \le k \le 1000, \\ 5 \times 10^{-4} + 4.5 \times 10^{-3} \left(1 + \frac{k-1000}{1000}\right)^{-1/2}, & k > 1000. \end{cases}$$

Each independent UD Phase 2 run starts from the same endpoint:

$$\eta_{\text{base},j}(s) = \eta_0 + (\eta_{\min} - \eta_0)\frac{s}{S_j}, \quad \eta_0 \approx 1.04 \times 10^{-3}, \quad \eta_{\min} = 10^{-5}.$$

- s: samples consumed after the shared Phase 1 checkpoint.
- S_j: branch-specific decay budget; schedules differ by horizon.
- Wrapped matrices use $10\times$ base LR in production. Ladder and WSD multipliers differ.

Paper, Eqs.(10)–(11), pp.40–41; Fig.13a, p.42; Appendices D/F.

Backup: curriculum coordinates and the exact average

For internal bin j of source c, let $T_{c,j}$ be its token count:

$$p(x) = \frac{\sum_{i<j(x)} T_{c,i} + u_{c,j(x)}(x)\, T_{c,j(x)}}{T_c}, \qquad b(x) = \min\{375, \lfloor 376 p(x) \rfloor\}.$$

Here $u(x) \in [0, 1)$ is a deterministic within-bin position. Each bucket draws approximately the same token fraction from each source.

$$\bar{\theta} = \frac{1}{6} \sum_{i=1}^{6} \theta_{t_i}, \qquad \{t_i\} = \left\{ \begin{array}{c} 222100, 222200, 222300, \\ 222400, 222500, 222569 \end{array} \right\}.$$

The average spans 469 steps (≈2.95B tokens); the constant-LR continuation lasts about 29B tokens. Only model parameters are averaged.

Paper, Appendix H, pp.54–56. Scored sources: up to 256 score bins ×64 hash shards; unscored sources: 4096 random shards.

Backup: MPL separates exposure and decay response

For signal q_t (scalar LR or effective LR), let S_q be cumulative post-warmup exposure and S_w the warmup exposure:

$$L(t) = L_0 + A[S_q(t) + S_w]^{-\alpha} + B \sum_{k > t_{\text{warm}}}^{t} \Delta q_k \, G(x_k(t)),$$
$$x_k(t) = C q_k^{-\gamma} [S_q(t) - S_q(k-1)],$$
$$G(x) = 1 - (1+x)^{-\beta}.$$

- A decay has $\Delta q_k < 0$; its loss effect arrives gradually through G.
- Effective-LR input reduces held-out mean RMSE: 0.0265 to 0.0210.
- The simpler $G = 1$ variant reaches 0.0207 in the same two-run diagnostic.

Paper, Appendix F.1 / Fig.16, pp.49–50. Last 20% of validation targets held out. C here is a response coefficient, not FLOPs.

Backup: all four math/code scores matter

Base model	GSM8K	MATH	MBPP	HumanEval
Qwen2-1.5B	59.82	23.70	50.19	27.44
Qwen2.5-1.5B	67.70	32.28	58.37	31.71
Puro UD, \$4.4K	56.03	27.86	49.81	20.73
Puro CMA, \$6.9K	59.67	30.30	52.92	31.10

- GSM8K and MATH use 4-shot generation.
- Sanitized-MBPP and HumanEval use 3-shot generation.
- SFT GSM8K uses a different zero-shot chat protocol; compare paired SFT initializations within each setup.

Paper, Table 4, p.21; Appendix E.1, p.46. All scores in percent, higher is better.