Symmetry-Compatible Principle for Optimizer Design

Embeddings, LM Heads, SwiGLU MLPs, and MoE Routers

Tim Tsz-Kit Lau

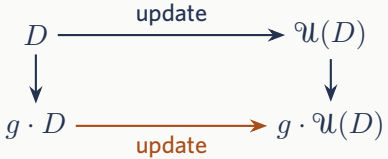
Joint work with Weijie Su (Penn Wharton and OpenAI)

NeurIPS 2026, Sydney

December 2026

The Symmetry-Compatible Principle

An allowed change of coordinates should transform the optimizer step along with the gradient or momentum matrix D.



$$\boxed{\mathcal{U}(g \cdot D) = g \cdot \mathcal{U}(D)}$$

Why can a coordinatewise step fail?

Rotate $d = (1, 0)^\top$ to $Qd = (3/5, 4/5)^\top$.

$$Q\,\mathrm{sgn}(d) = (3/5, 4/5)^\top,$$
$$\mathrm{sgn}(Qd) = (1, 1)^\top.$$

Taking signs in the new basis changes the direction of the step.

Lau and Su (2026), arXiv v5 · Section 3.1 and Appendix C. The displayed action is orthogonal and $\mathrm{sgn}(0) = 0$.

Vocabulary rows label tokens. Columns carry feature coordinates.

Embedding

Normalize each momentum row.

$$[\mathcal{U}_{\text{row}}(M)]_{i:} = \frac{M_{i:}}{\|M_{i:}\|_2 + \varepsilon}.$$

The hybrid then uses the feature Gram matrix $Z^\top Z$ of the normalized rows Z for its right-spectral step.

Untied softmax head

A common logit shift preserves probabilities.

$$\operatorname{softmax}(z + c\mathbf{1}_v) = \operatorname{softmax}(z).$$

Center across rows before and after normalization. With $\Pi = I_v - \frac{1}{v}\mathbf{1}_v\mathbf{1}_v^\top$, the head update is $\Pi\mathcal{U}_{\text{row}}(\Pi M)$.

Lau and Su (2026), arXiv v5 · Sections 3.3.2 and 3.7.2, equation (5) and Algorithm 2. The head uses ordinary, uncapped softmax logits.

A Symmetry-Compatible Optimizer Stack

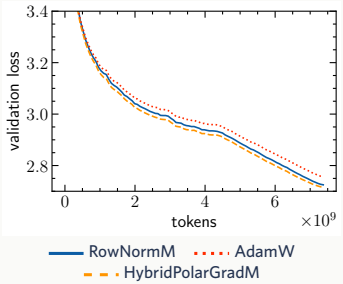
Block	Structure	Example update
Linear map	Input and output bases	Spectral update, e.g., MUON
Embedding	Token rows, features	Row norm, then right spectral
LM head	Token rows, shared shift	Centered row or hybrid update
Gated MLP	Shared neuron order	Normalize gate/up rows, down columns
MoE router	Expert rows, shared shift	Centered left spectral, then row norm

A 16-expert router can use a 16×16 Gram matrix.

Lau and Su (2026), arXiv v5 · Table 1 and Sections 3.2–3.7. AdamW handles biases and normalization gains in the reported runs.

Gemma 3 Vocabulary Optimizers

Only the embedding and untied LM-head optimizers varied. Together these matrices hold **604M of 1.087B parameters** (56%). Each run used 7.37B tokens and hybrid MLPs.



Final loss over three seeds

Optimizer	Mean ± SD
AdamW	2.7588 ± 0.0025
RowNormM	2.7254 ± 0.0012
HybridPolarGradM	$\mathbf{2.7166} \pm 0.0006$

0.0422 nats/token lower mean loss with the hybrid.

Lau and Su (2026), arXiv v5 · Figure 4(b), one run per curve. Tables G.1 and G.7, parameter counts and three-seed mean ± sample SD. Learning rates tuned per method.

The MLP Optimizer Changes the Ranking

At equal tokens, the vocabulary hybrid took 8.9% longer than AdamW with hybrid MLPs.

	Final validation loss		Wall time (h)
Embedding and head	Muon MLP	Hybrid MLP	Hybrid MLP
AdamW	2.7761	2.7562	4.191
RowNormM	**2.7434**	2.7251	4.216
HybridPolarGradM	2.7492	**2.7160**	4.566

One-seed ablation. Against 2.7562 for AdamW, the head-only hybrid lowered loss to 2.7216. The embedding-only hybrid raised it to 2.8131.

Lau and Su (2026), arXiv v5 · Figure 4 and Tables G.6 and G.9. Eight H200s. Mixed assignments kept their original learning rates.

Qwen Vocabulary and gpt-oss Routing

Qwen3-0.6B-style

Change the embedding and head. Keep hybrid MLPs fixed.

Vocabulary optimizer	Final loss
AdamW	2.8009
RowNormM	2.7927
HybridPolarGradM	**2.7710**

6.88B tokens.

Downsized gpt-oss

Fix RowNormM on the vocabulary pair. Vary the router with auxiliary losses off.

Router optimizer	Unused experts
AdamW	23.96%
HybridPolarGradM	0.52%
LeftPolarGradM	**0%**

3.93B tokens, 16 experts.

Lau and Su (2026), arXiv v5 · Figure 3(b) and Table G.14. One run per assignment. Unused-expert fractions are measured at final evaluation.

Questions

Symmetry-Compatible Principle for Optimizer Design

OpenReview · Extended paper (arXiv v5) · Code

Sydney Poster Session 4

Wednesday, December 9, 2026, 5:00–8:00 PM (AEDT)
Hall 1–4, board S1612

Supplementary Slides

A Coordinatewise Update Under Rotation

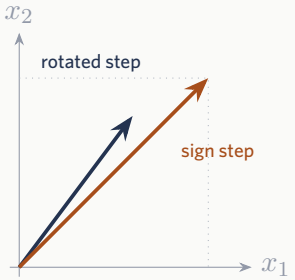
Let $d = (1, 0)^\top$ and $Qd = (3/5, 4/5)^\top$.

- Rotate the sign update

$$Q\,\mathrm{sgn}(d) = (3/5, 4/5)^\top.$$

- Rotate first, then take signs

$$\mathrm{sgn}(Qd) = (1, 1)^\top.$$



Lau and Su (2026), arXiv v5 · Section 3.1 and Appendix C. The sign rule is the zero-damping first-step limit of bias-corrected Adam with zero state and $\mathrm{sgn}(0) = 0$.

Spectral Maps and Orthogonal Equivariance

$$\mathcal{U}(PDQ^\top) = P\mathcal{U}(D)Q^\top, \qquad P \in \mathbb{O}^m,\, Q \in \mathbb{O}^n.$$

Write the singular value decomposition $D = U\,\mathrm{Diag}(\sigma(D))V^\top$.

$$\boxed{\mathcal{U}(D) = U\,\mathrm{Diag}(\psi(\sigma(D)))V^\top}$$

- Singular directions are preserved.
- The absolutely symmetric map ψ can use the whole spectrum.
- Equal singular values receive equal coefficients. Zeros stay zero.

Lau and Su (2026), arXiv v5 · Section 3.8, Theorem 3.7. The theorem characterizes continuous, single-valued maps of one direction. The ideal polar factor can be discontinuous at rank changes.

Compatible Geometries

For the compact SVD $D = U\Sigma V^\top$, the polar factor is $\mathrm{polar}(D) = UV^\top$.

Update	Action for $\sigma_i > 0$	Geometry
Gradient	$\sigma_i \mapsto \sigma_i$	Keeps relative magnitudes
Polar	$\sigma_i \mapsto 1$	Equalizes active directions
Nuclear-scaled polar	$\sigma_i \mapsto \sum_{j=1}^{m\wedge n} \sigma_j(D)$	Adds a global scale

- Ideal Muon uses the polar factor of momentum.
- Nuclear-scaled PolarGrad uses $\|\|\|D\|\|\|_{\mathrm{nuc}}\ \mathrm{polar}(D)$.

Curvature, noise, state, and compute still guide the choice.

Lau and Su (2026), Sections 3.8.1 and F.2.2. PolarGrad, Lau, Long and Su (2025). Muon, Jordan et al. (2024).

Token Relabeling and Feature Coordinates

$$D \in \mathbb{R}^{v \times d}, \qquad D \longmapsto PDR^\top$$

Token labels

$P \in \mathbb{P}^v$ relabels the rows and their token IDs together.

Feature basis

$R \in \mathbb{O}^d$ changes the orthonormal feature basis.

$$\boxed{\mathcal{U}_{\mathsf{LPRO}}(PDR^\top) = P\mathcal{U}_{\mathsf{LPRO}}(D)R^\top}$$

This permits more than a purely spectral update.

Lau and Su (2026), arXiv v5 · Section 3.3. Token labels admit permutations and feature coordinates admit orthogonal basis changes.

Row-Norm Updates Preserve Token Structure

$$M_k = \beta M_{k-1} + (1-\beta) G_k$$

$$\left[\mathcal{U}_{\text{row}}(M_k)\right]_{i:} = \frac{M_{k,i:}}{\|M_{k,i:}\|_2 + \varepsilon}$$

- A row permutation reorders the row norms.
- A right orthogonal transformation preserves every row norm.
- Each row is rescaled using one scalar, preserving its direction.

RowNormM adapts each row without a matrix inverse root.

Lau and Su (2026), arXiv v5 · Equation (4) and Section 3.7.2. The more general family uses any scalar factor $\eta(\|M_{k,i:}\|_2)$.

Hybrid Row-Norm and Right-Spectral Updates

$$Z_k = \mathcal{U}_{\text{row}}(M_k), \qquad C_k = Z_k^\top Z_k \in \mathbb{R}^{d\times d}$$
$$R_k = (C_k + \varepsilon I_d)^{-1/2}, \qquad U_k = \nu_k^\alpha Z_k R_k$$
$$\nu_k = \max\{\operatorname{tr}(C_k R_k), \varepsilon\}, \qquad \alpha \in [0,1]$$

- Normalize each row of the embedding momentum M_k.
- The right Gram matrix couples feature directions.
- Composition preserves the transformation law.

Lau and Su (2026), arXiv v5 · Section 3.7.2, Algorithm 2 (lines 15–21) and Table G.5. Gemma uses $\alpha = 1$ for the vocabulary hybrid.

Shared-Shift Redundancy in the LM Head

$$\mathrm{softmax}(z + c\mathbf{1}_v) = \mathrm{softmax}(z)$$
$$(W_{\mathrm{out}} + \mathbf{1}_v a^\top)h = W_{\mathrm{out}}h + (a^\top h)\mathbf{1}_v$$

Choose a centered update

$$\Pi_v^\perp = I_v - \frac{1}{v}\mathbf{1}_v\mathbf{1}_v^\top, \qquad \mathbf{1}_v^\top U_k = 0.$$

- Remove a shared change in all vocabulary logits.
- The input embedding does not have this softmax redundancy.

Lau and Su (2026), arXiv v5 · Section 3.3.2, equation (5). This argument assumes untied weights and ordinary, uncapped logits.

Center Again After Row Normalization

$$\mathcal{U}_{\text{row}}^{\text{LM}}(D) = \Pi_v^{\perp}\mathcal{U}_{\text{row}}(D_c), \qquad D_c = \Pi_v^{\perp} D$$

$$\underbrace{\begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}}_{\text{mean } 0} \xrightarrow{\mathcal{U}_{\text{row}}} \underbrace{\begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}}_{\text{mean } -1/3} \xrightarrow{\Pi_3^{\perp}} \underbrace{\begin{pmatrix} 4/3 \\ -2/3 \\ -2/3 \end{pmatrix}}_{\text{mean } 0}$$

- Row-dependent scaling can reintroduce a shared component.
- Project after rowwise nonlinear stages and on the final update.

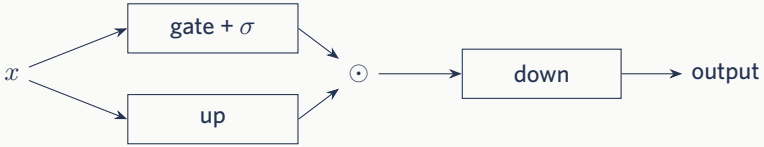
Lau and Su (2026), arXiv v5 · Sections 3.3.2 and 3.7.2. Equation (5) gives the projected row update. This example uses exact row normalization with zero damping.

SwiGLU Couples Three Projections

$$\mathrm{SwiGLU}(x; W_{\mathrm{gate}}, W_{\mathrm{up}}, W_{\mathrm{down}}) = W_{\mathrm{down}}\big(\sigma(W_{\mathrm{gate}}x) \odot (W_{\mathrm{up}}x)\big)$$

For one neuron permutation $P \in \mathbb{P}^{d_{\mathrm{ff}}}$,

$$\widetilde{W}_{\mathrm{gate}} = PW_{\mathrm{gate}}, \qquad \widetilde{W}_{\mathrm{up}} = PW_{\mathrm{up}}, \qquad \widetilde{W}_{\mathrm{down}} = W_{\mathrm{down}}P^{\top}.$$



Lau and Su (2026), arXiv v5 · Section 3.4, Proposition 3.4. The same reordering of the intermediate neurons leaves the SwiGLU output unchanged.

Softmax Routers Operate on Expert Contrasts

$$W \in \mathbb{R}^{e \times d}, \qquad p(x; W) = \mathrm{softmax}(Wx)$$

$$D_c = \Pi_e^{\perp} D, \qquad \Pi_e^{\perp} = I_e - \frac{1}{e}\mathbf{1}_e\mathbf{1}_e^{\top}$$

Projected row map

$$\mathcal{U}_{\mathrm{row}}^{\mathrm{router}}(D) = \Pi_e^{\perp}\mathcal{U}_{\mathrm{row}}(D_c)$$

Rescale expert rows, then center.

Left-spectral map

$$\mathcal{U}_{\mathsf{L}}^{\mathrm{router}}(D) = \Psi(D_c D_c^{\top})D_c$$

Use an $e \times e$ expert Gram matrix.

With 16 experts, the spectral statistic is only 16×16.

Lau and Su (2026), arXiv v5 · Section 3.5, Proposition 3.5. The map Ψ is permutation equivariant and preserves the centered expert subspace $\mathbf{1}_e^{\perp}$.

Gemma 3 Experimental Setup

- **Model** Gemma 3 1B-style, 18 layers, 1.087B parameters. Untied vocabulary matrices, each $262{,}144 \times 1{,}152$.
- **Budget** 7.37B FineWeb-Edu tokens, 50,000 steps, context length 1,024, BF16 on eight H200 GPUs.
- **Varied** embedding and head optimizer together, with method-specific base learning-rate sweeps.
- **Fixed** Muon on attention/linear matrices, hybrid MLP updates, AdamW on scalars and vectors.

Lau and Su (2026), arXiv v5 · Tables G.1 and G.4–G.7. The sweep uses a shared base learning rate for embedding and head within each method.

The Ordering Holds Across Three Seeds

Final validation cross-entropy at 7.37B tokens

Embedding and head	Seed 1	Seed 2	Seed 3	Mean $\pm$ SD
AdamW	2.7562	2.7612	2.7591	2.7588 ± 0.0025
RowNormM	2.7251	2.7267	2.7243	2.7254 ± 0.0012
HybridPolarGradM	2.7160	2.7172	2.7166	$\mathbf{2.7166} \pm 0.0006$

0.0334 nats/token
mean reduction with RowNormM.

0.0422 nats/token
mean reduction with HybridPolarGradM.

Lau and Su (2026), arXiv v5 · Table G.7. SD is the sample standard deviation across seeds. All other optimizer assignments are fixed.

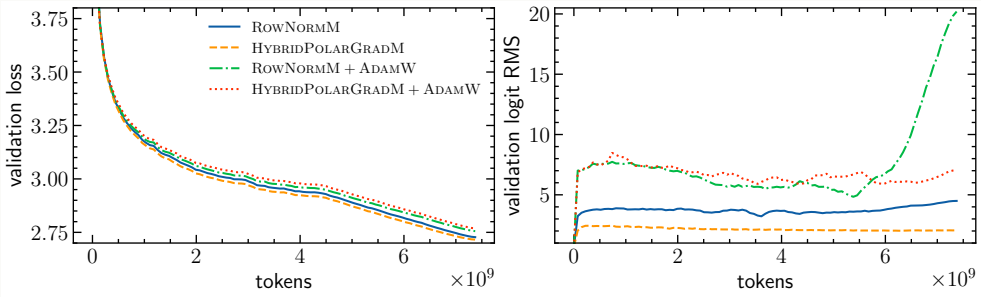
The LM Head Explains Most of the Paired Gain

Changed block	RowNormM	HybridPolarGradM
Neither (AdamW on both)	2.7562	2.7562
Embedding only	2.7497	2.8131
LM head only	**2.7275**	**2.7216**
Both	2.7251	2.7160

- The head-only hybrid recovers 0.0346 of the paired 0.0402 loss reduction.
- The embedding-only hybrid increases loss at the transferred learning rates.

Lau and Su (2026), arXiv v5 · Table G.9. One seed, untied weights. Original method-specific learning rates are retained for these swaps.

Centered Heads Reduce Shared-Logit Drift



AdamW head

Raw RMS 20.20 Centered RMS 2.57

Projected RowNormM head

Raw RMS 4.49 Centered RMS 4.49

Lau and Su (2026), arXiv v5 · Figure G.3(a) and Table G.8. Uncapped logits, RowNormM embedding for both highlighted heads. In the legend, + AdamW changes only the head.

Vocabulary Learning-Rate Sweep

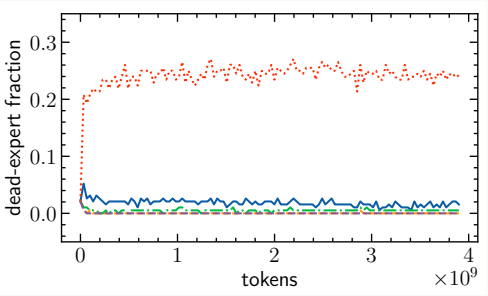
Vocabulary optimizer	Tested LR range	Final loss range	Selected LR
AdamW	0.00025–0.0025	2.7562–2.7790	0.001
RowNormM	0.001–0.01	2.7251–2.7467	0.005
HybridPolarGradM	0.00005–0.0005	2.7160–2.7362	0.00005

- Four tested rates per method; embedding and head share a base rate.
- The hybrid's selected rate is at the lower edge of its tested grid.
- The three-seed comparison reuses the selected rates.

Lau and Su (2026), arXiv v5 · Table G.6(b). Gemma 3 1B-style with hybrid MLPs.

Router Geometry Changes Expert Utilization

16 experts per layer · fixed RowNormM vocabulary pair · auxiliary losses off



Router	Loss	Dead frac.
RowNormM	2.8631	0.0156
LeftPolarGradM	2.8670	**0**
HybridPolarGradM	**2.8606**	0.0052
Uncentered left polar	2.8666	**0**
AdamW	2.8858	0.2396

Lau and Su (2026), arXiv v5 · Section G.4.1, Figure G.7(b) and Table G.14. Dead-expert fractions are measured at final evaluation, one run per router.

Router Geometry With Auxiliary Losses

	Validation cross-entropy		Expert-load CV	
Router	Aux off	Aux on	Aux off	Aux on
AdamW	2.8858	2.8693	1.0373	0.2398
RowNormM	2.8631	**2.8605**	0.6148	0.1954
HybridPolarGradM	**2.8606**	2.8612	0.5650	**0.1883**

- With auxiliary losses, every tested router uses all experts at final evaluation.
- The router z-loss changes under a common logit shift, although softmax probabilities do not.

Lau and Su (2026), arXiv v5 · Table G.14. CV is the coefficient of variation of expert load. Bold marks the best of the three methods shown.

Compatible Statistics and Systems Cost

Update	Statistic	Under row sharding
RowNormM	One norm per row	Local for complete rows
Vocabulary hybrid	$C_k = \sum_s Z_{k,s}^\top Z_{k,s}$	Reduce a $d \times d$ feature Gram matrix
Head centering	$\sum_i M_{k,i:}$	Reduce a length-d feature vector
Router left spectral	$M_{k,c} M_{k,c}^\top$	Cross-shard products between expert rows

Geometry, sharding, precision, and iteration count must be designed together.

Lau and Su (2026), arXiv v5 · Sections 3.7 and 4. The table derives row-sharding costs from the update formulas. The reported experiments used data parallelism. $Z_{k,s}$ contains the rows of Z_k on shard s.

Projected Hybrid Head Update

$$
\begin{aligned}
M_k &= \beta M_{k-1} + (1-\beta) G_k, \\
M_{k,c} &= \Pi_v^{\perp} M_k, \qquad Z_k = \Pi_v^{\perp} \mathfrak{U}_{\mathsf{row}}(M_{k,c}), \\
C_k &= Z_k^{\top} Z_k, \qquad R_k = (C_k + \varepsilon I_d)^{-1/2}, \\
\nu_k &= \max\{\operatorname{tr}(C_k R_k), \varepsilon\}, \qquad U_k = \Pi_v^{\perp}(\nu_k^{\alpha} Z_k R_k), \\
W_{k+1} &= (1-\gamma_k \lambda) W_k - \gamma_k U_k.
\end{aligned}
$$

- Use a finite matrix polynomial iteration for the spectral operation.
- Project the gradient-driven update. Weight decay also acts on the parameter representative.

Lau and Su (2026), arXiv v5 · Algorithm 2, lines 15–21, and Appendix E.1. The inverse square root shown is the target of the finite polynomial iteration.

Limits of Shared-Shift Invariance

Setting	Consequence
Untied head, ordinary softmax	Shared row shift leaves probabilities unchanged.
Tied embedding and output head	Shifting the common parameter also changes the input representation.
Nonlinear final-logit cap	A shared pre-cap shift generally becomes nonuniform after the cap.
Router z-loss	Gating probabilities are unchanged, but the auxiliary objective changes.

The quotient depends on the architecture and objective.

Lau and Su (2026), arXiv v5 · Sections 3.3.2, G.2.3 and G.4.1.

Two Compatible Norm Choices

For a nonzero direction D, both norm balls respect orthogonal changes of basis.

$$\langle\!\langle D, D/|||D|||_{\mathrm{F}} \rangle\!\rangle_{\mathrm{F}} = \max_{|||Z|||_{\mathrm{F}} \leqslant 1} \langle\!\langle D, Z \rangle\!\rangle_{\mathrm{F}} = |||D|||_{\mathrm{F}}$$

$$\langle\!\langle D, \mathrm{polar}(D) \rangle\!\rangle_{\mathrm{F}} = \max_{|||Z|||_{\mathrm{S}} \leqslant 1} \langle\!\langle D, Z \rangle\!\rangle_{\mathrm{F}} = |||D|||_{\mathrm{nuc}}$$

- The Frobenius ball keeps relative singular-value magnitudes.
- The spectral-norm ball permits a polar maximizer.
- At rank deficiency, the spectral-norm oracle can be set-valued.

Different invariant norms give different steepest directions.

Lau and Su (2026), arXiv v5 · Sections 2.1, 3.8.1 and A.1. The subscripts F, S, and nuc denote the Frobenius, spectral, and nuclear norms.

Shared Offsets and Finite Precision

$$\mathrm{softmax}(0, 1) = \mathrm{softmax}(1024, 1025) \quad \text{in exact arithmetic.}$$

$$\mathrm{BF16}(1024, 1025) = (1024, 1024).$$

- A common offset can make the difference between logits hard to resolve.
- A centered update contributes zero to the mean logit shift.
- Individual token logits can still be large after centering.

Lau and Su (2026), arXiv v5 · The BF16 calculation illustrates rounding. Section G.2.3 and Table G.8 report the measured raw and centered logit scales.

Architecture and Optimizer Co-Design

- **Identify equivalent descriptions.** For a new block, ask which row or column permutations, basis changes, or shared shifts preserve its meaning.
- **Choose the update.** Match the update to those transformations. For an ordinary softmax router, center the update across experts.
- **Choose the statistic.** For a 16-expert router, a left-spectral update uses a 16×16 Gram matrix. A projected row update uses one norm per expert.

The layer determines which updates are compatible. Their cost can inform the design of the layer itself.

Lau and Su (2026), arXiv v5 · Section 3.6 and Table 1. Compare candidate training recipes at matched token budgets and report their computational cost.

References — Principle and Optimizers

- **Lau and Su (2026).** Symmetry-Compatible Principle for Optimizer Design: Embeddings, LM Heads, SwiGLU MLPs, and MoE Routers. *Advances in Neural Information Processing Systems (NeurIPS).* arXiv:2605.18106v5.
- **Lau, Long and Su (2025).** PolarGrad: A Class of Matrix-Gradient Optimizers from a Unifying Preconditioning Perspective. *arXiv:2505.21799.*
- **Jordan et al. (2024).** Muon: An Optimizer for Hidden Layers in Neural Networks. Technical note and software.
- **Kingma and Ba (2015).** Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR).*
- **Loshchilov and Hutter (2019).** Decoupled Weight Decay Regularization. *International Conference on Learning Representations (ICLR).*

References — Models, Data, and Numerics

- **Gemma Team (2025).** Gemma 3 Technical Report. *arXiv:2503.19786.*
- **Yang et al. (2025).** Qwen3 Technical Report. *arXiv:2505.09388.*
- **OpenAI (2025).** gpt-oss-120b & gpt-oss-20b Model Card. *arXiv:2508.10925.*
- **Penedo et al. (2024).** The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track.*
- **Amsel, Persson, Musco and Gower (2026).** The Polar Express: Optimal Matrix Sign Methods and Their Application to the Muon Algorithm. *International Conference on Learning Representations (ICLR).*