Design architectures that are easier to optimize

Case studies in scaling LLMs

Bingcong Li

ETH Zurich

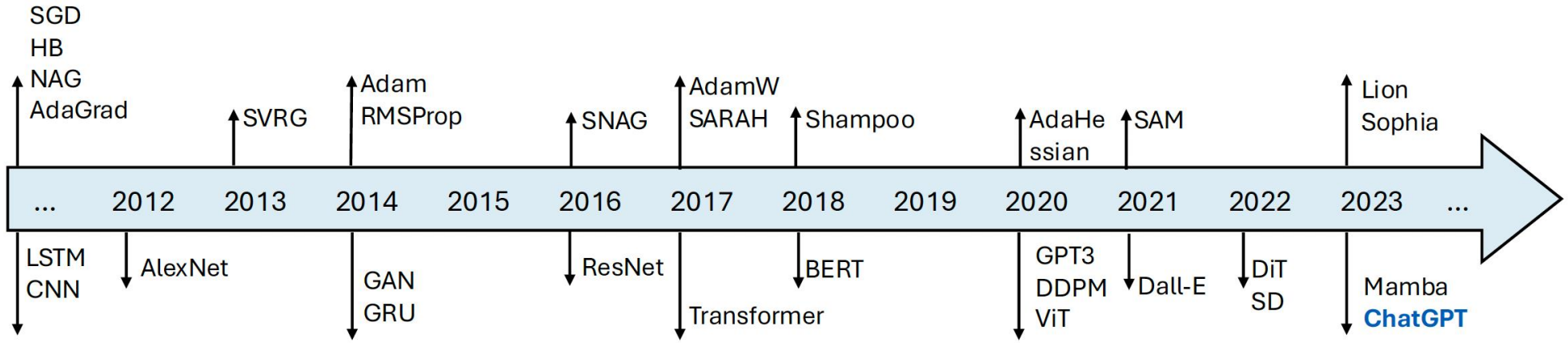
https://bingcongli.github.io/

Focus Period Lund

Lund University

April 22 2026

Two ways of making training easier



Improving optimizers

- Design smart update rules
- One optimizer, many architectures

Improving architectures

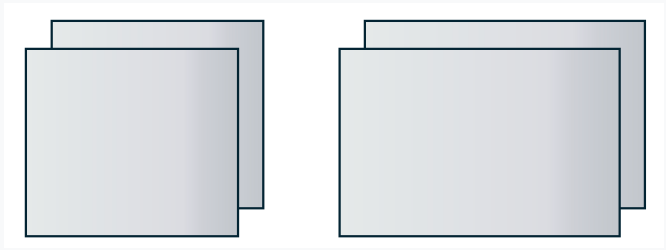
- Design new architectures
- Rarely designed ***for*** optimization

Architectures with equal expressiveness can be very different to optimize!

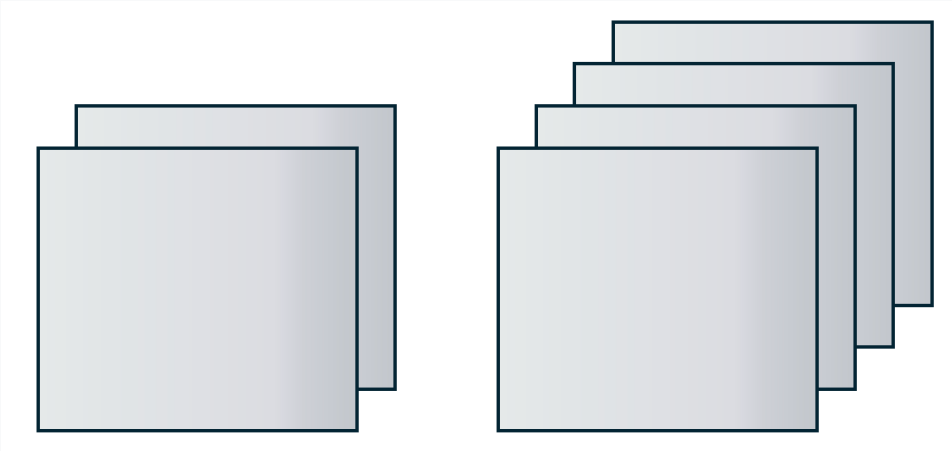
Roadmap

Scaling wider

- Bottleneck: linear layers
- Solution: more structures
- Application: Fine-tuning LLMs



Scaling deeper



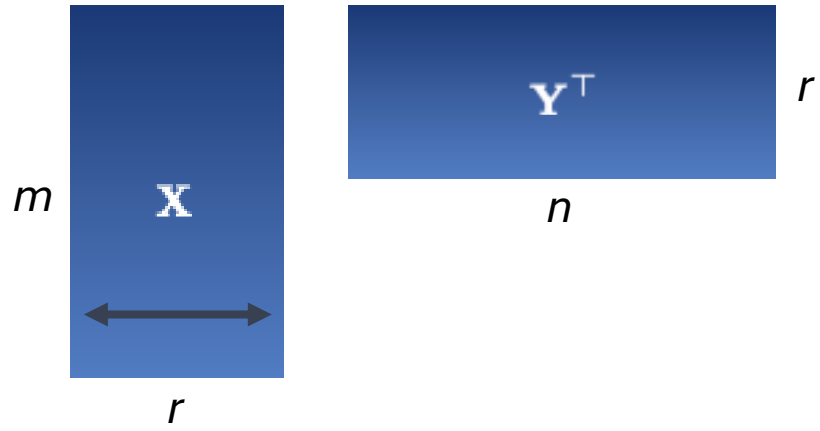
Yudong Wei, Liang Zhang, Bingcong Li* and Niao He*. On the benefits of weight normalization for overparameterized matrix sensing. *ICLR 2026*.
Kai Lion, Liang Zhang, Bingcong Li* and Niao He*. Polar: Polar-Decomposed Low-Rank Adapter Representation. *NeurIPS 2025*.

The scaling wider problem, in its simplest form

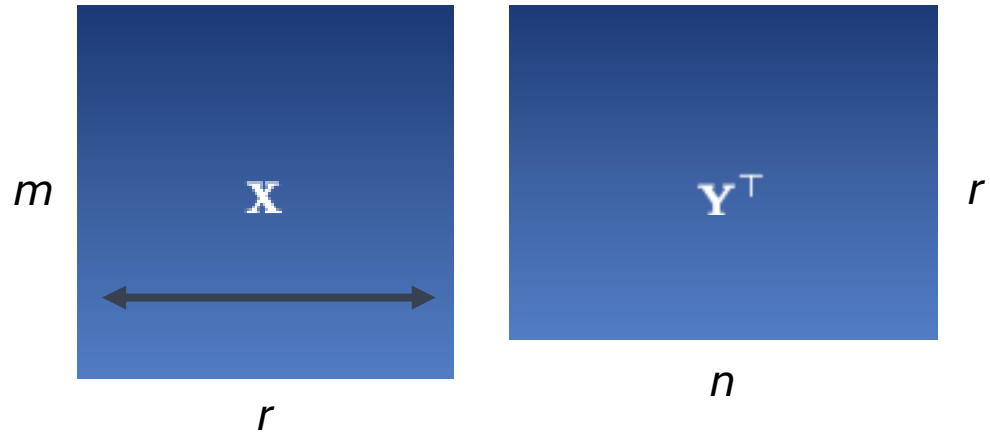
A nonconvex problem with known global minima

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times r},\mathbf{Y}\in\mathbb{R}^{n\times r}} \frac{1}{4}\|\mathcal{M}(\mathbf{X}\mathbf{Y}^\top)-\mathbf{b}\|^2$$

$$\mathcal{M}(\mathbf{X}\mathbf{Y}^\top)=\begin{bmatrix}\langle\mathbf{M}_1,\mathbf{X}\mathbf{Y}^\top\rangle\\ \langle\mathbf{M}_2,\mathbf{X}\mathbf{Y}^\top\rangle\\ \vdots\\ \langle\mathbf{M}_s,\mathbf{X}\mathbf{Y}^\top\rangle\end{bmatrix} \qquad \mathbf{b}=\mathcal{M}(\mathbf{A})=\begin{bmatrix}\langle\mathbf{M}_1,\mathbf{A}\rangle\\ \langle\mathbf{M}_2,\mathbf{A}\rangle\\ \vdots\\ \langle\mathbf{M}_s,\mathbf{A}\rangle\end{bmatrix}$$



- $\{\mathbf{M}_i\in\mathbb{R}^{m\times n}\}_{i=1}^s$ features
- $b_i=\langle\mathbf{M}_i,\mathbf{A}\rangle$ labels
- $\mathbf{A}\in\mathbb{R}^{m\times n}$ latent ground truth
- r_A latent rank



B. Recht, M. Fazel, P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, *52*(3), 471-501, 2010

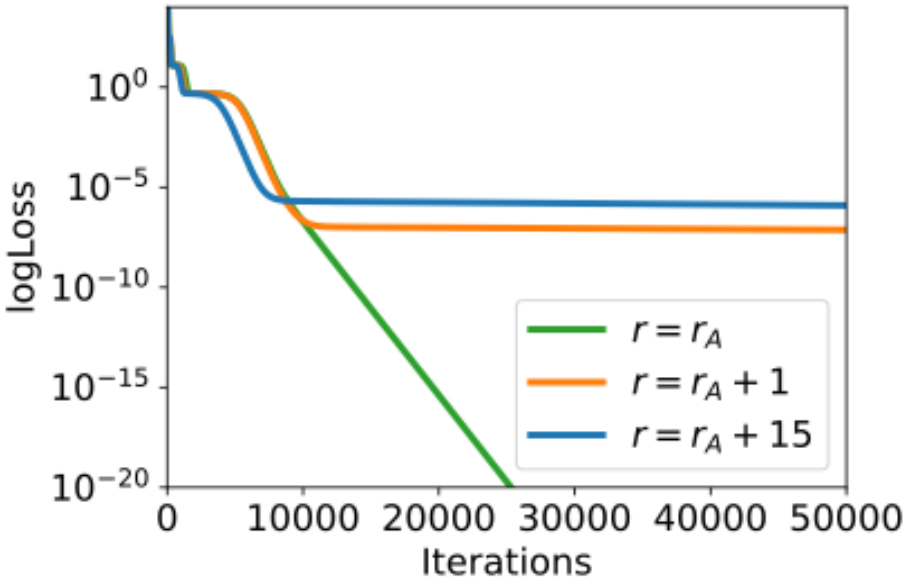
The width dilemma

- 0 training loss if $\mathbf{XY}^\top = \mathbf{A}$, i.e., wide NNs $r \geq r_A$ solves this problem
- NNs with $r > r_A$ converge **exponentially slower** than $r = r_A$

Theorem [XDD'24]

0 training loss cannot be attained via gradient flow (GF) faster than

$$\Omega\big((r - r_A)/t\big)$$



Q. How can we improve performance with larger *r* ?

N. Xiong, L. Ding, and S. S. Du. How over-parameterization slows down gradient descent in matrix sensing. *ICLR 2024*

A manifold based architecture

Manifold architecture

Polar decomposition $\mathbf{X} = \mathbf{U}\boldsymbol{\Theta}_1 \quad \mathbf{Y} = \mathbf{V}\boldsymbol{\Theta}_2$

- Direction $\mathbf{U} \in \mathrm{St}(m,r) := \{\mathbf{U} \in \mathbb{R}^{m\times r} | \mathbf{U}^\top \mathbf{U} = \mathbf{I}_r\}$
- Magnitude $\boldsymbol{\Theta}_1 \in \mathbb{S}_+^{r\times r}$

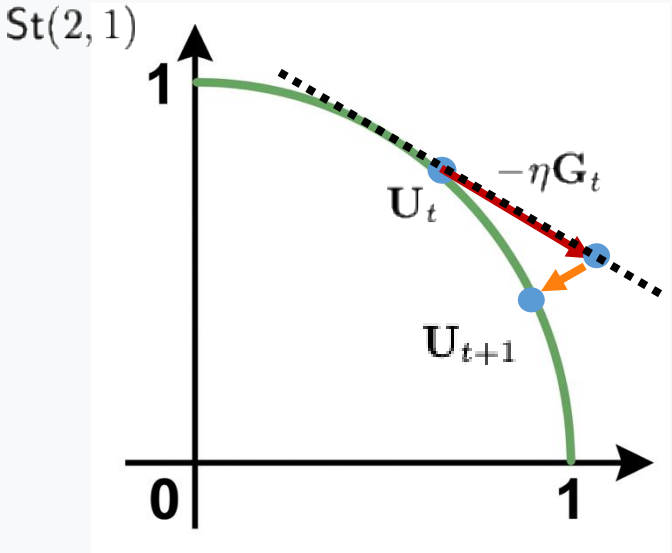
$$\min_{\mathbf{U},\mathbf{V},\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2} \quad \frac{1}{4}\left\|\mathcal{M}(\mathbf{U}\boldsymbol{\Theta}_1\boldsymbol{\Theta}_2^\top\mathbf{V}^\top) - \mathbf{b}\right\|^2$$
$$\text{s.t.} \quad \mathbf{U} \in \mathrm{St}(m,r), \quad \mathbf{V} \in \mathrm{St}(n,r),$$
$$\boldsymbol{\Theta}_1 \in \mathbb{S}_+^{r\times r}, \quad \boldsymbol{\Theta}_2 \in \mathbb{S}_+^{r\times r}$$

Merging magnitude together

$$\min_{\mathbf{U},\boldsymbol{\Theta},\mathbf{V}} \quad \frac{1}{4}\left\|\mathcal{M}(\mathbf{U}\boldsymbol{\Theta}\mathbf{V}^\top) - \mathbf{b}\right\|^2$$
$$\text{s.t.} \quad \mathbf{U} \in \mathrm{St}(m,r), \quad \mathbf{V} \in \mathrm{St}(n,r), \quad \boldsymbol{\Theta} \in \mathbb{R}^{r\times r}$$

Basic Riemannian optimization

- **Riemannian opt.** for $\mathbf{U}_t$ and $\mathbf{V}_t$
- **GD** for $\boldsymbol{\Theta}_t$



P.A. Absil, R. Mahony, and R. Sepulchre. Optimization algorithms on matrix manifolds. *Princeton University Press*, 2009.
N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

Wide NNs ⇒ fast convergence & less data

Theorem (width is a friend, provably)

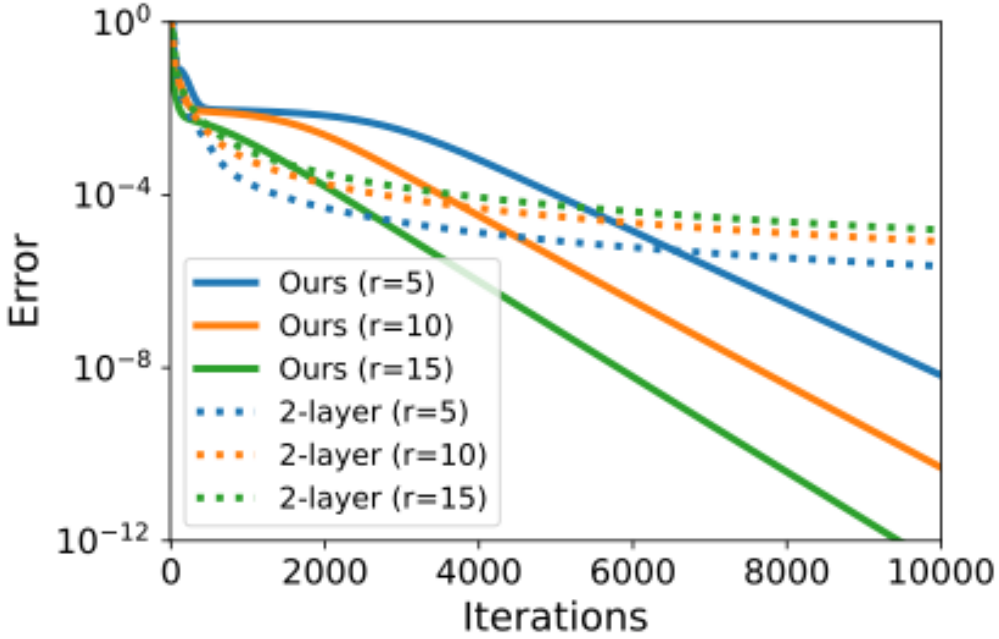
✓ **Wider ⇒ faster**. To achieve an ϵ - error globally, iterations needed are bounded by:

$$T = \mathcal{O}\left(\frac{\kappa^4 m^2 r_A}{r^2} \log \frac{1}{\epsilon}\right)$$

✓ **Wider ⇒ less data**. Samples needed for global convergence:

$$s = \mathcal{O}(\kappa^4 / r)$$

✓ Byproduct: exponentially faster convergence



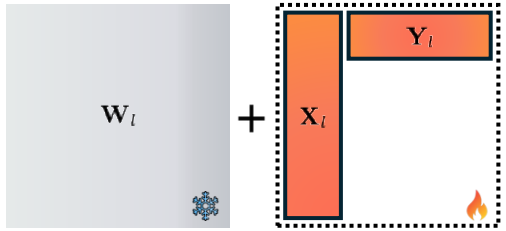
Manifold architectures turn width into a friend

Yudong Wei, Liang Zhang, Bingcong Li and Niao He*. On the benefits of weight normalization for overparameterized matrix sensing. *ICLR 2026*.

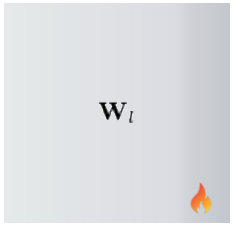
LoRA: an additive two-layer NN for fine-tuning

- LoRA for fine-tuning LLMs
  - Frozen pretrained weights $\mathbf{W}_l \in \mathbb{R}^{m\times n}$
  - Trainable two-layer nets $\mathbf{X}_l \in \mathbb{R}^{m\times r}$, $\mathbf{Y}_l \in \mathbb{R}^{n\times r}$
  - Number of variables: *O((m + n)r)* vs. *O(mn)*

$$\min_{\{\mathbf{X}_l,\mathbf{Y}_l\}} f(\{\mathbf{W}_l + \mathbf{X}_l\mathbf{Y}_l^\top\}_l)$$

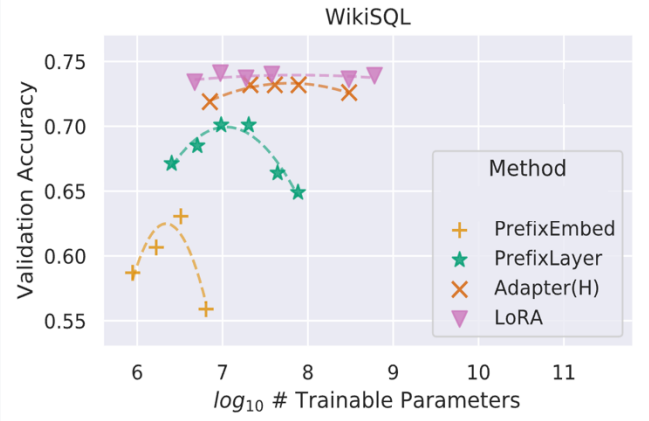


$$\min_{\{\mathbf{W}_l\}} f(\{\mathbf{W}_l\}_l)$$



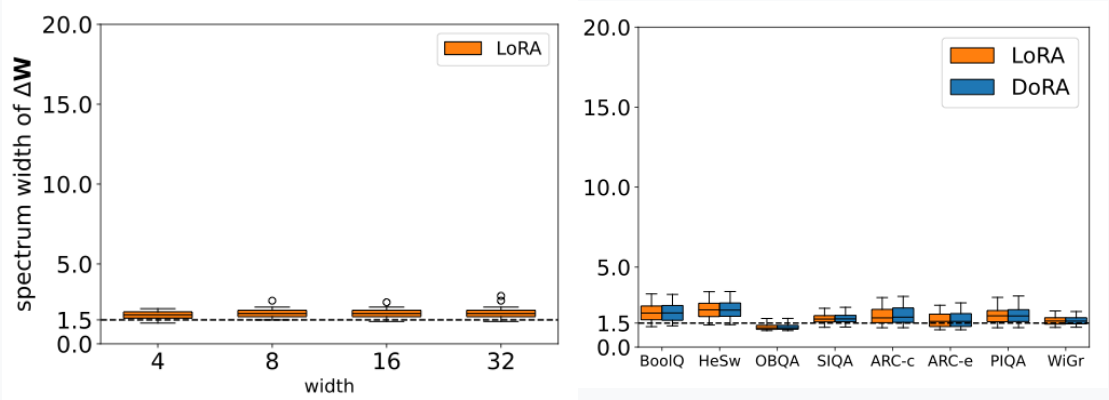
Larger width != better performance

Performance is agnostic to width *r*



LoRA wastes spectrum width

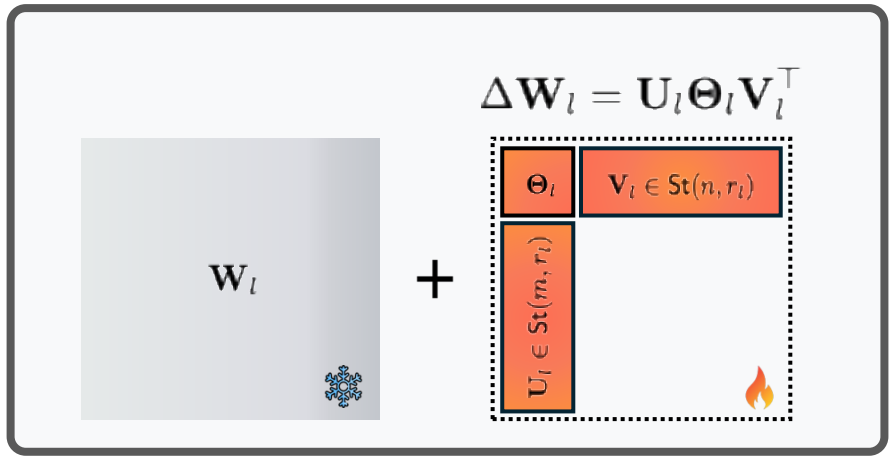
$$\Delta\mathbf{W}_l = \mathbf{X}_l\mathbf{Y}_l^\top \qquad \text{sw}(\Delta\mathbf{W}_l) = \|\Delta\mathbf{W}_l\|_F^2/\|\Delta\mathbf{W}_l\|^2$$



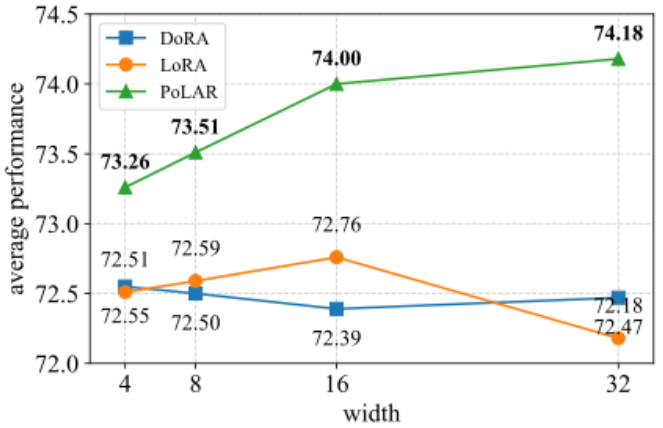
E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. *ICLR* 2022

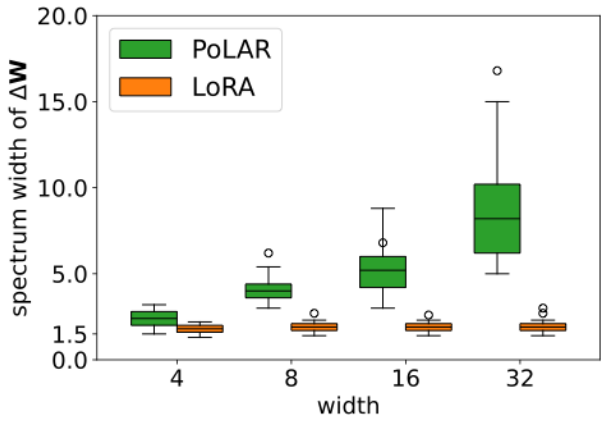
PoLAR: polar decomposed low-rank representation

PoLAR: Manifold arch



➢ LLaMA2-7B on commonsense reasoning





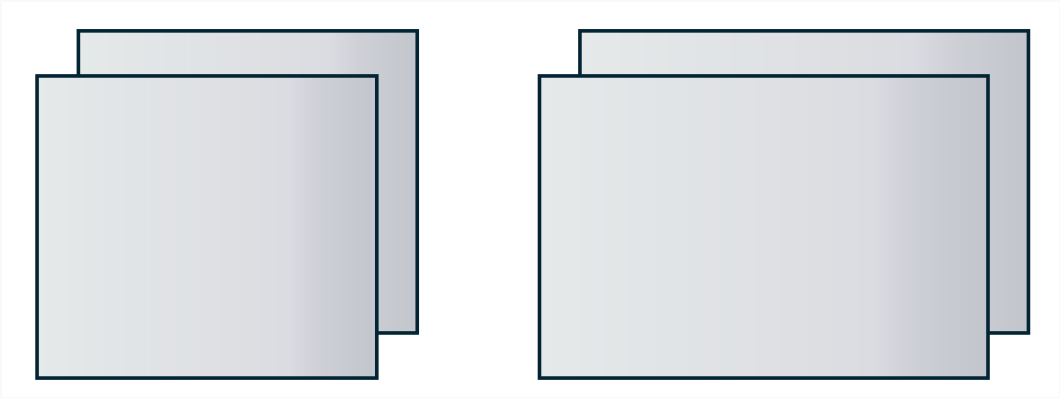
➢ Scale to larger models

	GSM8K primary school	**MATH** high school competition
GPT-4 (baseline)	92.00	42.50
Gemma-3-27B + LoRA	85.37	41.94
Gemma-3-27B + PoLAR	**85.67**	**42.70**

Kai Lion, Liang Zhang, Bingcong Li* and Niao He*. Polar: Polar-Decomposed Low-Rank Adapter Representation. *NeurIPS 2025*.

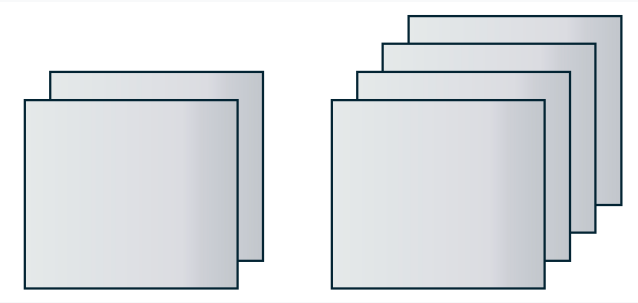
Roadmap

Scaling wider



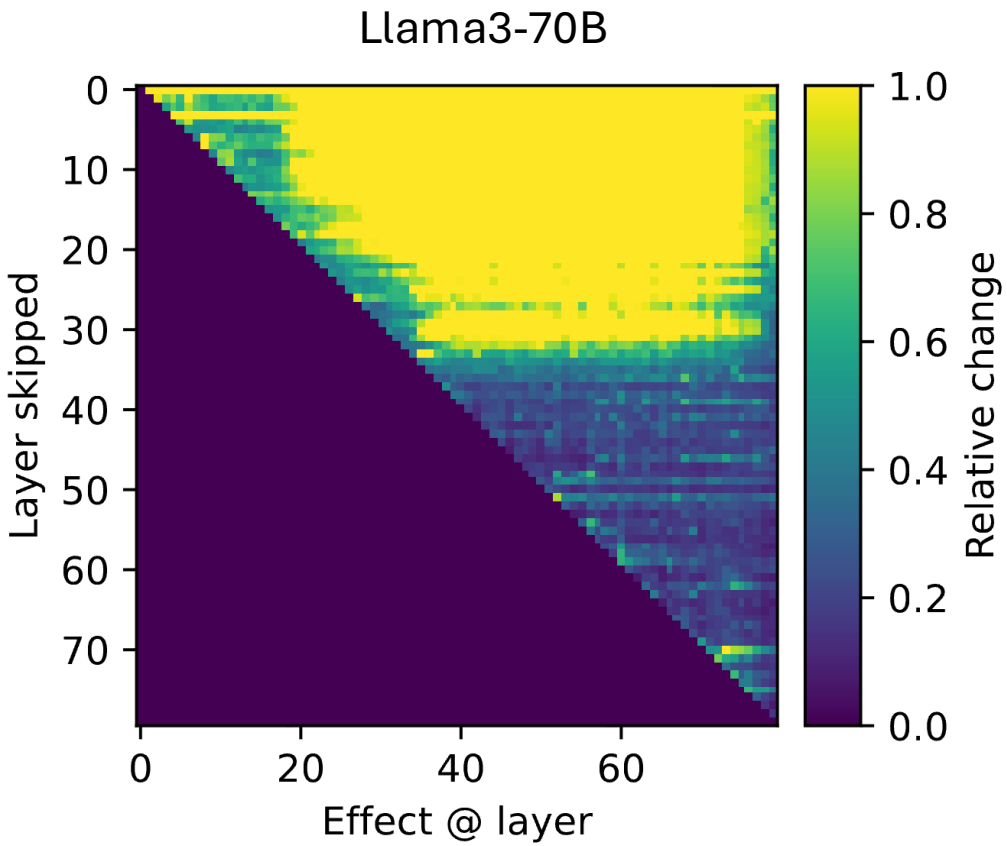
Scaling deeper

- Bottleneck: residual connections
- Solution: learnable topology
- Application: Pretraining LLMs



Yilang Zhang*, Bingcong Li*, Niao He, and Georgios B. Giannakis. Ancre: Adaptive neural connection reassignment for efficient depth scaling. *Arxiv 2602.09009*.

Scaling deeper can be inefficient

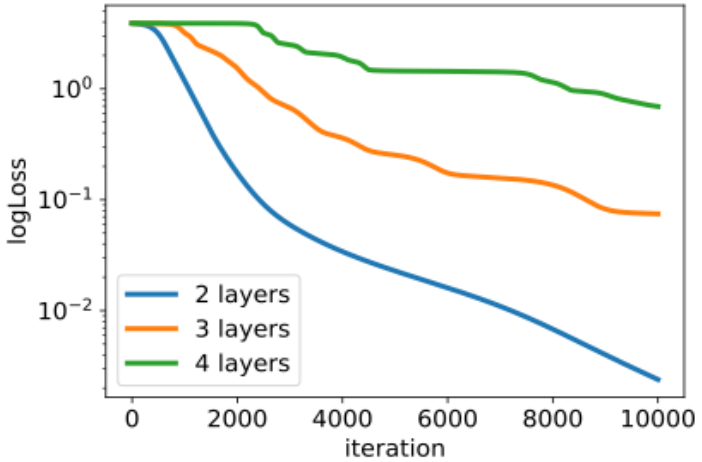


R. Csordás, C. Manning, C. Potts. Do language models Use their depth efficiently? *NeurIPS 2025*.

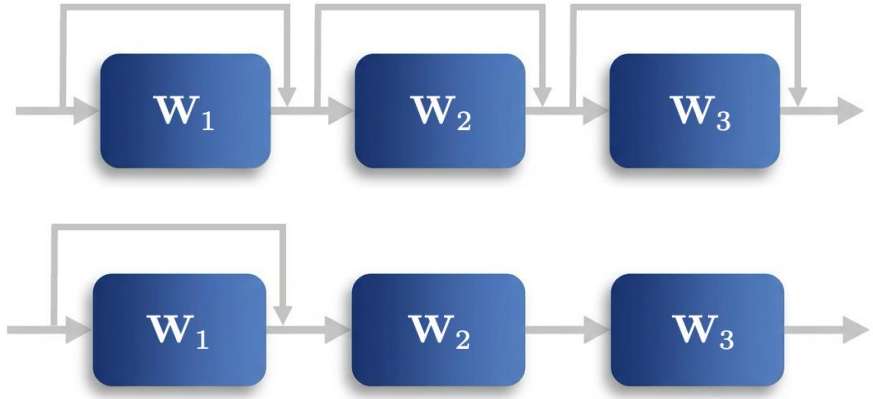
Residual connections revisited

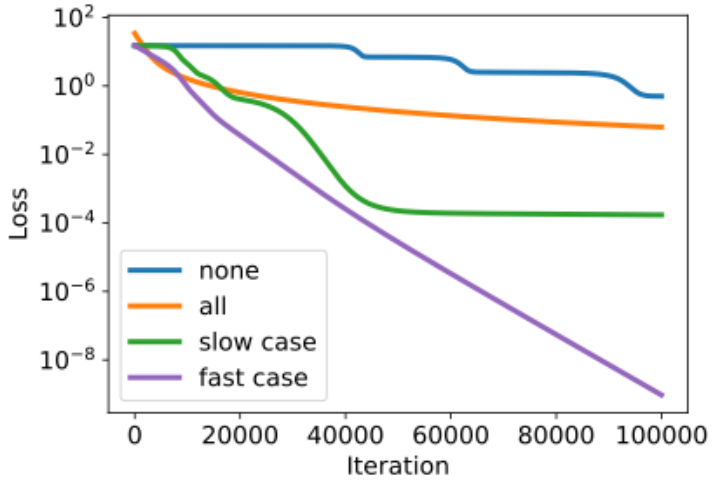
- Deep linear NNs
  - Deeper is exponentially harder

3-layer case $\|\mathbf{W}_3\mathbf{W}_2\mathbf{W}_1\mathbf{X} - \mathbf{A}\|_F^2$



- Which topology converges faster?





S. Arora, N. Cohen, N. Golowich, W. Hu. A convergence analysis of gradient descent for deep linear neural networks. *ICLR 2019*.
O. Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. *COLT 2019*.

Topology can make an exponential difference

Theorem (Lower bound)

GF with small initialization cannot converge faster than a **sublinear** rate

$$\mathcal{L}(t) \geq \Omega(1/t^2)$$

Theorem (Upper bound)

GF with small initialization ensures **linear** convergence

$$\mathcal{L}(t) \leq \mathcal{L}(0)e^{-2(1-\lambda)^2 t}$$

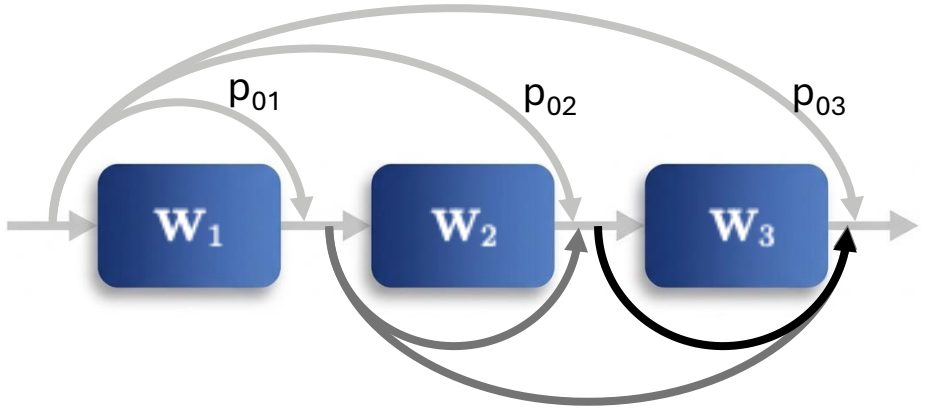
Q. How to find an optimal topology?

Yilang Zhang*, Bingcong Li*, Niao He, and Georgios B. Giannakis. Ancre: Adaptive neural connection reassignment for efficient depth scaling. *Arxiv 2602.09009*.

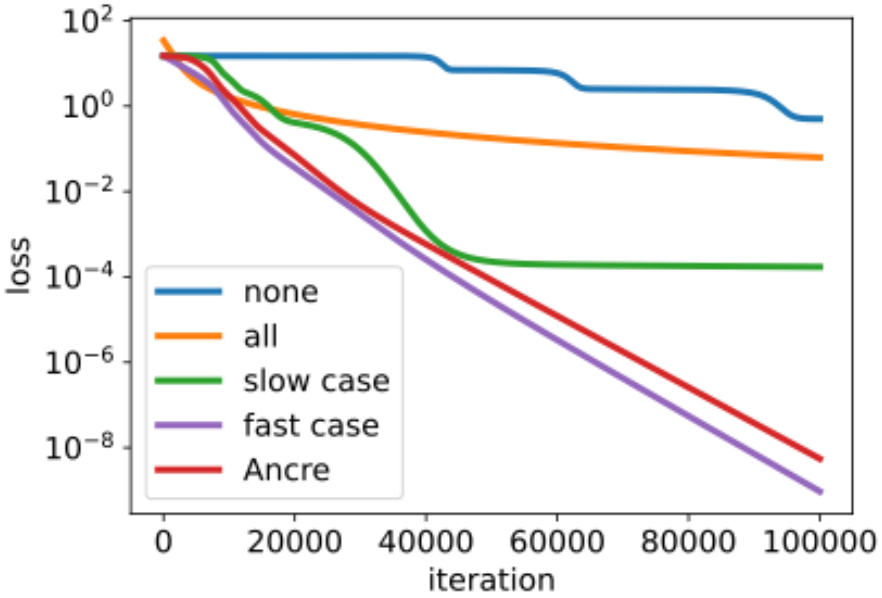
Ancre: Adaptive neural connection reassignment

- Reduce *O(L!)* patterns to $O(L^2)$ variables

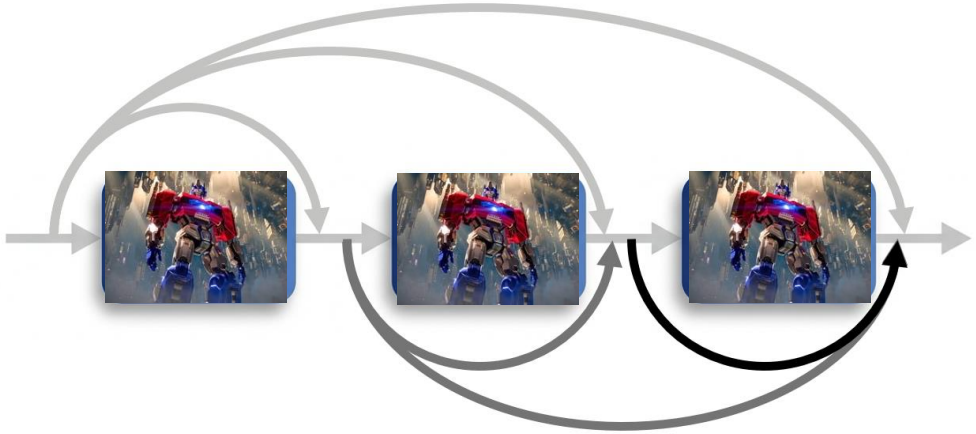
$p_{01} + p_{02} + p_{03} = 1, 0 <= p_{0i} <= 1$



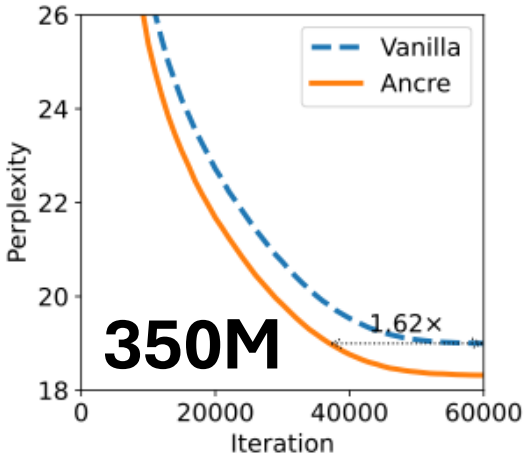
- Ancre is faster than all connections
- Ancre is exponentially faster than slow case
- Ancre is only slightly slower than fast case

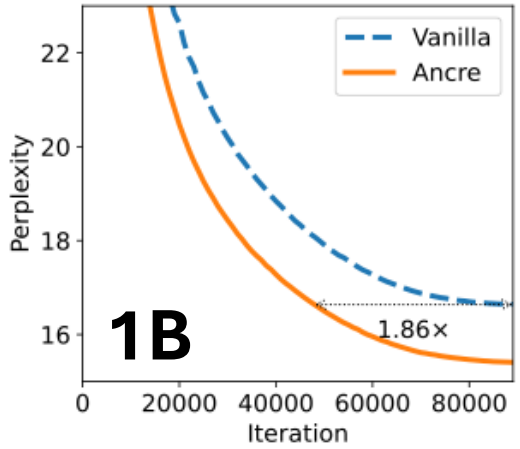


Ancre in practice



- Pretrain Llama series on the C4 dataset
  - > 1.5x gain





- DiT-B/S on ImageNet

Impv. Ancre / vanilla	FID ↓	SFID↓	IS↑
DiT-S/2 + Ancre	- 3.39 66.01 / 69.40	- 0.77 11.68 / 12.45	+ 1.04 20.70/19.66
DiT-S/2+ Ancre (cfg = 1.5)	- 2.79 42.99 / 45.78	- 0.47 8.61 / 9.08	+ 1.56 35.04 / 33.48
DiT-B/2 + Ancre	- 2.65 41.66 / 44.31	- 0.53 7.89/8.42	+ 1.51 34.40 / 32.89
DiT-B/2 + Ancre (cfg = 1.5)	- 1.88 20.53/ 22.41	- 0.54 5.81/6.35	+ 5.18 70.45 / 65.27

Concluding remarks

✓ **Scaling wider**

- Manifold arch.
- Wider models yield better scaling in FT

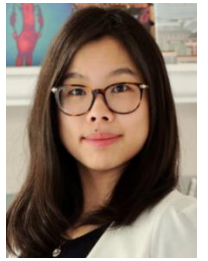
✓ **Scaling deeper**

- Position of residuals (exponentially) matters
- LLM and diffusion can benefit more from depth

❑ **Future directions**

- More general problems
- Arch and Opt co-design

Georgios B. Giannakis



Niao He

Kai Lion

Yudong Wei

Liang Zhang

Yilang Zhang

Thank You!