Towards Predictive Science for Controllable Learning Systems

Previous Work, Curiosity and Future Vision

Keyu Wang
September, 2026

Background

➢ Education

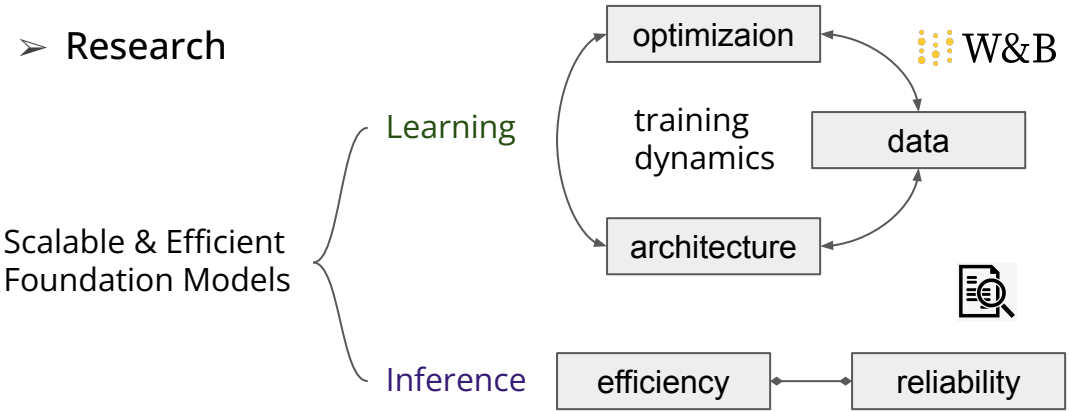
M.Sc. of Machine Learning — Oct. 2024 – May 2027 (Expected)
University of Tuebingen — *Tuebingen, Germany*

B.Eng of Artificial Intelligence — Sep. 2020 – Jun. 2024
Southeast University — *Nanjing, China*

➢ Experience

Research Assistant — Dec. 2025 – Jan. 2027 (Expected)
Max Planck Institute for Intelligent Systems — *Tuebingen, Germany*

➢ Research



what I worked on before (RA@MPI)

what I am thinking about next (MS Thesis)

observe

explain

predict

control

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

ellis INSTITUTE TÜBINGEN

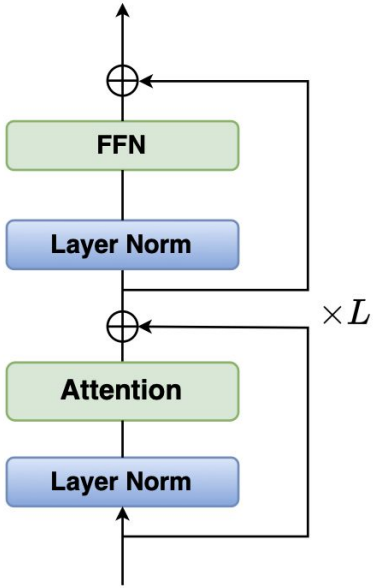
Towards Effective Depth Scaling in LLMs

My Recent Research Line

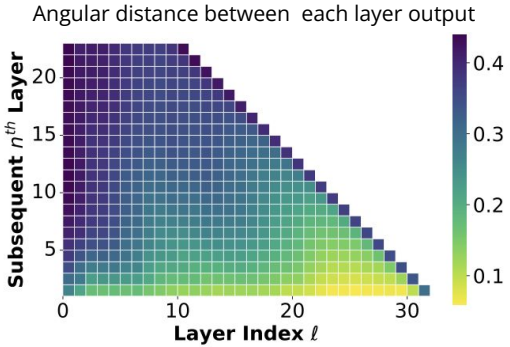
Keyu Wang
Advisor: Dr. Shiwei Liu

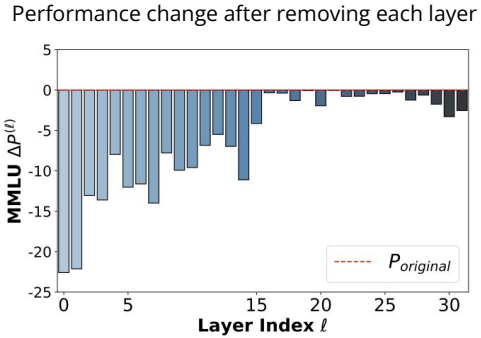
Motivation: depth scaling pays diminishing returns

- Pre-LN dominates in modern LLMs
  — simple yet stable at scale



- But, Pre-LN faces the Curse of Depth
  — deep layers become similar and ineffective



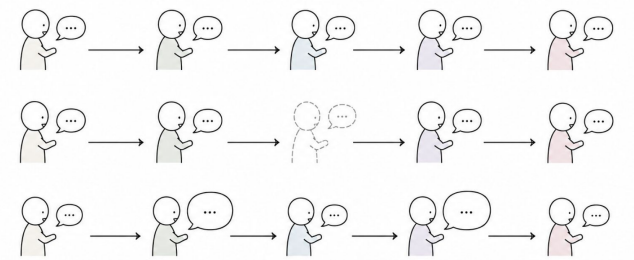


Outline

Proj. 1

Compression & Inference

Wang et al. COLM 2026*
Song, Wang et al. Report 2025



Proj. 2

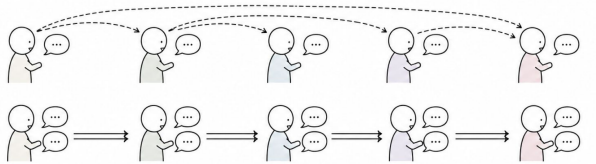
Pre-training Strategy

He, Tu, Wang et al. ICML 2026

Proj. 3

Architecture Choice

Wang et al. Paper in prep.*



Project 1: Conference on Language Modeling - 2026

When Fewer Layers Break More Chains: Layer Pruning Harms Test-Time Scaling in LLMs

Keyu Wang[1,2,3]* **Tian Lyu**[4]* **Guinan Su**[1,2,3] **Lu Yin**[5] **Marco Canini**[4]
Jonas Geiping[1,2,3] **Shiwei Liu**[1,2,3]
[1]Max Planck Institute for Intelligent Systems [2]ELLIS Institute Tübingen
[3]Tübingen AI Center [4]King Abdullah University of Science and Technology
[5]University of Surrey
keyu.wang@tuebingen.mpg.de sliu@tue.ellis.eu

Project 1: Motivation

- The curse of depth seems to be a blessing for layer pruning…

ShortGPT: Layers in Large Language Models are More Redundant Than You Expect

Xin Men* Baichuan Inc. | Mingyu Xu* Baichuan Inc. | Qingyu Zhang* ISCAS | Bingning Wang † Baichuan Inc.

Hongyu Lin ISCAS | Yaojie Lu ISCAS | Xianpei Han ISCAS | Weipeng Chen Baichuan Inc.

PRUNE&COMP: Free Lunch for Layer-Pruned LLMs via Iterative Pruning with Magnitude Compensation

Xinrui Chen[1], Hongxing Zhang[2], Fanyi Zeng[1], Yongxian Wei[1], Yizhi Wang[1], Xitong Ling[1], Guanghao Li[1], Chun Yuan[1]*

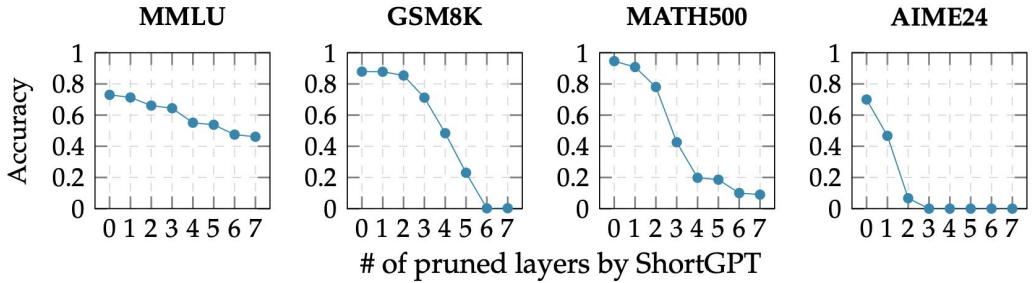
[1] Shenzhen International Graduate School, Tsinghua University

[2] School of Information Science and Technology, Guangdong University of Foreign Studies

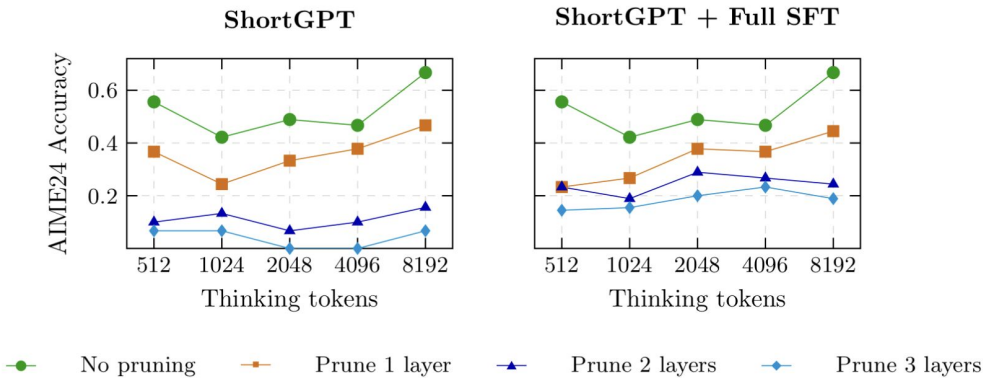
cxr22@tsinghua.org.cn, yuanc@sz.tsinghua.edu.cn

Recent studies reveal that the deeper layers (Transformer blocks) in modern LLMs tend to be less effective than the earlier ones (Gromov et al., 2024; Li et al., 2024b; Men et al., 2024; Yin et al., 2024). On the one hand, this interesting observation provides an effective indicator for LLM compression. For instance, we can compress deeper layers significantly more (Dumitru et al., 2024; Lu et al., 2024; Yin et al., 2024) to achieve high compression ratios. Even more aggressively, entire deep layers can be pruned completely without compromising performance (Muralidharan et al., 2024; Siddiqui et al., 2024).

- But layer pruning collapses complex reasoning!
  Were we deceived by previous benchmarks of multiple-choice questions?



Project 1: Key Findings



- Sequential test-time scaling is highly fragile under layer pruning
- SFT can't lasso the runaway stallion

[AIME24]
Let p be the least prime number for which there exists a positive integer n such that n^4+1 is divisible by p^2. Find the least positive integer m such that m^4+1 is divisible by p^2. **Answer:** 110

Before Finetuning (Pruned Only)
The smallest prime p congruent to 1 modulo 8 is 17... 17 divides 83521-17=83504. Therefore, 17 divides 83521-17, so 17 divides 83521-17. Therefore, 17 divides 83521-17. Therefore, 17 divides 83521-17...

After Finetuning (Pruned + SFT)
Let check small primes. If $p = 2$, $n^4+1 \equiv 2 \pmod 4$. If $p = 3$, $n^4+1 \equiv 2 \pmod 9$. ...If $p = 11$, $n^4 \equiv -1 \equiv 10 \pmod{11}$, if $n = 5$, $n^4+1 = 626 \equiv 10 \pmod{11}$...The least prime p for which there exists n such that $p^2 \mid n^4+1$ is $p = 11$. When $p = 11$, we need $n^4 \equiv 10 \pmod{11}$. If $n = 5$, $n^4 = 625 \equiv 10 \pmod{11}$... Now we need $m^4+1 \equiv 0 \pmod{121}$. If $m = 5$, $5^4+1 = 626 \equiv 120 \pmod{121}$. Final Answer: $\boxed{5}$.

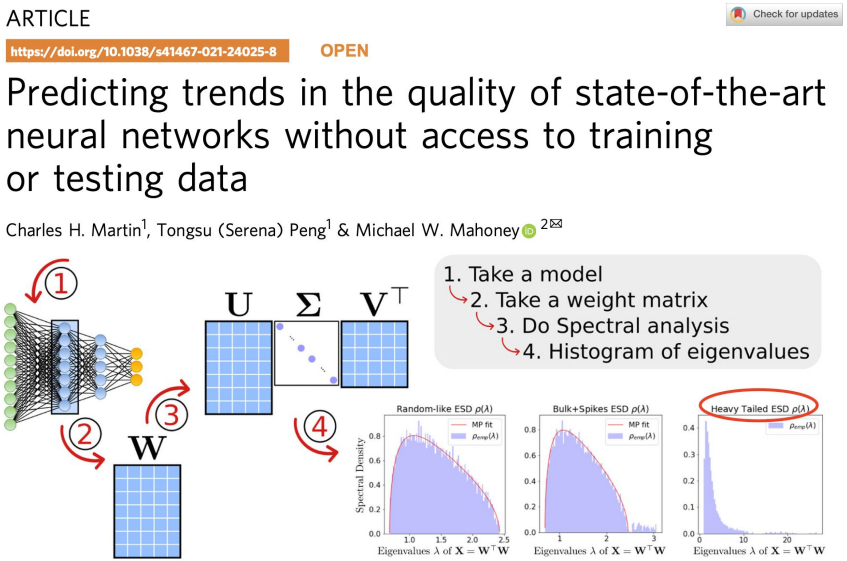
Project 2: International Conference on Machine Learning - 2026

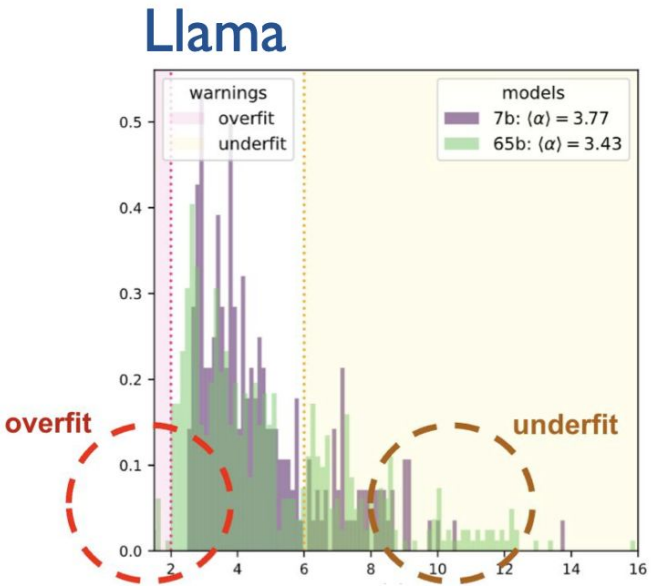
One LR Doesn't Fit All: Heavy-Tail Guided Layerwise Learning Rates for LLMs

Di He [1 2 3] **Songjun Tu** [2 3] **Keyu Wang** [4] **Lu Yin** [5 †] **Shiwei Liu** [6 7 8 †]

Project 2: Motivation

- Transformer layers are naturally trained heterogeneously





$$\rho_{emp}(\lambda) \sim \lambda^{-\alpha}$$

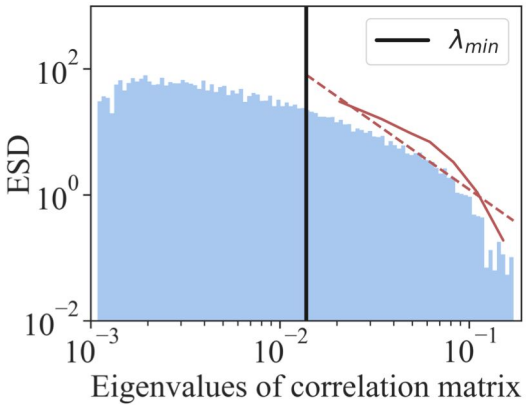
Project 2: Algorithm & Key Results

➢ Measure Heavy-tail structure per-layer

$$p(\lambda) \propto \lambda^{-\alpha}$$

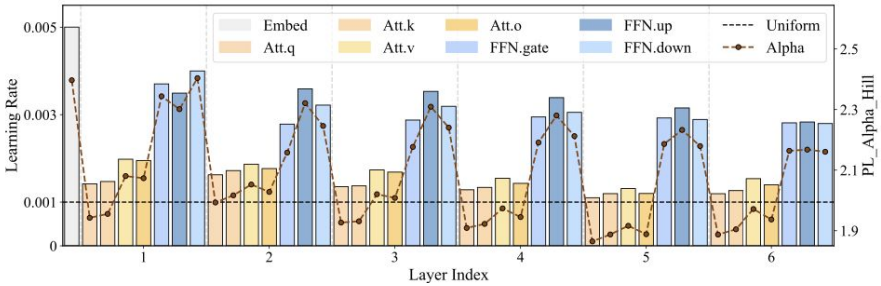
estimate

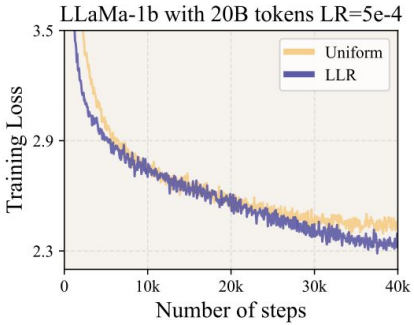
$$\text{PL\_Alpha\_Hill} = 1 + \frac{k}{\sum_{i=1}^{k} ln \frac{\lambda_{n-i+1}}{\lambda_{n-k}}}$$

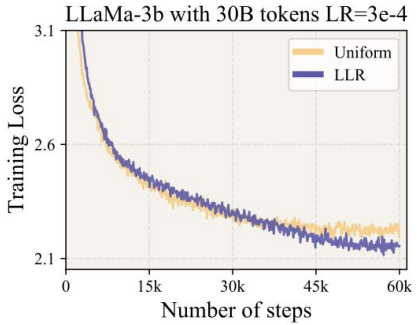


➢ Transfer to layerwise LR

$$\eta_i^T = \eta \left[1 + (s-1) \frac{\alpha_i^T - \alpha_{\min}^T}{\alpha_{\max}^T - \alpha_{\min}^T} \right]$$







Project 3: Paper is in cooking

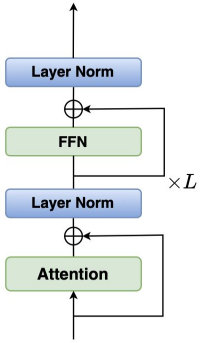
DepthBench: Measuring How Residual Connections Enable More Computational Depth

Keyu Wang[1,2,3*], **Yangyi Huang**[4*], **Jiale Kang**[4], **David González-Martínez**[1,2,3], **Weiyang Liu**[1,4], **Shiwei Liu**[1,2,3]

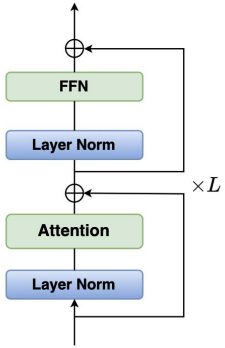
[1]Max Planck Institute for Intelligent Systems, [2]ELLIS Institute Tuebingen,
[3]Tuebingen AI Center, [4]The Chinese University of Hong Kong
[*]Equal contribution. Correspondence to: sliu@tue.ellis.eu.

Project 3: Motivation

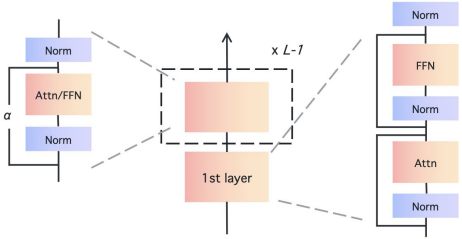
- Which one scales its depth better?



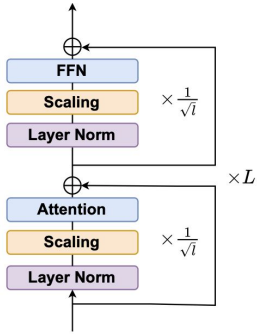
Post-LN



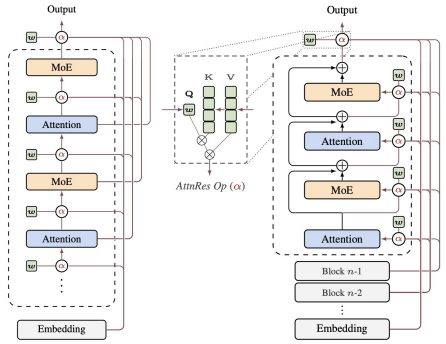
Pre-LN
Dominate in modern LLMs



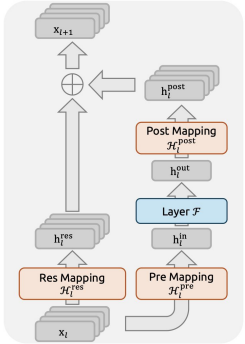
KEEL



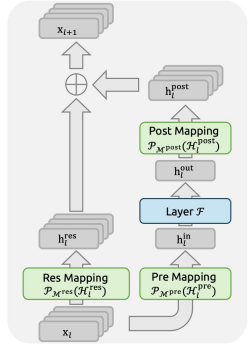
LayerNorm Scaling (LNS)



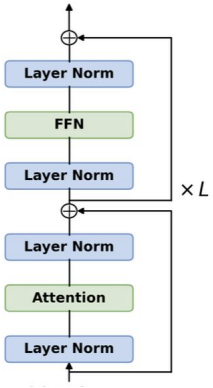
Attention Residuals (AttnRes, left: Full, right: Block)
Applied in Kimi K3



Hyper Connections (HC)
Applied (Modified) in Qwen 3.8

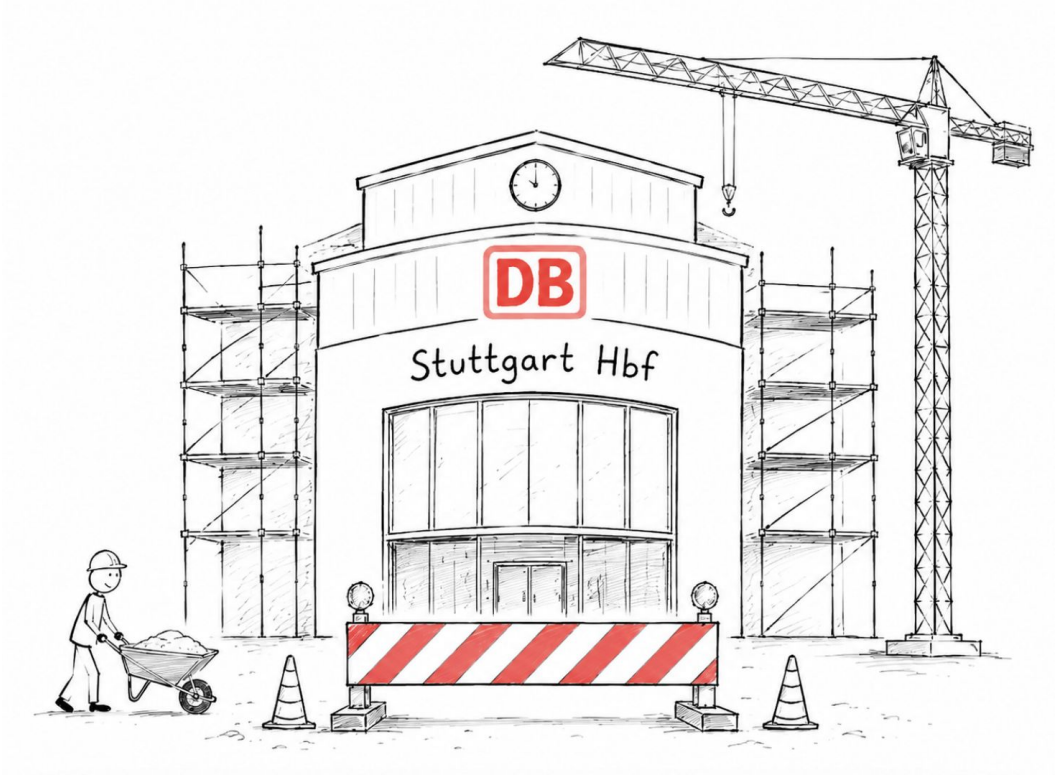


manifold-constrained HC (mHC)
Applied in Deepseek V4



Sandwich-LN
Applied in Gemma 3/4

Project 3: Paper is coming soon!



Towards Effective Depth Scaling in LLMs

2025.06 - 2026.12

Architecture Choice

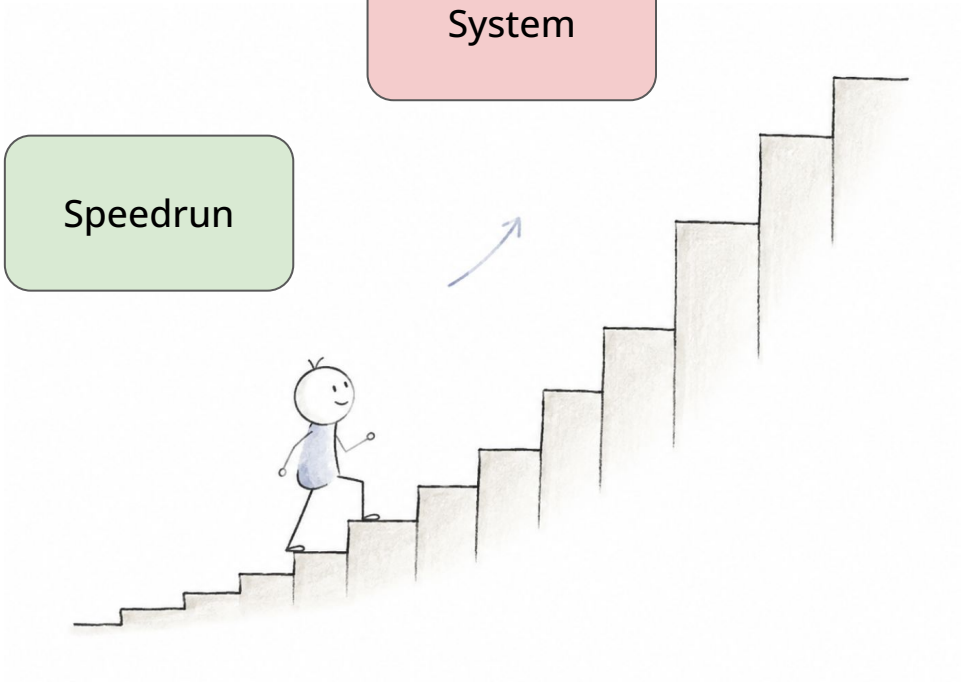
Pre-training Strategy

Compression & Inference

→

Speedrun

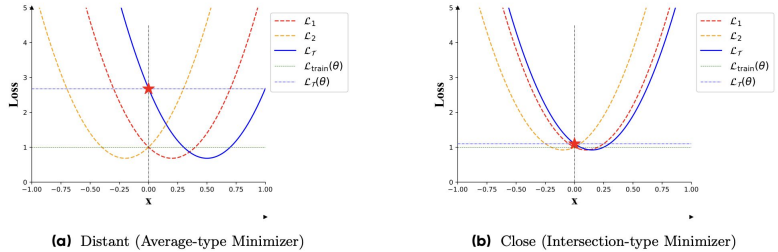
System



Future Vision

BTW I'm exploring my master's thesis direction about model dynamics & continual learning.
Happy to talk.

➢ Curiosity



(a) Distant (Average-type Minimizer) **(b)** Close (Intersection-type Minimizer)

observe

explain

predict

control

Figure from [Nexus Optimizer]

➢ Vision

What we can do to the system that would make the principle reveal whether it is true?
If we understand a dynamical mechanism well enough, we should eventually be able to control it.

01	02	03	04
Mechanics of Learning	Scaling of Learning	Dynamics of Evolution	System Support
What governs the tragectory?	Which principle is universal?	How does new learning interact with existing knowledge?	How should learning be engineered across the computing?

Thanks :)