

Mix-LN: Unleashing the Power of Deeper Layers by Combining Pre-LN and Post-LN

Pengxiang Li^{1*}, Lu Yin^{2*}, Shiwei Liu^{3†}

¹Dalian University of Technology, ²University of Surrey, ³University of Oxford



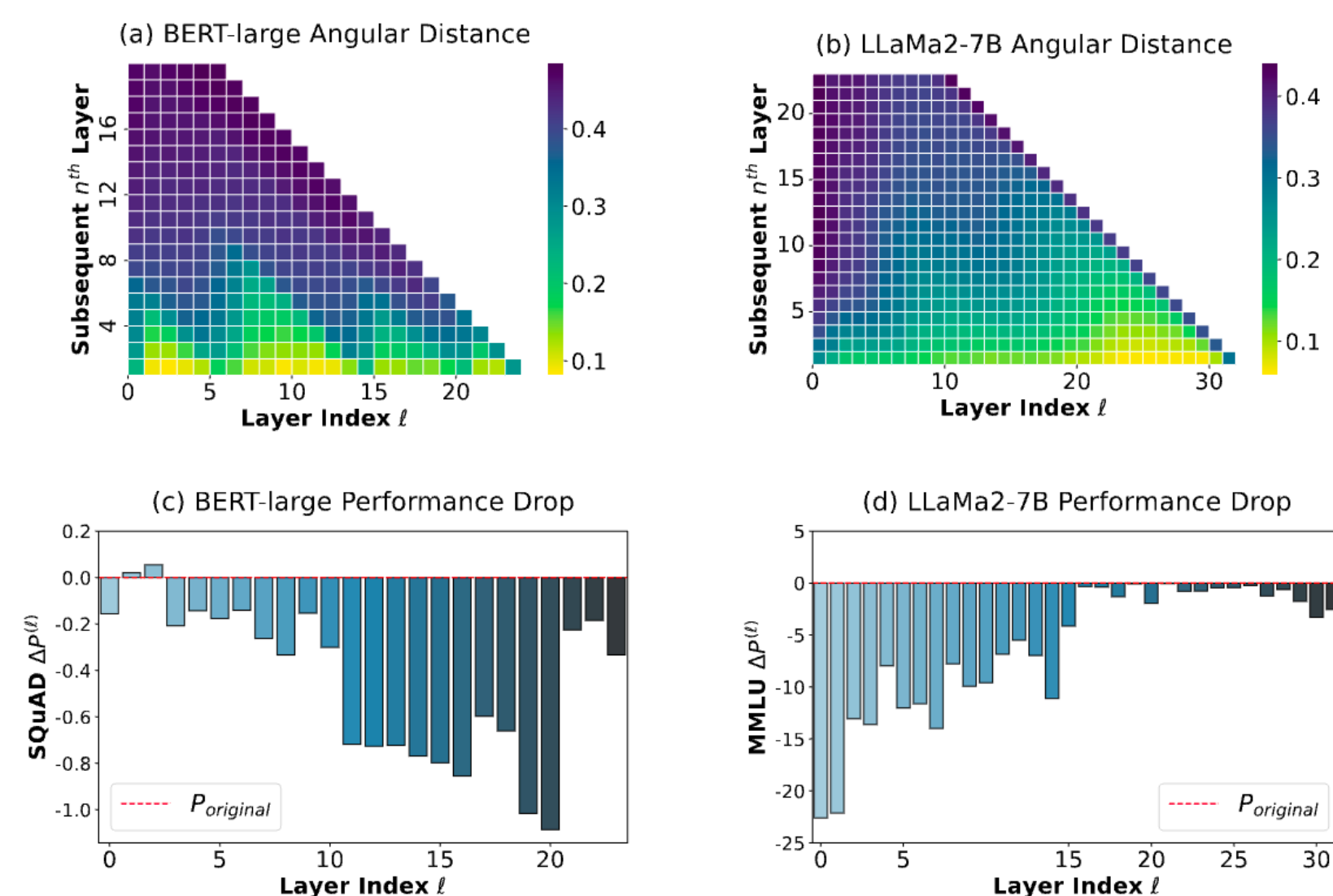
Paper Link 



Layer Pruning Analysis

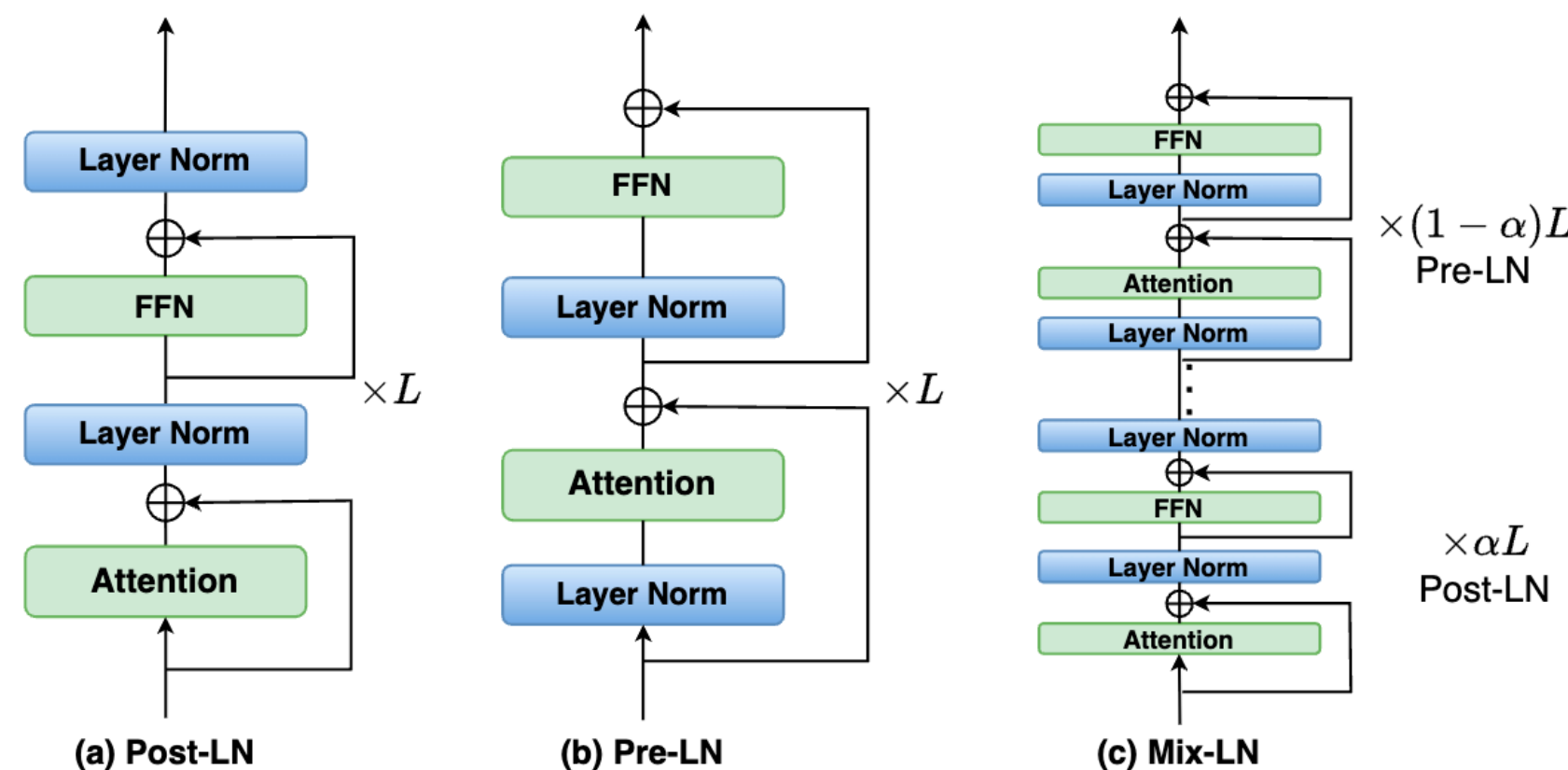
Which Layers Really Matter ? 🤔

A Pruning-Based Reality Check







- (a-b) Angular distance matrices show early layers are redundant
- (c-d) Removing any deep layer hurts much more

Method



We propose Mix-LN to makes deep layers indispensable.

-  Combines Pre-LN and Post-LN within the same model
-  Balances gradients and boosts deep-layer learning
-  Improves representation diversity and training efficiency
-  Simple, effective, and easy to integrate

Experiments

Perplexity (↓) comparison of various layer normalization methods

Training Tokens	LLaMA-71M 1.1B	LLaMA-130M 2.2B	LLaMA-250M 3.9B	LLaMA-1B 5B
Post-LN	35.18	26.95	1409.09	1411.54
DeepNorm	34.87	27.17	22.77	1410.94
Pre-LN	34.77	26.78	21.92	18.65
Mix-LN	33.12	26.07	21.39	18.18

Fine-tuning performance (↑) of LLaMA with various normalizations

Method	MMLU	BoolQ	ARC-e	PIQA	Hellaswag	OBQA	Winogrande	Avg.
LLaMA-250M								
Post-LN	22.95	37.83	26.94	52.72	26.17	11.60	49.56	32.54
DeepNorm	23.60	37.86	36.62	61.10	25.69	15.00	49.57	35.63
Pre-LN	24.93	38.35	40.15	63.55	26.34	16.20	49.01	36.93
Mix-LN	26.53	56.12	41.68	66.34	30.16	18.00	50.56	41.34
LLaMA-1B								
Post-LN	22.95	37.82	25.08	49.51	25.04	13.80	49.57	31.96
DeepNorm	23.35	37.83	27.06	52.94	26.19	11.80	49.49	32.67
Pre-LN	26.54	62.20	45.70	67.79	30.96	17.40	50.51	43.01
Mix-LN	27.99	61.93	48.11	68.50	31.35	18.80	55.93	44.66

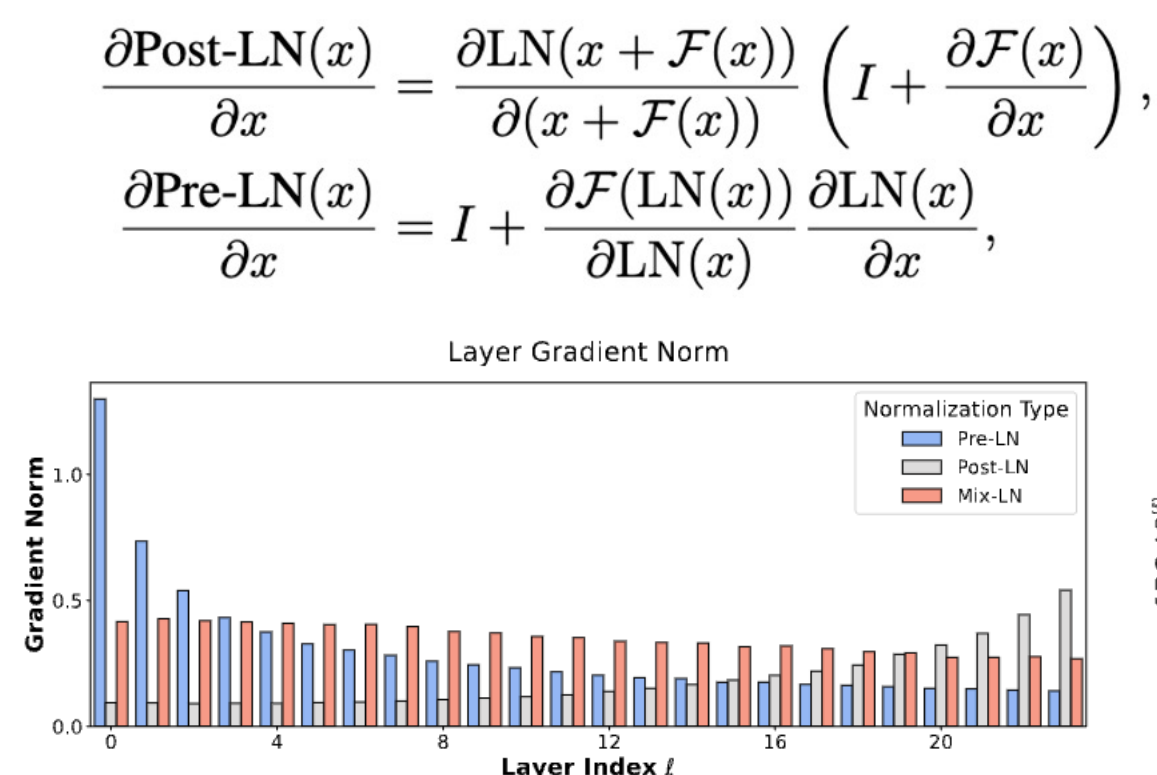
Accuracy (↑) comparison of Pre-LN and Mix-LN on ViT models

Model	ViT-Tiny	ViT-Small
Pre-LN	67.30	75.99
Mix-LN	67.34	76.40

Why Does It Work?

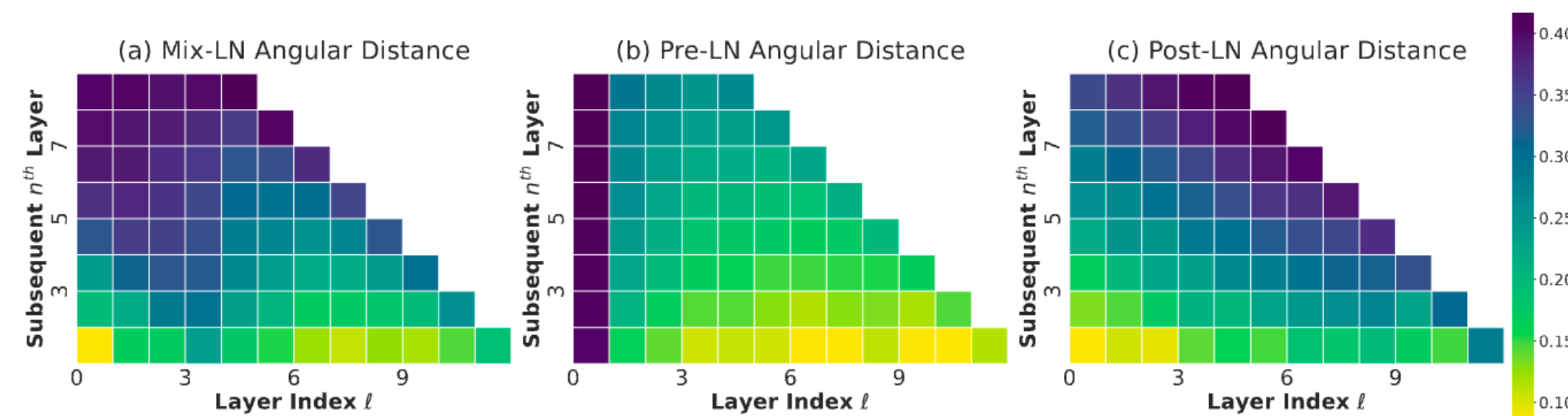
Vanishing derivative 

$$\frac{\partial \text{LN}(x)}{\partial x} = \frac{1}{\sigma_x} I \approx 0.$$



(a) Layer gradient norm of LLaMA-250M with various normalization techniques.

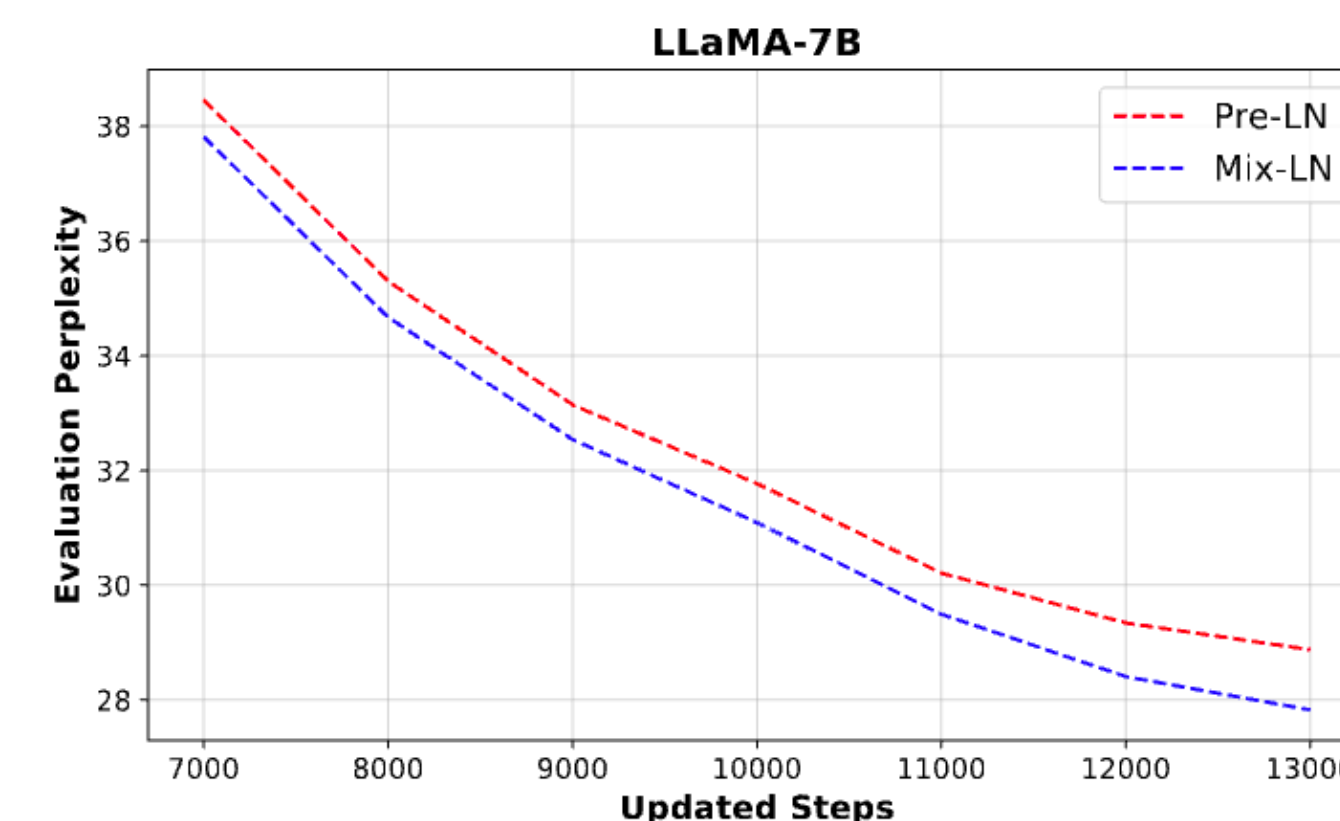
(b) Performance drop comparison of LLaMA-130M across layers for Pre-LN, Post-LN, and Mix-LN.



Angular distance from initial layer ℓ (x-axis) with block size n (y-axis) of LLaMA-130M

 **Mix-LN keeps deep-layer signal alive; vanilla Pre-/Post-LN waste it.**

Training curve (eval perplexity) of Mix-LN and Pre-LN with LLaMa-7B



Comparison of gradient norms and performance drops.