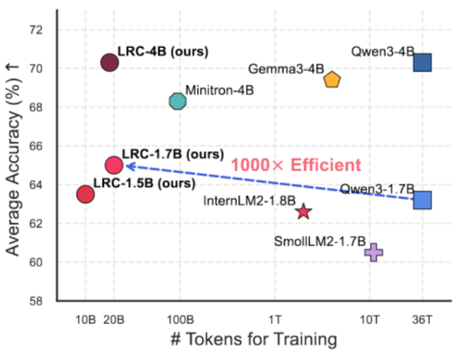
A Token is Worth over 1,000 Tokens: Efficient Knowledge Distillation through Low-Rank Clone

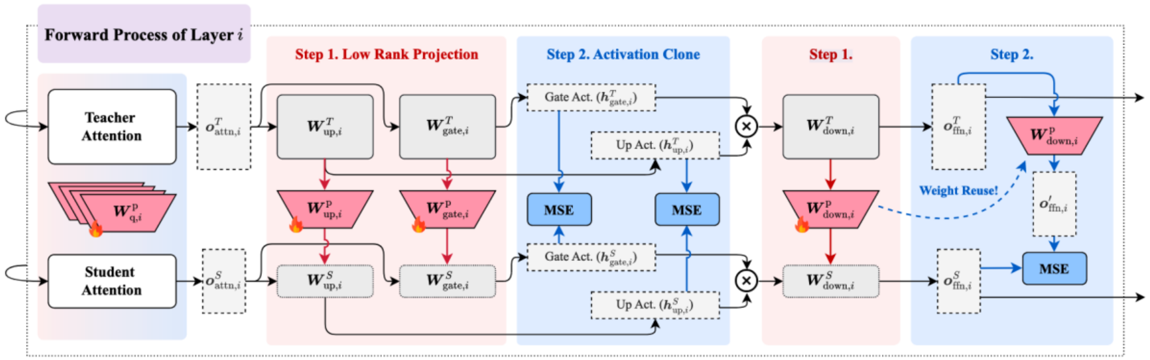
Spotlight

Jitai Hao, Qiang Huang†, Hao Liu, Xinyan Xiao, Zhaochun Ren, Jun Yu†

HITSZ, Baidu, Leiden, Pengcheng Lab



LRC surpasses SOTA models trained on trillions of tokens--while using only **20B** tokens **FROM SCRATCH**, achieving over **1,000x** training efficiency.

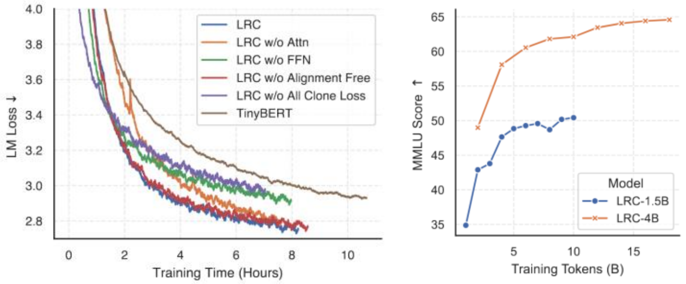


Generate the Student using Low-Rank Projection Instead of Training One!

Attention and normalization modules are omitted. LRC involves two main steps: (1) **Low-Rank Projection**: applying low-rank projection matrices to compress the teacher's weights into a lower-dimensional space, which are then assigned to the student. (2) **Activation Clone**, executing standard forward passes in both models to collect intermediate activations, which are aligned using Mean Squared Error (MSE) loss.

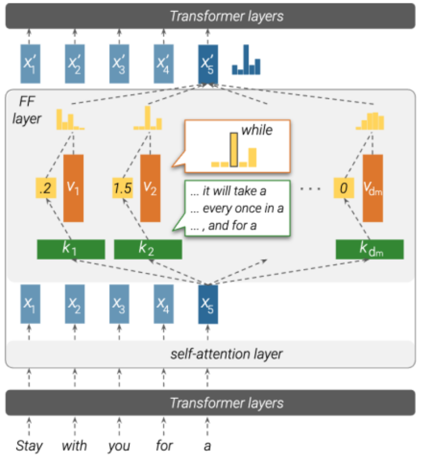
Model	Gemma3-4B	Minitron-4B	Qwen3-4B	**LRC-4B**	**LRC-2.7B-B**	Sheared-Llama-2.7B-B
Teacher	–	Nemotron4-15B	–	Qwen2.5-7B	Llama2-7B	Llama2-7B
# Tokens	4T	94B	36T	**18B**	**10B**	50B
Dataset	N/A	N/A	N/A	Mixed-2.0	Redpajama	Redpajama
ARC-E	82.53	79.59	80.47	78.37	58.59	67.30
ARC-C	57.08	54.35	53.58	52.47	29.61	33.58
LogiQA	33.03	30.26	33.64	34.10	29.03	28.26
CSQA	69.37	71.09	75.76	79.28	36.36	18.92
PIQA	76.44	77.64	75.08	76.82	66.97	76.17
WinoG	69.38	65.93	65.27	67.72	62.43	65.04
BoolQ	83.94	82.60	84.95	84.50	74.31	65.99
SciQ	95.50	96.60	95.50	95.00	85.50	91.10
MMLU	57.58	56.77	68.38	64.41	31.20	26.56
Avg. ↑	69.43	68.31	70.29	**70.30**	**52.67**	52.55

Main Perf. ↑



Ablation ↑

Insights: Clone the Key-Value Knowledge in FFN.



Algorithm 1: Overall Procedure of LRC

Input: Input token sequence $\mathcal{T}$; number of layers l; RMSNorm constant ϵ; teacher's weights $\{\boldsymbol{W}^{\mathrm{T}}_{m,i}\}, \boldsymbol{W}^{\mathrm{T}}_{\mathrm{emb}}, \boldsymbol{W}^{\mathrm{T}}_{\mathrm{lm}}$; low-rank projection matrices $\{\boldsymbol{W}^{\mathrm{p}}_{m,i}\}, \boldsymbol{W}^{\mathrm{p}}_{\mathrm{emb}}, \boldsymbol{W}^{\mathrm{p}}_{\mathrm{lm}}$;

Output: Clone loss $\mathcal{L}_{\mathrm{clone}}$;

▷ Step 1: Low-Rank Projection

for $i = 1$ **to** l **do**

  foreach $m \in \{\mathrm{q,k,v,o,up,gate,down}\}$ **do**

    $\boldsymbol{W}^{\mathrm{S}}_{m,i} \leftarrow \boldsymbol{W}^{\mathrm{T}}_{m,i}\boldsymbol{W}^{\mathrm{p}}_{m,i}$; ▷ Generate student weights

$\boldsymbol{W}^{\mathrm{S}}_{\mathrm{emb}} \leftarrow \boldsymbol{W}^{\mathrm{T}}_{\mathrm{emb}}\boldsymbol{W}^{\mathrm{p}}_{\mathrm{emb}}$; $\boldsymbol{W}^{\mathrm{S}}_{\mathrm{lm}} \leftarrow \boldsymbol{W}^{\mathrm{T}}_{\mathrm{lm}}\boldsymbol{W}^{\mathrm{p}}_{\mathrm{lm}}$;

▷ Step 2: Activation Clone

$\mathcal{L}_{\mathrm{clone}} \leftarrow 0$;

$\boldsymbol{h}^{\mathrm{T}}, \boldsymbol{o}^{\mathrm{T}}_{\mathrm{attn}}, \boldsymbol{o}^{\mathrm{T}}_{\mathrm{ffn}} \leftarrow \mathrm{Forward}(\mathcal{T}, l, \epsilon, \{\boldsymbol{W}^{\mathrm{T}}_{m,i}\}, \boldsymbol{W}^{\mathrm{T}}_{\mathrm{emb}}, \boldsymbol{W}^{\mathrm{T}}_{\mathrm{lm}})$; ▷ Get teacher act. dict.

$\boldsymbol{h}^{\mathrm{S}}, \boldsymbol{o}^{\mathrm{S}}_{\mathrm{attn}}, \boldsymbol{o}^{\mathrm{S}}_{\mathrm{ffn}} \leftarrow \mathrm{Forward}(\mathcal{T}, l, \epsilon, \{\boldsymbol{W}^{\mathrm{S}}_{m,i}\}, \boldsymbol{W}^{\mathrm{S}}_{\mathrm{emb}}, \boldsymbol{W}^{\mathrm{S}}_{\mathrm{lm}})$; ▷ Get student act. dict.

for $i = 1$ **to** l **do**

  foreach $m \in \{\mathrm{q,k,v,gate,up}\}$ **do** ▷ Compute clone loss of interm. states

    $\mathcal{L}_{\mathrm{clone}} \leftarrow \mathcal{L}_{\mathrm{clone}} + \mathcal{E}(\boldsymbol{h}^{\mathrm{S}}_{m,i}, \boldsymbol{h}^{\mathrm{T}}_{m,i})$;

  $\mathcal{L}_{\mathrm{clone}} \leftarrow \mathcal{L}_{\mathrm{clone}} + \mathcal{E}(\boldsymbol{o}^{\mathrm{S}}_{\mathrm{attn},i}, \boldsymbol{o}^{\mathrm{T}}_{\mathrm{attn},i}\boldsymbol{W}^{\mathrm{p}}_{\mathrm{o},i}) + \mathcal{E}(\boldsymbol{o}^{\mathrm{S}}_{\mathrm{ffn},i}, \boldsymbol{o}^{\mathrm{T}}_{\mathrm{ffn},i}\boldsymbol{W}^{\mathrm{p}}_{\mathrm{down},i})$;

return $\mathcal{L}_{\mathrm{clone}}$;

Score Type	Teacher	Student
Original Score	0.85	0.48
Important Neurons Masked	0.62 (-27%)	0.33 (-31%)
Random Neurons Masked	0.85	0.49

Student FFN clones Teacher's ↑

Method	# Tokens/Sec
LRC	84K
Sheared Llama (Prune)	30K
Ordinary Training	146K
TinyBERT	65K

Training Efficiency ←

Model	LRC-1.5B		
Teacher	Llama3-3B		
# Tokens	20B	10B	10B
Dataset	Mixed-2.0	Mixed-1.0	Mixed-1.1
Avg. ↑	62.12	61.35	62.48

Paper **Github**