TECHNISCHE UNIVERSITÄT DARMSTADT

BASIS SHARING: CROSS-LAYER PARAMETER SHARING FOR LARGE LANGUAGE MODEL COMPRESSION

Jingcun Wang[1], Yu-Guang Chen[2], Ing-Chao Lin[3], Bing Li[4], Grace Li Zhang[1]

[1]Technical University of Darmstadt

[2] National Central University

[3] National Cheng Kung University

[4] University of Siegen

MOTIVATION

- Scaling LLMs with More Parameters
  - Example: Llama-3.1-405B
- Reducing the Number of Parameters
  - Techniques: Pruning, Parameter Sharing, Matrix Decomposition
- Avoiding Training from Scratch
  - Not affordable

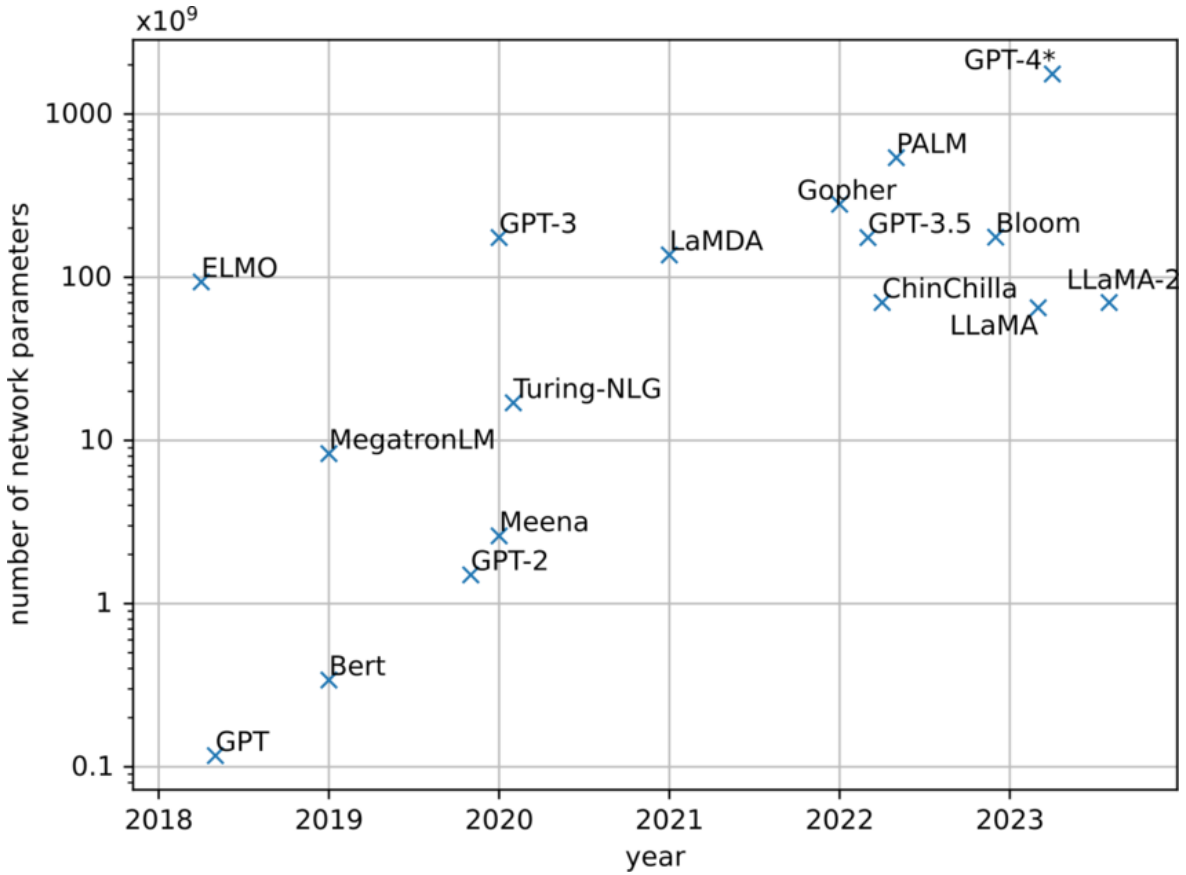
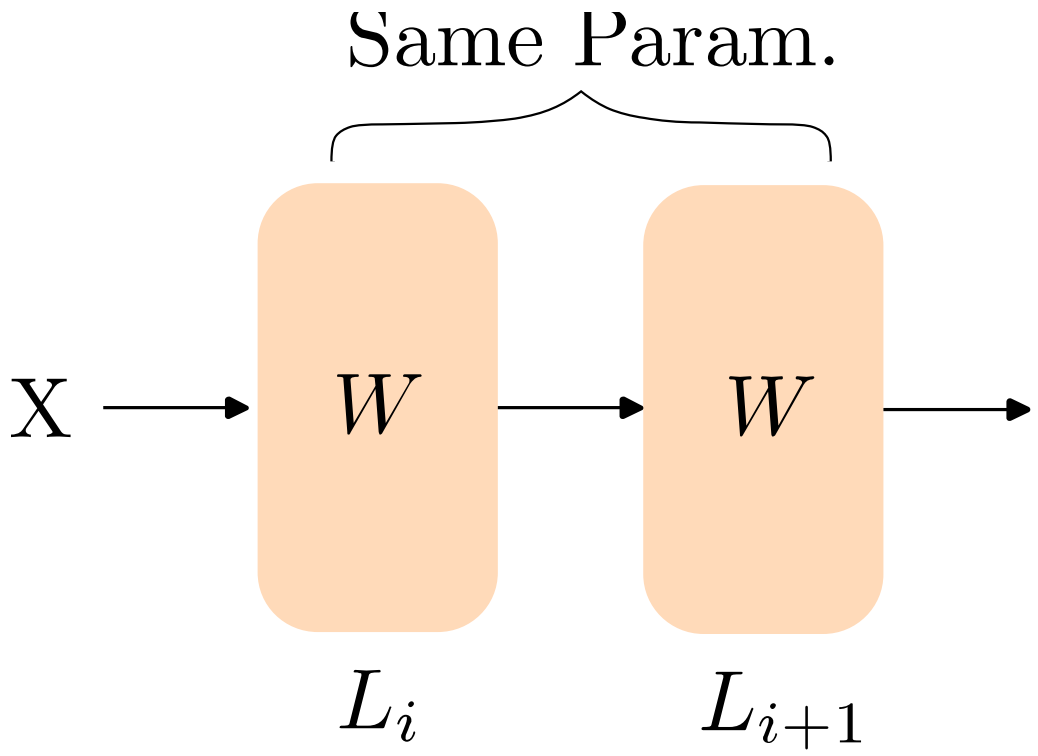


Image source: Gerstmayr, J., Manzl, P. & Pieber, M., 2024

PARAMETER SHARING

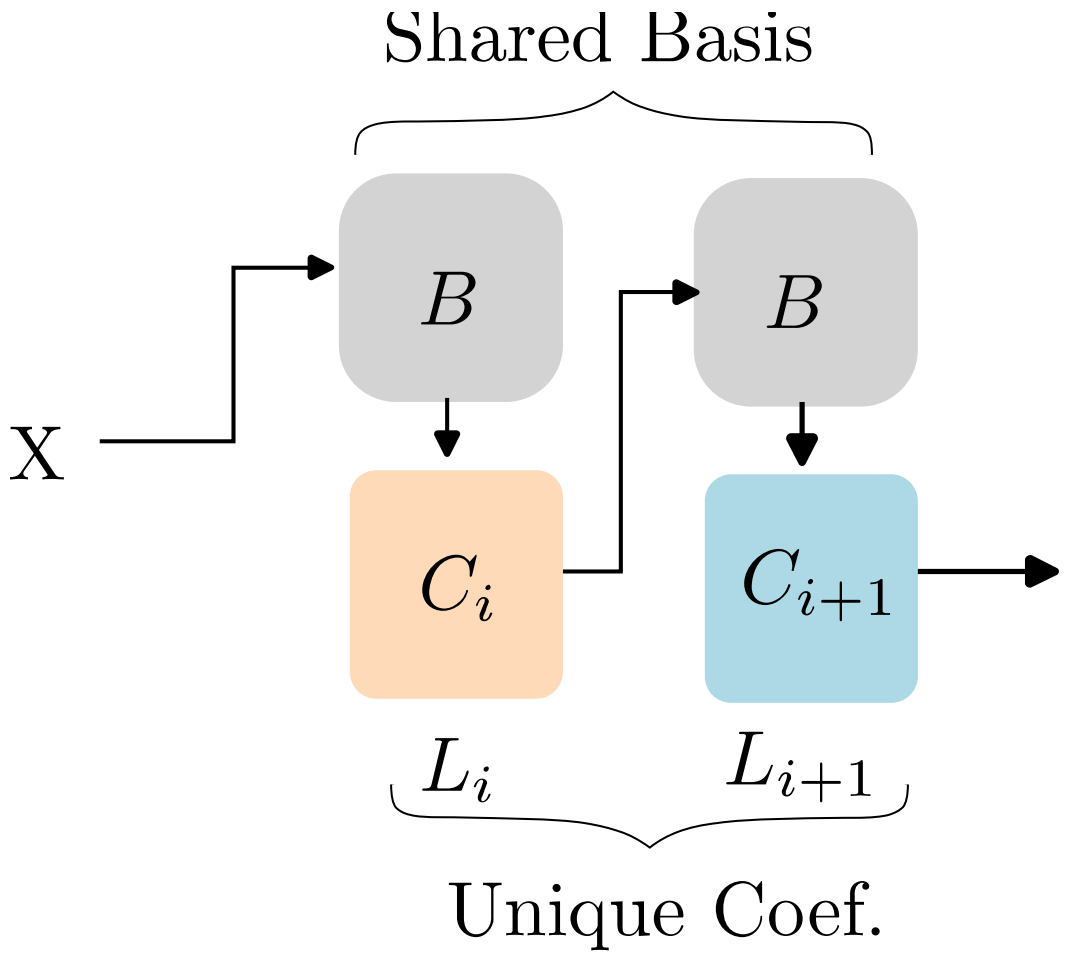
- Parameter Sharing → Reduces Count but Limits Flexibility
- Less Flexibility → Potential Performance Drop
- Performance Drop → Model Retraining



TECHNISCHE UNIVERSITÄT DARMSTADT

BASIS SHARING

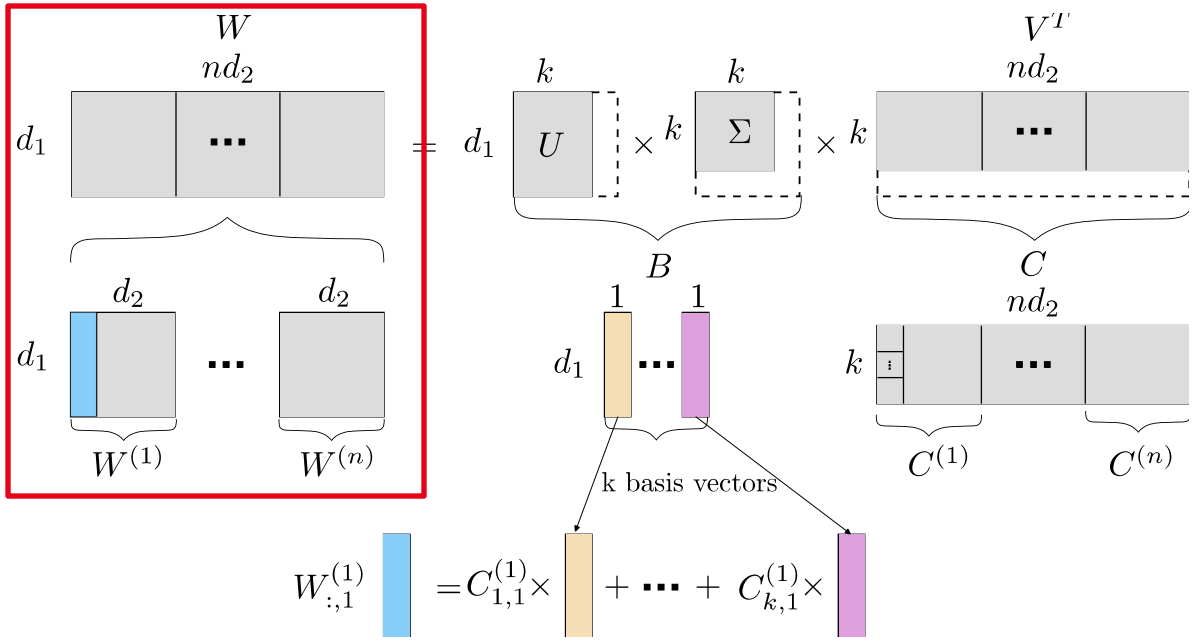
- Identify Commonalities Across Layers
- Share the Commonalities (Basis), Keep Unique Parts (Coefficients)
- Reduce Parameters without Losing Flexibility & Performance
- No Need for Retraining



TECHNISCHE UNIVERSITÄT DARMSTADT

HOW TO

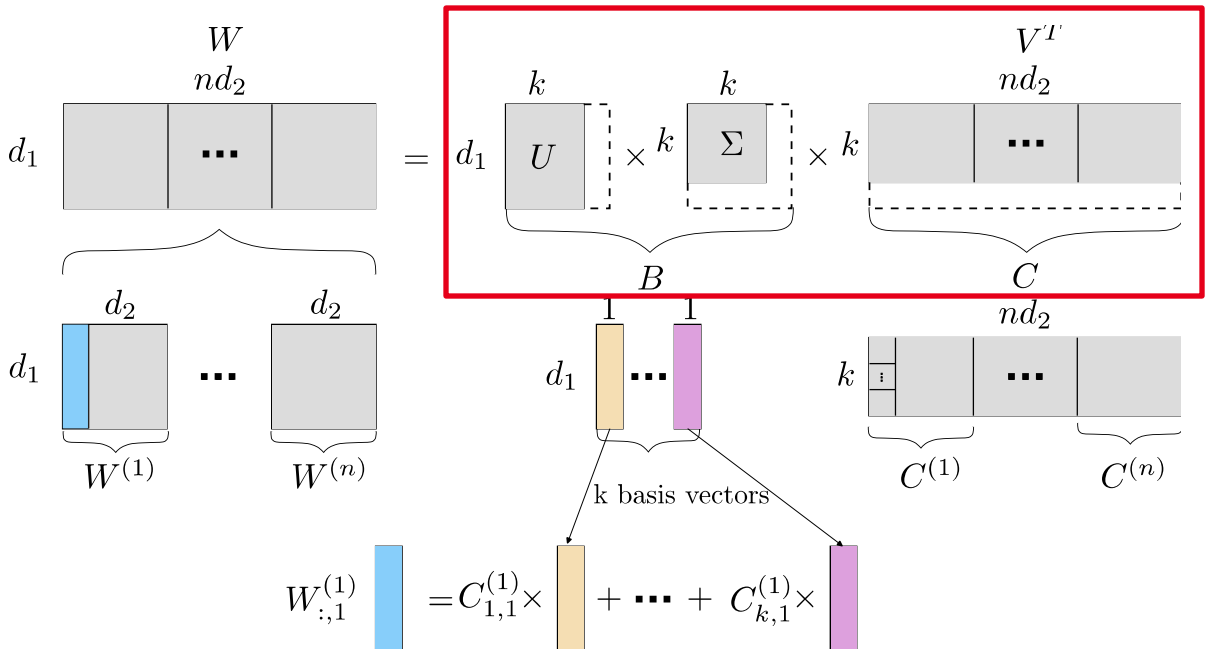
- Concatenate all matrices to be shared.



TECHNISCHE UNIVERSITÄT DARMSTADT

HOW TO

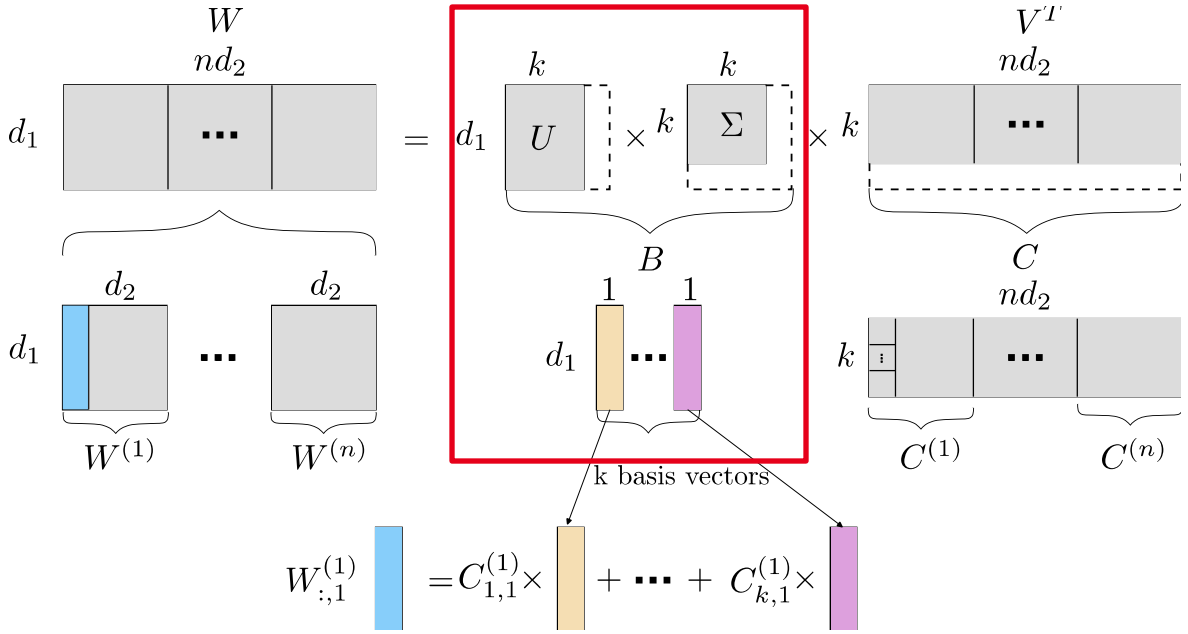
- Concatenate all matrices to be shared.
- Do SVD to delete redundant component.



TECHNISCHE UNIVERSITÄT DARMSTADT

HOW TO

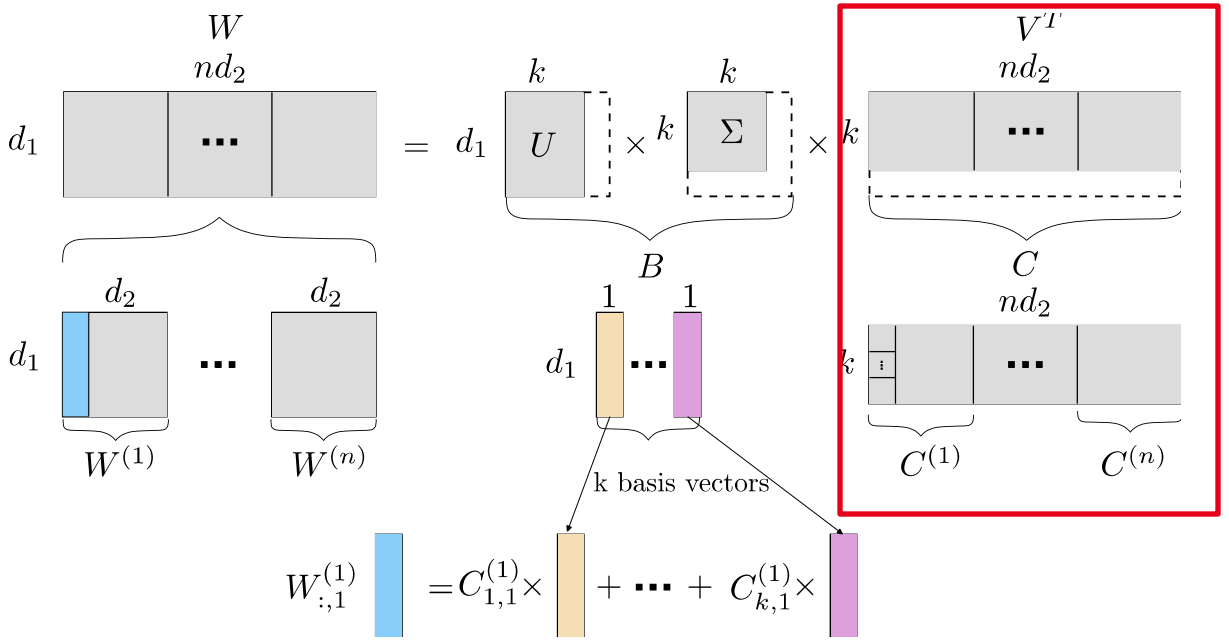
- Concatenate all matrices to be shared.
- Do SVD to delete redundant component.
- This part is shared between different matrices.



TECHNISCHE UNIVERSITÄT DARMSTADT

HOW TO

- Concatenate all matrices to be shared.
- Do SVD to delete redundant component.
- This part is shared between different matrices.
- This part is hold by each matrix to keep its diversity.



TECHNISCHE UNIVERSITÄT DARMSTADT

MATRIX CALIBRATION

- Take outliers in the input data into account.
- Scale weight matrices based on the presence of outliers in the input data before concatenation.
- Minimize $||X(W - W')||_F$ instead of $||W - W'||_F$ to improve accuracy and account for data-dependent variations.

ΔW_1

1.0	0.5	0.4
0.2	0.6	0.8
1.1	0.7	0.6

X

0.01
10
0.0

$$||\Delta W_1||_F = 2.1$$
$$||X\Delta W_1||_F = 10.8$$

ΔW_2

1.0	0.01	1.0
0.4	0.02	2.0
5.0	0.03	8.0

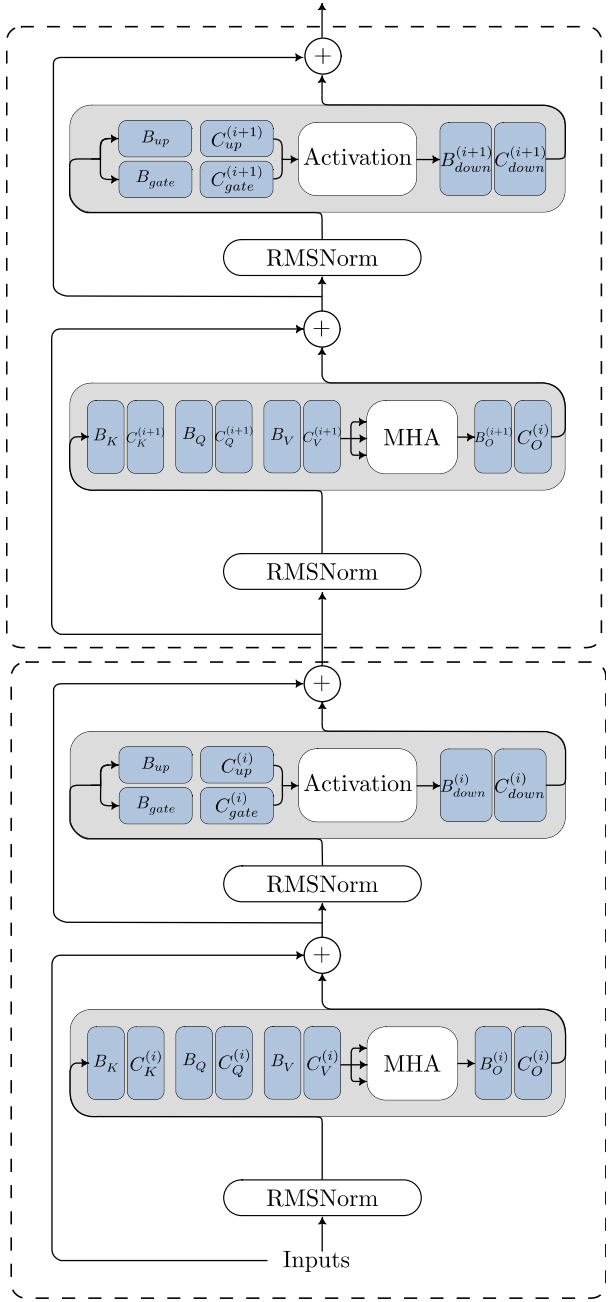
X

0.01
10
0.0

$$||\Delta W_2||_F = 9.8$$
$$||X\Delta W_2||_F = 0.8$$

MODEL OVERVIEW

- Two consecutive layers in a LLaMA model are considered.
- The weight matrices W_Q, W_K, W_V, $\mathrm{W_{up}}$, $\mathrm{W_{gate}}$, $\mathrm{W_{down}}$ share a common basis, and keep own coefficient at the same time.
- W_O, W_{down} are compressed using SVD only, as sharing them leads to higher reconstruction errors.

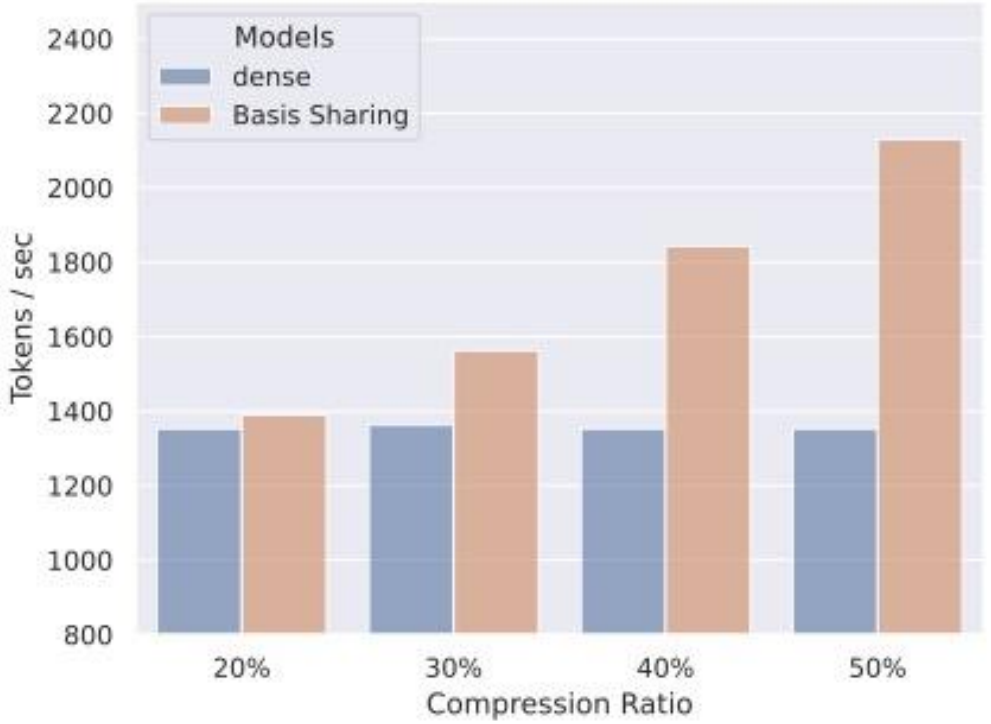


TECHNISCHE UNIVERSITÄT DARMSTADT

RESULTS

Ratio	Method	WikiText-2↓	PTB↓	C4↓	Openb.	ARC_e	WinoG.	HellaS.	ARC_c	PIQA	MathQA	Average↑
0%	Original	5.68	8.35	7.34	0.28	0.67	0.67	0.56	0.38	0.78	0.27	0.52
20%	SVD	20061	20306	18800	0.14	0.27	0.51	0.26	0.21	0.53	0.21	0.31
	FWSVD	1727	2152	1511	0.15	0.31	0.50	0.26	0.23	0.56	0.21	0.32
	ASVD	11.14	16.55	15.93	0.25	0.53	0.64	0.41	0.27	0.68	0.24	0.43
	SVD-LLM	7.94	18.05	15.93	0.22	0.58	0.63	0.43	0.29	0.69	0.24	0.44
	Basis Sharing	**7.74**	**17.35**	**15.03**	**0.28**	**0.66**	**0.66**	**0.46**	**0.36**	**0.71**	**0.25**	**0.48**
30%	SVD	13103	17210	20871	0.13	0.26	0.51	0.26	0.21	0.54	0.22	0.30
	FWSVD	20127	11058	7240	0.17	0.26	0.49	0.26	0.22	0.51	0.19	0.30
	ASVD	51	70	41	0.18	0.43	0.53	0.37	0.25	0.65	0.21	0.38
	SVD-LLM	9.56	29.44	25.11	0.20	0.48	0.59	0.40	0.26	0.65	0.22	0.40
	Basis Sharing	**9.25**	**29.12**	**22.46**	**0.27**	**0.63**	**0.63**	**0.40**	**0.30**	**0.68**	**0.24**	**0.45**
40%	SVD	52489	59977	47774	0.15	0.26	0.52	0.26	0.22	0.53	0.20	0.30
	FWSVD	18156	20990	12847	0.16	0.26	0.51	0.26	0.22	0.53	0.21	0.30
	ASVD	1407	3292	1109	0.13	0.28	0.48	0.26	0.22	0.55	0.19	0.30
	SVD-LLM	13.11	63.75	49.83	0.19	0.42	0.58	0.33	0.25	0.60	0.21	0.37
	Basis Sharing	**12.39**	**55.78**	**41.28**	**0.22**	**0.52**	**0.61**	**0.35**	**0.27**	**0.62**	**0.23**	**0.40**
50%	SVD	131715	87227	79815	0.16	0.26	0.50	0.26	0.23	0.52	0.19	0.30
	FWSVD	24391	28321	23104	0.12	0.26	0.50	0.26	0.23	0.53	0.20	0.30
	ASVD	15358	47690	27925	0.12	0.26	0.51	0.26	0.22	0.52	0.19	0.30
	SVD-LLM	23.97	150.58	118.57	0.16	0.33	0.54	0.29	0.23	0.56	0.21	0.33
	Basis Sharing	**19.99**	**126.35**	**88.44**	**0.18**	**0.42**	**0.57**	**0.31**	**0.23**	**0.58**	**0.22**	**0.36**

Basis Sharing outperforms other decomposition methods at the same compression ratio.



Basis Sharing can improve the throughput of a model without any specific optimization.