Global PCA for SliceGPT: A Micro-Study on Simplifying Layer-wise Rotations

Danish Ahmed
GIKI
danishahmed3232@gmail.com

***Abstract*—This study focuses on the trade-offs between per-layer and global PCA rotation strategies in SliceGPT model compression. Through comprehensive experiments on OPT models (125M–6.7B) and LLaMA-2-7B across sparsity levels of 10–30%, this study demonstrate that global PCA achieves 94–98% shortcut memory savings with only 0.02–5.3 perplexity increase, while providing approximately 2× compression speedup and 5–12% inference acceleration. The study further analyzes the impact of calibration sample count and definitively characterize the catastrophic failure modes of intermediate K-block hybrid strategies across all configurations.**

I. Introduction

SliceGPT [1] employs orthogonal transformations to enable structured pruning of transformer models. The original per-layer PCA approach computes distinct rotation matrices Q_ℓ for each layer ℓ, which requires shortcut matrices to align residual connections. These shortcut matrices represent a memory overhead and computational cost during inference.

This study investigates whether a single global rotation matrix Q applied across all layers can simplify this architecture while maintaining acceptable model quality. Global PCA offers a theoretical advantage: since the same rotation is applied everywhere, shortcut matrices become identity ($Q^\top Q = I$) and can be eliminated entirely.

Contributions:

- Comprehensive evaluation of per-layer vs global PCA across 5 model scales (OPT + LLaMA-2) and 4 sparsity levels
- Quantification of memory savings (94–98%), inference speedup (5–12%), and compression speedup ($\sim$2×)
- Analysis of calibration sample requirements for both methods
- Definitive characterization of K-block hybrid failure modes

II. Experimental Setup

- **Models:** OPT-125M, OPT-1.3B, OPT-2.7B, OPT-6.7B from the OPT family, and LLaMA-2-7B.
- **Sparsity Levels:** 10%, 20%, 25%, and 30% embedding dimension reduction.
- **Dataset:** WikiText-2 for calibration and perplexity evaluation.
- **Calibration:** 128 samples with 2048 token context length (unless otherwise noted).
- **Hardware:** Single NVIDIA RTX4090 GPU for all experiments.

III. Results

A. Perplexity Comparison

Table I presents the core perplexity comparison between per-layer and global PCA across all configurations. It report both absolute perplexity difference (ΔPPL) and relative increase (Δ%) to provide complete context.

TABLE I
Perplexity Comparison: Per-Layer vs Global PCA Across All Sparsities

Model	Sparsity	Per-Layer	Global	ΔPPL	Δ%
OPT-125M	10%	**29.33**	30.27	+0.94	+3.2%
	20%	**34.52**	37.59	+3.07	+8.9%
	25%	**38.13**	43.01	+4.88	+12.8%
	30%	**44.61**	52.37	+7.76	+17.4%
OPT-1.3B	10%	**15.14**	15.14	+0.00	+0.0%
	20%	**16.58**	16.95	+0.37	+2.2%
	25%	**17.77**	18.71	+0.94	+5.3%
	30%	**19.60**	21.62	+2.02	+10.3%
OPT-2.7B	10%	**12.82**	12.84	+0.02	+0.2%
	20%	**13.90**	14.08	+0.18	+1.3%
	25%	**14.84**	15.28	+0.44	+3.0%
	30%	**16.30**	17.37	+1.07	+6.6%
OPT-6.7B	10%	**11.11**	11.20	+0.09	+0.8%
	20%	**11.89**	12.10	+0.21	+1.8%
	25%	**12.50**	12.79	+0.29	+2.3%
	30%	**13.36**	13.84	+0.48	+3.6%
LLaMA-2-7B	10%	**5.96**	6.16	+0.20	+3.5%
	20%	**6.86**	7.84	+0.98	+14.3%
	25%	**7.56**	9.79	+2.23	+29.5%
	30%	**8.63**	13.92	+5.29	+61.3%

Key insight: For 7B-class models, global PCA works well at low sparsities (10–20%) but degrades significantly at higher sparsities. LLaMA-2-7B shows larger gaps than OPT-6.7B, showing architecture-dependent sensitivity.

B. Memory Efficiency Analysis

Table II presents shortcut matrix memory consumption across all configurations. Global PCA eliminates 94–98% of shortcut storage because shortcut matrices become identity when using a single rotation matrix.

C. Inference Throughput Analysis

Table III compares prefill and decode throughput across all configurations. Global PCA achieves consistent speedups by eliminating shortcut matrix multiplications during inference.

TABLE II
SHORTCUT MEMORY: PER-LAYER VS GLOBAL PCA (IN MB)

Model	Sparsity	Per-Layer	Global	Saved	Reduction
OPT-125M	10%	21.77	1.01	20.76	**95.4%**
	20%	17.11	0.89	16.22	**94.8%**
	25%	15.40	0.84	14.56	**94.5%**
	30%	13.39	0.79	12.60	**94.1%**
OPT-1.3B	10%	310.69	7.19	303.50	**97.7%**
	20%	245.14	6.38	238.76	**97.4%**
	25%	217.50	6.00	211.50	**97.2%**
	30%	189.42	5.59	183.83	**97.0%**
OPT-2.7B	10%	649.13	11.25	637.88	**98.3%**
	20%	514.00	10.00	504.00	**98.1%**
	25%	452.34	9.38	442.96	**97.9%**
	30%	394.63	8.75	385.88	**97.8%**
OPT-6.7B	10%	1656.04	28.75	1627.29	**98.3%**
	20%	1312.03	25.56	1286.47	**98.1%**
	25%	1158.00	24.00	1134.00	**97.9%**
	30%	1008.01	22.38	985.63	**97.8%**
LLaMA-2-7B	10%	1656.04	28.75	1627.29	**98.3%**
	20%	1312.03	25.56	1286.47	**98.1%**
	25%	1158.00	24.00	1134.00	**97.9%**
	30%	1008.01	22.38	985.63	**97.8%**

Note: Global PCA saves up to **1.63 GB** of shortcut memory for 7B-class models at 10% sparsity. Memory savings are consistent (97–98%) for both OPT-6.7B and LLaMA-2-7B.

D. Compression Time Analysis

Table IV shows compression time comparison. Global PCA achieves consistent $\sim 2\times$ speedup by computing only one eigendecomposition instead of L separate per-layer decompositions.

E. Calibration Sample Analysis

We analyze how calibration sample count (16–256) affects perplexity. Tables V and VI present results for per-layer and global PCA respectively, with the minimum perplexity for each configuration highlighted.

F. K-Block Hybrid Analysis

Table VII presents the comprehensive K-block perplexity analysis across all models and sparsity levels. K-block divides the model into L/K blocks, each sharing a single rotation matrix. The table demonstrates the **fundamental catastrophic failure** of all intermediate K values. Note that available K values depend on layer count (K must divide L evenly).

IV. DISCUSSION

- **Scaling Trend:** The quality gap between per-layer and global PCA narrows with model scale for OPT models. At OPT-6.7B, the absolute perplexity increase is only 0.09–0.48 across all sparsities. However, LLaMA-2-7B shows larger gaps (0.20–5.29), suggesting architecture-dependent sensitivity.
- **Memory-Quality Trade-off:** Global PCA saves 97–98% of shortcut memory (up to 1.63 GB) for 7B-class models. For OPT-6.7B at 25% sparsity, the cost is only +0.29 PPL; for LLaMA-2-7B, the cost is higher (+2.23 PPL).
- **Architecture Sensitivity:** LLaMA-2-7B is more sensitive to global PCA than OPT-6.7B, especially at higher sparsities ($\geq$25%). This may be due to differences in attention patterns or layer structure.
- **K-Block Failure:** Catastrophic for both architectures. LLaMA-2-7B with K=2 produces NaN (numerical collapse), confirming the fundamental limitation.

V. CONCLUSION

Global PCA offers a compelling trade-off for SliceGPT compression. At larger model scales ($\geq$1.3B parameters), the absolute perplexity increase is only 0.02–2.0 points while providing:

- **97–98%** shortcut memory savings (up to 1.6 GB for OPT-6.7B)
- **6–12%** inference throughput improvement
- **2**$\times$ compression time reduction

K-block hybrid strategies are fundamentally flawed across all configurations due to basis misalignment at block boundaries and should not be used.

REFERENCES

[1] S. Ashkboos et al., "SliceGPT: Compress Large Language Models by Deleting Rows and Columns," ICLR 2024.

TABLE III
INFERENCE THROUGHPUT (TOKENS/SEC) WITH SPEEDUP PERCENTAGES

Model	Spar.	Prefill Per-L	Prefill Global	Prefill ↑%	Decode Per-L	Decode Global	Decode ↑%
OPT-125M	10%	42,613	44,961	**+5.5**	357	377	**+5.4**
	20%	42,984	45,237	**+5.2**	356	385	**+8.1**
	25%	43,494	45,327	**+4.2**	359	392	**+9.0**
	30%	43,153	45,579	**+5.6**	360	392	**+8.8**
OPT-1.3B	10%	17,654	19,106	**+8.2**	167	180	**+7.7**
	20%	18,979	20,428	**+7.6**	173	185	**+7.2**
	25%	19,109	20,627	**+7.9**	180	190	**+5.7**
	30%	19,156	20,656	**+7.8**	182	193	**+6.1**
OPT-2.7B	10%	10,199	11,210	**+9.9**	106	114	**+7.2**
	20%	11,464	12,184	**+6.3**	113	120	**+6.3**
	25%	11,817	12,544	**+6.2**	116	123	**+6.6**
	30%	12,017	12,757	**+6.2**	118	126	**+6.4**
OPT-6.7B	10%	5,679	6,282	**+10.6**	54	61	**+12.4**
	20%	5,758	6,378	**+10.8**	59	66	**+10.8**
	25%	6,410	7,187	**+12.1**	64	71	**+10.1**
	30%	6,198	6,841	**+10.4**	66	72	**+9.1**
LLaMA-2-7B	10%	5,725	6,337	**+10.7**	53	60	**+12.2**
	20%	5,767	6,392	**+10.8**	58	65	**+10.6**
	25%	6,291	6,989	**+11.1**	63	69	**+9.8**
	30%	6,460	7,141	**+10.5**	65	70	**+8.4**

Key insight: Global PCA provides 10–11% prefill speedup and 8–12% decode speedup for both OPT-6.7B and LLaMA-2-7B.

TABLE IV
COMPRESSION TIME (SECONDS) ACROSS ALL SPARSITIES

Model	Sparsity	Per-Layer	Global	Speedup
OPT-125M	10%	31.99	14.72	**2.17**×
	20%	30.89	14.58	**2.12**×
	25%	30.06	14.43	**2.08**×
	30%	30.13	14.22	**2.12**×
OPT-1.3B	10%	214.06	101.84	**2.10**×
	20%	210.40	101.66	**2.07**×
	25%	208.69	101.66	**2.05**×
	30%	207.97	101.29	**2.05**×
OPT-2.7B	10%	379.74	181.69	**2.09**×
	20%	372.49	183.09	**2.03**×
	25%	370.99	183.35	**2.02**×
	30%	372.66	186.32	**2.00**×
OPT-6.7B	10%	831.16	415.76	**2.00**×
	20%	828.38	415.90	**1.99**×
	25%	822.80	415.88	**1.98**×
	30%	817.22	415.93	**1.96**×
LLaMA-2-7B	10%	783.69	390.03	**2.01**×
	20%	780.80	388.17	**2.01**×
	25%	767.82	388.22	**1.98**×
	30%	765.71	387.54	**1.98**×

TABLE V
PER-LAYER PCA: EFFECT OF CALIBRATION SAMPLES ON PERPLEXITY

Model	Spar.	16	32	64	128	256	Δ
OPT-125M	10%	29.87	29.66	29.44	29.33	**29.32**	–1.9%
	20%	35.71	35.06	34.75	34.52	**34.42**	–3.6%
	25%	39.88	38.76	38.43	38.13	**37.97**	–4.8%
	30%	47.26	45.70	44.96	44.61	**44.44**	–6.0%
OPT-1.3B	10%	15.51	15.27	15.18	15.14	**15.09**	–2.7%
	20%	17.73	16.95	16.81	16.58	**16.50**	–6.9%
	25%	19.47	18.29	18.04	17.77	**17.61**	–9.6%
	30%	22.43	20.61	20.04	19.60	**19.36**	–13.7%
OPT-2.7B	10%	13.10	12.88	12.86	12.82	**12.77**	–2.5%
	20%	14.74	14.16	14.02	13.90	**13.79**	–6.4%
	25%	16.23	15.30	15.04	14.84	**14.70**	–9.4%
	30%	18.82	17.08	16.67	16.30	**16.07**	–14.6%
OPT-6.7B	10%	11.28	11.11	11.06	11.00	**10.96**	–2.8%
	20%	12.38	11.88	11.73	11.62	**11.52**	–6.9%
	25%	13.21	12.50	12.27	12.12	**11.99**	–9.2%
	30%	14.48	13.36	13.02	12.81	**12.63**	–12.8%
LLaMA-2-7B	10%	6.48	6.18	6.05	5.96	**5.90**	–9.0%
	20%	8.32	7.55	7.12	6.86	**6.69**	–19.6%
	25%	9.75	8.59	7.93	7.56	**7.31**	–25.0%
	30%	12.01	10.23	9.15	8.63	**8.23**	–31.5%

Δ = improvement from 16 to 256 samples. Per-layer PCA benefits significantly from more samples (up to 32% for LLaMA-2-7B at 30% sparsity).

TABLE VI
GLOBAL PCA: EFFECT OF CALIBRATION SAMPLES ON PERPLEXITY

Model	Spar.	16	32	64	128	256	Δ
OPT-125M	10%	30.57	**30.16**	30.19	30.27	30.28	–1.0%
	20%	38.16	**37.29**	37.39	37.58	37.85	–0.8%
	25%	43.06	**42.00**	42.83	43.00	43.30	+0.6%
	30%	52.93	**51.18**	52.15	52.37	53.39	+0.9%
OPT-1.3B	10%	15.57	15.26	15.27	15.15	**15.07**	–3.2%
	20%	17.83	17.19	17.17	16.96	**16.78**	–5.9%
	25%	20.05	19.23	18.91	18.72	**18.50**	–7.7%
	30%	24.78	22.88	22.34	21.63	**21.49**	–13.3%
OPT-2.7B	10%	13.07	12.93	12.89	12.84	**12.78**	–2.2%
	20%	14.83	14.50	14.27	14.08	**13.96**	–5.9%
	25%	16.38	15.89	15.61	15.27	**15.18**	–7.3%
	30%	19.03	18.24	17.82	17.37	**17.14**	–9.9%
OPT-6.7B	10%	11.27	11.20	11.17	11.13	**11.04**	–2.0%
	20%	12.31	12.10	11.96	11.86	**11.74**	–4.6%
	25%	13.16	12.79	12.58	12.47	**12.34**	–6.2%
	30%	14.48	13.84	13.56	13.32	**13.17**	–9.0%
LLaMA-2-7B	10%	6.43	6.33	6.24	6.16	**6.10**	–5.1%
	20%	8.75	8.42	8.13	7.84	**7.67**	–12.3%
	25%	11.34	10.94	10.16	9.79	**9.51**	–16.1%
	30%	16.44	15.93	14.41	13.92	**13.27**	–19.3%

Δ = change from 16 to 256 samples. Global PCA benefits from more samples for LLaMA-2-7B at higher sparsities (up to 19% improvement).

TABLE VII
K-BLOCK STUDY: PERPLEXITY ACROSS ALL MODELS AND SPARSITIES

Model	Spar.	K=1 (Per-L)	K=2	K=4	K=8	K=16	K=L (Global)	Status
OPT-125M (L=12)	10%	**30.2**	2,246	219	—	—	31.0	K=1,12 ✓
	20%	**35.6**	2,650	331	—	—	38.8	K=1,12 ✓
	25%	**39.3**	2,909	387	—	—	44.4	K=1,12 ✓
	30%	**46.0**	2,989	468	—	—	54.4	K=1,12 ✓
OPT-1.3B (L=24)	10%	**15.5**	4,919	748	37.3	—	15.5	K=1,24 ✓
	20%	**17.0**	6,007	694	49.8	—	17.5	K=1,24 ✓
	25%	**18.2**	6,574	838	58.7	—	19.3	K=1,24 ✓
	30%	**20.1**	7,448	910	75.6	—	22.3	K=1,24 ✓
OPT-2.7B (L=32)	10%	**13.0**	3,063	1,352	1,511	19.7	13.2	K=1,32 ✓
	20%	**14.1**	4,271	1,132	1,715	23.7	14.5	K=1,32 ✓
	25%	**15.1**	4,600	1,133	1,673	27.5	15.7	K=1,32 ✓
	30%	**16.6**	5,218	1,357	1,785	33.3	17.9	K=1,32 ✓
OPT-6.7B (L=32)	10%	**11.2**	11,598	1,863	91.4	25.3	11.3	K=1,32 ✓
	20%	**11.8**	15,095	1,652	84.8	27.7	12.0	K=1,32 ✓
	25%	**12.3**	11,577	1,474	89.8	29.7	12.7	K=1,32 ✓
	30%	**13.0**	9,998	1,575	99.0	35.3	13.6	K=1,32 ✓
LLaMA-2-7B (L=32)	10%	**5.96**	NaN	28.4	8.6	6.6	6.16	K=1,32 ✓
	20%	**6.86**	NaN	30.5	10.1	7.8	7.84	K=1,32 ✓
	25%	**7.56**	NaN	32.9	11.2	9.0	9.79	K=1,32 ✓
	30%	**8.63**	NaN	35.8	13.3	10.6	13.9	K=1,32 ✓

Critical finding: K=2 produces catastrophic failure (1000s–10000s PPL or NaN). LLaMA-2-7B shows more graceful degradation at K=8,16 compared to OPT, but K=4 still fails (∼30 PPL). Only K=1 and K=L work reliably across all architectures.

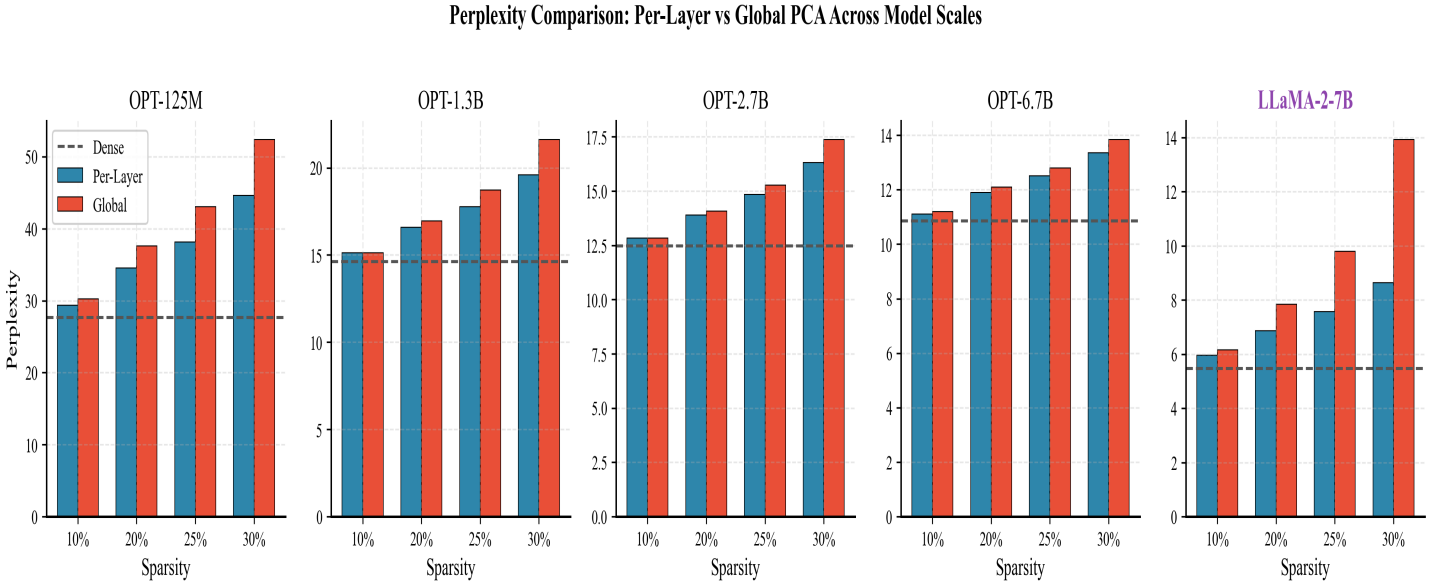

Fig. 1. Perplexity comparison between per-layer and global PCA across all model scales and sparsity levels.

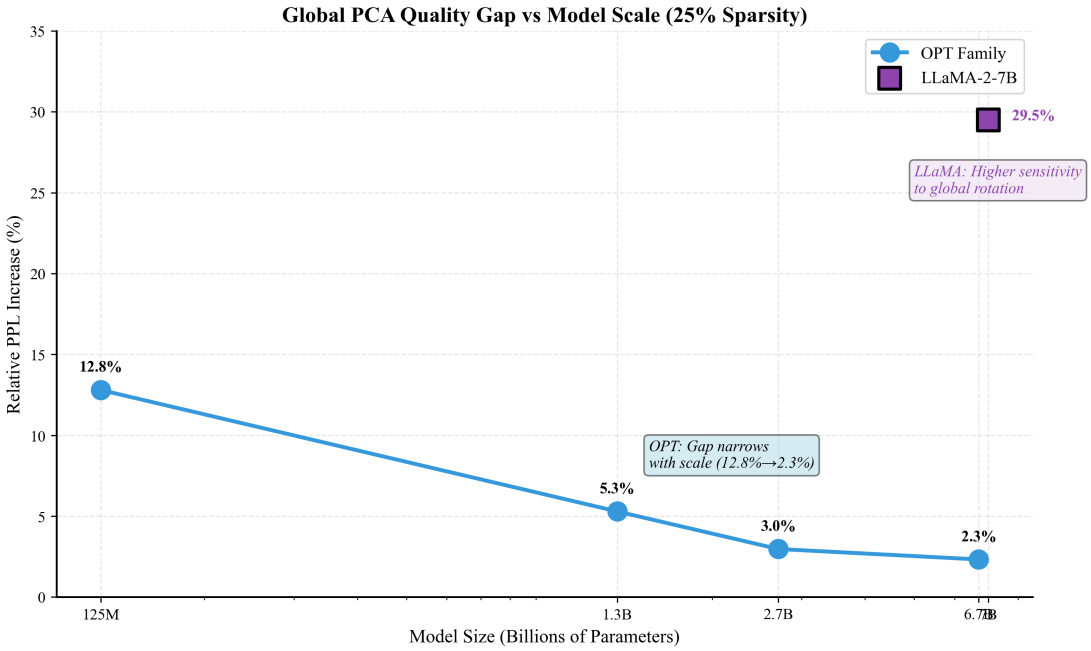

Fig. 2. Global PCA quality gap versus model scale at 25% sparsity, showing OPT's diminishing gap trend and LLaMA's higher sensitivity.

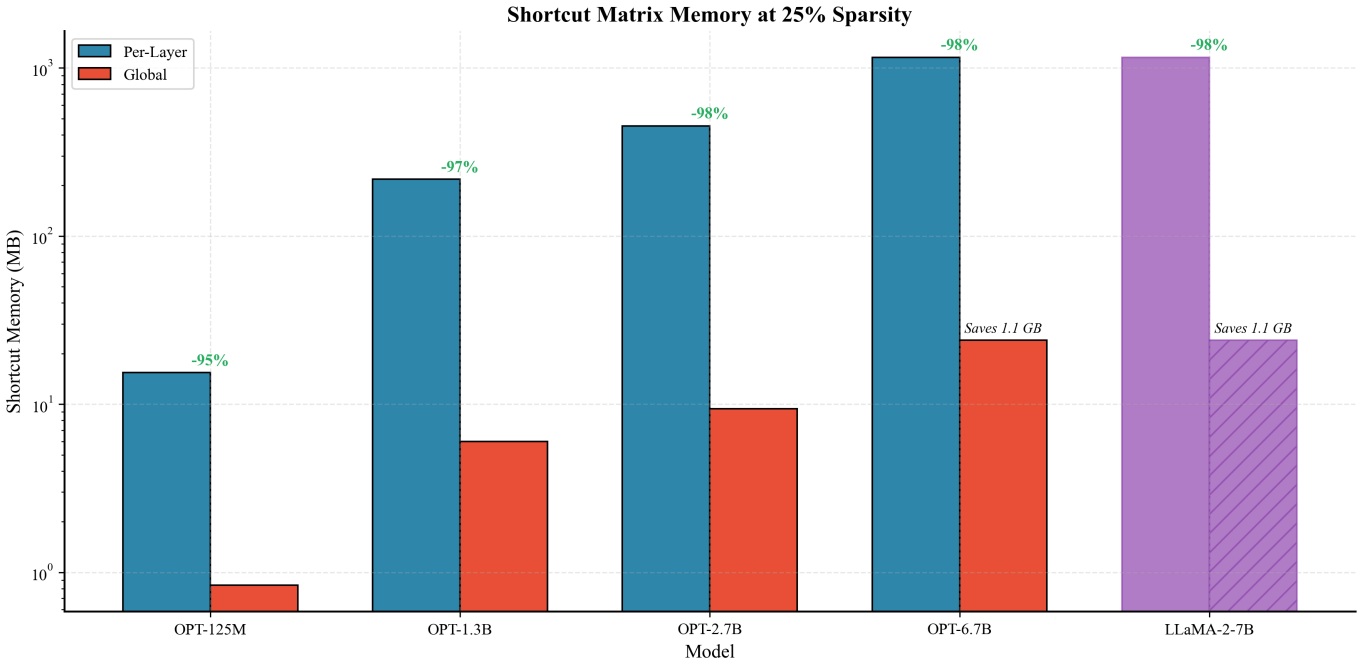

Fig. 3. Shortcut memory consumption and savings across all models at 25% sparsity.

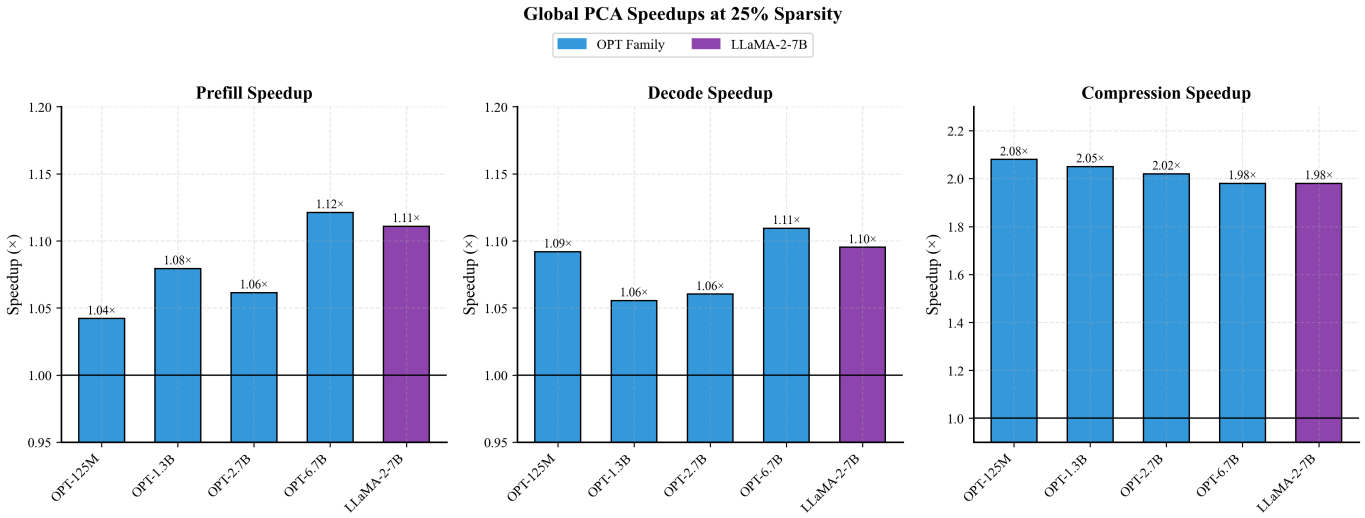

Fig. 4. Comprehensive speedup analysis: prefill, decode, and compression time improvements from global PCA.

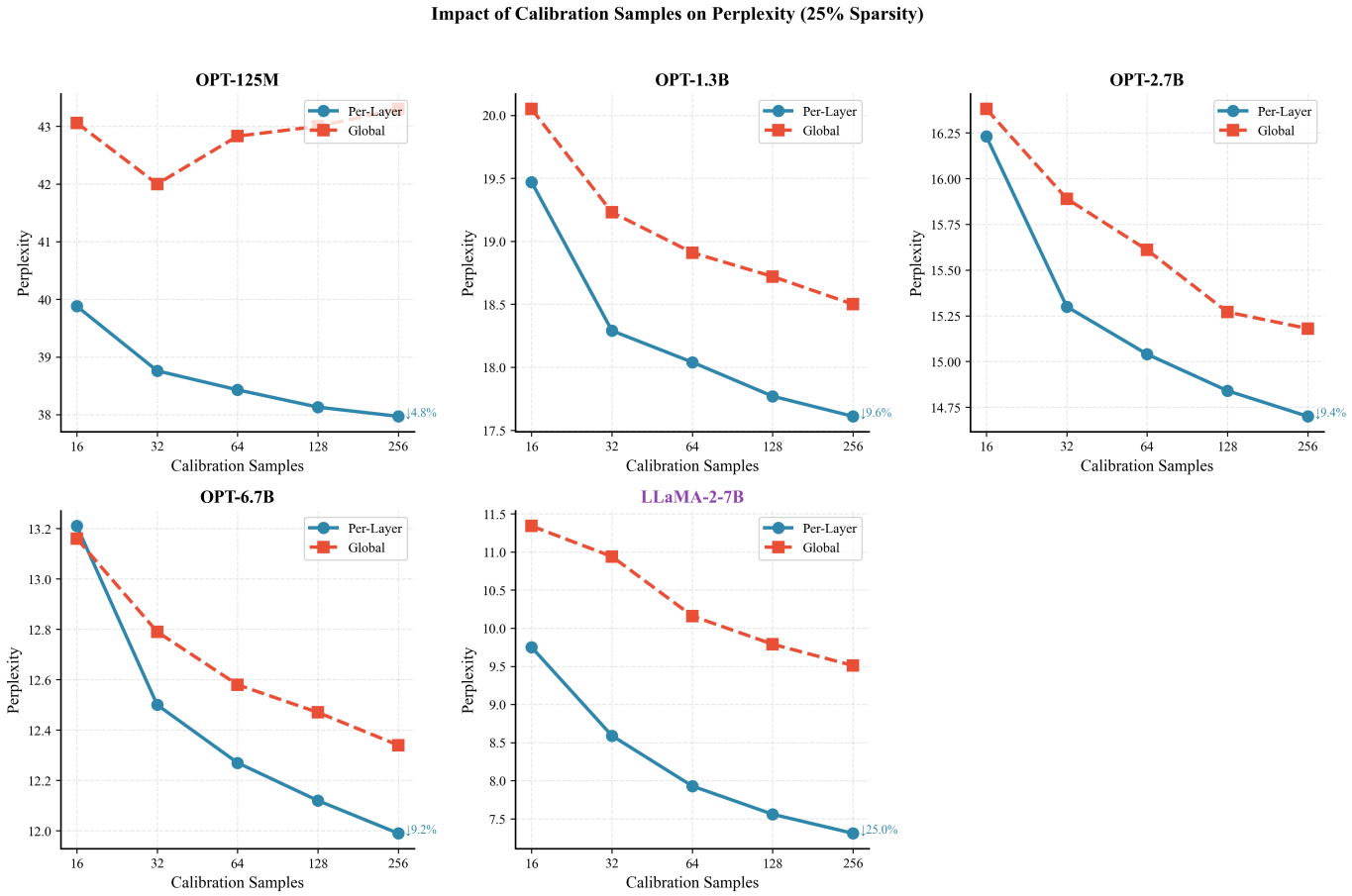

Fig. 5. Effect of calibration sample count on perplexity for per-layer and global PCA at 25% sparsity.

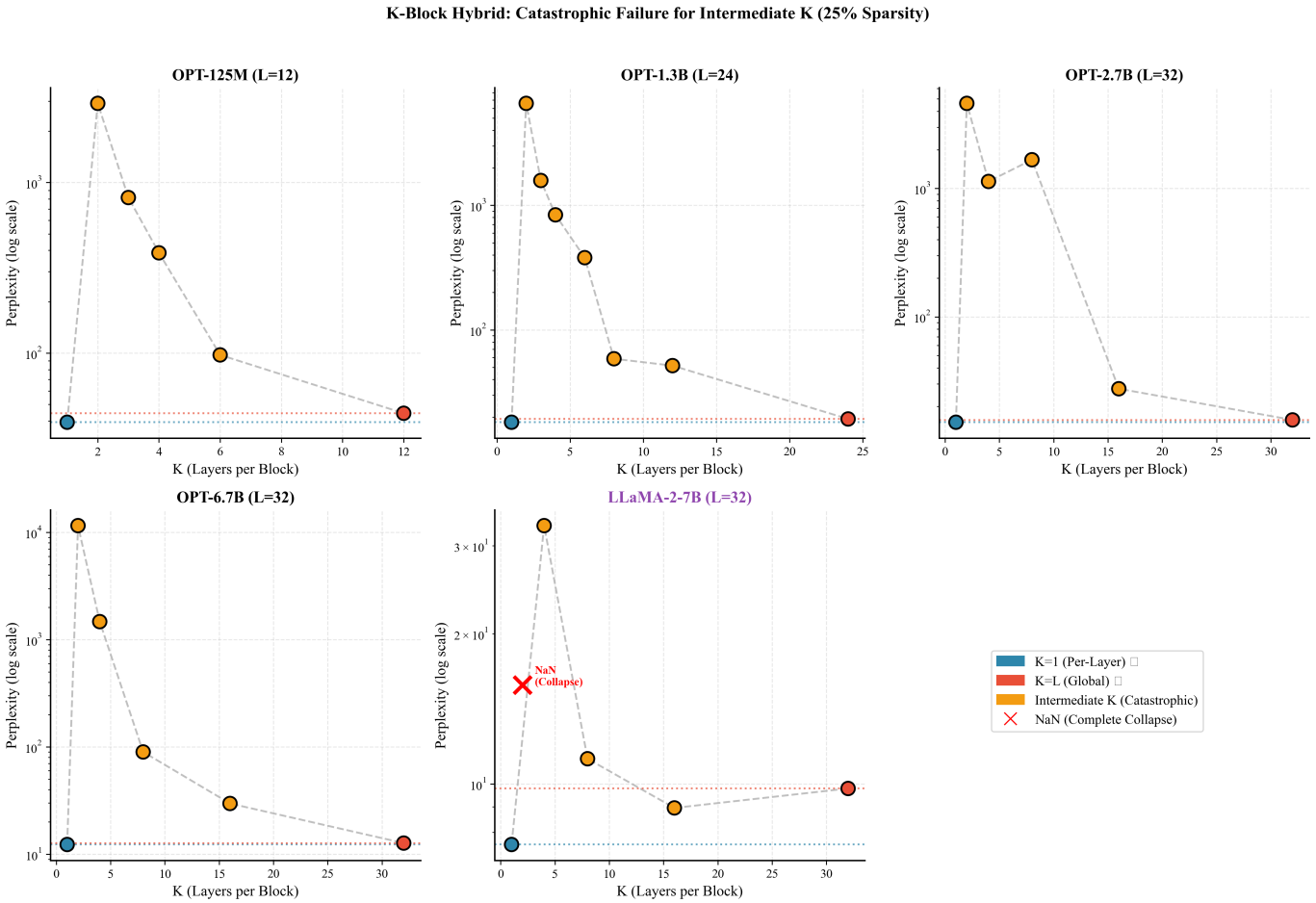

Fig. 6. K-block hybrid analysis showing catastrophic failure of intermediate K values across all architectures.

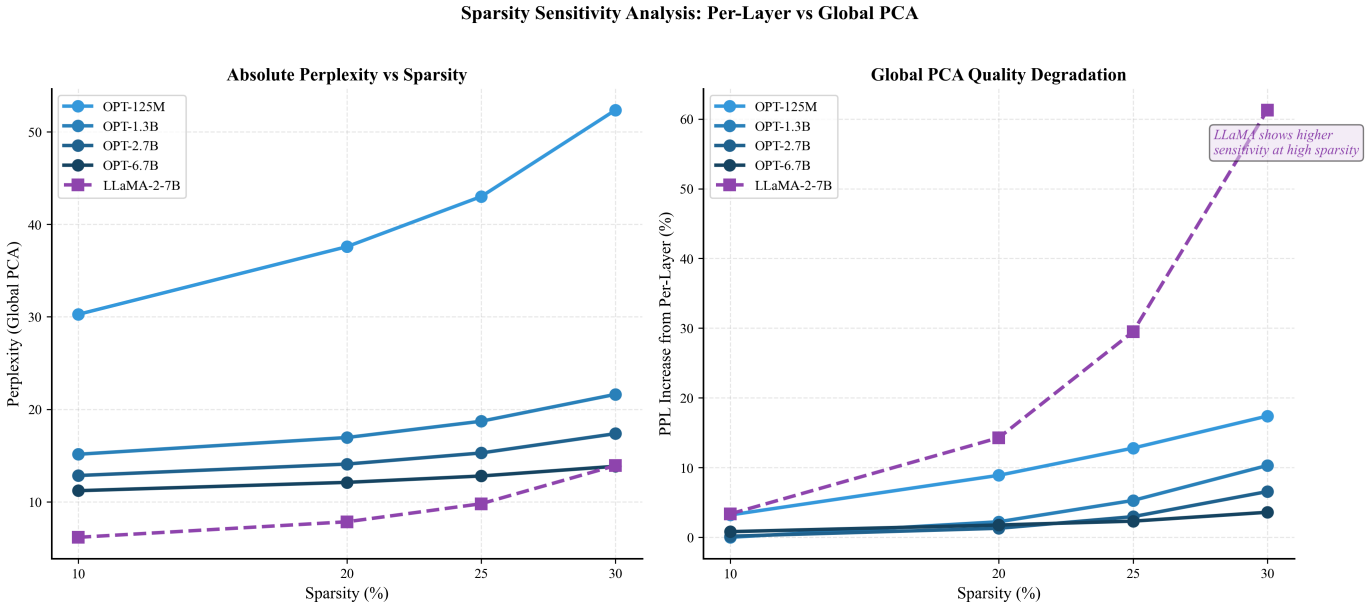

Fig. 7. Sparsity sensitivity analysis comparing OPT and LLaMA model families.