

Table 1: Perplexity for LLaMA-3 8b; the lower the perplexity, the better.

LLaMA-3 8b (WikiText2 (WT2): 6.14; PTB: 11.18; C4: 9.45)												
Sparsity	10%			20%			50%			2:4		
Dataset	WT2	PTB	C4	WT2	PTB	C4	WT2	PTB	C4	WT2	PTB	C4
SparseGPT	6.16	11.18	9.47	6.23	11.23	9.57	7.62	13.19	11.71	11.17	19.38	16.99
<i>SparseLLM</i>	6.15	11.18	9.47	6.23	11.21	9.56	7.61	13.07	11.70	11.15	19.33	16.94
Sparsity	70%			80%			90%			3:4		
Dataset	WT2	PTB	C4	WT2	PTB	C4	WT2	PTB	C4	WT2	PTB	C4
SparseGPT	22.37	36.56	30.53	72.87	113.95	79.86	214.68	261.18	198.34	96.75	107.52	102.11
<i>SparseLLM</i>	20.98	33.78	28.94	57.83	85.98	72.18	197.47	241.68	181.69	76.33	99.54	93.68

Table 2: Perplexity in low sparsity regimes; the lower the perplexity, the better.

OPT-1.3b (WikiText2 (WT2): 14.62; PTB: 20.29; C4: 16.07)									
Sparsity	10%			20%			50%		
Dataset	WT2	PTB	C4	WT2	PTB	C4	WT2	PTB	C4
Magnitude	14.63	20.35	16.10	15.13	20.83	16.27	944.04	1638.46	307.64
Wanda	14.61	20.31	16.07	14.91	20.55	16.09	15.25	21.25	16.67
SparseGPT	14.63	20.29	16.07	14.76	20.54	16.09	15.16	20.85	16.50
<i>SparseLLM</i>	14.63	20.29	16.07	14.76	20.53	16.09	15.06	20.68	16.37
OPT-2.7b (WikiText2 (WT2): 12.47; PTB: 17.97; C4: 14.32)									
Sparsity	10%			20%			50%		
Dataset	WT2	PTB	C4	WT2	PTB	C4	WT2	PTB	C4
Magnitude	12.49	18.14	14.37	12.72	18.49	14.49	66.44	87.85	33.74
Wanda	12.49	17.97	14.34	12.49	17.97	14.35	13.24	18.49	14.69
SparseGPT	12.49	17.96	14.34	12.48	18.01	14.35	12.55	18.14	14.54
<i>SparseLLM</i>	12.49	17.96	14.34	12.48	17.99	14.32	12.48	18.05	14.43

Table 3: Computation time in seconds of OPT-1.3b and OPT-2.7b

Sparsity	OPT-1.3b			OPT-2.7b		
	10%	20%	50%	10%	20%	50%
SparseGPT	122.03	121.75	122.15	222.63	220.22	221.81
<i>SparseLLM</i>	273.18	272.22	274.07	514.28	515.41	514.31

Table 4: Ablations of the hyperparameters α and β on OPT-1.3b with 70% sparsity (in perplexity)

α / β	0.01	0.1	1	5	10	100
0.01	18.01	17.97	17.97	-	-	-
0.1	18.04	17.82	17.96	18.04	18.40	-
1	18.20	18.02	18.11	17.87	17.96	18.22
5	18.06	18.02	18.03	17.92	17.96	18.04
10	18.03	18.01	17.96	17.96	17.96	18.03
100	18.04	18.04	17.98	18.01	18.01	18.03

Table 5: Results for 2:4 sparsity pruning; the lower the perplexity, the better.

Dataset	OPT-1.3b (2:4)			OPT-2.7b (2:4)			OPT-6.7b (2:4)			OPT-13b (2:4)		
	WT2	PTB	C4	WT2	PTB	C4	WT2	PTB	C4	WT2	PTB	C4
Magnitude	96.68	133.92	48.08	272.34	308.55	267.70	64.11	92.23	82.67	67.07	110.77	52.61
Wanda	15.63	24.04	18.23	13.66	21.67	16.10	11.86	18.54	14.77	10.33	15.35	12.54
SparseGPT	15.11	23.71	17.88	12.62	19.28	15.12	11.30	16.90	13.51	10.20	15.14	12.48
<i>SparseLLM</i>	14.77	23.40	17.57	12.52	19.10	14.93	11.07	16.73	13.42	10.11	15.03	12.41