Appendix

A Additional visualizations of Layer-wise Magnitude Gain

Figure 1 presents the channel-wise averaged magnitude gain ratio for each layer across different models. As illustrated, every layer contributes to the magnitude gain, with variations in intensity observed across layers and models. This phenomenon is observed across a diverse set of models, including LLaMA-2-7B, LLaMA-2-13B, LLaMA-3-8B, DeepSeek-R1-Distill-Llama-8B and Qwen-3-8B. The results suggest that layer-wise hidden state amplification is a common structural characteristic of LLMs.

B More Results on PPL Benchmark

Table 1 presents perplexity results for LLaMA-2-13B and Qwen3-8B across diverse pruning configurations

For LLaMA-2-13B under 8/40 sparsity, Mag+ metric combined with PRUNE&COMP achieves 28.03 average perplexity, representing an 86.03% reduction from the baseline of 200.74. This trend holds across higher sparsity, with 12/40 layers pruned, where PRUNE&COMP maintains an average perplexity of 38.38 for the Mag+ metric, outperforming the baseline by 94.34%. For Qwen3-8B, metrics with PRUNE&COMP demonstrate similar improvement. Under 5/36 sparsity, CosSim(CL) metric with PRUNE&COMP reduces average perplexity from 32.16 to 25.23. Notably, under 8/40 sparsity, Taylor+ metric with PRUNE&COMP achieves an average perplexity of 30.71, a 19.7% reduction from the baseline of 126.45.

C More Results on QA Benchmark

Table 2 presents a comprehensive analysis of the question answering (QA) benchmark of LLaMA-2-7B and LLaMA-2-13B across diverse pruning configurations.

Across model scales, PRUNE&COMP demonstrates consistent effectiveness. For LLaMA-2-7B under 7/32 sparsity, Taylor+ metric with PRUNE&COMP achieves 93.48 average score, representing improvements of 6.81 over baseline, while under 9/32 sparsity, Mag+ metric average score improves from 65.92 to 72.14. For LLaMA-2-13B under 8/40 sparsity, where Mag+ metric combined with PRUNE&COMP increases average score from 66.85 to 78.13. Similar gains are observed under 12/40 sparsity, achieving 73.99 average score for Mag+ metric compared to baseline 59.95.

D The Pruning Settings

Tables 3 and Tables 4 present layer pruning configurations for LLaMA-2-7B, LLaMA-3-8B, LLaMA-2-13B, and Qwen3-8B under varying sparsity levels.

E Further Ablation Analysis

Table 5 reports the ablation results of iterative pruning (+IterPrune) and magnitude compression (+MagComp). Both techniques individually lead to notable gains over naive one-shot pruning across QA benchmarks. Specifically, naive one-shot pruning yields an average accuracy of 65.50%, which is improved to 66.86% by iterative pruning and further to 68.35% by magnitude compression. Importantly, their joint application achieves the highest average accuracy of 68.44%, highlighting a clear synergistic effect that surpasses the contributions of each component in isolation.

Table 6 presents the impact of calibration set size on performance for LLaMA-3-8B, pruned 5 layers using the CosSim(BI) metric. Results show consistent performance and indicate calibration set size within the tested range does not significantly affect pruning outcomes, suggesting that even a small calibration set suffices to stabilize performance metrics for the evaluated pruning configuration.

LLaMA-2-13B						**Qwen3-8B**					
Sparsity	**Metric**	**WikiText-2**	**C4**	**PTB**	**Average**	**Sparsity**	**Metric**	**WikiText-2**	**C4**	**PTB**	**Average**
-	Dense	4.88	6.47	28.87	13.4	-	Dense	9.72	13.29	16.90	13.30
8/40	PPL	7.43	9.39	48.4	21.7	5/36	PPL	12.17	16.83	23.02	17.34
	+PRUNE&COMP	**6.71**	**8.59**	**38.46**	**17.92**		+PRUNE&COMP	**12.00**	**16.54**	**22.75**	**17.10**
	CosSim(CL)	8.30	10.36	44.96	21.21		CosSim(CL)	27.17	28.41	40.89	32.16
	+PRUNE&COMP	**7.75**	**9.63**	**36.14**	**17.84**		+PRUNE&COMP	**20.29**	**23.19**	**32.20**	**25.23**
	Mag+	226.45	114.05	261.71	200.74		Mag+	**14.55**	**18.30**	**26.23**	**19.69**
	+PRUNE&COMP	**10.22**	**12.05**	**61.83**	**28.03**		+PRUNE&COMP	15.99	20.59	29.38	21.99
	Taylor+	16.92	16.24	**49.74**	**27.63**		Taylor+	18.42	22.09	30.29	23.60
	+PRUNE&COMP	**9.43**	**11.21**	68.57	29.74		+PRUNE&COMP	**16.50**	**20.12**	**27.41**	**21.34**
	CosSim(BI)	8.30	10.36	44.96	21.21		CosSim(BI)	**11.80**	**16.27**	23.65	**17.24**
	+PRUNE&COMP	**7.11**	**9.03**	**34.92**	**17.02**		+PRUNE&COMP	12.33	16.68	**23.44**	17.48
12/40	PPL	10.51	13.50	65.23	29.75	7/36	PPL	17.34	21.16	**26.67**	21.72
	+PRUNE&COMP	**8.22**	**10.27**	**45.28**	**21.26**		+PRUNE&COMP	**13.97**	**18.87**	27.71	**20.18**
	CosSim(CL)	34.71	28.06	77.38	46.72		CosSim(CL)	60.41	49.70	80.43	63.51
	+PRUNE&COMP	**12.18**	**13.91**	**46.94**	**24.34**		+PRUNE&COMP	**34.37**	**30.17**	**43.71**	**36.08**
	Mag+	592.13	307.72	1136.36	678.74		Mag+	**19.11**	**22.55**	**31.44**	**24.37**
	+PRUNE&COMP	**15.71**	**16.21**	**83.23**	**38.38**		+PRUNE&COMP	24.90	28.99	46.43	33.44
	Taylor+	39.63	26.19	**78.89**	48.24		Taylor+	85.91	44.84	248.61	126.45
	+PRUNE&COMP	**12.12**	**13.59**	86.17	**37.29**		+PRUNE&COMP	**24.37**	**26.74**	**41.03**	**30.71**
	CosSim(BI)	34.71	28.06	77.38	46.72		CosSim(BI)	15.38	19.99	27.27	20.88
	+PRUNE&COMP	**10.67**	**12.14**	**41.53**	**21.45**		+PRUNE&COMP	**14.34**	**19.20**	**27.05**	**20.20**

Table 1: Performance comparison on perplexity (PPL) benchmark. Sparsity is indicated by layers pruned/total layers.

Model	Sparsity	Metric	ARC-c	ARC-e	BoolQ	CoPa	HeSw	PIQA	Race-h	WG	WSC	Average	RP
LLaMA-2-7B	**0/32**	Dense	46.25	74.58	77.74	87.00	75.97	79.11	39.62	68.98	80.59	69.98	100.00
	7/32 (21.02%)	PPL	30.20	55.64	62.14	76.00	59.81	**74.54**	32.25	55.72	64.47	56.75	81.10
		+PRUNE&COMP	**30.80**	**56.94**	**63.70**	**77.00**	**60.08**	74.21	**32.44**	**56.99**	**72.16**	**58.26**	**83.25**
		CosSim(CL)	36.18	55.89	**62.17**	81.00	62.66	70.04	33.78	**77.29**	77.29	61.81	88.32
		+PRUNE&COMP	**38.48**	**60.27**	**62.17**	**85.00**	**63.45**	**70.78**	**36.36**	65.51	**79.49**	**62.39**	**89.15**
		Mag+	24.23	44.07	57.40	72.00	40.22	65.89	24.78	52.96	61.17	49.19	70.29
		+PRUNE&COMP	**27.82**	**46.42**	**63.21**	**77.00**	**53.75**	**70.84**	**29.47**	**55.41**	**62.64**	**54.06**	**77.25**
		Taylor+	36.26	55.89	62.17	81.00	62.66	70.40	33.78	**66.06**	**77.66**	60.65	86.67
		+PRUNE&COMP	**37.03**	**59.26**	**89.00**	**89.00**	**63.37**	**72.96**	**37.32**	65.04	75.82	**65.42**	**93.48**
		CosSim(BI)	36.18	55.89	62.17	**81.00**	62.66	70.04	33.78	**77.29**	**77.29**	**61.81**	**88.32**
		+PRUNE&COMP	**36.95**	**61.70**	**63.49**	80.00	**65.47**	**72.03**	**35.60**	65.67	73.26	61.57	87.98
	9/32 (27.03%)	PPL	26.11	49.12	62.11	75.00	53.90	70.35	28.33	**54.78**	60.07	53.31	76.17
		+PRUNE&COMP	**29.01**	**50.08**	**62.42**	**78.00**	**54.17**	**72.47**	**31.29**	54.14	**66.30**	**55.32**	**79.05**
		CosSim(CL)	32.76	48.61	**62.17**	77.00	56.17	64.36	32.25	64.33	71.06	56.52	80.77
		+PRUNE&COMP	**34.13**	**54.59**	**62.17**	**78.00**	**57.93**	**67.52**	**35.31**	**64.72**	**75.82**	**58.91**	**84.18**
		Mag+	22.87	37.92	**61.28**	67.00	35.17	61.43	25.55	50.83	53.11	46.13	65.92
		+PRUNE&COMP	**26.28**	**41.00**	58.04	**72.00**	**46.73**	**66.32**	**30.05**	**53.90**	**60.07**	**50.49**	**72.14**
		Taylor+	32.59	48.65	62.17	77.00	56.15	64.31	32.34	**64.40**	71.43	56.56	80.82
		+PRUNE&COMP	**32.76**	**54.38**	**70.09**	**80.00**	**59.10**	**69.97**	**36.27**	63.54	**77.66**	**60.42**	**86.34**
		CosSim(BI)	32.76	48.61	**62.17**	77.00	56.17	64.36	32.25	64.33	71.06	**56.52**	**80.77**
		+PRUNE&COMP	**32.85**	**52.86**	44.83	**80.00**	**57.47**	**66.76**	**35.69**	**65.27**	**72.89**	56.51	80.75
LLaMA-2-13B	**0/32**	Dense	49.06	77.40	80.61	91.00	79.35	80.52	40.48	72.38	86.81	73.07	100.00
	8/40 (19.50%)	PPL	**42.83**	**67.13**	**62.87**	**88.00**	**72.34**	75.73	**39.14**	**69.77**	**84.98**	**66.98**	**91.66**
		+PRUNE&COMP	37.20	65.28	62.26	82.00	68.37	**77.31**	36.27	63.77	76.92	63.26	86.58
		CosSim(CL)	44.03	67.38	57.00	89.00	72.38	75.24	38.18	69.61	79.85	65.85	90.12
		+PRUNE&COMP	**44.62**	**68.90**	**66.12**	**91.00**	**73.31**	**75.63**	**39.43**	**69.85**	**82.78**	**67.96**	**93.01**
		Mag+	23.55	46.13	60.98	74.00	37.74	66.10	24.11	50.99	56.04	48.85	66.85
		+PRUNE&COMP	**31.40**	**51.39**	**66.67**	**76.00**	**55.66**	**73.99**	**32.44**	**56.99**	**69.23**	**57.09**	**78.13**
		Taylor+	38.74	60.48	37.90	83.00	65.35	74.65	34.64	70.09	79.49	60.48	82.77
		+PRUNE&COMP	**41.38**	**67.13**	**73.24**	**89.00**	**67.18**	**76.22**	**38.76**	**65.75**	**82.42**	**66.79**	**91.40**
		CosSim(BI)	44.03	67.38	57.00	89.00	72.38	75.24	38.18	69.61	79.85	65.85	90.12
		+PRUNE&COMP	**45.05**	**68.98**	**63.15**	**92.00**	**73.73**	**75.84**	**38.95**	**70.64**	**83.52**	**67.98**	**93.04**
	12/40 (29.24%)	PPL	**38.48**	**59.60**	**62.39**	82.00	**66.50**	71.76	33.40	**66.22**	**74.73**	**61.67**	**84.41**
		+PRUNE&COMP	34.30	58.21	62.17	**85.00**	62.63	**74.32**	**33.88**	62.83	72.53	60.65	83.01
		CosSim(CL)	39.42	55.72	61.68	83.00	64.52	69.53	**37.99**	69.77	73.26	61.65	84.38
		+PRUNE&COMP	**40.02**	**60.69**	**62.60**	**86.00**	**67.70**	**72.14**	36.75	**70.40**	**80.95**	**64.14**	**87.78**
		Mag+	22.87	33.25	59.79	64.00	30.36	55.88	24.02	51.70	52.38	43.81	59.95
		+PRUNE&COMP	**28.33**	**44.02**	**64.95**	**80.00**	**48.34**	**68.12**	**31.48**	**56.12**	**65.20**	**54.06**	**73.99**
		Taylor+	35.75	52.82	37.83	81.00	57.92	69.86	30.72	**69.06**	78.39	57.04	78.06
		+PRUNE&COMP	**38.65**	**60.94**	**63.64**	**87.00**	**71.93**	**72.91**	**36.94**	67.25	**79.85**	**64.35**	**88.06**
		CosSim(BI)	39.42	55.72	61.68	83.00	64.52	69.53	37.99	69.77	73.26	61.65	84.38
		+PRUNE&COMP	**40.27**	**60.14**	**62.63**	**87.00**	**67.55**	**71.82**	**38.66**	**70.17**	**77.29**	**63.95**	**87.52**

Table 2: Performance comparison on question answering (QA) benchmark. Sparsity is indicated by layers pruned/total layers (compression rate). RP denotes the relative performance (%).

Model	**Sparsity**	**Metric**	**Layer Index**
LLaMA-2-7B	-	Dense	-
	7/32	PPL	[14, 12, 11, 24, 23, 10, 7]
		+PRUNE&COMP	[14, 11, 24, 10, 23, 12, 13]
		CosSim(CL)	[23, 24, 25, 26, 27, 28, 29]
		+PRUNE&COMP	[23, 24, 25, 26, 27, 28, 29]
		Mag+	[7, 6, 11, 8, 9, 4, 10]
		+PRUNE&COMP	[7, 11, 12, 14, 15, 16, 17]
		Taylor+	[27, 29, 28, 26, 25, 24, 23]
		+PRUNE&COMP	[27, 4, 29, 26, 25, 24, 12]
		CosSim(BI)	[27, 26, 24, 28, 29, 25, 23]
		+PRUNE&COMP	[27, 26, 24, 23, 22, 21, 29]
	9/32	PPL	[14, 12, 11, 24, 23, 10, 7, 18, 15]
		+PRUNE&COMP	[14, 11, 24, 10, 23, 12, 13, 27, 9]
		CosSim(CL)	[21, 22, 23, 24, 25, 26, 27, 28, 29]
		+PRUNE&COMP	[21, 22, 23, 24, 25, 26, 27, 28, 29]
		Mag+	[7, 6, 11, 8, 9, 4, 10, 12, 14]
		+PRUNE&COMP	[7, 11, 12, 14, 15, 16, 17, 19, 20]
		Taylor+	[27, 29, 28, 26, 25, 24, 23, 22, 21]
		+PRUNE&COMP	[27, 4, 29, 26, 25, 24, 12, 23, 22]
		CosSim(BI)	[27, 26, 24, 28, 29, 25, 23, 22, 21]
		+PRUNE&COMP	[27, 26, 24, 23, 22, 21, 29, 25, 19]
LLaMA-3-8B	-	Dense	-
	5/32	PPL	[10, 11, 8, 9, 12]
		+PRUNE&COMP	[10, 25, 11, 26, 12]
		CosSim(CL)	[23, 24, 25, 26, 27]
		+PRUNE&COMP	[23, 24, 25, 26, 27]
		Mag+	[5, 8, 7, 4, 11]
		+PRUNE&COMP	[5, 8, 11, 13, 14]
		Taylor+	[29, 28, 27, 26, 25]
		+PRUNE&COMP	[29, 28, 4, 25, 23]
		CosSim(BI)	[25, 24, 26, 27, 23]
		+PRUNE&COMP	[25, 26, 24, 23, 22]
	7/32	PPL	[10, 11, 8, 9, 12, 25, 26]
		+PRUNE&COMP	[10, 25, 11, 26, 12, 9, 8]
		CosSim(CL)	[23, 24, 25, 26, 27, 28, 29]
		+PRUNE&COMP	[23, 24, 25, 26, 27, 28, 29]
		Mag+	[5, 8, 7, 4, 11, 6, 10]
		+PRUNE&COMP	[5, 8, 11, 13, 14, 16, 18]
		Taylor+	[29, 28, 27, 26, 25, 24, 23]
		+PRUNE&COMP	[29, 28, 4, 25, 23, 26, 24]
		CosSim(BI)	[25, 24, 26, 27, 23, 28, 22, 29, 20]
		+PRUNE&COMP	[25, 26, 24, 23, 22, 20, 28]

Table 3: Layer pruning configurations of LLaMA-2-7B and LLaMA-3-8B.

Model	Sparsity	Metric	Layer Index
LLaMA-2-13B	-	Dense	-
	8/40	PPL	[31, 28, 27, 29, 30, 33, 32, 25]
		+PRUNE&COMP	[31, 27, 13, 12, 33, 14, 28, 11]
		CosSim(CL)	[28, 29, 30, 31, 32, 33, 34, 35]
		+PRUNE&COMP	[28, 29, 30, 31, 32, 33, 34, 35]
		Mag+	[5, 4, 6, 10, 7, 8, 9, 13]
		+PRUNE&COMP	[5, 6, 10, 13, 14, 16, 17, 19]
		Taylor+	[33, 35, 34, 32, 36, 37, 30, 31]
		+PRUNE&COMP	[33, 4, 37, 5, 35, 32, 30, 29]
		CosSim(BI)	[33, 32, 31, 30, 29, 34, 35, 28]
		+PRUNE&COMP	[33, 32, 31, 30, 29, 28, 27, 26]
	12/40	PPL	[31, 28, 27, 29, 30, 33, 32, 25, 26, 34, 13, 19]
		+PRUNE&COMP	[31, 27, 13, 12, 33, 14, 28, 11, 35, 10, 26, 29]
		CosSim(CL)	[25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36]
		+PRUNE&COMP	[25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36]
		Mag+	[5, 4, 6, 10, 7, 8, 9, 13, 12, 11, 14, 16]
		+PRUNE&COMP	[5, 6, 10, 13, 14, 16, 17, 19, 21, 23, 24, 25]
		Taylor+	[33, 35, 34, 32, 36, 37, 30, 31, 29, 27, 28, 26]
		+PRUNE&COMP	[33, 4, 37, 5, 35, 32, 30, 29, 27, 31, 28, 26]
		CosSim(BI)	[33, 32, 31, 30, 29, 34, 35, 28, 27, 36, 26, 25]
		+PRUNE&COMP	[33, 32, 31, 30, 29, 28, 27, 26, 25, 35, 24, 34]
Qwen3-8B	-	Dense	-
	5/36	PPL	[21, 20, 15, 17, 16]
		+PRUNE&COMP	[21, 17, 16, 15, 18]
		CosSim(CL)	[28, 29, 30, 31, 32]
		+PRUNE&COMP	[28, 29, 30, 31, 32]
		Mag+	[4, 19, 21, 18, 20]
		+PRUNE&COMP	[4, 19, 21, 22, 23]
		Taylor+	[4, 29, 26, 27, 30]
		+PRUNE&COMP	[4, 29, 26, 27, 30]
		CosSim(BI)	[20, 21, 2, 17, 16]
		+PRUNE&COMP	[20, 21, 2, 17, 18]
	7/36	PPL	[21, 20, 15, 17, 16, 19, 18]
		+PRUNE&COMP	[21, 17, 16, 15, 18, 19, 2]
		CosSim(CL)	[26, 27, 28, 29, 30, 31, 32]
		+PRUNE&COMP	[26, 27, 28, 29, 30, 31, 32]
		Mag+	[4, 19, 21, 18, 20, 15, 17]
		+PRUNE&COMP	[4, 19, 21, 22, 23, 24, 25]
		Taylor+	[4, 29, 26, 27, 30, 28, 31]
		+PRUNE&COMP	[4, 29, 26, 27, 30, 31, 28]
		CosSim(BI)	[20, 21, 2, 17, 16, 18, 19]
		+PRUNE&COMP	[20, 21, 2, 17, 18, 19, 16]

Table 4: Layer pruning configurations of LLaMA-2-13B and Qwen3-8B.

Method	**ARC-c**	**ARC-e**	**BoolQ**	**CoPa**	**HeSw**	**PIQA**	**Race-h**	**WG**	**WSC**	**Average**
Naive (One-shot pruning)	45.56	63.51	73.12	79.00	70.13	74.92	36.94	71.19	75.09	65.50
+IterPrune	46.67	67.93	73.30	84.00	70.36	74.59	35.31	70.48	79.12	66.86
+MagComp	**48.63**	69.99	**74.07**	**85.00**	**72.62**	**76.01**	37.51	72.53	78.75	68.35
+MagComp +IterPrune	46.50	**70.54**	71.31	84.00	72.43	**76.01**	**37.99**	**73.32**	**83.88**	**68.44**

Table 5: Ablation Study on QA benchmarks shows effectiveness of the proposed PRUNE&COMP. 7 layers of LLaMA-3-8B are pruned with CosSim(BI) metric.

Size	**WikiText-2**	**C4**	**PTB**
64	11.87	14.87	16.93
128	11.87	14.87	16.93
256	11.87	14.87	16.93

Table 6: Ablation Study on Size of Calibration Set. 5 layers of LLaMA-3-8B are pruned with CosSim(BI) metric.

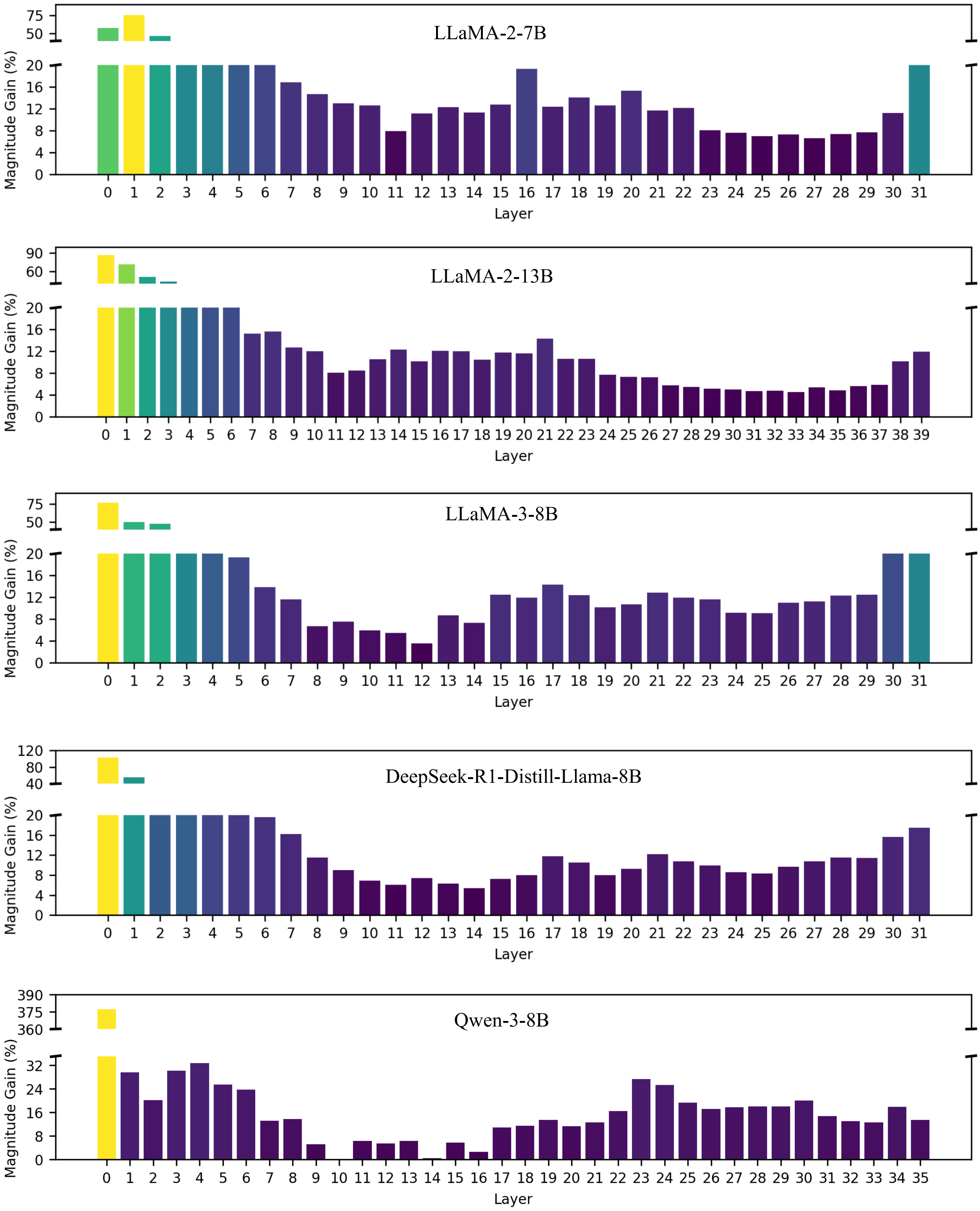

Figure 1: Visualization on the channel-wise averaged magnitude gain ratio of each layer. All layers produce magnitude gain.