further unlock the benefits of MatQuant, even during large-scale pretraining.

6. Conclusions

In this work, we presented MatQuant, a novel multi-scale training technique that leverages the nested structure of integer data types to simultaneously optimize model weight quantization across multiple precisions (int8, int4, and int2) within a single model. This general-purpose method, applicable to learning-based quantization techniques like OmniQuant and QAT, produces models with comparable accuracy to baselines for int8 and int4, while achieving significant improvements, up to 10% (using co-distillation), for int2 models. MatQuant further enables bit-width interpolation and layer-wise mix-and-match for flexible accuracy-cost trade-offs, promising more efficient deployment of large models across various hardware settings. Finally, MatQuant also helped discover Single Precison MatQuant, which significantly improves standalone low-bit quantization.

Acknowledgments

We are grateful to Varun Yerram, Shreya Pathak and Devvrit for assistance in setting up inference pipelines, Praneeth Netrapalli, Rakesh Shivanna, Tom Duerig, Abhijit Ogale, Jon Shlens, Ali Farhadi and Rahul Sukthankar for helpful discussions, support and feedback.

References

A. Abdolrashidi, L. Wang, S. Agrawal, J. Malmaud, O. Rybakov, C. Leichner, and L. Lew. Pareto-optimal quantized resnet is mostly 4-bit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3091–3099, 2021.

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984.

H. Adepu, Z. Zeng, L. Zhang, and V. Singh. Framequant: Flexible low-bit quantization for transformers. *arXiv preprint arXiv:2403.06082*, 2024.

S. Ashkboos, A. Mohtashami, M. L. Croci, B. Li, M. Jaggi, D. Alistarh, T. Hoefler, and J. Hensman. Quarot: Outlier-free 4-bit inference in rotated llms. *CoRR*, abs/2404.00456, 2024. doi: 10.48550/ARXIV.2404.00456. URL https://doi.org/10.48550/arXiv.2404.00456.

Y. Bengio, N. Léonard, and A. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Y. Bisk, R. Zellers, R. L. Bras, J. Gao, and Y. Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6239. URL https://doi.org/10.1609/aaai.v34i05.6239.

J. Chee, Y. Cai, V. Kuleshov, and C. M. De Sa. Quip: 2-bit quantization of large language models with guarantees. *Advances in Neural Information Processing Systems*, 36, 2024.

M. Chen, W. Shao, P. Xu, J. Wang, P. Gao, K. Zhang, Y. Qiao, and P. Luo. Efficientqat: Efficient quantization-aware training for large language models. *CoRR*, abs/2407.11062, 2024. doi: 10.48550/ARXIV.2407.11062. URL https://doi.org/10.48550/arXiv.2407.11062.

C. Clark, K. Lee, M. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In J. Burstein, C. Doran, and T. Solorio,

editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2924–2936. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1300. URL https://doi.org/10.18653/v1/n19-1300.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018. URL http://arxiv.org/abs/1803.05457.

E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems*, 28, 2015.

T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.

F. Devvrit, S. Kudugunta, A. Kusupati, T. Dettmers, K. Chen, I. Dhillon, Y. Tsvetkov, H. Hajishirzi, S. Kakade, A. Farhadi, P. Jain, et al. Matformer: Nested transformer for elastic inference. *arXiv preprint arXiv:2310.07707*, 2023.

D. Du, Y. Zhang, S. Cao, J. Guo, T. Cao, X. Chu, and N. Xu. Bitdistiller: Unleashing the potential of sub-4-bit llms via self-distillation. In L. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 102–116. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.7. URL https://doi.org/10.18653/v1/2024.acl-long.7.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

G. G Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Gemma-Team. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118, 2024. URL https://api.semanticscholar.org/CorpusID:270843326.

B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.2310.06825.

S. Kim, C. Hooper, A. Gholami, Z. Dong, X. Li, S. Shen, M. W. Mahoney, and K. Keutzer. Squeezellm: Dense-and-sparse quantization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=0jpbpFia8m.

A. Kusupati, G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

J. Lin, J. Tang, H. Tang, S. Yang, X. Dang, and S. Han. Awq: Activation-aware weight quan-

tization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2023.

T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

Z. Liu, B. Oguz, C. Zhao, E. Chang, P. Stock, Y. Mehdad, Y. Shi, R. Krishnamoorthi, and V. Chandra. LLM-QAT: data-free quantization aware training for large language models. In L. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 467–484. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.FINDINGS-ACL.26. URL https://doi.org/10.18653/v1/2024.findings-acl.26.

Z. Liu, C. Zhao, I. Fedorov, B. Soran, D. Choudhary, R. Krishnamoorthi, V. Chandra, Y. Tian, and T. Blankevoort. Spinquant: LLM quantization with learned rotations. *CoRR*, abs/2405.16406, 2024b. doi: 10.48550/ARXIV.2405.16406. URL https://doi.org/10.48550/arXiv.2405.16406.

Y. Ma, H. Li, X. Zheng, F. Ling, X. Xiao, R. Wang, S. Wen, F. Chao, and R. Ji. Affinequant: Affine transformation quantization for large language models. *arXiv preprint arXiv:2403.12544*, 2024.

P. A. Nair and A. S. Suggala. Cdquant: Accurate post-training weight quantization of large pre-trained models using greedy coordinate descent. *CoRR*, abs/2406.17542, 2024. doi: 10.48550/ARXIV.2406.17542. URL https://doi.org/10.48550/arXiv.2406.17542.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

O. Rippel, M. Gelbart, and R. Adams. Learning ordered representations with nested dropout. In *International Conference on Machine Learning*, pages 1746–1754. PMLR, 2014.

K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi. Winogrande: An adversarial winograd schema challenge at scale. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8732–8740. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6399. URL https://doi.org/10.1609/aaai.v34i05.6399.

W. Shao, M. Chen, Z. Zhang, P. Xu, L. Zhao, Z. Li, K. Zhang, P. Gao, Y. Qiao, and P. Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*, 2023.

A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:13756489.

G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, pages 38087–38099. PMLR, 2023.

J. Yu, L. Yang, N. Xu, J. Yang, and T. Huang. Slimmable neural networks. *arXiv preprint arXiv:1812.08928*, 2018.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence? In A. Korhonen, D. R. Traum, and L. Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1472. URL https://doi.org/10.18653/v1/p19-1472.

A. Addition Training Details

We run all our experiments on TPUv5e chips. For OmniQuant experiments, we use a constant learning rate of $1e-3$ and for QAT experiments, we linearly warmup the learning rate to $1e-5$ for 150 and use a consine decay schedule thereafter. For OmniQuant experiments, we sample 128 examples with a sequence length of 2048 from the C4 dataset (Raffel et al., 2020) and train using a batch size of 4. We train for a total of 10M tokens for all models except the int2 baseline, where we train the model for 20M tokens (Shao et al., 2023). For QAT experiments, we sample a fixed set of 100M tokens from the C4 dataset and train all our models using a batch size of 16 and a sequence length of 8192 for a single epoch. For Attn + FFN experiments with QAT, we sample a fixed set of 300M tokens from C4 and train with a batch size of 16 for a single epoch.

Mix'n'Match For a fixed effective bits-per-FFN layer, where each layer was quantized to either int2, int4, or int8, we explored four different quantization strategies: Pyramid, Reverse Pyramid, Increasing, and Decreasing. In the Pyramid strategy, the initial and final layers were quantized to int2, the central layers to int8, with int4 serving as an intermediate step. The Reverse Pyramid strategy followed the opposite approach, assigning int8 to the initial and final layers, int2 to the central layers, and int4 in between. The Increasing and Decreasing strategies assigned bit precision in ascending and descending order, respectively, across the layers. Our experimental results demonstrated that, for a given effective bits per FFN layer, the Pyramid strategy consistently outperformed the others. Allocating higher precision (int8) to the middle layers helped preserve critical information, while the initial and final layers performed adequately with lower bit precision (int2 and int4), leading to a more efficient and effective quantization scheme.

B. Detailed Downstream Evaluations for OmniQuant ad QAT

Tables 7, 8, 9, 10, 11, and 12 present downstream evaluation results on Gemma-2 2B, Gemma-2 9B and Mistral 7B with OmniQuant and QAT.

C. Detailed Downstream Evaluations for MatQuant Re-weighting

Tables 13, 14, and 15 present downstream evaluation results for OmniQuant reweighting experiments on Gemma-2 2B, Gemma-2 9B and Mistral 7B.

D. Detailed Downstream Evaluations for Co-Distillation

Tables 16 and 17 present the downstream evaluation and perplexity results and for MatQuant co-distillation on Gemma-2 9B with OmniQuant and QAT.

E. Detailed Evaluations for FFN + Attention Quantization

Tables 18 and 19 present the downstream evaluation and perplexity results for FFN + Attention quantization on Gemma-2 9B and Mistral 7B with OmniQuant and QAT.

F. Detailed Evaluation for Single Precison MatQuant

Tables 20, 21, 22, and 23 present the downstream evaluation results comparing Single Precison MatQuant to MatQuant and the *Baseline* for int2 quantization of Gemma-2 2B, Gemma-2 9B and Mistral 7B with OmniQuant and QAT. Since Single Precison MatQuant slices 2 bits from an 8-bit model and computes loss only over the first two bits, we can evaluate the Single Precison MatQuant model trained for 2-bits on int4 and int8. Downstream evaluation and perplexity results for this are presented in Tables 21 and 22. We also plot the weight distribution for Single Precison MatQuant in Figure 3.

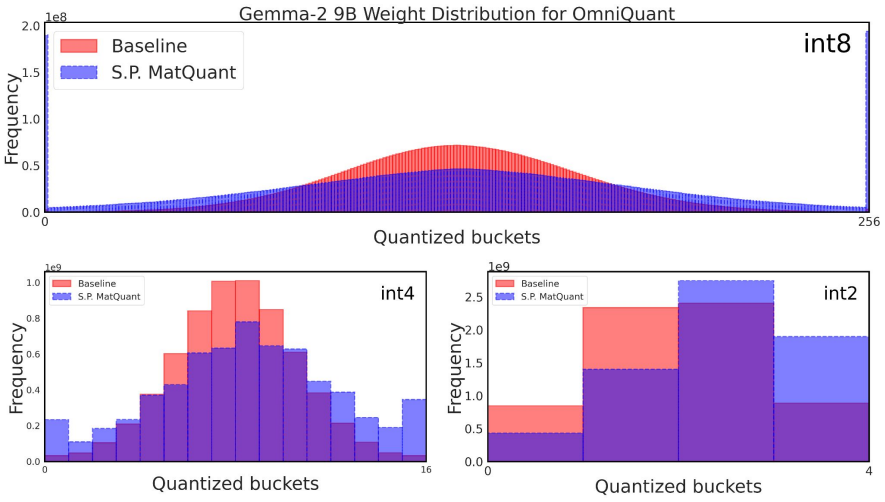

Figure 3 | The Figure presents the weight distribution for Gemma-2 9B when trained with Single Precison MatQuant for int2 quantization. The right-shifted quantized weight distribution is a consequence of Single Precison MatQuant's training mechanism that heavily optimizes for the first 2 MSBs of the int8 representation.

Table 7 | Table presents the downstream evaluation results for MatQuant when applied to OmniQuant on Gemma-2 2B.

Data type	Method	Gemma-2 2B						
	OmniQuant	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average
bfloat16		50.09	71.59	76.45	69.69	78.29	63.14	68.21
int8	Baseline	50	71.46	76.36	69.76	78.24	63.69	68.25
	MatQuant	48.04	71.8	75.78	67.64	78.07	63.22	67.42
int4	Sliced int8	41.81	66.2	71.35	62.64	75.95	59.91	62.98
	Baseline	48.46	70.96	74.22	67.66	77.26	63.61	67.03
	MatQuant	45.65	70.29	74.8	66.07	77.58	62.27	66.11
int2	Sliced int8	23.81	23.53	53.06	24.78	51.8	49.09	37.68
	Baseline	31.31	53.58	62.2	40.78	66.05	54.06	51.33
	MatQuant	34.39	59.64	62.69	52.11	69.86	55.56	55.71
int6	Sliced int8	48.55	71.25	75.87	69.18	78.35	62.75	67.66
	Baseline	49.32	71.76	76.48	69.52	78.56	62.75	68.06
	MatQuant	47.1	71.46	76.02	67.47	77.91	63.61	67.26
int3	Sliced int8	23.21	34.43	58.2	30.48	56.69	49.01	42
	Baseline	46.25	68.64	72.97	62.24	76.06	60.06	64.37
	MatQuant	44.45	68.56	69.11	62.28	75.95	62.59	63.82

Table 8 | Table presents the downstream evaluation results for MatQuant when applied to OmniQuant on Gemma-2 9B.

Data type	Method	Gemma-2 9B						
	OmniQuant	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average
bfloat16		58.96	77.57	83.33	77.31	81.12	67.96	74.38
int8	Baseline	59.47	77.31	83.94	77.35	81.39	68.11	74.59
	MatQuant	58.11	78.03	83.27	76.17	81.18	67.09	73.97
int4	Sliced int8	55.97	75.04	81.19	73.81	80.52	66.61	72.19
	Baseline	58.79	78.37	83.55	76.71	81.45	67.09	74.33
	MatQuant	57.25	77.36	84.86	75.52	81.5	66.77	73.88
int2	Sliced int8	23.21	24.92	38.13	25.37	51.36	51.54	35.75
	Baseline	39.16	63.43	72.11	52.24	72.63	61.88	60.24
	MatQuant	48.72	72.18	79.2	68.11	76.17	66.77	68.52
int6	Sliced int8	59.04	77.53	84.68	77.1	81.23	68.11	74.61
	Baseline	59.22	77.27	83.21	77.1	81.12	67.48	74.23
	MatQuant	58.87	78.03	83.61	76.18	81.45	67.09	74.21
int3	Sliced int8	35.84	57.32	67.61	48.58	68.61	56.59	55.76
	Baseline	57.17	77.06	83.79	74.45	80.36	66.54	73.23
	MatQuant	55.46	76.14	84.04	74.49	80.14	67.32	72.93

Table 9 | Table presents the downstream evaluation results for MatQuant when applied to OmniQuant on Mistral 7B.

Data type	Method	Mistral 7B						
	OmniQuant	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average
bfloat16		49.57	73.74	84.4	80.61	81.18	74.43	73.99
int8	Baseline	49.23	73.19	83.88	80.41	81.39	74.51	73.77
	MatQuant	48.04	73.44	84.13	79.37	81.12	74.66	73.46
int4	Sliced int8	27.65	46.72	49.17	36.88	64.09	55.01	46.59
	Baseline	49.23	73.23	83.94	79.9	81.34	74.11	73.62
	MatQuant	48.21	72.69	83.49	78.82	81.12	74.43	73.13
int2	Sliced int8	23.72	25.29	43.21	25.45	50.49	49.33	36.25
	Baseline	36.69	61.36	70.06	57.47	70.67	62.19	59.74
	MatQuant	41.38	67.42	71.62	71.98	77.86	65.67	65.99
int6	Sliced int8	48.98	72.01	83.46	79.95	81.72	74.9	73.5
	Baseline	50.26	73.65	84.04	80.55	81.66	74.43	74.1
	MatQuant	48.46	72.98	84.07	79.64	81.18	75.22	73.59
int3	Sliced int8	22.78	24.66	37.86	24.12	49.24	48.93	34.6
	Baseline	46.33	70.71	82.72	77.74	80.74	71.82	71.68
	MatQuant	45.65	71.21	80.43	78.31	81.07	72.61	71.55

Table 10 | Table presents the downstream evaluation results for MatQuant when applied to QAT on Gemma-2 2B.

Data type	Method	Gemma-2 2B						
	QAT	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average
bfloat16		50.09	71.59	76.45	69.69	78.29	63.14	68.21
int8	Baseline	47.78	70.66	75.08	69.92	78.35	65.11	67.82
	MatQuant	46.25	71.21	75.6	69.97	78.4	64.64	67.68
int4	Sliced int8	46.08	69.36	75.78	68.05	78.18	65.75	67.2
	Baseline	46.16	71.59	73.73	68.72	78.62	63.38	67.03
	MatQuant	44.37	70.45	75.81	68.43	78.35	64.88	67.05
int2	Sliced int8	25.6	26.3	57.98	25.82	52.12	50.2	39.67
	Baseline	24.66	43.22	62.17	38.39	64.42	53.59	47.74
	MatQuant	28.24	51.73	64.19	46.76	68.66	55.01	52.43
int6	Sliced int8	47.78	70.79	74.25	69.73	77.64	65.11	67.55
	Baseline	47.7	70.88	74.92	69.72	78.07	65.19	67.75
	MatQuant	46.5	70.71	75.72	69.69	78.02	64.96	67.6
int3	Sliced int8	38.74	63.13	65.57	58.86	74.81	60.3	60.23
	Baseline	39.68	65.28	67.03	62.68	77.04	58.8	61.75
	MatQuant	38.65	67.34	70.49	61.47	75.41	61.72	62.51

Table 11 | Table presents the downstream evaluation results for MatQuant when applied to QAT on Gemma-2 9B.

Data type	Method	Gemma-2 9B						
	QAT	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average
bfloat16		58.96	77.57	83.33	77.31	81.12	67.96	74.38
int8	Baseline	58.11	75.38	80.12	78.7	81.5	71.19	74.17
	MatQuant	58.19	76.18	81.5	79.57	82.15	71.03	74.77
int4	Sliced int8	57.42	75.08	78.1	76.97	81.23	70.72	73.25
	Baseline	56.91	75.42	75.38	78.06	81.39	72.38	73.26
	MatQuant	57.94	76.64	75.2	78.71	81.66	72.14	73.71
int2	Sliced int8	23.89	27.61	57.95	30.16	54.68	47.83	40.35
	Baseline	33.45	55.43	62.26	54.8	70.51	59.67	56.02
	MatQuant	39.85	65.66	65.93	64.08	75.68	62.75	62.32
int6	Sliced int8	57.85	75.13	80.67	78.63	81.56	70.88	74.12
	Baseline	57.94	76.14	79.63	78.93	82.1	71.11	74.31
	MatQuant	58.02	75.63	81.31	79.43	81.66	71.27	74.55
int3	Sliced int8	50	68.1	75.2	71.31	79.43	67.4	68.57
	Baseline	53.07	75.04	66.61	74.94	80.03	69.69	69.9
	MatQuant	51.62	71.93	78.78	73.99	80.14	67.64	70.68

Table 12 | Table presents the downstream evaluation results for MatQuant when applied to QAT on Mistral 7B.

Data type	Method	Mistral 7B						
	QAT	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average
bfloat16		49.57	73.74	84.4	80.61	81.18	74.43	73.99
int8	Baseline	48.89	71.63	82.42	81.69	81.18	75.06	73.48
	MatQuant	46.76	70.37	82.51	79.73	80.9	74.19	72.41
int4	Sliced int8	47.18	70.41	80.37	79.84	80.25	72.93	71.83
	Baseline	47.27	70.62	81.28	78.95	81.12	73.56	72.13
	MatQuant	45.65	68.64	82.02	79	81.07	73.4	71.63
int2	Sliced int8	25.34	26.47	54.95	25.18	48.48	49.96	38.4
	Baseline	29.78	48.23	64.5	55.11	70.84	61.25	54.95
	MatQuant	34.3	55.09	71.83	65.89	75.52	65.11	61.29
int6	Sliced int8	48.21	71.51	82.42	81.67	81.72	74.27	73.3
	Baseline	47.7	71.3	82.23	79.84	80.79	74.43	72.71
	MatQuant	47.53	71	81.9	79.73	81.28	74.74	72.7
int3	Sliced int8	40.1	61.49	72.91	68.72	77.97	70.56	65.29
	Baseline	44.54	67.97	73.98	76.31	79.65	70.48	68.82
	MatQuant	38.82	62.42	77.74	71.1	78.07	70.48	66.44

Table 13 | Tables presents the downstream evaluation results on Gemma-2 2B for MatQuant loss reweighting when applied to OmniQuant. Weightings: $(x, y, z) \rightarrow (\lambda_8, \lambda_4, \lambda_2)$ (from Equation 7).

		Gemma-2 2B						
Data type	Weightings	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average
int8	$(1, 1, 1)$	48.04	71.8	75.78	67.64	78.07	63.22	67.42
	$(1\sqrt{2}, \sqrt{2})$	47.35	71.34	75.66	67.99	78.07	63.38	67.3
	$(\sqrt{2}, 1, \sqrt{2})$	47.44	72.43	76.02	67.45	78.02	63.85	67.54
	$(1, 1\sqrt{2})$	47.7	71.89	75.63	67.21	78.07	63.38	67.31
	$(2, 2, 1)$	48.38	72.31	76.3	68.32	78.35	63.46	67.85
	$(\sqrt{2}, 2, 1)$	48.46	71.84	75.93	68.35	77.91	63.14	67.6
	$(2, \sqrt{2}, 1)$	47.95	71.72	75.26	68.13	78.07	62.75	67.31
	$(\sqrt{2}, \sqrt{2}, 1)$	47.35	71.34	75.66	67.99	78.07	63.38	67.3
int4	$(1, 1, 1)$	45.65	70.29	74.8	66.07	77.58	62.27	66.11
	$(1\sqrt{2}, \sqrt{2})$	46.33	70.92	73.7	67.67	77.26	62.9	66.46
	$(\sqrt{2}, 1, \sqrt{2})$	46.42	70.96	74.71	65.78	77.58	63.14	66.43
	$(1, 1\sqrt{2})$	45.56	71.55	75.75	66.18	77.48	63.69	66.7
	$(2, 2, 1)$	46.84	70.88	74.92	66.48	77.91	62.19	66.54
	$(\sqrt{2}, 2, 1)$	47.35	71.68	72.69	66.79	77.26	63.38	66.52
	$(2, \sqrt{2}, 1)$	45.9	70.83	75.11	66.97	77.37	62.27	66.41
	$(\sqrt{2}, \sqrt{2}, 1)$	46.33	70.92	73.7	67.67	77.26	62.9	66.46
int2	$(1, 1, 1)$	34.39	59.64	62.69	52.11	69.86	55.56	55.71
	$(1\sqrt{2}, \sqrt{2})$	32.76	56.99	63.46	51.99	70.29	56.27	55.29
	$(\sqrt{2}, 1, \sqrt{2})$	35.07	62.04	65.78	54.26	71.65	56.27	57.51
	$(1, 1\sqrt{2})$	34.22	60.4	64.98	54.3	71.38	57.22	57.08
	$(2, 2, 1)$	34.47	57.95	63.94	51.84	69.75	56.27	55.7
	$(\sqrt{2}, 2, 1)$	33.45	57.49	65.02	52.22	70.4	55.64	55.7
	$(2, \sqrt{2}, 1)$	34.04	58.84	65.11	51.77	70.89	57.14	56.3
	$(\sqrt{2}, \sqrt{2}, 1)$	32.76	56.99	63.46	51.99	70.29	56.27	55.29
int6	$(1, 1, 1)$	47.1	71.46	76.02	67.47	77.91	63.61	67.26
	$(1\sqrt{2}, \sqrt{2})$	47.44	71.42	74.95	67.85	77.86	63.3	67.14
	$(\sqrt{2}, 1, \sqrt{2})$	47.61	71.89	75.9	67.37	78.24	63.77	67.46
	$(1, 1\sqrt{2})$	47.78	71.63	75.47	67.2	77.86	63.61	67.26
	$(2, 2, 1)$	48.55	72.69	76.3	68.02	78.67	63.85	68.01
	$(\sqrt{2}, 2, 1)$	48.29	71.76	75.72	68.42	78.02	63.38	67.6
	$(2, \sqrt{2}, 1)$	48.38	71.51	75.84	68.24	78.18	63.85	67.67
	$(\sqrt{2}, \sqrt{2}, 1)$	47.44	71.42	74.95	67.85	77.86	63.3	67.14
int3	$(1, 1, 1)$	44.45	68.56	69.11	62.28	75.95	62.59	63.82
	$(1\sqrt{2}, \sqrt{2})$	43.17	68.73	64.74	61.31	76.39	61.48	62.64
	$(\sqrt{2}, 1, \sqrt{2})$	41.98	68.6	70.34	61.95	75.9	63.3	63.68
	$(1, 1\sqrt{2})$	41.64	66.71	71.62	61.94	76.01	61.09	63.17
	$(2, 2, 1)$	41.98	68.35	68.41	63.74	76.17	60.77	63.24
	$(\sqrt{2}, 2, 1)$	42.66	66.54	70.46	63.61	75.63	62.98	63.65
	$(2, \sqrt{2}, 1)$	43.17	66.71	60.03	62.71	76.77	61.64	61.84
	$(\sqrt{2}, \sqrt{2}, 1)$	43.17	68.73	64.74	61.31	76.39	61.48	62.64

Table 14 | Tables presents the downstream evaluation results on Gemma-2 9B for MatQuant loss reweighting when applied to OmniQuant. Weightings: $(x, y, z) \rightarrow (\lambda_8, \lambda_4, \lambda_2)$ (from Equation 7).

		Gemma-2 9B						
Data type	Weightings	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average
int8	$(1, 1, 1)$	58.11	78.03	83.27	76.17	81.18	67.09	73.97
	$(1\sqrt{2}, \sqrt{2})$	57.68	77.4	83.73	76.1	81.18	67.64	73.95
	$(\sqrt{2}, 1, \sqrt{2})$	58.11	77.86	81.04	76	81.18	67.09	73.55
	$(1, 1\sqrt{2})$	56.91	77.1	82.39	75.93	81.18	67.17	73.45
	$(2, 2, 1)$	58.79	77.48	82.66	76.55	81.23	67.4	74.02
	$(\sqrt{2}, 2, 1)$	58.53	77.31	82.63	76.54	80.96	67.56	73.92
	$(2, \sqrt{2}, 1)$	58.62	77.27	84.31	76.54	81.34	66.85	74.16
	$(\sqrt{2}, \sqrt{2}, 1)$	59.13	78.07	84.16	76.46	80.9	67.25	74.33
int4	$(1, 1, 1)$	57.25	77.36	84.86	75.52	81.5	66.77	73.88
	$(1\sqrt{2}, \sqrt{2})$	56.74	77.74	85.08	75.5	80.85	66.85	73.79
	$(\sqrt{2}, 1, \sqrt{2})$	57.42	78.28	82.51	75.97	81.34	67.56	73.85
	$(1, 1\sqrt{2})$	57.59	77.82	84.28	75.32	81.12	66.38	73.75
	$(2, 2, 1)$	58.62	78.28	83.67	76.01	81.5	67.88	74.33
	$(\sqrt{2}, 2, 1)$	58.19	77.82	83.91	76.62	81.99	67.72	74.37
	$(2, \sqrt{2}, 1)$	58.28	78.16	84.53	76.41	81.72	67.09	74.36
	$(\sqrt{2}, \sqrt{2}, 1)$	57.94	78.11	84.98	76.5	81.01	67.01	74.26
int2	$(1, 1, 1)$	48.72	72.18	79.2	68.11	76.17	66.77	68.52
	$(1\sqrt{2}, \sqrt{2})$	49.83	73.91	78.75	67.27	77.2	66.46	68.9
	$(\sqrt{2}, 1, \sqrt{2})$	48.55	74.24	81.5	68.44	76.5	65.9	69.19
	$(1, 1\sqrt{2})$	48.29	72.94	74.74	68.34	77.58	65.67	67.93
	$(2, 2, 1)$	46.76	73.27	71.96	67.98	76.77	63.61	66.72
	$(\sqrt{2}, 2, 1)$	46.76	73.7	77.65	67.01	77.58	65.98	68.11
	$(2, \sqrt{2}, 1)$	46.76	72.35	75.35	67.51	76.39	67.56	67.65
	$(\sqrt{2}, \sqrt{2}, 1)$	46.59	72.6	79.3	67.58	77.69	65.75	68.25
int6	$(1, 1, 1)$	58.87	78.03	83.61	76.18	81.45	67.09	74.21
	$(1\sqrt{2}, \sqrt{2})$	57.51	77.53	83.55	75.98	80.9	67.17	73.77
	$(\sqrt{2}, 1, \sqrt{2})$	58.79	77.82	81.38	76.21	81.07	67.72	73.83
	$(1, 1\sqrt{2})$	57.34	77.23	82.57	75.89	81.12	67.17	73.55
	$(2, 2, 1)$	59.04	77.4	82.66	76.55	81.56	68.03	74.21
	$(\sqrt{2}, 2, 1)$	59.22	77.65	82.17	76.62	81.23	67.8	74.11
	$(2, \sqrt{2}, 1)$	58.36	77.82	83.79	76.47	81.23	67.25	74.15
	$(\sqrt{2}, \sqrt{2}, 1)$	59.3	78.37	84.5	76.57	80.85	67.4	74.5
int3	$(1, 1, 1)$	55.46	76.14	84.04	74.49	80.14	67.32	72.93
	$(1\sqrt{2}, \sqrt{2})$	56.23	76.05	82.6	74.85	80.9	67.01	72.94
	$(\sqrt{2}, 1, \sqrt{2})$	56.4	77.86	80.64	75.11	79.87	68.51	73.06
	$(1, 1\sqrt{2})$	55.63	76.05	82.39	74.21	80.3	67.17	72.62
	$(2, 2, 1)$	55.2	76.56	84.19	74.87	80.2	67.72	73.12
	$(\sqrt{2}, 2, 1)$	54.44	75.63	80.55	74.97	80.96	67.72	72.38
	$(2, \sqrt{2}, 1)$	56.14	75.67	83.33	74.96	80.52	67.72	73.06
	$(\sqrt{2}, \sqrt{2}, 1)$	56.31	77.4	83.24	75.62	80.41	66.54	73.25