Table 15 | Tables presents the downstream evaluation results on Mistral 7B for MatQuant loss reweighting when applied to OmniQuant. Weightings: $(x, y, z) \rightarrow (\lambda_8, \lambda_4, \lambda_2)$ (from Equation 7).

		Mistral 7B						
Data type	Weightings	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average
int8	$(1, 1, 1)$	48.04	73.44	84.13	79.37	81.12	74.66	73.46
	$(1\sqrt{2}, \sqrt{2})$	48.46	73.19	84.28	79.19	81.12	74.74	73.5
	$(\sqrt{2}, 1, \sqrt{2})$	47.95	73.4	84.46	79.11	81.34	74.51	73.46
	$(1, 1\sqrt{2})$	48.21	73.02	84.34	79.03	81.28	74.59	73.41
	$(2, 2, 1)$	49.06	73.48	84.74	79.73	81.56	74.35	73.82
	$(\sqrt{2}, 2, 1)$	49.06	73.57	84.56	79.64	81.39	74.27	73.75
	$(2, \sqrt{2}, 1)$	48.98	73.95	84.50	79.60	81.61	74.90	73.92
	$(\sqrt{2}, \sqrt{2}, 1)$	48.98	73.86	84.56	79.55	81.23	74.74	73.82
int4	$(1, 1, 1)$	48.21	72.69	83.49	78.82	81.12	74.43	73.13
	$(1\sqrt{2}, \sqrt{2})$	49.15	72.81	83.39	78.71	80.79	74.66	73.25
	$(\sqrt{2}, 1, \sqrt{2})$	47.95	72.43	83.43	79.24	81.01	74.03	73.01
	$(1, 1\sqrt{2})$	48.46	73.44	84.07	78.9	81.01	73.88	73.29
	$(2, 2, 1)$	49.15	72.81	83.88	79.8	81.88	73.48	73.5
	$(\sqrt{2}, 2, 1)$	48.89	72.69	82.72	79.53	81.66	73.88	73.23
	$(2, \sqrt{2}, 1)$	47.87	72.05	83	79.56	81.23	74.27	73
	$(\sqrt{2}, \sqrt{2}, 1)$	48.29	72.47	82.84	79.52	81.07	73.64	72.97
int2	$(1, 1, 1)$	41.38	67.42	71.62	71.98	77.86	65.67	65.99
	$(1\sqrt{2}, \sqrt{2})$	40.78	66.2	73.61	72.68	77.75	67.4	66.4
	$(\sqrt{2}, 1, \sqrt{2})$	40.36	67.09	75.35	72.46	77.48	65.9	66.44
	$(1, 1\sqrt{2})$	40.36	67.17	74.83	71.64	77.53	66.14	66.28
	$(2, 2, 1)$	37.2	62.46	67.74	70.29	76.55	66.69	63.49
	$(\sqrt{2}, 2, 1)$	37.29	64.35	61.1	68.88	74.86	65.19	61.94
	$(2, \sqrt{2}, 1)$	39.68	65.24	68.93	66.64	75.19	64.09	63.29
	$(\sqrt{2}, \sqrt{2}, 1)$	34.56	61.24	60.61	58.07	72.63	59.98	57.85
int6	$(1, 1, 1)$	48.46	72.98	84.07	79.64	81.18	75.22	73.59
	$(1\sqrt{2}, \sqrt{2})$	49.06	73.44	84.59	79.51	81.28	74.74	73.77
	$(\sqrt{2}, 1, \sqrt{2})$	47.95	73.48	84.43	79.28	81.45	75.14	73.62
	$(1, 1\sqrt{2})$	48.38	72.94	84.34	79.15	81.18	74.59	73.43
	$(2, 2, 1)$	48.46	72.94	84.13	79.89	81.5	74.9	73.64
	$(\sqrt{2}, 2, 1)$	48.81	73.48	84.34	79.67	81.34	74.9	73.76
	$(2, \sqrt{2}, 1)$	49.4	73.65	84.4	79.68	81.28	74.74	73.86
	$(\sqrt{2}, \sqrt{2}, 1)$	49.23	73.57	84.43	79.55	81.12	74.66	73.76
int3	$(1, 1, 1)$	45.65	71.21	80.43	78.31	81.07	72.61	71.55
	$(1\sqrt{2}, \sqrt{2})$	47.7	72.05	82.81	78.74	81.12	72.77	72.53
	$(\sqrt{2}, 1, \sqrt{2})$	46.33	72.43	81.8	79.03	82.1	73.4	72.51
	$(1, 1\sqrt{2})$	45.99	71.09	80.73	78.77	80.85	72.53	71.66
	$(2, 2, 1)$	47.95	73.36	82.57	79.31	81.39	74.9	73.25
	$(\sqrt{2}, 2, 1)$	44.45	69.7	82.11	77.68	80.2	71.74	70.98
	$(2, \sqrt{2}, 1)$	46.84	72.73	80.95	78.79	81.56	73.01	72.31
	$(\sqrt{2}, \sqrt{2}, 1)$	47.01	71.59	81.96	78.89	81.39	72.45	72.22

Table 16 | Table presents the downstream evaluation and perplexity results for our MatQuant co-distillation experiments on Gemma-2 9B with OmniQuant.

OmniQuant		Gemma-2 9B							
Data type	Config.	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average	log pplx.
int8	$[8, 4, 8 \rightarrow 2]$	57.59	77.27	81.83	75.48	81.01	67.25	73.4	2.467
	$[8, 4, 2, 8 \rightarrow 2]$	57.17	77.36	82.2	75.82	80.96	67.25	73.46	2.466
	$[8, 4, 2, 8 \rightarrow 4; 2]$	56.4	77.82	82.32	75.02	80.63	67.72	73.32	2.466
int4	$[8, 4, 8 \rightarrow 2]$	57.68	78.45	82.97	75.5	80.85	67.56	73.84	2.488
	$[8, 4, 2, 8 \rightarrow 2]$	57.51	77.61	80.46	74.74	81.12	66.61	73.01	2.495
	$[8, 4, 2, 8 \rightarrow 4; 2]$	56.57	77.99	82.54	74.77	80.58	66.3	73.12	2.518
int2	$[8, 4, 8 \rightarrow 2]$	48.81	74.03	81.65	68.1	77.48	65.11	69.2	2.796
	$[8, 4, 2, 8 \rightarrow 2]$	49.15	75.34	83.12	68.79	77.64	67.01	70.17	2.778
	$[8, 4, 2, 8 \rightarrow 4; 2]$	49.83	75.04	79.79	68.38	77.86	67.4	69.72	2.804
int6	$[8, 4, 8 \rightarrow 2]$	57.42	77.19	81.87	75.42	81.01	67.8	73.45	2.468
	$[8, 4, 2, 8 \rightarrow 2]$	57.51	77.48	82.32	75.88	81.07	66.61	73.48	2.467
	$[8, 4, 2, 8 \rightarrow 4; 2]$	56.4	78.03	82.63	75.14	80.79	67.4	73.4	2.498
int3	$[8, 4, 8 \rightarrow 2]$	55.63	75.88	80.12	74.01	80.36	67.96	72.33	2.549
	$[8, 4, 2, 8 \rightarrow 2]$	54.35	76.85	79.33	74.6	80.47	67.4	72.17	2.543
	$[8, 4, 2, 8 \rightarrow 4; 2]$	55.2	76.98	82.45	73.59	80.41	68.43	72.84	2.58

Table 17 | Table presents the downstream evaluation and perplexity results for our MatQuant co-distillation experiments on Gemma-2 9B with QAT.

QAT		Gemma-2 9B							
Data type	Config.	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average	log pplx.
int8	$[8, 4, 8 \rightarrow 2]$	58.11	76.43	81.25	79.12	82.05	71.35	74.72	2.298
	$[8, 4, 2, 8 \rightarrow 2]$	57.51	76.43	81.53	78.95	82.1	71.19	74.62	2.299
	$[8, 4, 2, 8 \rightarrow 4; 2]$	58.11	76.14	81.68	79.12	82.26	71.51	74.8	2.302
int4	$[8, 4, 8 \rightarrow 2]$	57.42	76.35	77.55	78.06	81.61	71.59	73.76	2.328
	$[8, 4, 2, 8 \rightarrow 2]$	56.91	75.8	78.44	77.76	81.39	72.38	73.78	2.329
	$[8, 4, 2, 8 \rightarrow 4; 2]$	57.51	75.76	75.96	77.96	81.72	71.98	73.48	2.33
int2	$[8, 4, 8 \rightarrow 2]$	39.51	65.03	66.88	63.37	75.08	61.01	61.81	2.74
	$[8, 4, 2, 8 \rightarrow 2]$	40.78	66.5	67.55	63.67	75.95	60.62	62.51	2.746
	$[8, 4, 2, 8 \rightarrow 4; 2]$	40.19	65.7	65.57	63.83	75.3	62.12	62.12	2.746
int6	$[8, 4, 8 \rightarrow 2]$	57.85	76.09	81.47	78.98	81.88	71.27	74.59	2.301
	$[8, 4, 2, 8 \rightarrow 2]$	57.17	75.97	82.2	79	81.83	71.9	74.68	2.302
	$[8, 4, 2, 8 \rightarrow 4; 2]$	57.42	76.09	82.29	78.95	82.10	71.27	74.69	2.305
int3	$[8, 4, 8 \rightarrow 2]$	51.96	71.55	78.07	73.17	79.43	66.93	70.18	2.485
	$[8, 4, 2, 8 \rightarrow 2]$	50.94	71.76	78.78	73.09	79.05	66.77	70.06	2.486
	$[8, 4, 2, 8 \rightarrow 4; 2]$	51.45	72.39	78.84	73.46	79.6	67.96	70.62	2.731

Table 18 | Table presents the downstream evaluation results for MatQuant FFN + Attention quantization on Gemma-2 9B with QAT.

Data type	Method	Gemma-2 9B						
		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	58.62	77.02	83.43	79.01	81.34	68.27	74.61
	MatQuant	59.04	77.9	84.4	78.76	81.12	69.22	75.07
int4	Sliced int8	57.42	76.73	81.62	76.02	80.58	68.98	73.56
	Baseline	56.06	74.96	79.27	77.83	80.25	69.53	72.98
	MatQuant	57.34	76.77	84.19	77.51	80.74	68.11	74.11
int2	Sliced int8	24.74	25.63	58.53	25.5	50.71	49.17	39.05
	Baseline	-	-	-	-	-	-	-
	S.P. MatQuant	24.91	41.62	62.26	40.87	63.38	53.67	47.78
	MatQuant	28.24	39.23	62.17	39.13	63.49	50.75	47.17
int6	Sliced int8	58.53	77.15	82.48	79.04	81.5	68.67	74.56
	Baseline	58.87	77.06	83.12	78.81	81.23	68.82	74.65
	MatQuant	59.81	77.9	84.8	78.68	81.07	67.96	75.04
int3	Sliced int8	43.6	64.98	72.66	66	75.95	62.19	64.23
	Baseline	-	-	-	-	-	-	-
	S.P. MatQuant	50.85	73.11	71.13	72.01	79.38	65.67	68.69
	MatQuant	45.22	69.32	78.5	68.72	76.01	63.85	66.94

Table 19 | Table presents the downstream evaluation results for MatQuant FFN + Attention quantization on Mistral 7B with QAT.

Data type	Method	Mistral 7B						
		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	72.9	83.49	80.26	81.28	75.22	73.73
	MatQuant	49.32	72.31	83.76	80.2	81.18	74.74	73.58
int4	Sliced int8	45.99	71.76	81.41	76.95	80.41	71.98	71.42
	Baseline	48.04	71.72	78.87	78.93	80.36	73.32	71.87
	MatQuant	47.01	69.95	82.02	76.81	80.25	72.93	71.5
int2	Sliced int8	22.78	24.03	58.75	24.63	50.54	49.64	38.39
	Baseline	-	-	-	-	-	-	-
	S.P. MatQuant	23.21	23.82	37.83	24.67	49.02	49.57	34.69
	MatQuant	22.27	32.49	62.02	32.43	59.3	51.46	43.33
int6	Sliced int8	49.32	73.53	82.66	80.16	81.12	75.45	73.71
	Baseline	49.32	73.4	82.48	80.24	81.28	75.61	73.72
	MatQuant	49.15	71.76	83.73	80.13	81.18	74.19	73.36
int3	Sliced int8	20.65	31.57	44.34	28.79	59.41	51.38	39.36
	Baseline	-	-	-	-	-	-	-
	S.P. MatQuant	41.98	65.53	79.39	74.42	79.22	69.93	68.41
	MatQuant	34.64	55.13	70.43	58.61	73.39	64.48	59.45

Table 20 | Table presents downstream evaluation and perplexity results for Single Precison MatQuant, comparing it with MatQuant and the *Baseline* for int2 quatization of Gemma-2 2B with OmniQuant and QAT.

	int2	Gemma2-2B							
	Method	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Task Avg.	log pplx.
OmniQuant	S.P. MatQuant	34.64	64.06	65.69	53.07	69.7	57.14	57.38	3.185
	Baseline	31.31	53.58	62.2	40.78	66.05	54.06	51.33	3.835
	MatQuant	34.39	59.64	62.69	52.11	69.86	55.56	55.71	3.292
QAT	S.P. MatQuant	28.92	53.79	62.84	48.41	69.86	55.25	53.18	3.090
	Baseline	24.66	43.22	62.17	38.39	64.42	53.59	47.74	3.433
	MatQuant	28.24	51.73	64.19	46.76	68.66	55.01	52.43	3.153

Table 21 | Table presents downstream evaluation and perplexity results for Single Precison MatQuant, comparing it with MatQuant and the *Baseline* for int2, int4, int8 quatization of Gemma-2 9B with OmniQuant. Note that the model was trained with Single Precison MatQuant for int2, the int4 and int8 model were sliced post training.

		Gemma-2 9B							
Data type	Method	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average	log pplx.
int8	S.P. MatQuant	56.48	76.85	73.36	74.87	80.74	66.77	71.51	2.525
	OmniQuant	59.47	77.31	83.94	77.35	81.39	68.11	74.59	2.418
	MatQuant	58.11	78.03	83.27	76.17	81.18	67.09	73.97	2.451
int4	S.P. MatQuant	57.17	77.02	74.28	74.41	80.69	67.56	71.85	2.543
	OmniQuant	58.79	78.37	83.55	76.71	81.45	67.09	74.33	2.451
	MatQuant	57.25	77.36	84.86	75.52	81.5	66.77	73.88	2.481
int2	S.P. MatQuant	49.74	74.66	80.92	66.57	76.06	63.54	68.58	2.857
	OmniQuant	39.16	63.43	72.11	52.24	72.63	61.88	60.24	3.292
	MatQuant	48.72	72.18	79.2	68.11	76.17	66.77	68.52	2.809

Table 22 | Table presents downstream evaluation and perplexity results for Single Precison MatQuant, comparing it with MatQuant and the *Baseline* for int2, int4, int8 quatization of Gemma-2 9B with QAT. Note that the model was trained with Single Precison MatQuant for int2, the int4 and int8 model were sliced post training.

		Gemma-2 9B							
Data type	Method	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Average	log pplx.
int8	S.P. MatQuant	55.97	76.18	80.09	75.43	80.69	68.9	72.88	2.429
	QAT	47.78	70.66	75.08	69.92	78.35	65.11	67.82	2.29
	MatQuant	46.25	71.21	75.6	69.97	78.4	64.64	67.68	2.301
int4	S.P. MatQuant	55.2	76.01	74.74	74.19	80.41	68.9	71.57	2.429
	QAT	46.16	71.59	73.73	68.72	78.62	63.38	67.03	2.324
	MatQuant	44.37	70.45	75.81	68.43	78.35	64.88	67.05	2.332
int2	S.P. MatQuant	41.21	66.2	65.02	64.31	76.06	62.35	62.53	2.706
	QAT	33.45	55.43	62.26	54.8	70.51	59.67	56.02	2.923
	MatQuant	39.85	65.66	65.93	64.08	75.68	62.75	62.32	2.756

Table 23 | Table presents downstream evaluation and perplexity results for Single Precison MatQuant, comparing it with MatQuant and the *Baseline* for int2 quatization of Mistral 7B with OmniQuant and QAT.

	int2	Mistral 7B							
	Method	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	Winogrande	Task Avg.	log pplx.
OmniQuant	S.P. MatQuant	39.93	66.25	76.97	72.99	78.07	69.93	67.36	2.464
	Baseline	36.69	61.36	70.06	57.47	70.67	62.19	59.74	3.931
	MatQuant	41.38	67.42	71.62	71.98	77.86	65.67	65.99	2.569
QAT	S.P. MatQuant	34.64	56.19	70.73	66.77	75.52	65.43	61.55	2.435
	Baseline	29.78	48.23	64.5	55.11	70.84	61.25	54.95	2.694
	MatQuant	34.3	55.09	71.83	65.89	75.52	65.11	61.29	2.474