Table 1: Comparison of WikiText-2 Perplexity for each method with and without finetuning for quantizing Llama 2 7B model. For each method, **PPL no FT** denotes its perplexity without fine-tuning, whereas **PPL w/ FT** is perplexity with fine-tuning. We use the same setup as in Section 4.1.

Method	Avg bits	PPL no FT↓	PPL w/ FT↓
–	16	5.12	–
VQ/AQ	2.01	6.64	6.17
VQ/AQ	2.29	6.31	5.92
VQ/AQ	3.04	5.46	5.39
QuIP#	2.00	8.22	6.19
QuIP#	3.00	5.60	5.41

Method	Avg bits	PPL no FT↓	PPL w/ FT↓
GPTQ	2	3290	16.77
GPTQ	3	8.52	7.26
GPTQ	4	5.87	5.74
SpQR	2.09	12.19	9.90
SpQR	3.45	5.48	5.37
SpQR	3.98	5.29	5.21

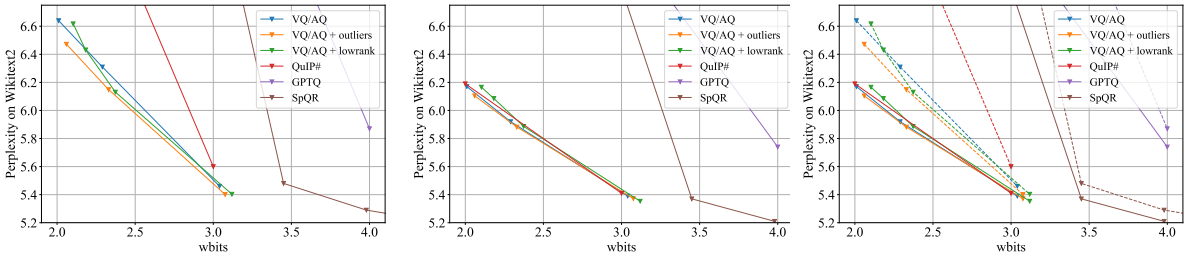

Figure 1: Llama2 7B perplexity on WikiText-2 after compression without finetuning **(left)** and with finetuning **(middle)**. On the **(right)**, there is a plot that combines the first two, allowing for a better comparison of each method with and without fine-tuning. **Compression algorithms without finetuning are represented with dashed lines, while algorithms with finetuning are represented with continuous lines.**

Table 2: Evaluation of quantized LLAMA 2 models for 2.x bits per weight. We use the same setup as in the main paper (Section 4.2 with an extra baseline). As requested, we finetune the model in 16-bit precision for the same number of steps, then quantize it with AQLM, reported as **"FT+AQLM"**. Finally, the "Finetuned" row corresponds to an uncompressed 16-bit model finetuned without quantization. We hypothesize that finetuning the model has little effect since we train on a dataset resembling its original pretraining data.

Size	Method	Avg bits	Wiki2↓	C4↓	ArcC↑	ArcE↑	HellaSwag↑	PiQA↑	WinoGrande↑	Average↑
7B	–	16.00	5.12	6.63	43.43	76.3	57.14	78.07	69.06	64.80
	AQLM	2.02	6.64	8.56	33.28	61.87	49.49	73.56	64.17	56.47
	PV-Tuning	2.02	5.84	7.62	38.40	71.17	53.50	76.99	66.69	61.35
	Finetuned	16	5.13	6.59	43.46	76.47	56.96	78.14	68.91	64.79
	FT + AQLM	2.02	6.48	8.36	33.29	62.45	49.31	73.49	64.82	56.67

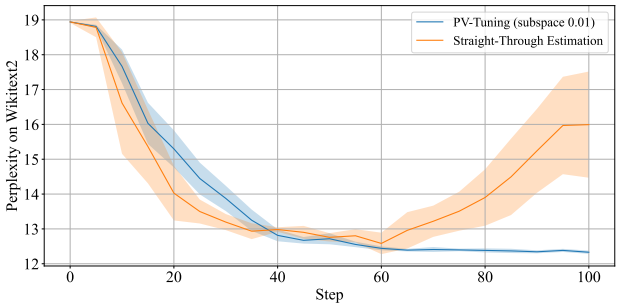

Figure 2: Learning curve for PV-tuning and STE algorithms, when tuning tinyllama model with 2x8g8 AQLM quantization (2 codebooks with 8 bits per code and input groupsize equal to 8).