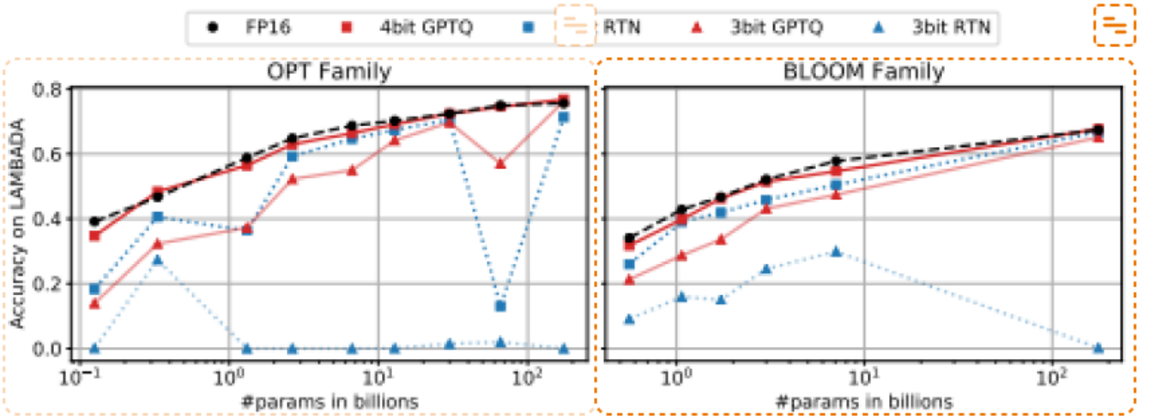
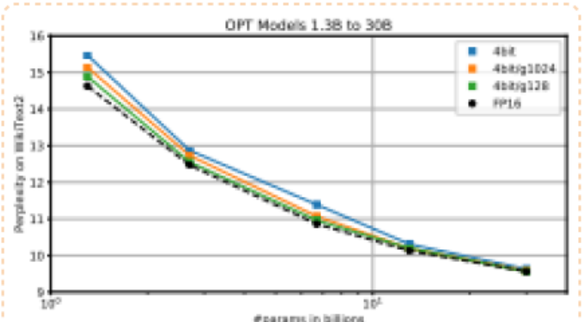


Figure 3: The accuracy of OPT and BLOOM models post-GPTQ, measured on LAMBADA.

Additional Tricks. While our experiments so far have focused exclusively on vanilla row-wise quantization, we want to emphasize that GPTQ is *compatible with essentially any choice of quantization grid*. For example, it is easily combined with standard *grouping* (Alistarh et al., 2017; Park et al., 2022), i.e. applying independent quantization to groups of g consecutive weights. As shown in the last rows of Table 5, this can bring noticeable extra accuracy for the largest models at 3-bit. Further, as visualized in Figure 4, it significantly reduces the accuracy losses for medium sized models at 4-bit precision.

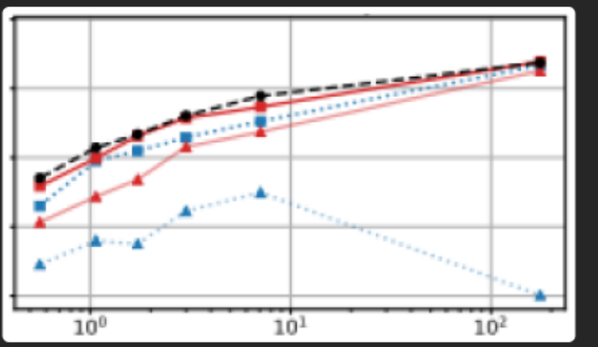
Model	FP16	g128	g64	g32	3-bit
OPT-175B	8.34	9.58	9.18	8.94	8.68
BLOOM	8.11	9.55	9.17	8.83	8.64

Table 7: 2-bit GPTQ quantization results with varying group-sizes; perplexity on WikiText2.



Chat with Paper

Алгоритмы и модели	Разделение данных
DL модели (RNN, LSTM)	80/5/15
DistilRuBERT	80/5/15



Ask a document...