BLOOM Model Family

- ▲ 3bit RTN
- ▲ 3bit GPTQ
- ● FP16

Y-axis: Perplexity on WikiText2

X-axis: #params in billions

571.