

Perplexity corroborates: NF4 explodes on Qwen

WikiText-2 perplexity ($\downarrow$, log)

fp16 NF4 asym-NF4

6×10^1
 4×10^1
 3×10^1
 2×10^1
 10^1
 6×10^0

Llama-2-7B
(MHA 1:1)

Llama-2-13B
(MHA 1:1)

Mistral-7B
(GQA 4:1)

Qwen2.5-7B
(GQA 7:1)

4.2
(10x fp16)

