

4-bit KV quantization: NF4 collapses on high-GQA; asym-NF4 is robust

