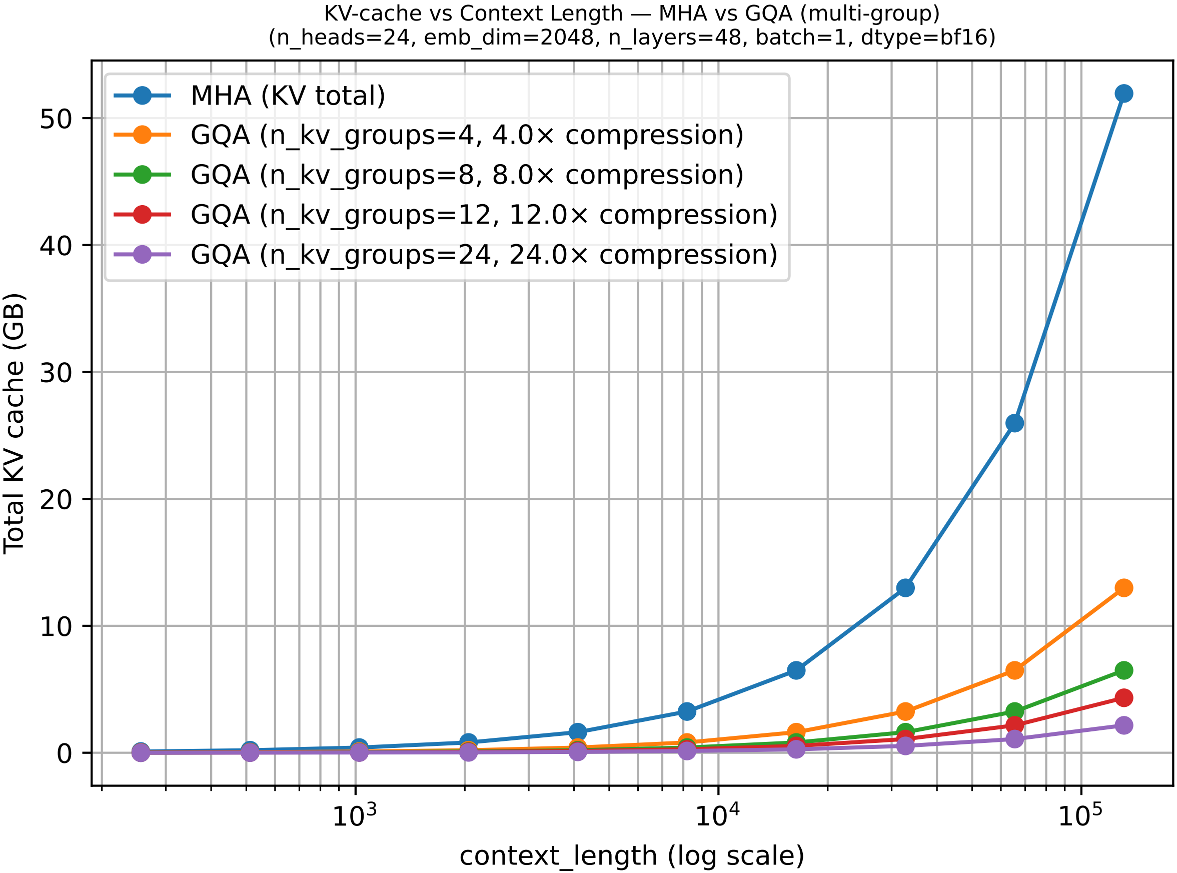
KV-cache vs Context Length — MHA vs GQA (multi-group)
(n_heads=24, emb_dim=2048, n_layers=48, batch=1, dtype=bf16)
MHA (KV total)
GQA (n_kv_groups=4, 4.0× compression)
GQA (n_kv_groups=8, 8.0× compression)
GQA (n_kv_groups=12, 12.0× compression)
GQA (n_kv_groups=24, 24.0× compression)
Total KV cache (GB)
0
10
20
30
40
50
10^3
10^4
10^5
context_length (log scale)