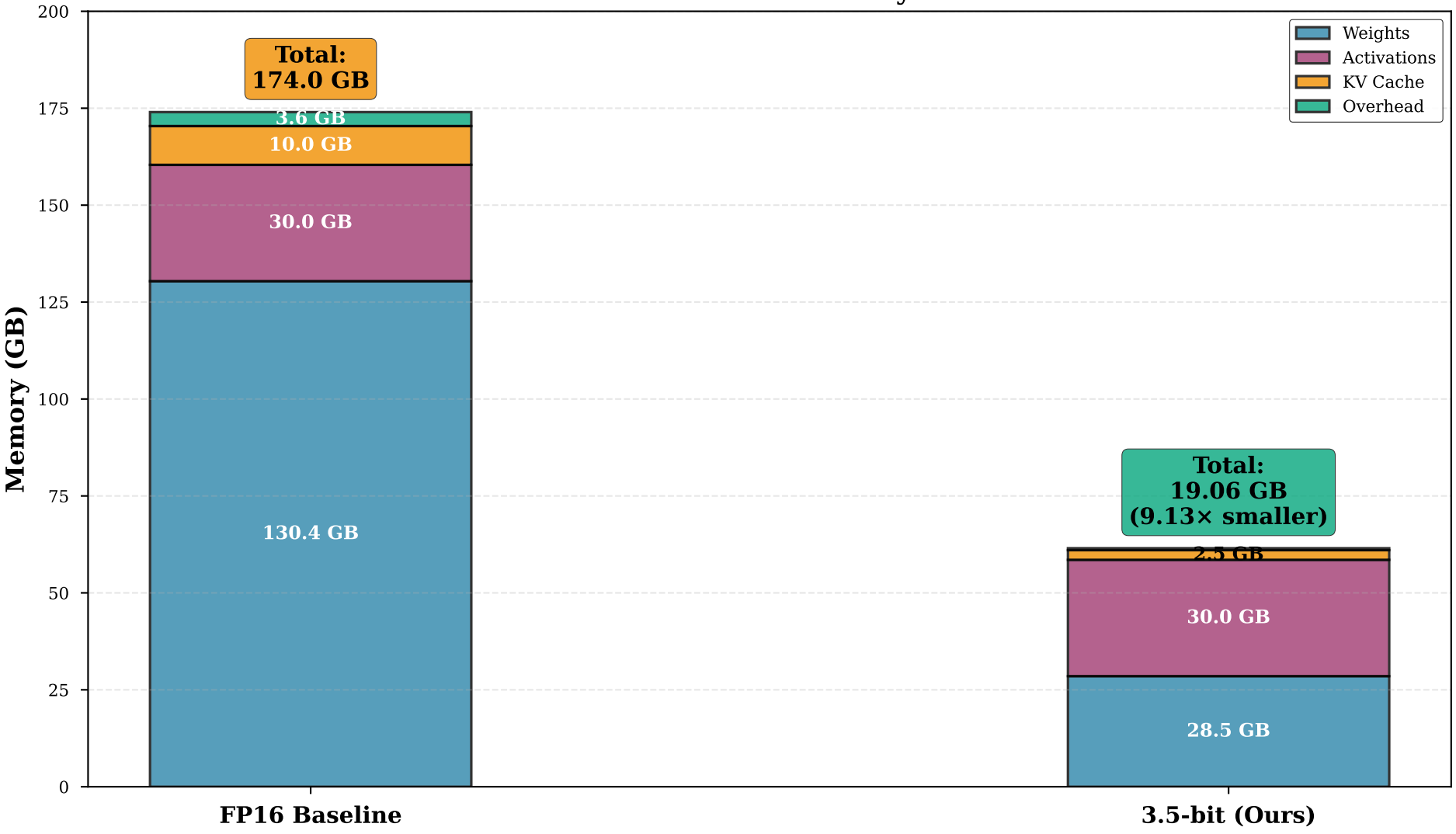
LLaMA 70B Inference Memory Breakdown
Memory (GB)
200
175
150
125
100
75
50
25
0
Weights
Activations
KV Cache
Overhead
Total:
174.0 GB
3.6 GB
10.0 GB
30.0 GB
130.4 GB
Total:
19.06 GB
(9.13× smaller)
2.5 GB
30.0 GB
28.5 GB
FP16 Baseline
3.5-bit (Ours)