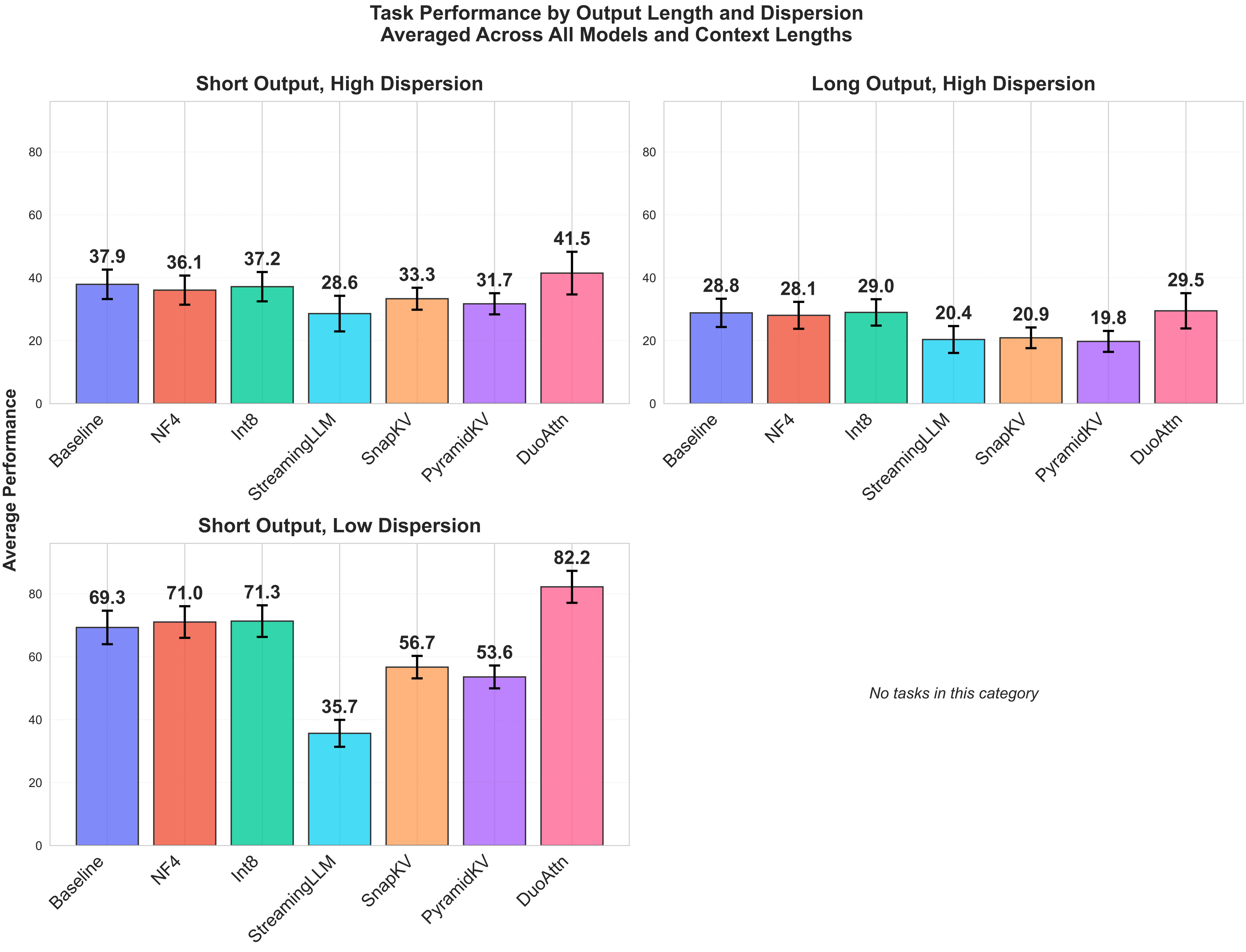
Task Performance by Output Length and Dispersion
Averaged Across All Models and Context Lengths
Short Output, High Dispersion
Long Output, High Dispersion
Short Output, Low Dispersion
Average Performance
0
20
40
60
80
Baseline
NF4
Int8
StreamingLLM
SnapKV
PyramidKV
DuoAttn
37.9
36.1
37.2
28.6
33.3
31.7
41.5
28.8
28.1
29.0
20.4
20.9
19.8
29.5
69.3
71.0
71.3
35.7
56.7
53.6
82.2
No tasks in this category