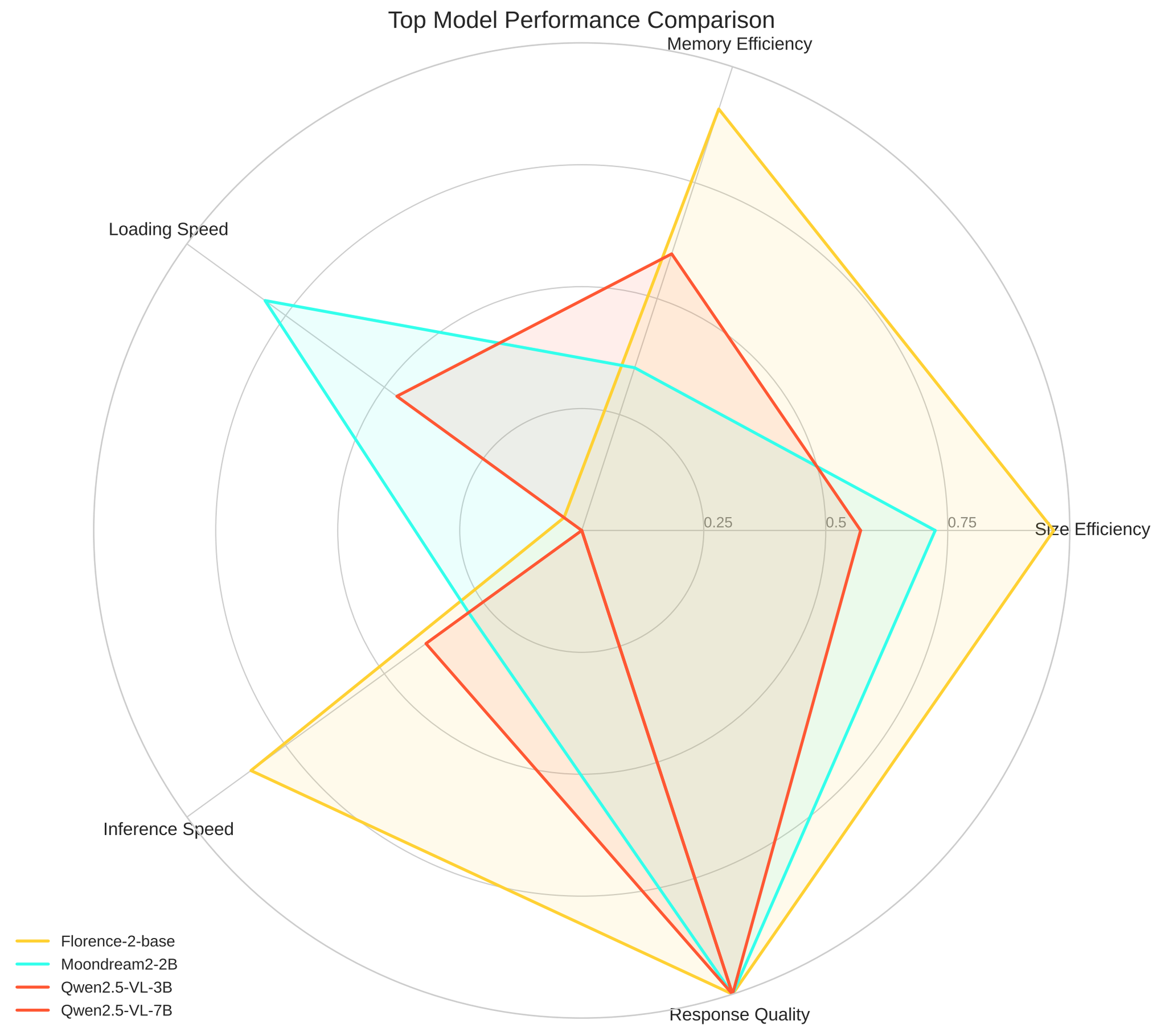
Top Model Performance Comparison
Memory Efficiency
Size Efficiency
Response Quality
Inference Speed
Loading Speed
0.25
0.5
0.75
Florence-2-base
Moondream2-2B
Qwen2.5-VL-3B
Qwen2.5-VL-7B