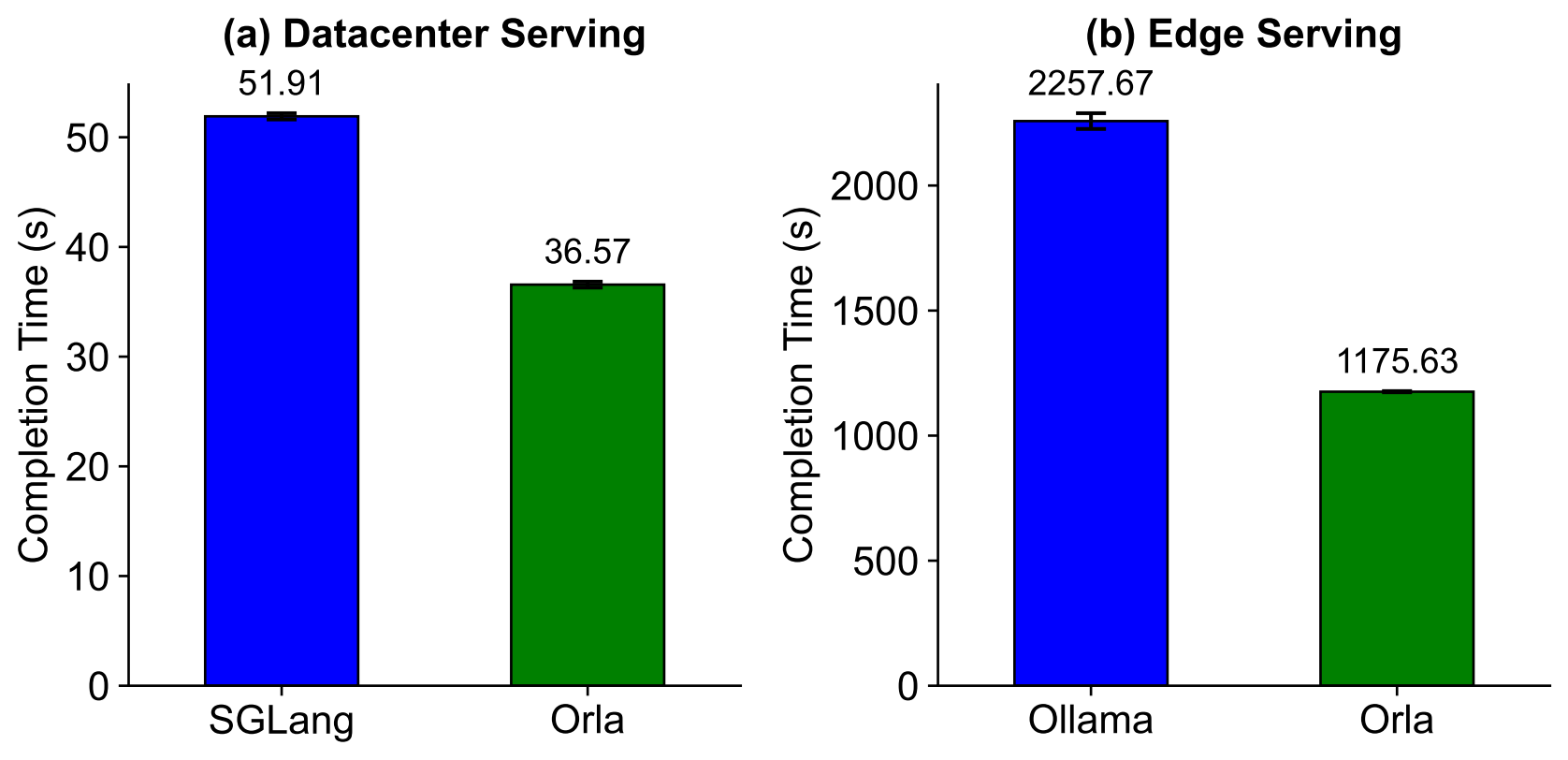
(a) Datacenter Serving
Completion Time (s)
0
10
20
30
40
50
51.91
36.57
SGLang
Orla
(b) Edge Serving
Completion Time (s)
0
500
1000
1500
2000
2257.67
1175.63
Ollama
Orla