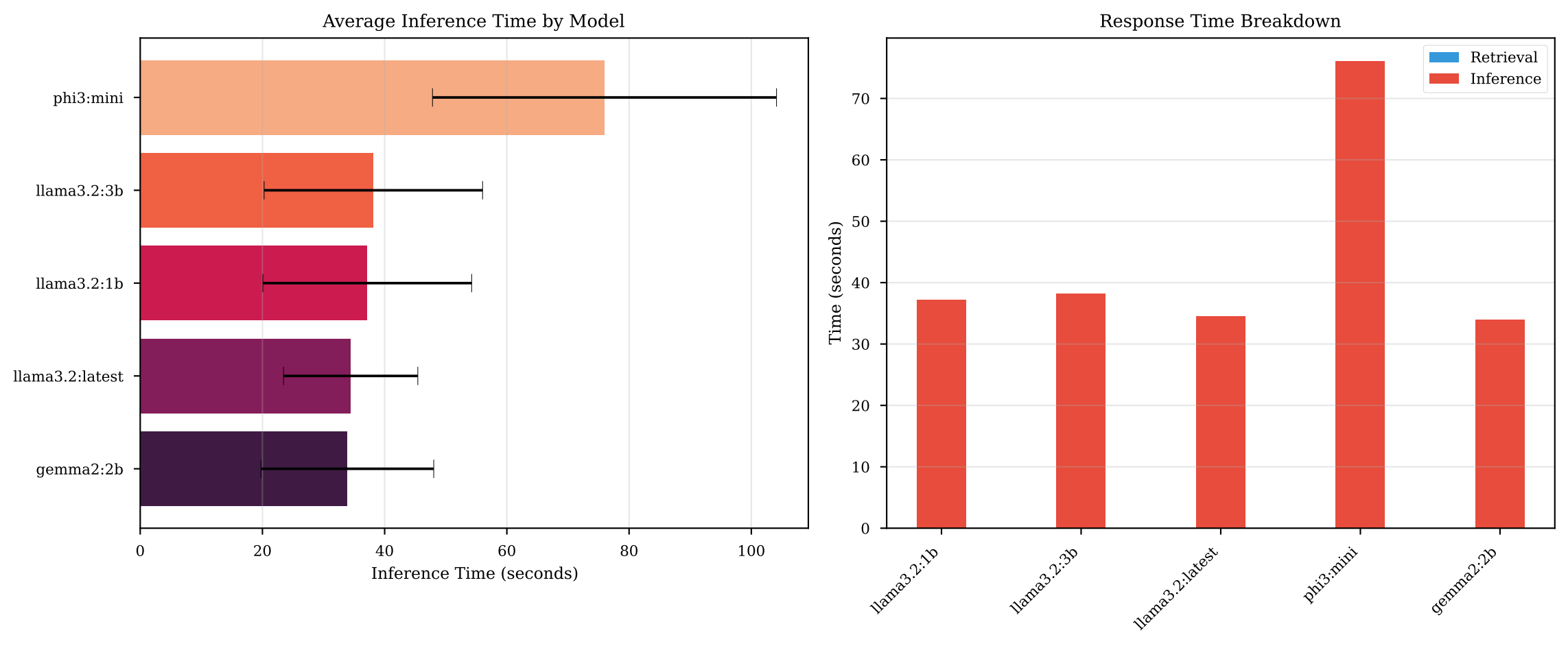
Average Inference Time by Model
phi3:mini
llama3.2:3b
llama3.2:1b
llama3.2:latest
gemma2:2b
0
20
40
60
80
100
Inference Time (seconds)
Response Time Breakdown
Retrieval
Inference
Time (seconds)
0
10
20
30
40
50
60
70
llama3.2:1b
llama3.2:3b
llama3.2:latest
phi3:mini
gemma2:2b