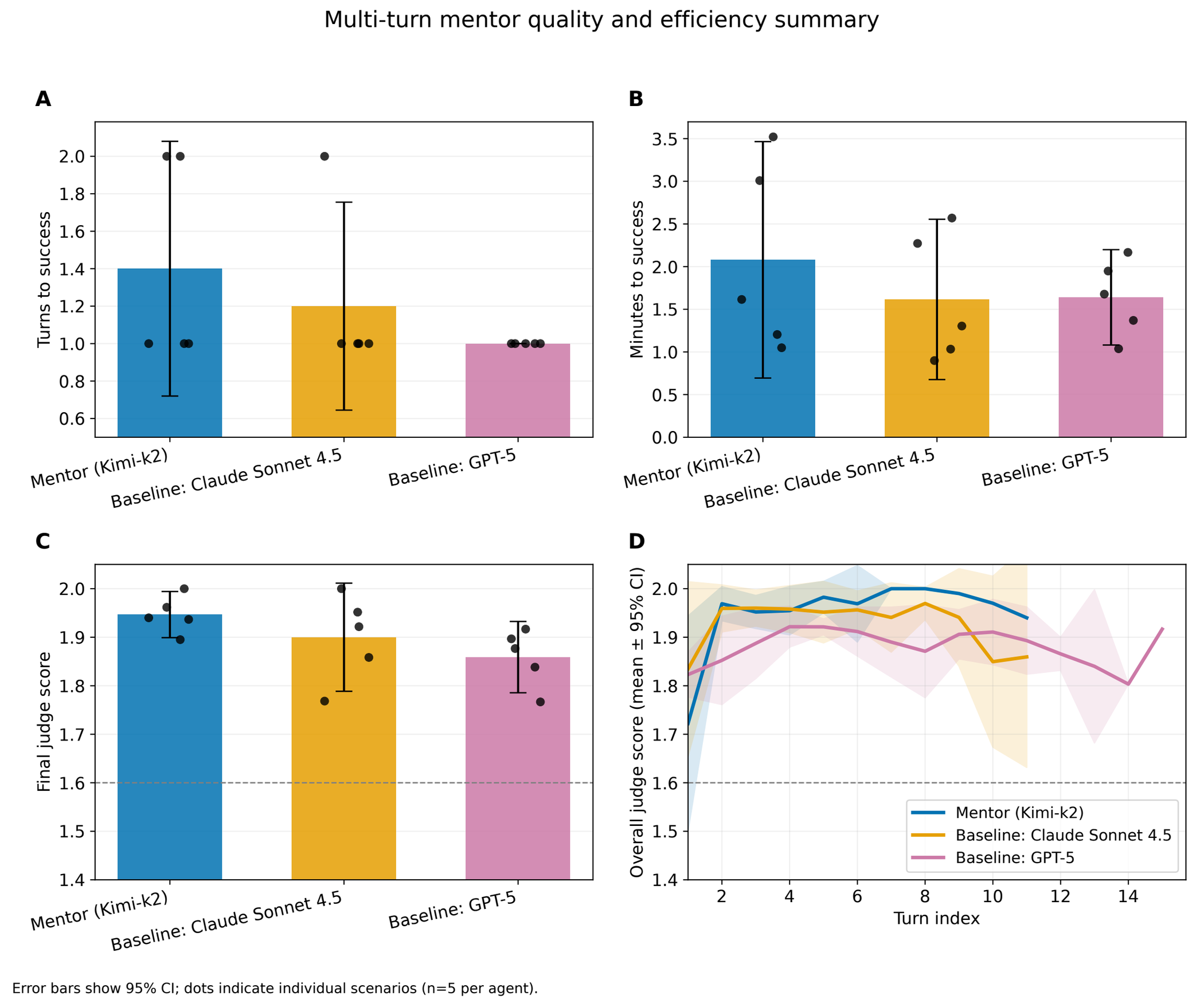
Multi-turn mentor quality and efficiency summary
A
Turns to success
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
Mentor (Kimi-k2)
Baseline: Claude Sonnet 4.5
Baseline: GPT-5
B
Minutes to success
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Mentor (Kimi-k2)
Baseline: Claude Sonnet 4.5
Baseline: GPT-5
C
Final judge score
2.0
1.9
1.8
1.7
1.6
1.5
1.4
Mentor (Kimi-k2)
Baseline: Claude Sonnet 4.5
Baseline: GPT-5
D
Overall judge score (mean ± 95% CI)
2.0
1.9
1.8
1.7
1.6
1.5
1.4
2
4
6
8
10
12
14
Turn index
Mentor (Kimi-k2)
Baseline: Claude Sonnet 4.5
Baseline: GPT-5
Error bars show 95% CI; dots indicate individual scenarios (n=5 per agent).