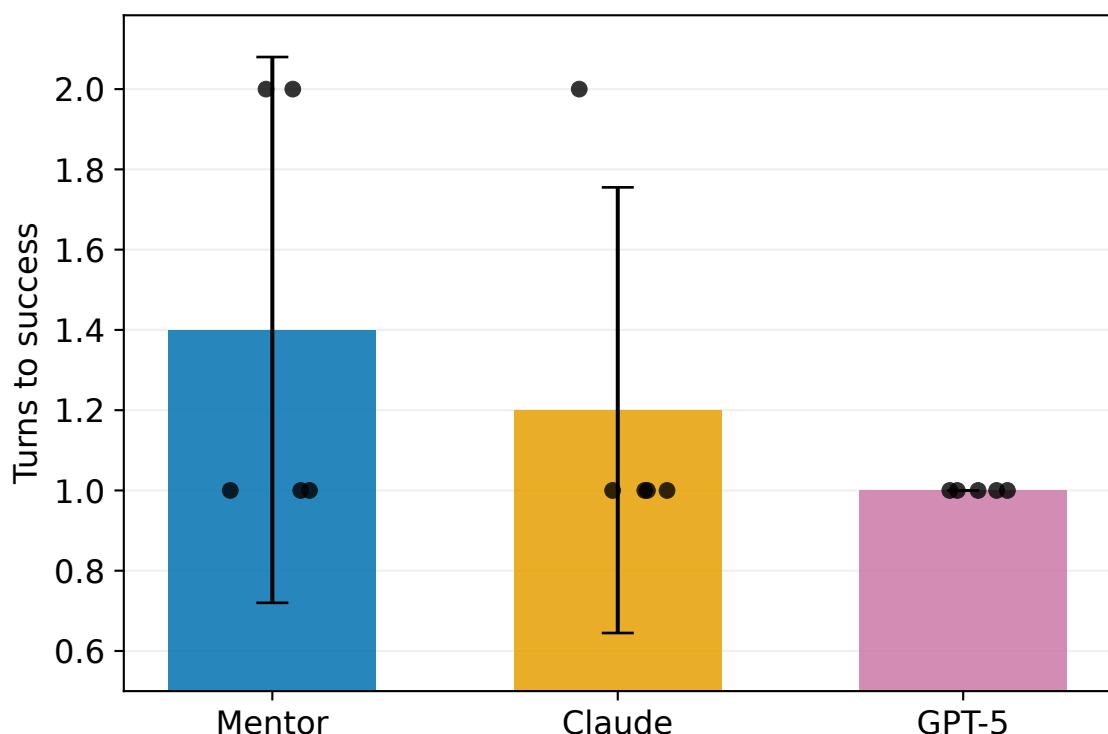


Multi-turn conversation quality and efficiency

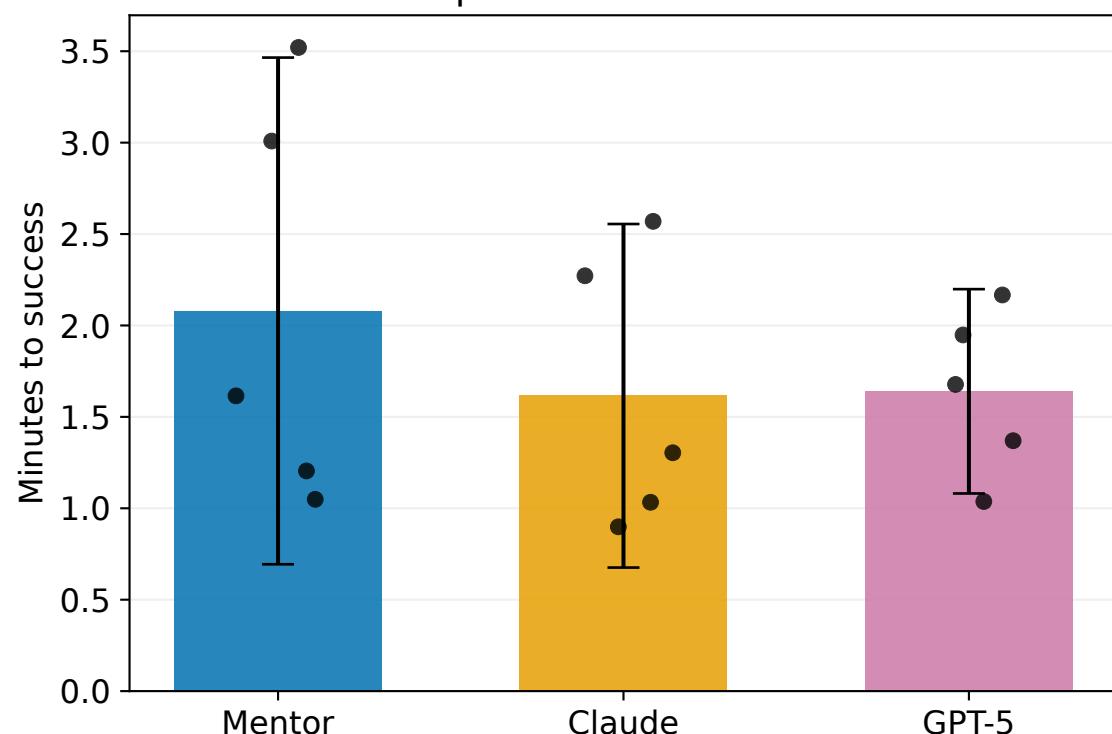
A

Conversation turns to success



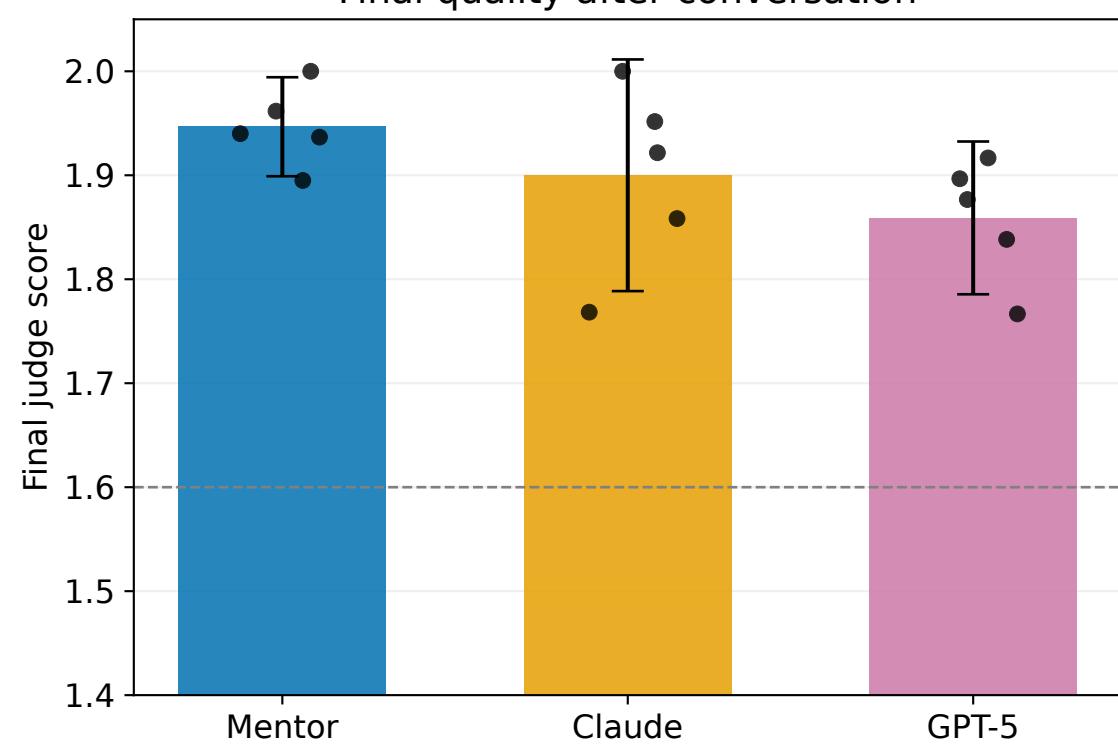
B

Elapsed time to success



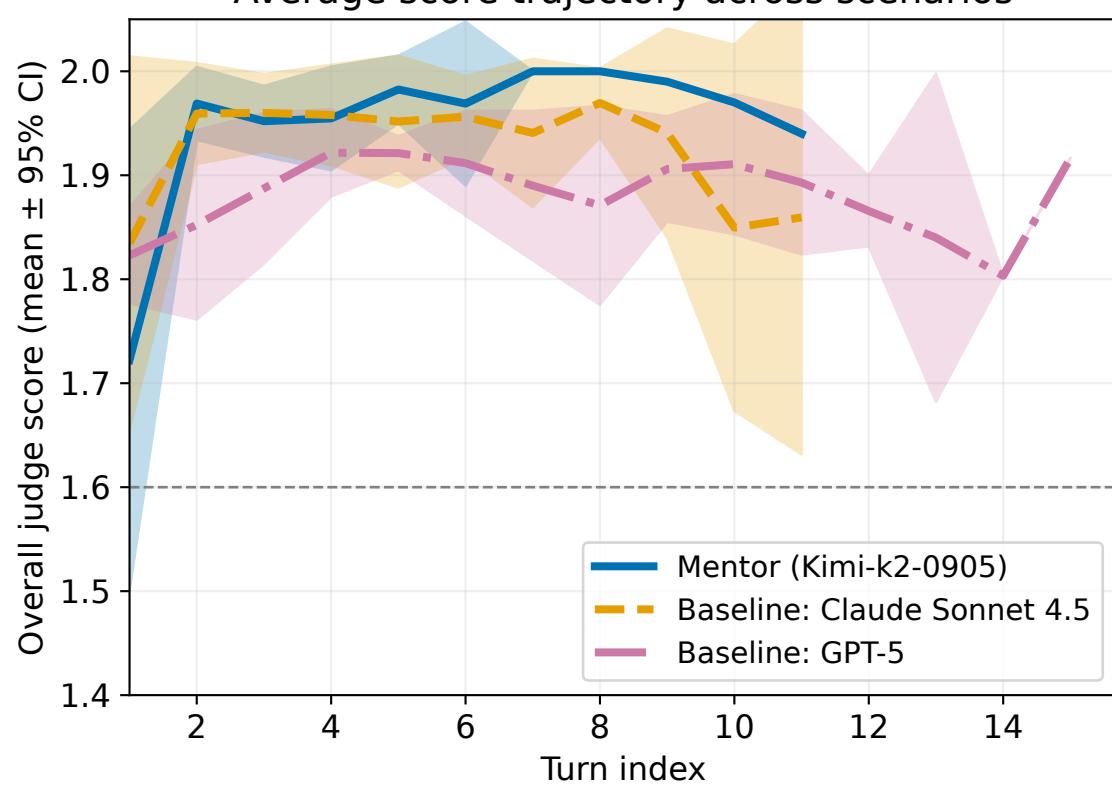
C

Final quality after conversation



D

Average score trajectory across scenarios



Error bars show 95% CI; dots mark individual scenarios (n=5 per agent). Success threshold (score ≥ 1.6) shown as dashed gray line.