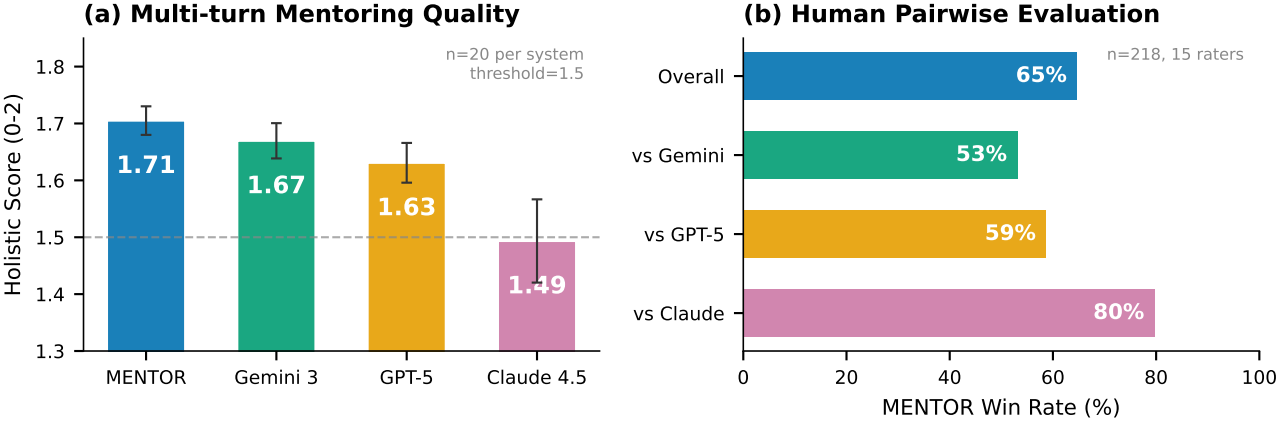
(a) Multi-turn Mentoring Quality
n=20 per system
threshold=1.5
Holistic Score (0-2)
1.8
1.7
1.6
1.5
1.4
1.3
1.71
1.67
1.63
1.49
MENTOR
Gemini 3
GPT-5
Claude 4.5
(b) Human Pairwise Evaluation
n=218, 15 raters
Overall
vs Gemini
vs GPT-5
vs Claude
65%
53%
59%
80%
0
20
40
60
80
100
MENTOR Win Rate (%)