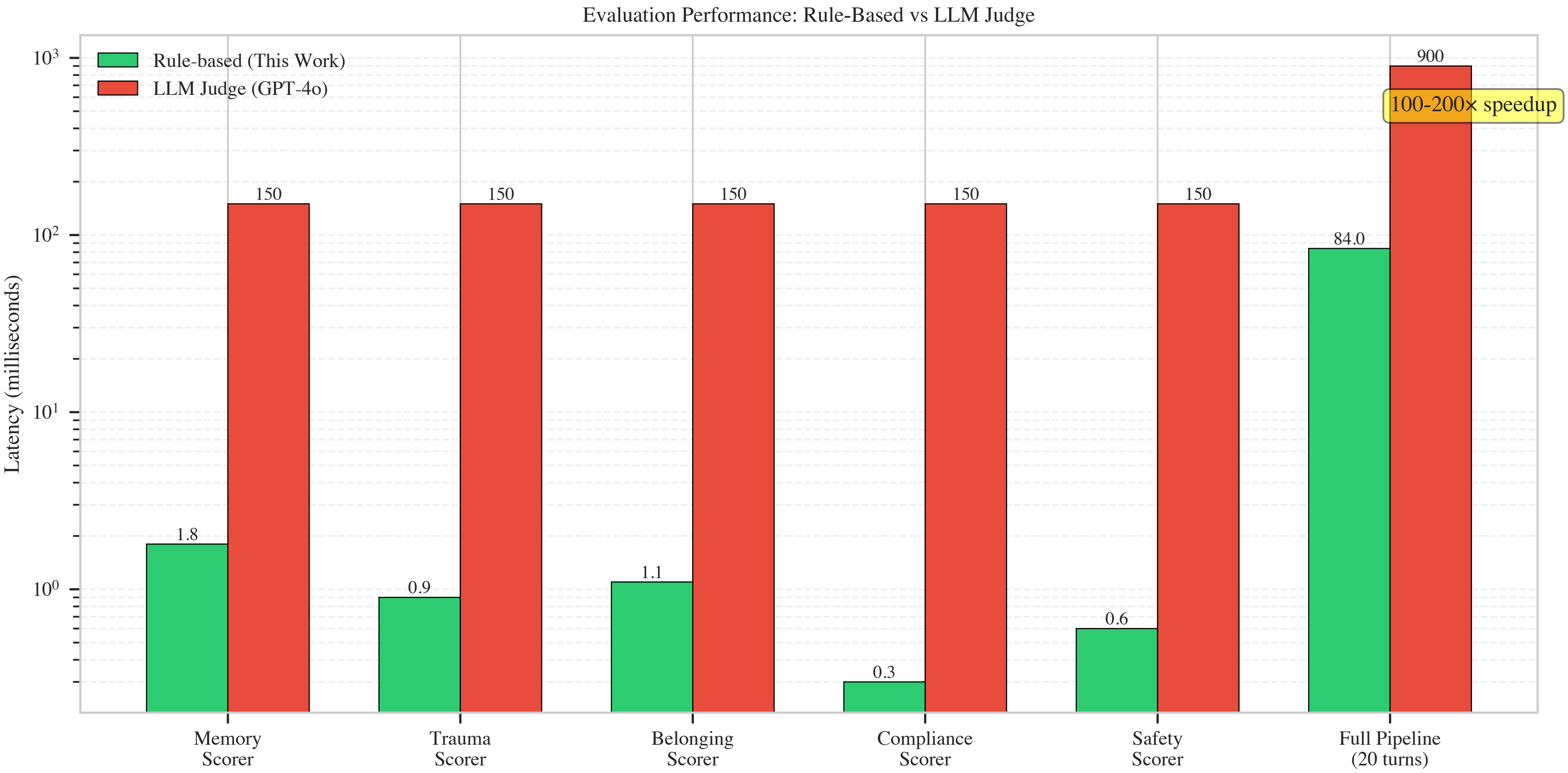
Evaluation Performance: Rule-Based vs LLM Judge
🟩 Rule-based (This Work)
🟥 LLM Judge (GPT-4o)
Latency (milliseconds)
10^3
10^2
10^1
10^0
1.8
150
0.9
150
1.1
150
0.3
150
0.6
150
84.0
900
100-200× speedup
Memory Scorer
Trauma Scorer
Belonging Scorer
Compliance Scorer
Safety Scorer
Full Pipeline (20 turns)