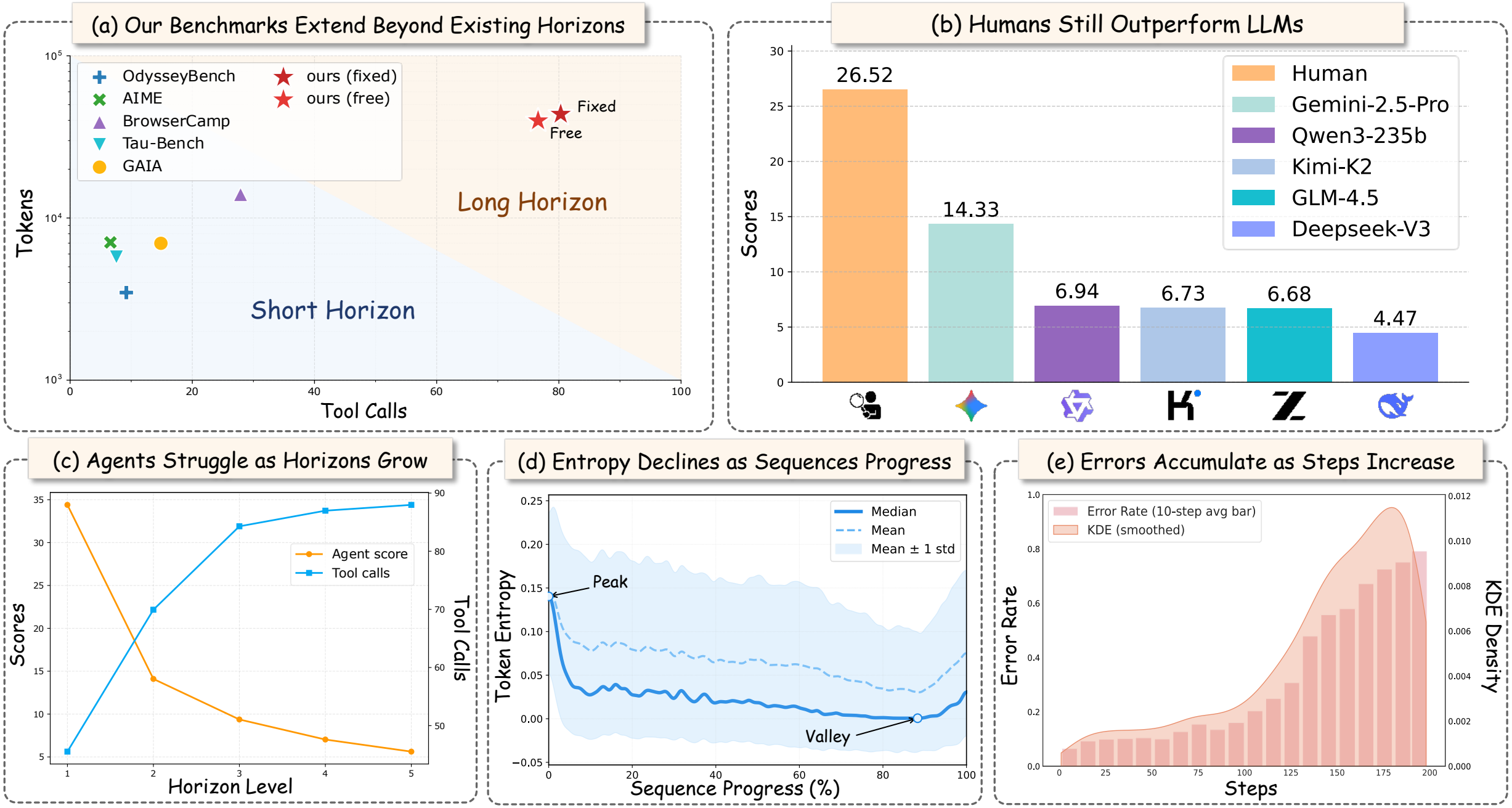
(a) Our Benchmarks Extend Beyond Existing Horizons
OdysseyBench
AIME
BrowserCamp
Tau-Bench
GAIA
ours (fixed)
ours (free)
Fixed
Free
Long Horizon
Short Horizon
Tokens
Tool Calls
(b) Humans Still Outperform LLMs
Scores
26.52
14.33
6.94
6.73
6.68
4.47
Human
Gemini-2.5-Pro
Qwen3-235b
Kimi-K2
GLM-4.5
Deepseek-V3
(c) Agents Struggle as Horizons Grow
Agent score
Tool calls
Scores
Tool Calls
Horizon Level
(d) Entropy Declines as Sequences Progress
Median
Mean
Mean ± 1 std
Peak
Valley
Token Entropy
Sequence Progress (%)
(e) Errors Accumulate as Steps Increase
Error Rate (10-step avg bar)
KDE (smoothed)
Error Rate
KDE Density
Steps