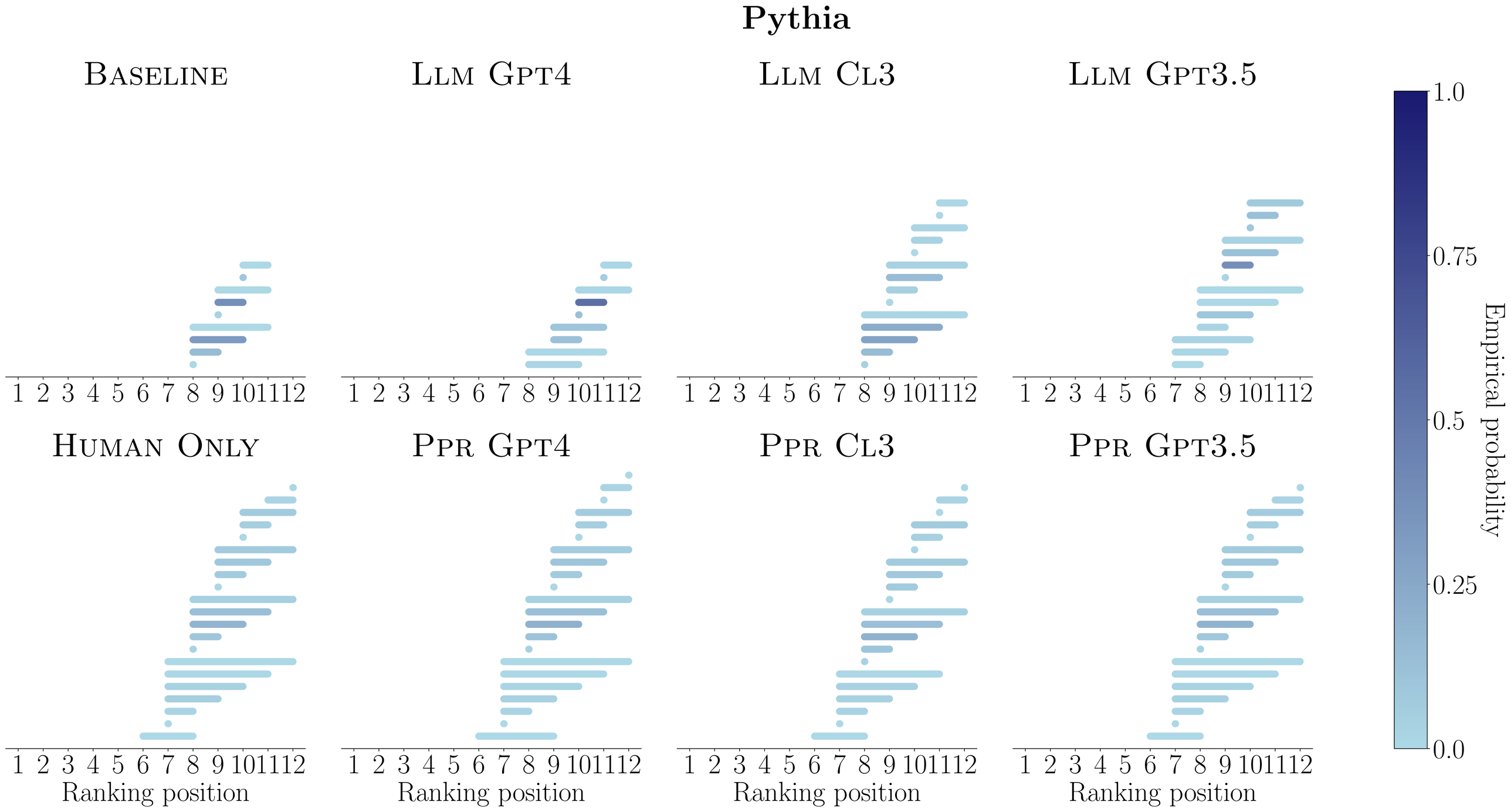
Pythia
Baseline
Llm Gpt4
Llm Cl3
Llm Gpt3.5
Human Only
Ppr Gpt4
Ppr Cl3
Ppr Gpt3.5
1 2 3 4 5 6 7 8 9 10 11 12
Ranking position
Empirical probability
0.0
0.25
0.5
0.75
1.0