

FastChat

BASELINE

LLM GPT4

LLM CL3

LLM GPT3.5

HUMAN ONLY

PPR GPT4

PPR CL3

PPR GPT3.5

Empirical probability

1.0

0.75

0.5

0.25

0.0

1 2 3 4 5 6 7 8 9 10 11 12

1 2 3 4 5 6 7 8 9 10 11 12

1 2 3 4 5 6 7 8 9 10 11 12

1 2 3 4 5 6 7 8 9 10 11 12

1 2 3 4 5 6 7 8 9 10 11 12

1 2 3 4 5 6 7 8 9 10 11 12

1 2 3 4 5 6 7 8 9 10 11 12

1 2 3 4 5 6 7 8 9 10 11 12

Ranking position

Ranking position

Ranking position

Ranking position