Expected LLM faceoff win rate (Terry-Bradley). claude-3-haiku ' 0.50 0.40 0.53 0.60 0.24 0.41 0.38 claude-3-opus 0.63 0.70 0.60 0.50 0.32 0.51 0.25 0.48 - 0.8 mistral-medium 0.37 0.50 0.47 0.57 0.15 0.38 0.35 gpt-4-0613 0.30 0.40 0.43 0.50 0.12 0.31 0.13 0.29 - 0.6 Completer qwen1.5-110B-Chat gpt-4o-2024-05-13 0.83 0.61 0.85 0.88 0.76 0.50 0.53 - 0.4 0.76 0.68 0.50 0.69 0.67 0.42 0.83 0.39 qwen1.5-32B-Chat 0.59 0.62 0.31 0.24 0.49 0.69 0.50 0.47 llama-3-70b-chat - 0.2 0.82 0.75 0.84 0.87 0.47 0.58 0.76 0.50 0.74 llama-3-8b-chat 0.33 0.62 0.65 0.24 0.26 0.52 0.71 0.53 0.50 claude-3-haiku gpt-4-0613 llama-3-70b-chat llama-3-8b-chat gpt-40-2024-05-13 qwen1.5-32B-Chat claude-3-opus vs. Completer