Accuracy vs. Cost

Overall Accuracy

Mistral Small
Gemini 2.0 Flash (OpenRouter)
Llama 3.1 8B (Groq)
Llama 4 Scout 17B (Groq)
50%
25%
Llama 3.3 70B (Groq)
Pareto frontier
10,000   100,000
Total Tokens (cost proxy)