LLM Performance: Speed vs Quality Trade-off

Quality Score (1-10)

Speed (Tokens per Second)

Google
Openai

Fastest: models/gemini-2.5-fl...