LLM API Providers Leaderboard - Comparison of over 100 LLM endpoints

Comparison and ranking of API provider performance for over 100 AI LLM Model endpoints across performance key metrics including price, output speed, latency, context window & others. For more details including relating to our methodology, see our FAQs.

API providers compared: OpenAI, Playground AI, Microsoft Azure, Mistral, Ideogram, DeepSeek, Amazon Bedrock, Hyperbolic, Groq, FriendliAI, Together.ai, Anthropic, Black Forest Labs, Perplexity, Google, Lambda Labs, Fireworks, Cerebras, Leonardo.Ai, Cohere, Recraft AI, Upstage, Simplismart, Speechmatics, Deepinfra, Fish Audio, Replicate, , Genmo, Nebius, Adobe, MiniMax, CentML, StepFun, Runpod, Zyphra, Murf AI, Speechify, Rev AI, AssemblyAI, fal.ai, Rime, kluster.ai, Prodia, Hume AI, Reka AI, Deepgram, Gladia, Baseten, Stability.ai, Midjourney, Reve, Databricks, ElevenLabs, IBM, Vivago AI, SambaNova, Dreamina, xAI, Cartesia, LMNT, PlayAI, 01.AI, Alibaba Cloud, Novita, AI21 Labs, and WaveSpeed.

PROMPT OPTIONS ⌄

Filter, e.g. GPT, Meta

EXPAND COLUMNS |→

API PROVIDER ⇅	MODEL ⇅	FEATURES |→ CONTEXT WINDOW ⇅	MODEL INTELLIGENCE |→ ARTIFICIAL ANALYSIS INTELLIGENCE INDEX ⇅	PRICE |→ BLENDED ⇅ *USD/1M Tokens*	OUTPUT TOKENS/S |→ MEDIAN ⇅ *Tokens/s*	LATENCY |→ MEDIAN ⇅ *First Chunk (s)*	END-TO-END RESPONSE TIME |→ TOTAL ⇅ *Response (s)*	REASONING ⇅ *Time (s)*	FURTHER ANALYSIS
🟡 OpenAI	🔵 o4-mini (high)	200k	70	$1.93	125.2	31.35	35.35	N/A	⧉ Model | ⧉ Providers
▦ Microsoft Azure	🔵 o4-mini (high)	200k	70	$1.93	34.9	147.40	161.73	N/A	⧉ Model | ⧉ Providers
Google	G Gemini 2.5 Pro	1m	68	$3.44	148.1	39.57	42.94	N/A	⧉ Model | ⧉ Providers
🟡 OpenAI	🔵 o3	128k	67	$17.50	195.6	13.24	15.80	N/A	⧉ Model | ⧉ Providers
▦ Microsoft Azure	🔵 o3	128k	67	$17.50	56.3	57.97	66.86	N/A	⧉ Model | ⧉ Providers
xI	xI Grok 3 mini Reasoning (high)	131k	67	$0.35	95.0	0.24	26.54	21.04	⧉ Model | ⧉ Providers