

	Pricing (per 1M tokens)	Context window	Multimodal	Deployment	Privacy alignment	Best business use case
GPT-5.5 (OpenAI)	\$5 / \$30	1M tokens	Yes (text, image)	API	No training on API inputs by default	Complex analytics, long-document review
Gemini 3.1 Pro (Google)	\$2 / \$12 (≤200K tokens); \$4 / \$18 (>200K tokens)	1M tokens	Yes (text, image, audio, video)	API, Vertex AI	Google Cloud data controls	Data-heavy enterprises in the Google ecosystem
Claude Opus 4.8 (Anthropic)	\$5 / \$25	1M tokens	Yes	API, Amazon Bedrock, Google Cloud's Vertex AI, Microsoft Foundry	No training on API inputs by default	Reasoning depth, regulated and safety- critical work
GPT-5.4 (OpenAI)	\$2.50 / \$15	1M tokens	Yes (text, image)	API	No training on API inputs by default	General-purpose workflows, content generation
DeepSeek V4 (DeepSeek)	\$0.435 / \$0.87 (V4- Pro, cache miss)	1M tokens	No (text only)	API, self-hosted	Full data control when self-hosted; see Open Platform terms for API data handling	On-prem reasoning, budget-constrained deployments
Llama 4 Maverick (Meta)	\$0.19–\$0.49 per 1M tokens (3:1 blended, self- hosted infrastructure estimate)	1M tokens	Yes (text, image)	Self-hosted, cloud partners	Self-hosted deployments are not subject to vendor data handling policies; data does not leave your own infrastructure.	Research, custom fine- tuning, niche verticals
Grok 4.1 Fast (xAI)	\$0.20 / \$0.50	2M tokens	Text, image	API	Live X platform data access via Agent Tools API	Real-time trend monitoring, marketing intelligence