Provider	Key models	Free requests/day	Token cap	Latency	Data policy
OpenRouter	25+ free models available via the :free model variant suffix, across providers including Meta, DeepSeek, Qwen, and others	50 requests/day on the free plan; 1,000 requests/day on free models after purchasing at least $10 in credits	Free-tier usage may be subject to additional rate limiting by the underlying provider, especially during peak times	Varies by model and provider; latency data is available in the OpenRouter model catalog	OpenRouter does not train on customer data; provider-side retention can be disabled at the account level or per API call
Google AI Studio	Gemini 2.5 Flash and Gemini 2.5 Flash-Lite on the free tier	Free-tier rate limits apply per project and are viewable in AI Studio	1,000,000 tokens per minute on Flash models	Fast; Flash-Lite optimized for high throughput	Free-tier inputs may be used to improve Google products; this does not apply in the EU, UK, or Switzerland
HuggingFace	Hundreds of models via Inference Providers, spanning text generation, image generation, embeddings, speech, and more across multiple partner providers	Credit-limited; free users receive $0.10 in monthly credits, subject to change	Routes to the fastest available provider by default	Actual latency varies by model and provider	HuggingFace does not store request bodies or responses when routing; debug logs are kept for up to 30 days but contain no user data or tokens; provider-side policies apply for each routed provider
Groq	Llama 3.3 70B Versatile, Llama 4 Scout, Whisper Large v3, and others running on Groq's LPU infrastructure	~1,000 RPD on most models; 14,400 RPD on Llama 3.1 8B Instant	Model-specific TPM limits; e.g., 12,000 TPM on Llama 3.3 70B Versatile	TTFT scales with input size and model architecture, with smaller models delivering the lowest latency at short context lengths	Groq does not retain customer inputs or outputs by default; Zero Data Retention available to all customers
Cloudflare Workers AI	Llama 3.3 70B, Llama 4 Scout, Mistral Small 3.1 24B, embeddings, image, and audio models	10,000 Neurons per day; limits reset daily at 00:00 UTC	Neuron-based compute cap; each model consumes neurons at a different rate per token	Low latency; runs on Cloudflare's global edge network	Cloudflare does not use customer content to train AI models or improve services without explicit consent