

LAKEER.AI

Free LLM API Resource Directory

*A curated, professionally organized reference of free and trial-credit LLM
inference APIs — with verified, clickable links.*

PREPARED FOR

Lakeer.AI

August 2026

About This Directory

This directory catalogues services offering free or trial-credit access to API-based LLM (Large Language Model) inference. Every provider name is a clickable hyperlink to its official console or homepage, and every documentation reference — rate limits, pricing, model catalogues — links directly to the primary source.

USAGE NOTE — Please use these free services responsibly. Excessive or automated abuse risks providers withdrawing free access for the wider community.

SCOPE — This directory excludes any service that is not a legitimate, officially sanctioned API (e.g. reverse-engineered or unofficial wrappers around an existing chatbot).

Contents

PART ONE

Free Providers

OpenRouter · Google AI Studio · NVIDIA NIM · Mistral (La Plateforme) · Mistral (Codestral) · HuggingFace Inference Providers · Vercel AI Gateway · Kilo Gateway · OpenCode Zen · Cerebras · Groq · Cohere · Cloudflare Workers AI

PART TWO

Providers with Trial Credits

Fireworks · Baseten · Nebius · Novita · AI21 · Upstage · NLP Cloud · Alibaba Cloud (International) Model Studio · Modal · Inference.net · Hyperbolic · SambaNova Cloud · Scaleway Generative APIs

Free Providers

Ongoing free tiers — not one-time trial credits.

| OpenRouter

Models share a common quota.

Limits 20 requests/minute · 50 requests/day · up to 1,000 requests/day with a \$10 lifetime top-up

Free Models

- Cohere North Mini Code
- Ling 3.0 Flash
- NVIDIA Nemotron 3 Nano Omni 30B A3B (Reasoning)
- NVIDIA Nemotron 3 Super 120B A12B
- NVIDIA Nemotron 3 Ultra 550B A55B
- NVIDIA Nemotron 3.5 Content Safety
- Poolside Laguna S 2.1
- Poolside Laguna XS 2.1
- google/gemma-4-26b-a4b-it:free
- google/gemma-4-31b-it:free
- nvidia/nemotron-3-nano-30b-a3b:free
- nvidia/nemotron-nano-12b-v2-vl:free
- nvidia/nemotron-nano-9b-v2:free
- openai/gpt-oss-20b:free

Google AI Studio

Data is used for training when used outside of the UK / CH / EEA / EU.

Model Name	Model Limits
Gemini 3.6 Flash	250,000 tok/min · 20 req/day · 5 req/min
Gemini 3.5 Flash	250,000 tok/min · 20 req/day · 5 req/min
Gemini 3 Flash	250,000 tok/min · 20 req/day · 5 req/min
Gemini 3.5 Flash-Lite	250,000 tok/min · 500 req/day · 15 req/min
Gemini 3.1 Flash-Lite	250,000 tok/min · 500 req/day · 15 req/min
Gemini 2.5 Flash	250,000 tok/min · 20 req/day · 5 req/min
Gemini 2.5 Flash-Lite	250,000 tok/min · 20 req/day · 10 req/min
Gemini 3.1 Flash TTS	10,000 tok/min · 10 req/day · 3 req/min
Gemini 2.5 Flash TTS	10,000 tok/min · 10 req/day · 3 req/min
Gemini Robotics-ER 1.6	250,000 tok/min · 20 req/day · 5 req/min
Gemini Robotics-ER 1.5	250,000 tok/min · 20 req/day · 10 req/min
Gemma 4 31B Instruct	16,000 tok/min · 14,400 req/day · 30 req/min
Gemma 4 26B A4B Instruct	16,000 tok/min · 14,400 req/day · 30 req/min
Gemma 3 27B Instruct	15,000 tok/min · 14,400 req/day · 30 req/min
Gemma 3 12B Instruct	15,000 tok/min · 14,400 req/day · 30 req/min
Gemma 3 4B Instruct	15,000 tok/min · 14,400 req/day · 30 req/min
Gemma 3 1B Instruct	15,000 tok/min · 14,400 req/day · 30 req/min

■ NVIDIA NIM

Phone number verification required.

Models tend to be context-window limited.

Limits 40 requests/minute

Models

- [Various open models](#)

■ Mistral (La Plateforme)

Free tier (Experiment plan) requires opting into data training.

Requires phone number verification.

Limits Set per-model and per-organization — see your limits page. As of July 2026, a new free account sees 25,000–20,000,000 tokens/minute and 0.03–12.5 requests/second, depending on the model.

Models

- [Open and Proprietary Mistral models](#)

■ Mistral (Codestral)

Currently free to use.

Monthly subscription based.

Requires phone number verification.

Limits 30 requests/minute, 2,000 requests/day

Models

- [Codestral](#)

■ HuggingFace Inference Providers

Serverless Inference limited to models smaller than 10GB. Some popular models are supported even if they exceed 10GB.

Limits [\\$0.10/month in credits](#)

Models

- [Various open models across supported providers](#)

■ Vercel AI Gateway

Routes to various supported providers.

The free tier covers a subset of the model catalogue, with per-model rate limits.

Limits \$5/month

■ Kilo Gateway

OpenAI-compatible gateway routing to various providers. Free models work without an account.

All free models may use your prompts for training.

Limits 200 requests/hour per IP, shared across all free models

Free Models

- Cohere North Mini Code
- Kilo Auto Free (Router)
- Ling 3.0 Flash
- NVIDIA Nemotron 3 Nano Omni 30B A3B (Reasoning)
- NVIDIA Nemotron 3 Super 120B A12B
- NVIDIA Nemotron 3 Ultra 550B A55B
- NVIDIA Nemotron 3.5 Content Safety
- OpenRouter Free Models (Router)
- Poolside Laguna S 2.1
- Poolside Laguna XS 2.1
- StepFun Step 3.7 Flash

OpenCode Zen

AI gateway with curated models.

Free models may use data for improvement.

Free Models

- Big Pickle
- DeepSeek V4 Flash Free
- MiMo-V2.5 Free
- Laguna S 2.1 Free
- Ling-3.0-flash Free
- North Mini Code Free
- Nemotron 3 Ultra Free

Cerebras

Model Name	Model Limits
gpt-oss-120b	5 req/min · 30,000 tok/min · 1M tok/hour · 1M tok/day
zai-glm-4.7	5 req/min · 30,000 tok/min · 1M tok/hour · 1M tok/day
gemma-4-31b	5 req/min · 30,000 tok/min · 1M tok/hour · 1M tok/day

Model Name	Model Limits
Allam 2 7B	7,000 req/day · 6,000 tok/min
Llama 3.1 8B	14,400 req/day · 6,000 tok/min
Llama 3.3 70B	1,000 req/day · 12,000 tok/min
Whisper Large v3	2,000 req/day
Whisper Large v3 Turbo	2,000 req/day
canopylabs/orpheus-arabic-saudi	—
canopylabs/orpheus-v1-english	—
groq/compound	250 req/day · 70,000 tok/min
groq/compound-mini	250 req/day · 70,000 tok/min
meta-llama/llama-prompt-guard-2-22m	—
meta-llama/llama-prompt-guard-2-86m	—
openai/gpt-oss-120b	1,000 req/day · 8,000 tok/min
openai/gpt-oss-20b	1,000 req/day · 8,000 tok/min
openai/gpt-oss-safeguard-20b	1,000 req/day · 8,000 tok/min
qwen/qwen3.6-27b	1,000 req/day · 8,000 tok/min

| Cohere

Models share a common monthly quota.

Limits 20 requests/minute · 1,000 requests/month

Models

- c4ai-aya-expanse-32b
- c4ai-aya-vision-32b
- command-a-03-2025
- command-a-plus-05-2026
- command-a-reasoning-08-2025
- command-a-translate-08-2025
- command-a-vision-07-2025
- command-r-08-2024
- command-r-plus-08-2024
- command-r7b-12-2024
- command-r7b-arabic-02-2025

Cloudflare Workers AI

Limits 10,000 neurons/day

Models

- @cf/aisingapore/gemma-sea-lion-v4-27b-it
- @cf/google/gemma-4-26b-a4b-it
- @cf/ibm-granite/granite-4.0-h-micro
- @cf/moonshotai/kimi-k2.6
- @cf/moonshotai/kimi-k2.7-code
- @cf/nvidia/nemotron-3-120b-a12b
- @cf/openai/gpt-oss-120b
- @cf/openai/gpt-oss-20b
- @cf/qwen/qwen3-30b-a3b-fp8
- @cf/zai-org/glm-4.7-flash
- @cf/zai-org/glm-5.2
- DeepSeek R1 Distill Qwen 32B
- Gemma 2B Instruct (LoRA)
- Gemma 7B Instruct (LoRA)
- Llama 2 7B Chat (LoRA)
- Llama 3.1 8B Instruct (FP8)
- Llama 3.2 11B Vision Instruct
- Llama 3.2 1B Instruct
- Llama 3.2 3B Instruct
- Llama 3.3 70B Instruct (FP8)
- Llama 4 Scout Instruct
- Llama Guard 3 8B
- Mistral 7B Instruct v0.2 (LoRA)
- Mistral Small 3.1 24B Instruct
- Qwen 2.5 Coder 32B Instruct
- Qwen QwQ 32B

Providers with Trial Credits

One-time free credit balance on signup — not an ongoing free tier.

Fireworks

Credits \$1

Models

- [Various open models](#)

Baseten

Credits \$30

Models

- [Any supported model — pay by compute time](#)

Nebius

Credits \$1

Models

- [Various open models](#)

Novita

Credits \$0.50 for 1 year

Models

- [Various open models](#)

AI21

Credits \$10 for 3 months

Models

- [Jamba family of models](#)

| [Upstage](#)

Credits \$10 for 3 months

Models

- Solar Pro / Mini

| [NLP Cloud](#)

Credits \$15

Requirements Phone number verification

Models

- Various open models

| [Alibaba Cloud \(International\) Model Studio](#)

Credits 1 million tokens per model, valid for 90 days (Singapore endpoint only)

Models

- [Various open and proprietary Qwen models](#)

| [Modal](#)

Credits \$30/month on the Starter plan

Models

- Any supported model — pay by compute time

| [Inference.net](#)

Credits \$1 (up to \$25 for responding to an email survey)

Models

- Various open models

Hyperbolic

Credits \$1

Models

- DeepSeek V3 0324
- Llama 3.3 70B Instruct
- deepseek-ai/deepseek-r1-0528
- qwen/qwen3-coder-480b-a35b-instruct

SambaNova Cloud

Credits \$5 for 3 months

Models

- deepseek-v3.1
- deepseek-v3.2
- gemma-4-31b-it
- gpt-oss-120b
- meta-llama-3.3-70b-instruct
- minimax-m2.7

| Scaleway Generative APIs

Credits 1,000,000 free tokens, plus 60 minutes of audio transcription

Models

- BGE-Multilingual-Gemma2
- Gemma 3 27B Instruct
- Llama 3.3 70B Instruct
- Pixtral 12B (2409)
- Whisper Large v3
- devstral-2-123b-instruct-2512
- gemma-4-26b-a4b-it
- glm-5.2
- gpt-oss-120b
- holo2-30b-a3b
- mistral-medium-3.5-128b
- mistral-small-3.2-24b-instruct-2506
- qwen3-235b-a22b-instruct-2507
- qwen3-coder-30b-a3b-instruct
- qwen3-embedding-8b
- qwen3.5-397b-a17b
- qwen3.6-35b-a3b
- voxtral-small-24b-2507

Closing Notes

Rate limits, credit amounts, and model availability change frequently at each provider's sole discretion. Please verify current terms directly on each provider's website before relying on a free tier for production workloads.