

Free AI API Platforms

Ongoing Free Tiers for a Unified LLM Routing Service

April 2026

Only platforms with ongoing monthly free access — no expiring trial credits.

Contents

1	Executive Summary	3
2	Platform-by-Platform Analysis	3
2.1	Google AI Studio (Gemini API)	3
2.2	Groq	3
2.3	Cerebras	4
2.4	SambaNova	4
2.5	NVIDIA NIM	5
2.6	Mistral (Experiment Plan)	5
2.7	OpenRouter (Free Models)	5
2.8	GitHub Models	6
2.9	Other Platforms	6
3	Comprehensive Rankings	7
3.1	By Intelligence (Best Free Model Per Platform)	7
3.2	By Monthly Token Budget	7
3.3	Final Composite Ranking	7
4	Architecture for Unified Routing Service	9
4.1	Routing Priority (by intelligence)	9
4.2	Key Pooling Multiplier	9
4.3	Recommended Architecture	9
4.4	Excluded Platforms	9

1 Executive Summary

This report catalogs every major platform offering **ongoing** free API access to LLMs (not one-time expiring trial credits). The goal: a service where users contribute their free API keys, and a unified endpoint routes requests to the best available free LLM, ranked by intelligence.

Key Findings:

- **13 platforms** offer genuinely ongoing free tiers. None require a credit card.
- **Google AI Studio** (Gemini 2.5 Pro) offers the highest-intelligence model for free.
- **Cerebras** and **NVIDIA NIM** offer the most generous throughput.
- **Groq** and **Cerebras** offer the fastest inference speeds.

2 Platform-by-Platform Analysis

2.1 Google AI Studio (Gemini API)

Free Tier Type	Ongoing, no expiration
Credit Card	No
Best Free Model	Gemini 2.5 Pro
Other Models	Gemini 2.5 Flash, Gemini 2.5 Flash-Lite

Rate Limits:

Model	RPM	RPD	TPM
Gemini 2.5 Pro	5	100	250,000
Gemini 2.5 Flash	10	250	250,000
Gemini 2.5 Flash-Lite	15	1,000	250,000

Monthly Token Budget: 12M tokens (Pro), 30M (Flash), 120M (Flash-Lite)

Benchmarks (Gemini 2.5 Pro):

Benchmark	Score
Global MMLU	89.8%
MMLU-Pro	86.0%
AIME	88.0%
GPQA	84.0%
SWE-Bench Verified	63.8%
Chatbot Arena ELO	1450+

Speed: 80–150 tokens/sec

Limitations: Free tier data may be used for training. Rate limits reduced 50–80% in Dec 2025 due to abuse. Limits are per-project.

2.2 Groq

Free Tier Type	Ongoing, no expiration
-----------------------	------------------------

Credit Card	No
Best Free Model	Llama 3.3 70B Versatile
Other Models	Llama 4 Scout, Qwen3 32B, Llama 3.1 8B, Kimi K2, 15+ more

Rate Limits:

Model	RPM	RPD	TPM	TPD
Llama 3.3 70B	30	1,000	6,000	500K
Llama 4 Scout 17B	30	1,000	30,000	1M
Qwen3 32B	60	1,000	6,000	500K
Llama 3.1 8B	30	14,400	6,000	500K

Monthly Token Budget: 15M/month per model, 45–60M combined

Benchmarks (Llama 3.3 70B): MMLU 82.0%, HumanEval 88.4%, Arena ELO 1250

Speed: 276–316 tok/sec (standard), up to 1,665 tok/sec (speculative decoding)

Limitations: Cached tokens don't count toward limits (advantage). Only open-source models.

2.3 Cerebras

Free Tier Type	Ongoing, no expiration
Credit Card	No
Best Free Model	Qwen3 235B-A22B Instruct
Other Models	Llama 3.1 8B/70B, Llama 4 Scout, GPT-OSS 120B

Rate Limits:

RPM	30
TPM	60,000
Tokens/Day	1,000,000
Context Window (free)	8,192 tokens

Monthly Token Budget: 30M tokens/month

Benchmarks (Qwen3 235B): MMLU 88.4%, HumanEval 79.2%, AIME '24 85.7%, Arena ELO 1422

Speed: 1,400 tok/sec (Qwen3 235B), 2,600 tok/sec (Scout), 1,800 tok/sec (8B)

Limitations: Context window capped at 8,192 tokens on free tier (major limitation).

2.4 SambaNova

Free Tier Type	Ongoing (after initial \$5 credit expires)
Credit Card	No
Best Free Model	Llama 3.1 405B / MiniMax-M2.5
Other Models	Llama 3.3 70B, Qwen3 32B, DeepSeek V3.1, Llama 4 Maverick

Rate Limits:

Model	RPM	TPD
Llama 3.1 405B	10	200K
Llama 3.3 70B	20	200K
Llama 3.1 8B	30	200K

Monthly Token Budget: 6M tokens/month

Benchmarks (Llama 3.1 405B): MMLU 88.6%, HumanEval 89.0%, MATH 73.8%, Arena ELO 1320

Speed: 114 tok/sec (405B)

2.5 NVIDIA NIM

Free Tier Type	Ongoing, rate-limited (no token cap)
Credit Card	No (requires NVIDIA Developer signup)
Best Free Model	Nemotron 3 Super 120B, Kimi K2.5, GLM-5 744B, DeepSeek-R1 671B
Catalog	100+ models

Rate Limits: 40 RPM, no daily token cap

Monthly Token Budget: 50–100M tokens/month (practically)

Speed: Varies by model; NIM-optimized for throughput

Limitations: Intended for prototyping/evaluation. Heavy models may be slow at peak times.

2.6 Mistral (Experiment Plan)

Free Tier Type	Ongoing (Experiment plan)
Credit Card	No (requires phone verification)
Best Free Model	Mistral Large 3
Other Models	Codestral, Mistral Small, all Mistral models

Rate Limits: 2 RPM, 500K TPM, 1B monthly token cap

Monthly Token Budget: 50–100M tokens/month (2 RPM is the bottleneck)

Benchmarks (Mistral Large 3): MMLU 85.5%, Arena ELO 1280

Limitations: Only 2 RPM is extremely restrictive. Data may be used for training.

2.7 OpenRouter (Free Models)

Free Tier Type	Ongoing, free model variants
Credit Card	No
Best Free Model	DeepSeek R1 (free), Qwen3 Coder 480B (free)
Free Models	29 total, including Gemma 3, Nemotron 3 Super

Rate Limits:

Tier	RPM	RPD
No credits purchased	20	50

\$10+ credits purchased	20	1,000
-------------------------	----	-------

Monthly Token Budget: 6M (no credits) / 120M (\$10 purchase)

Benchmarks (DeepSeek R1 free): MMLU 90.8%, AIME ‘24 79.8%, Arena ELO 1398

2.8 GitHub Models

Free Tier Type	Ongoing
Credit Card	No (requires GitHub account)
Best Free Model	GPT-4o, DeepSeek-R1, Llama 3.3 70B

Rate Limits:

Tier	RPM	RPD	Input Tok/Req	Output Tok/Req
High (GPT-4o)	10	50	8,000	4,000
Low (smaller)	15	150	8,000	4,000

Monthly Token Budget: 18M (high), 54M (low)

Benchmarks (GPT-4o): MMLU 88.7%, HumanEval 90.2%, Arena ELO 1350

2.9 Other Platforms

Platform	Best Free Model	Monthly Tokens	Notes
Hugging Face	Various (1000s)	5–10M	100K inference credits/mo
Cohere	Command R+	4M	1,000 calls/mo, 20 RPM
Cloudflare Workers AI	Llama 3.1 70B	18–45M	10K neurons/day
Fireworks AI	Open-source	5–10M	10 RPM (after \$1 credit)

3 Comprehensive Rankings

3.1 By Intelligence (Best Free Model Per Platform)

#	Platform	Best Free Model	MMLU	Human Eval	Arena ELO	Tier
1	Google AI Studio	Gemini 2.5 Pro	89.8%	92%	1450	Frontier
2	OpenRouter	DeepSeek R1 (free)	90.8%	85%	1398	Frontier
3	Cerebras	Qwen3 235B	88.4%	79.2%	1422	Near-Frontier
4	SambaNova	Llama 3.1 405B	88.6%	89.0%	1320	Near-Frontier
5	GitHub Models	GPT-4o	88.7%	90.2%	1350	Near-Frontier
6	Cohere	Command R+	88.2%	–	1200	Strong
7	Mistral	Mistral Large 3	85.5%	–	1280	Strong
8	NVIDIA NIM	Nemotron 3 / GLM-5	–	–	1300	Strong
9	Groq	Llama 3.3 70B	82.0%	88.4%	1250	Good
10	Cloudflare	Llama 3.1 70B	82.0%	88.4%	1250	Good

3.2 By Monthly Token Budget

#	Platform	Est. Monthly Tokens	Budget Tier
1	NVIDIA NIM	50–100M+	Excellent
2	Mistral	50–100M	Excellent
3	Google AI Studio (Flash-Lite)	120M	Excellent
4	Cloudflare Workers AI	18–45M	Very Good
5	Cerebras	30M	Very Good
6	GitHub Models	18–54M	Good
7	Groq	15–60M	Good
8	Hugging Face	5–10M	Moderate
9	SambaNova	6M	Moderate
10	OpenRouter (no credits)	6M	Moderate
11	Fireworks AI	5–10M	Moderate
12	Cohere	4M	Limited

3.3 Final Composite Ranking

Scoring: Intelligence (0–40) + Generosity (0–30) + Usability (0–20) + Reliability (0–10)

#	Platform	Best Model	Intel	Gener.	Usab.	Rel.	Total
1	Google AI Studio	Gemini 2.5 Pro	40	18	14	8	80
2	Cerebras	Qwen3 235B	35	22	16	7	80
3	NVIDIA NIM	100+ models	32	28	15	5	80
4	Groq	Llama 3.3 70B	28	20	20	8	76

5	Cloudflare	Llama 3.1 70B	28	22	12	6	68
6	OpenRouter	DeepSeek R1	38	12	12	6	68
7	GitHub Models	GPT-4o	34	16	10	7	67
8	Mistral	Mistral Large 3	30	24	6	6	66
9	SambaNova	Llama 3.1 405B	34	10	12	7	63
10	Cohere	Command R+	30	8	14	7	59
11	Hugging Face	Various	25	12	10	5	52
12	Fireworks AI	Open-source	25	10	10	5	50

4 Architecture for Unified Routing Service

4.1 Routing Priority (by intelligence)

1. **Gemini 2.5 Pro** (Google AI Studio) — highest intelligence, 100 RPD/key
2. **DeepSeek R1** (OpenRouter free) — near-frontier reasoning, 50 RPD/key
3. **Qwen3 235B** (Cerebras) — near-frontier, 1M tokens/day, 8K context limit
4. **GPT-4o** (GitHub Models) — strong, 50 RPD/key
5. **Llama 3.1 405B** (SambaNova) — strong, 10 RPM
6. **Mistral Large 3** (Mistral) — good, 2 RPM bottleneck
7. **Llama 3.3 70B** (Groq) — good intelligence, fastest speed, 1,000 RPD
8. **Any NIM model** (NVIDIA) — huge variety, no daily token cap

4.2 Key Pooling Multiplier

If 100 users each contribute one API key per platform:

Platform	Per Key RPD	100 Keys RPD
Google AI Studio (Pro)	100	10,000
Groq (70B)	1,000	100,000
Cerebras	33K tok/hr	3.3M tok/hr
OpenRouter (R1)	50	5,000
GitHub Models (GPT-4o)	50	5,000
NVIDIA NIM	40 RPM	4,000 RPM

4.3 Recommended Architecture

“**Quality burst**” backends (highest intelligence, low per-key limits):

- Gemini 2.5 Pro, DeepSeek R1, GPT-4o

“**Workhorse**” backends (high throughput, good intelligence):

- Cerebras Qwen3 235B (30M tok/mo/key)
- NVIDIA NIM (no daily cap, 100+ models)
- Groq Llama 3.3 70B (fast, reliable)

“**Speed**” backends (real-time chat):

- Groq: 276–1,665 tok/sec
- Cerebras: 1,400–2,600 tok/sec

4.4 Excluded Platforms

Platform	Reason
OpenAI	One-time \$5 trial credit, expires
Anthropic	One-time trial credits only
Together AI	\$25 signup credit, no confirmed ongoing free tier
DeepSeek	5M free tokens expire in 30 days (but API is near-free at \$0.28/M)

Free tier details change frequently. Verify current limits on each platform's pricing page. Benchmark scores from published papers, LMSYS Chatbot Arena, and OpenLLM Leaderboard as of April 2026.