@thefutureguyy.ai

FREE TIER · 100+ MODELS · NO CREDIT CARD

@nvidia FREE AI MODELS

How to Claim Your Free API Access

Step-by-step guide to 100+ hosted AI models on NVIDIA NIM

WHAT YOU GET

100+ AI Models	via hosted API — LLMs, vision, speech, embeddings
FREE	No billing, no credit card, no trial limits
OpenAI-Compatible	Drop-in replacement — works with any OpenAI SDK
40 RPM	Free tier rate limit — enough for active development
DGX Cloud GPUs	Runs on NVIDIA Blackwell & Hopper infrastructure

MODELS FEATURED

`MiniMax M2.7`	`GLM 5.1`	`Kimi K2.5`	`DeepSeek R1`	`GPT-OSS-120B`
`Sarvam-M`	`Nemotron-Super-120B`	`Llama 3.3 70B`	`Mistral Small 4`	`Gemma 4 31B`

PLUG INTO YOUR FAVOURITE TOOLS

Open Claude	**OpenCode**	**Zed IDE**	**Hermes Agent**	**Cursor IDE**	**Any OpenAI SDK**

Claim Your Free NVIDIA NIM API

Follow these steps exactly — takes under 5 minutes.

01 Go to NVIDIA Build Platform

Open your browser and visit:

- `https://build.nvidia.com`

This is NVIDIA's official API catalog for NIM (Inference Microservices). You'll see 100+ models from DeepSeek, Kimi, MiniMax, Llama, Nemotron, and more.

02 Create a Free Account

- Click 'Login' at the top-right corner of the page.
- Sign up using your email address or Google / GitHub SSO.
- No credit card required. The Developer Program is completely free.
- Once signed in, you'll land on the model catalog dashboard.

03 Generate Your API Key

Navigate to:

- `https://build.nvidia.com/settings/api-keys`
- Click '+ Generate API Key'.
- Your key will look like: nvapi-xxxxxxxxxxxxxxxxxxxx
- Copy it immediately and store it safely — it won't be shown again.

Treat it like a password. Do NOT commit it to GitHub.

04 Browse & Pick a Model

Return to the main catalog at build.nvidia.com/models

- Filter by category: LLM, Vision, Speech, Embedding, etc.
- Click on any model (e.g. 'MiniMax M2.7') to open its card.
- Click 'View Code' to get the exact model ID you'll use in API calls.

Recommended free models to start with:

`moonshotai/kimi-k2.5`	(reasoning + coding)
`deepseek-ai/deepseek-r1`	(math + logic)
`nvidia/nemotron-super-120b`	(agentic tasks)
`minimax/minimax-m2.7`	(general purpose)

05 Make Your First API Call

Use the OpenAI Python SDK — no special NVIDIA SDK needed.

```
$ pip install openai
```

```
from openai import OpenAI

client = OpenAI(
    base_url='https://integrate.api.nvidia.com/v1',
    api_key='nvapi-YOUR_KEY_HERE'
)

resp = client.chat.completions.create(
    model='moonshotai/kimi-k2.5',
    messages=[{'role':'user','content':'Hello!'}]
)
print(resp.choices[0].message.content)
```

Plug Into Your Dev Tools

Open Claude / OpenCode	Set base_url + api_key in model config. Choose 'OpenAI-compatible' endpoint.
Cursor IDE	Settings > Models > Add Custom. Use integrate.api.nvidia.com/v1 as base.
Zed IDE	Edit ~/.config/zed/settings.json. Add nvidia as custom LLM provider.
Hermes Agent	Pass base_url and api_key in the agent config YAML or env vars.
Any OpenAI SDK app	Just swap base_url and api_key. Zero other code changes needed.

Pro Tips & Limits to Know

- **Rate Limit:** 40 requests per minute on the free tier. Space out heavy workloads.
- **Model IDs:** Always copy model IDs from 'View Code' on the model card — they're case-sensitive.
- **Streaming:** Add stream=True to your API call for real-time token-by-token output.
- **Large Models:** DeepSeek R1 (671B) and GLM-5 (744B) may have slower response during peak hours.
- **No lock-in:** Same OpenAI-compatible code works on OpenRouter, Groq, Together.ai — just swap keys.
- **Cost:** Rate-limit based billing — no per-token charges on the free tier as of 2025.

You're all set. Start building.

100+ free AI models · OpenAI-compatible · No billing

build.nvidia.com/models