

Free AI API Keys From NVIDIA

NVIDIA runs a model catalogue where a large slice of the models have a free endpoint you can call from your own code. Here is the real count, the exact flow, and the two catches nobody mentions. Checked on build.nvidia.com, July 2026.

01
The real number is 73Not 80+

On **build.nvidia.com/models** the header says **138 models** and the Filters panel breaks them down: **Free Endpoint 73**, Partner Endpoint 42, Download Available 106. Every viral post quotes "80+". Today it is **73** — the number moves, so check the facet before you trust anyone's screenshot.

02
Pick a model with a free endpointStep one
Open the catalogue and switch on the **Free Endpoint** filter so you only see models you can actually call for nothing. The home page shelf is titled “Use Inference Endpoints — **Free inference with leading models**” and lists real ones: **z-ai glm-5.2**, **nvidia nemotron-3-ultra-550b-a55b**, **deepseek-ai deepseek-v4-pro**, thinkingmachines inkling.

03
Generate the keyStep two
Open the model you want and generate an API key from its page. This is where the first catch bites: you need to be **signed in with an NVIDIA account**. Free does not mean anonymous.

04
Point your agent at itStep three
The endpoints are OpenAI-compatible, so you set the base URL and the key and carry on. That means **Claude Code**, **Cursor**, or **your own scripts** keep working unchanged — you are only swapping who serves the tokens.
Swapping the endpoint, not the code

Copy

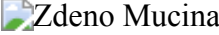
```
export OPENAI_BASE_URL="https://integrate.api.nvidia.com/v1"
export OPENAI_API_KEY="nvapi-..." # from the model page
```

then run your existing tool as normal — same SDK, same calls.
verify the model id on the model's own page before you hardcode it.

The Stack · everything you need

- [Claude Code](#)
- [Claude Code](#)
- [The Agent](#)
- [The AI agent that writes, designs, cuts and ships everything in this guide. Free to install.](#)
- [Install →](#)
- [Zernio](#)
- [Zernio](#)
- [Posting + DMs](#)
- [Posts to IG, TikTok and story, and runs the comment-to-DM automation that delivered this page.](#)
- [Start posting →](#)
- [ElevenLabs](#)
- [ElevenLabs](#)
- [Your Voice](#)
- [Clones your voice from 60s of audio and returns word timings your edits can sync to.](#)
- [Get your voice →](#)
- [HeyGen](#)
- [HeyGen](#)
- [Your Avatar](#)
- [Your talking avatar for hook shots and full videos, driven by script or API.](#)
- [Try it →](#)

Want my full agent setup?
How I wire cheap and free endpoints into a working agent stack, inside the community.
[Try It Free →](#)

Zdeno Mucina

Not ready to join yet?

Take one system like this per week by email instead.

Send it weekly →

[Download as PDF ↓](#)
[Start My Free Trial →](#)