Build and deploy clinical AI agents with NVIDIA Nemotron and Nebius Token Factory

Nemotron gives you the model. Token Factory ships it. You own the result.
Open models with the infrastructure to deploy them — built for the complexity of healthcare.

Nemotron family models at Token Factory

NVIDIA NemotronTM is a family of open models, weights, and libraries purpose-built for customization, giving digital health companies the foundation to deploy specialized clinical agents across ambient care, decision support, and patient outcome modeling.

Nemotron 3 Nano 30B

Compact MoE model optimized for efficient reasoning, chat, and coding with strong multilingual support and long-context RAG/ agent workflows

Nemotron 3 Nano Omni

The most open, efficient, and accurate omni-modal reasoning model for agentic AI.

Nemotron 3 Super 120b

Hybrid MoE model optimized for efficient multi-agent AI and complex reasoning tasks.

Nemotron 3 Ultra 550b

Frontier hybrid MoE model optimized for longrunning autonomous agents, deep research, and high throughput workflows.

Why digital health teams use Token Factory

Own your model foundation

Healthcare moves too fast for a one-size-fits-all approach. Nemotron open weights mean you can fine tune on your own clinical datasets, adapt to your patient population, and build a model that compounds in value over time.

Built for regulated industries like healthcare

From a startup scaling across markets to a health system with strict data residency requirements, Nemotron deploys on the infrastructure you own. No shared cloud dependencies, no data leaving your environment.

From evaluation to production, with urgency

Nebius Token Factory removes the infrastructure bottleneck — serverless fine-tuning, dedicated endpoints, and pertoken pricing so you can move from model selection to shipped agent without standing up your own compute.

Choose your path for a quick start

Playground

Experiment, compare models and tune prompts side-by-side.

OpenAI-compatible API

Move your endpoint to Token Factory and switch models in seconds.

Welcome to Nebius Token Factory

Get API key

Build, fine-tune, and run open models at production scale

Start building

1. Create an API key and save it into the `NEBIUS_API_KEY` environment variable.
2. Use the code below to start working with model, its parameters and the chat so far in your application.

Python (OpenAI) | cURL | JavaScript

```python
import os
from openai import OpenAI

client = OpenAI(
    base_url="https://api.studio.nebius.com/v1/",
    api_key=os.environ.get("NEBIUS_API_KEY")
)
```

Post-training

Run serverless fine-tuning with per-token pricing to build custom text-to-text models.

Data Lab

Create, explore, and manage datasets in one place, and seamlessly reuse them across Batch Inference and Fine-tuning workflows.

Dedicated Endpoints

Isolated model deployments with reserved GPU capacity, autoscaling, and OpenAI-compatible inference. Your region, your configuration.

Create endpoint

Featured models

Handpicked models optimized for production, reasoning, and developer workflows.

New

NVIDIA
Nemotron-3-Ultra-550b-a55b

$1.00 / 1M In | **$3.00** / 1M Out | **59** Tok/s

Base | us-central1 | Text-to-text

Nemotron 3 Ultra is a 550B hybrid MoE model from NVIDIA, optimized for the most demanding multi-agent AI and complex reasoning tasks.

NVIDIA
Nemotron-3-Nano-Omni

$0.06 / 1M In | **$0.24** / 1M Out | **90** Tok/s

Base | eu-north1 | Text-to-text

The most open, efficient, and accurate omni-modal reasoning model for agentic AI.

NVIDIA
Nemotron-3-Super-120b-a12b

$0.30 / 1M In | **$0.90** / 1M Out | **127** Tok/s

Base | us-central1 | Text-to-text

Nemotron 3 Super is a 120B hybrid MoE model optimized for efficient multi-agent AI and complex reasoning tasks.

NVIDIA
NVIDIA Nemotron-3-Nano-30B-A3B

$0.06 / 1M In | **$0.24** / 1M Out | **60** Tok/s

Base | eu-north1 | Text-to-text

Compact MoE model optimized for efficient reasoning, chat, and coding with strong multilingual support and long-context RAG/ agent workflows.

Trusted by AI practitioners and partners

sword

Sword Health delivered 10 million AI-guided care sessions in 2025 using Nebius infrastructure.

slingshot AI

Slingshot AI uses Nebius to build and scale Ash, a psychology-focused foundation model.

PRIMA MENTE

Prima Mente uses Nebius to train Pleiades, a model to treat Alzheimer's and other neurodegenerative diseases.

YN² · Genesis Therapeutics · The Sequencing Center · QUANTORI · compugen FROM CODE TO CURE® · elixir · SieveStack · Stanford · Utrecht University · XYME · lynxanalytics · xAID · TrialHub

Start building today!

tokenfactory.nebius.com

ISO/IEC 27001 | SOC 2 Type II | HIPAA