A Practical Guide to AI-Powered Catalog Enrichment

Blueprints from NVIDIA & Nebius

Andrew Eliseev
Head of Retail & Commerce AI,
Nebius

Antonio Martinez
GenAI Evangelist Product Engineer,
NVIDIA

Manuel Gomez
Solution Architect,
Nebius

Poll 1:
Let's get to know each other

Housekeeping

Use the Q&A tab for questions

The recording and materials will be sent after the session

Poll 1: Results

Agenda

Retail's AI execution challenge

1

NVIDIA Retail AI Blueprints

2

Demo & Architecture

3

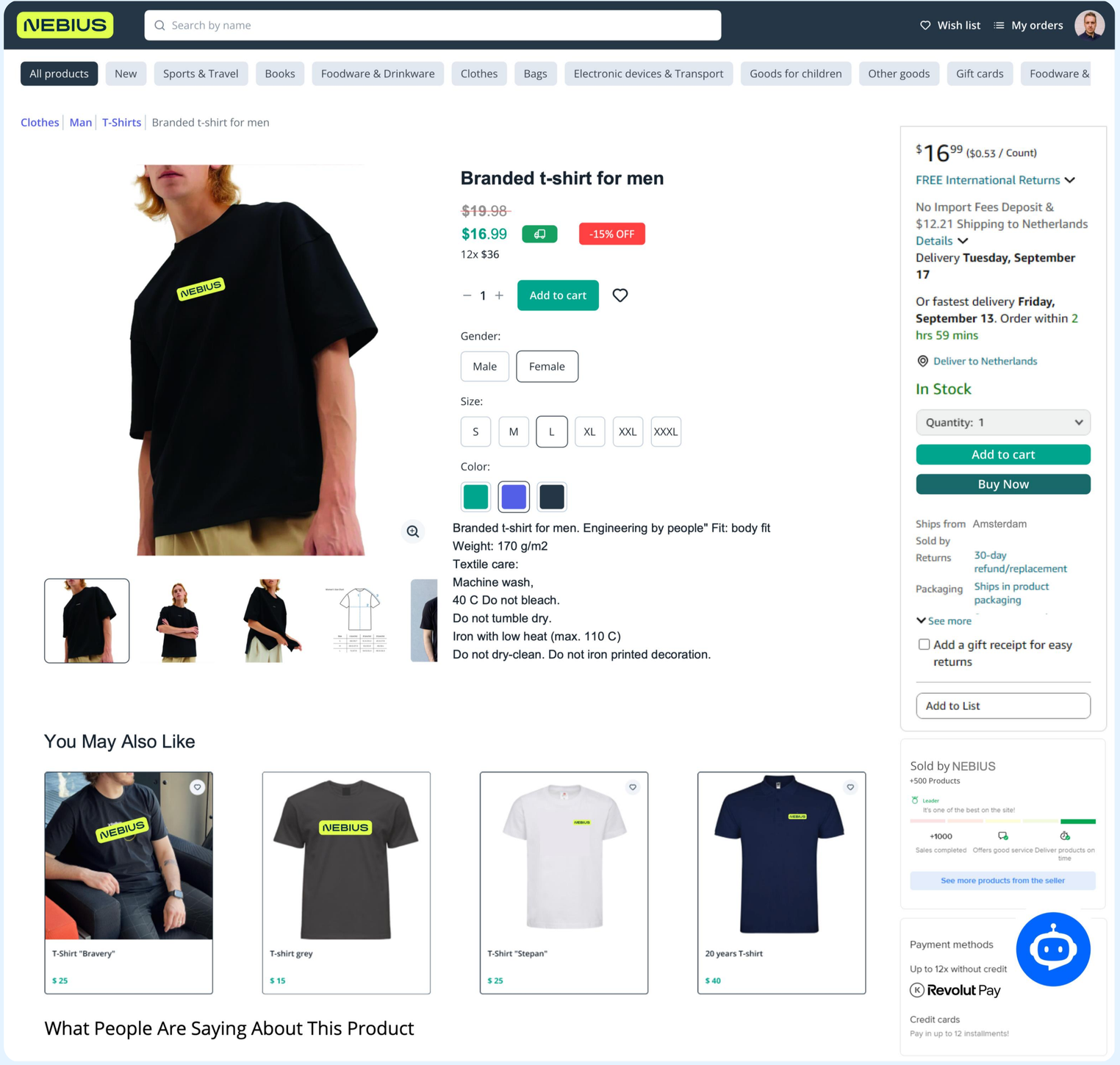
Q&A

4

Poll 2:
How many AI models are running on a single product detail page?

How many AI models are running on a single product detail page?

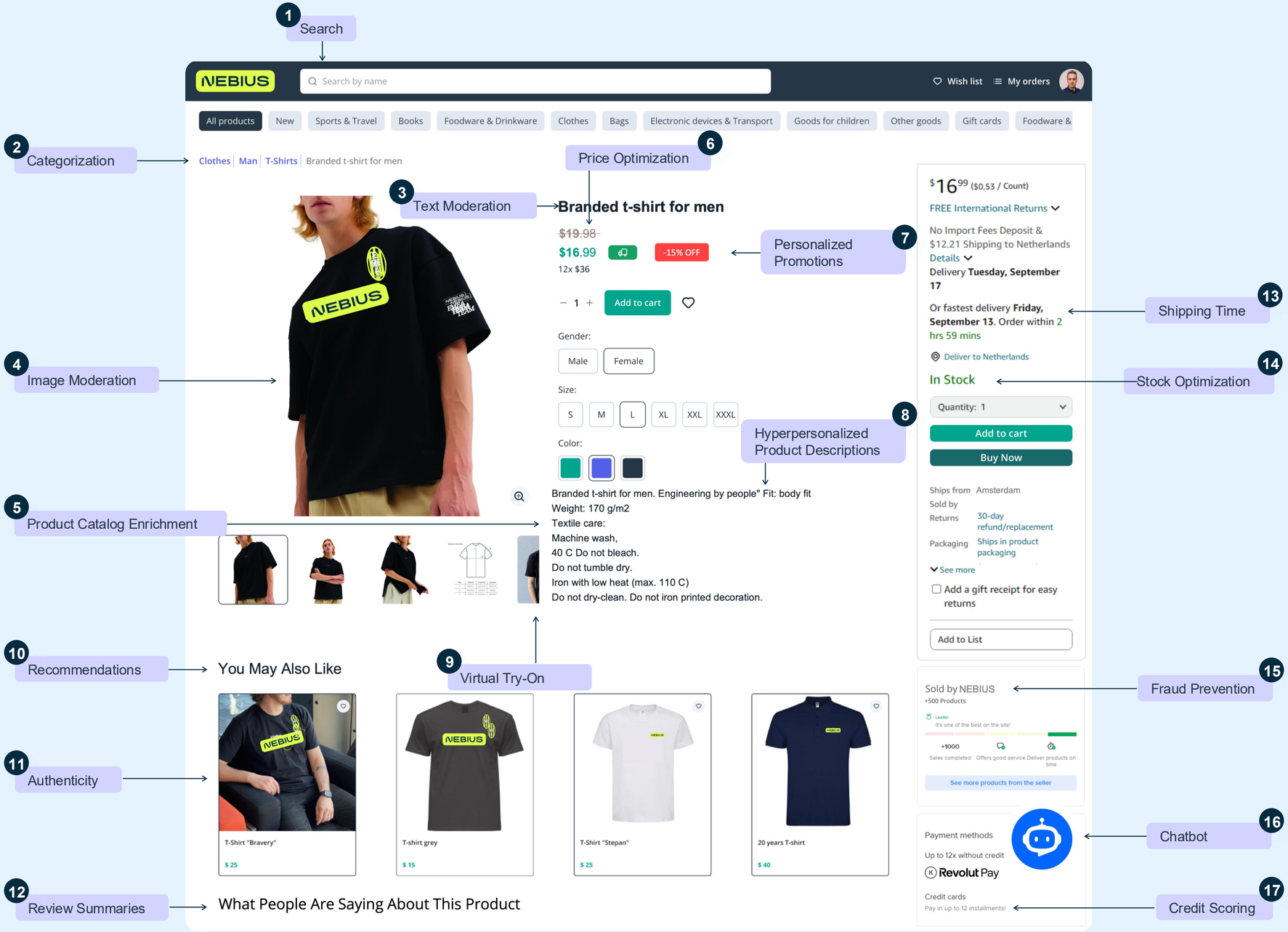
Modern, high performing PDPs infuse AI & GenAI throughout the user experience



Poll 2: Results

How many AI models are running on a single product detail page?

Modern, high performing PDPs infuse AI & GenAI throughout the user experience



Retail's ambition and ROI are proven.

Execution is the wall.

The industry is sold on AI and seeing real returns — but most still can't get agents into production.

WHERE THE MOMENTUM IS

91%
using or assessing AI
up from 75% in 2024

89%
say AI is increasing revenue

9 in 10
increasing AI budgets this year

95%
say AI is reducing costs

BUT EXECUTION STALLS

47% → 20%
explore agentic AI — only 1 in 5 in production

46%
cite the AI talent gap as a #1 barrier
up from 31% last year

Digital commerce is now the #1 focus area (**61%**) — and catalog enrichment, today's topic, is a top AI investment (**42%**).

The gap isn't strategy — it's execution infrastructure.
Retailers need production-ready architectures they can deploy now — on infrastructure built for open-model inference at scale.

Source: NVIDIA — 2026 State of AI in Retail and CPG survey

Nebius is the purpose-built AI cloud that agentic commerce runs on

Our mission

Fulfill the potential of AI by accelerating innovation globally and at scale with purpose-built infrastructure, serving the ecosystem of AI innovators of all sizes.

Engineering-led by design

Decades of infrastructure experience, now fueling next-gen AI workloads.

Nasdaq

NASDAQ-listed as NBIS

Backed by deep capital and built for long-term scale.

NVIDIA

Strategic NVIDIA partner

Nebius is a Reference Platform NVIDIA Cloud Partner, with validated reference architectures and close collaboration with NVIDIA for enablement.

Vertically integrated

Full-stack control, from hardware manufacturing to platform software.

Retail & Commerce AI

Retail AI is inference-heavy, latency-sensitive and seasonal — the specific workload profile for production token economics Nebius is built for

Agentic commerce & conversational AI

Shopping assistants, customer-service agents and autonomous checkout, grounded in real-time inference.

LLM inference · Agentic orchestration · Agentic search

Personalisation & recommendation

Recommendation and dynamic-pricing models that adapt in real time, served at catalogue scale.

Real-time inference · Ranking · Dynamic pricing

Demand forecasting & supply chain

Time-series and graph models that forecast demand and optimise inventory across millions of SKUs.

Time-series · Graph models · Large-scale training

Computer Vision for online, in-store & warehouse

Shelf monitoring, loss prevention, virtual try-on and visual search on multimodal pipelines.

Multimodal training · Vision inference · Video analytics

Why Nebius for Retail & CPG

Inference economics at scale
Sub-second latency and materially lower cost-per-token than hyperscaler list pricing.

Peak-season elasticity
Scale to Black Friday capacity in minutes, pay-as-you-go, with no GPU reservation queues.

Proven, and built with NVIDIA
NVIDIA Reference Platform Cloud Partner, SOC 2 Type II and ISO 27001.

Retail Clients

shopify · Photoroom · CATCHES · ZEBRA · prosus

End-to-end AI Cloud Platform

AIOps

Token Factory

Model Hub | NVIDIA NIMS | NVIDIA Blueprints

MLOps

Applications | Managed MLflow | JupyterLab | vLLM | ComfyUI | …

Serverless AI | DevPods* | Jobs | Endpoints

DataOps | Managed PostgreSQL | Toloka

AI orchestration | Managed Soperator | SkyPilot | Ray/Anyscale

*available in Q2'26

IaaS

Accelerated compute | Virtual machines | Managed Kubernetes | Block volumes

AI storage | Object storage | Shared filesystem | Data Transfer Service

Networking | NVIDIA NDR/XDR InfiniBand | VPC routing

Hardware

NVIDIA GPU servers | GB300 NVL72 | HGX B300 | HGX B200 | HGX H200 | HGX H100 | RTX PRO 6000

NEBIUS Enterprise platform

Access and experience

- Self-service
- GUI, API, IaC, SDKs

Security and compliance

- IAM (SSO, GBAC)
- Audit Logs and secrets mgmt.
- SOC 2, ISO, GDPR, HIPAA

Cost control and billing

- FOCUS-compliant data export
- Capacity and quota mgmt.

Operations and reliability

- Observability
- Health checks and auto-healing
- Multi-stage hardware tests

NVIDIA Retail AI Blueprints

NVIDIA Provides the Building Blocks for Agentic AI

NVIDIA AI Blueprints


Shopping Assistant


Catalog Enrichment


Agentic Commerce


Supply Chain


Video Analytics

NVIDIA NeMo

Curator | Customizer | Evaluator | Guardrails | Retriever

NVIDIA NIM


Understanding & Reasoning


Information Retrieval


AI Safety


Digital Humans


Visual Content Generation


Digital Biology


Physical AI

Accelerated Infrastructure


GPU


CPU


DPU


Networking

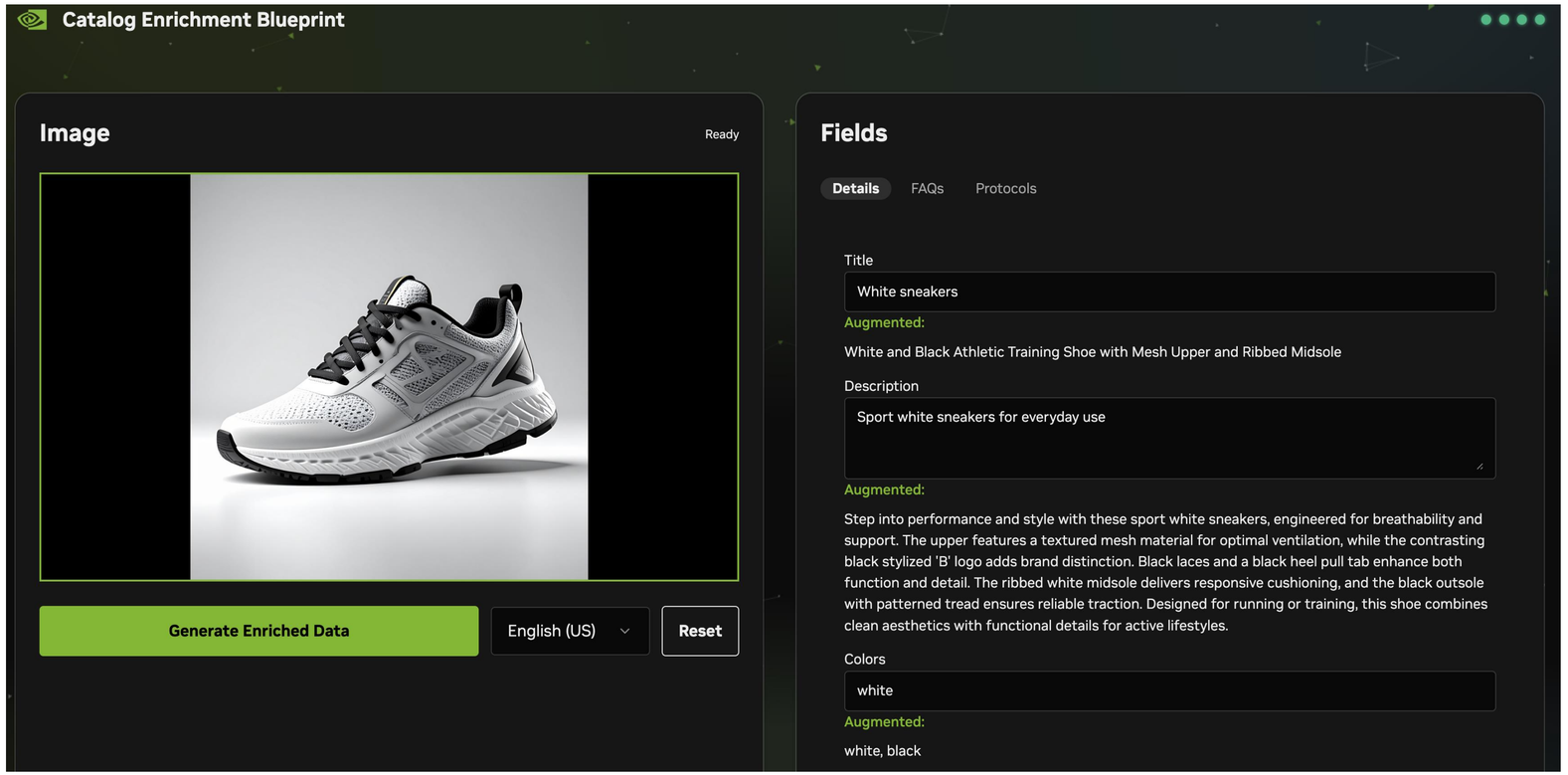
Agentic Commerce Journey

NVIDIA

Catalog Enrichment

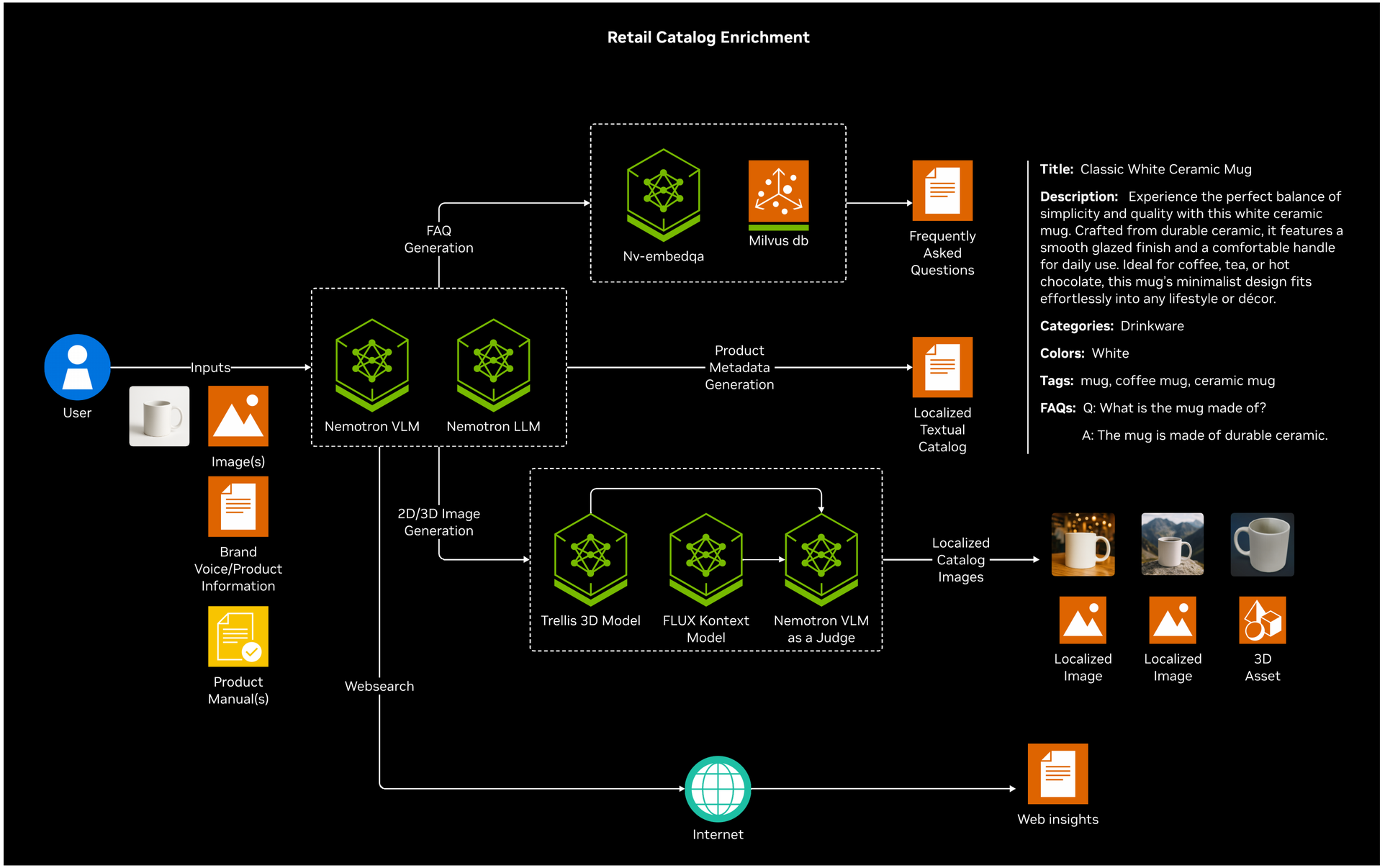
Strong data foundation for the Agentic Commerce

1. **Reduce** manual effort and cost
2. **Stronger data foundation** for high-margin services (ads, personalization, targeting)
3. **Improve discoverability and traffic**
   - Boost organic SEO/GEO, on-site search relevance.
4. **Reduce return rate** by delivering more accurate product information.
5. **Accelerate localization**
   - Description tailored for targeted markets
   - Multi-lingual
6. **Social media product enrichment**
   - Real consumer language, pain points, and trending features.



Arch Diagram

Catalog Enrichment



Demo & Architecture: Retail Catalog Enrichment

What’s left in my hands

Three things before Q&A

See it run, live

A bare product image + sparse data → enriched copy, FAQs, localized imagery, a 3D asset and agentic-commerce protocols – one pipeline.

1

The 5 models, up close

One slide per NIM: who built it, size, context, fine-tuning fit, and what you'd swap it for.

2

Make it yours on Nebius

Watch one architecture evolve – VM → Kubernetes → serverless – then batch and training options.

3

LIVE DEMO

Switching to the Catalog Enrichment console

Five NIM microservices, one node

Role	Model (NIM)	Does	
VLM	**Nemotron-3 Nano Omni 30B-A3B**	Visual analysis · VLM-judge	GPU 0
LLM	**Nemotron-3 Nano 30B-A3B**	Copy · planning · protocols	GPU 1
Embeddings	**NV-EmbedQA-E5-v5**	Policy / manual RAG	GPU 2
3D	**Microsoft TRELLIS**	Image → 3D .glb	GPU 2
Image	**FLUX.1-Kontext-Dev**	Localized 2D imagery	GPU 3

The reference node
1 × HGX H100
(4× H100 80GB)

- Multi-container Docker Compose
- Each NIM is single-GPU
- Embeddings + TRELLIS share GPU 2
- Milvus (CPU) for policy RAG
- Web search: Exa – Tavily on Nebius

Model 1 of 5 | Vision Language

Nemotron-3 Nano Omni 30B-A3B

NVIDIA , open weights. The eyes – extracts product attributes and grades generated images.

~31B (3B active)
Parameters

256K tokens
Context

Multimodal to text
In and out

~1× H100
GPU

How it works

- Hybrid Mamba-2 + Transformer, MoE
- Open recipe – fine-tune via NeMo / LoRA

NVIDIA Open Model – commercial OK

Swap options

Nemotron Nano 12B-v2-VL	Lighter
Qwen3-VL-30B-A3B	Open / OCR
Llama-3.2-90B-Vision	Bigger

Nemotron-3 Nano 30B-A3B

NVIDIA, open weights. The writer – copy, prompt planning, FAQs, ACP/UCP attributes.

~31.6B (3.2B active)	Up to 1M	Text to text	~1× H100
Parameters	Context	In and out	GPU

How it works

- Hybrid Mamba-2 + MoE (128 experts)
- Fine-tune via NeMo / LoRA · on Token Factory

NVIDIA Open Model – commercial OK

Swap options

Qwen3-30B-A3B-Instruct	Open
Llama-3.3-Nemotron-49B	Stronger
Nemotron-Nano-9B-v2	Faster

FLUX.1-Kontext-Dev

Black Forest Labs (not NVIDIA). Localized 2D lifestyle imagery – product into cultural backgrounds.

12B	Rectified-flow	1024×1024	1× H100 (FP8)
Parameters	Type	Output	GPU

How it works

- 12B rectified-flow transformer (image editing)
- LoRA fine-tune on ~15–20 image pairs

FLUX.1 [dev] NON-COMMERCIAL – license for production

Swap options

FLUX.1-Kontext [pro]/[max]	Commercial
Qwen-Image-Edit / SD 3.5	Open
Imagen / Gemini image	Managed

Microsoft TRELLIS

Microsoft Research (MIT license). One photo → interactive, textured 3D .glb asset.

~1.2B	.glb mesh	1 image to 3D	L40S / H100
Parameters	Output	In and out	GPU

How it works

- Sparse Flow Transformer on Structured Latents
- Customize via input conditioning (no LoRA)

MIT — commercial OK

Swap options

TRELLIS.2-4B	Higher fidelity
Hunyuan3D 2.x	Open
Stable Fast 3D / TripoSR	Fastest

NV-EmbedQA-E5-v5

NVIDIA NeMo Retriever (NIM). The librarian – embeds policy PDFs & manuals for retrieval.

~335M	512 tokens	1024-d	Shares a GPU
Parameters	MAX input	Vector	GPU

How it works

- Transformer bi-encoder (fine-tune of E5-large)
- RAG only here — no in-pipeline fine-tune

MIT — commercial OK

Swap options

llama-3.2-nv-embedqa-1b-v2	Multilingual
nv-embedqa-mistral-7b-v2	Most accurate
BGE-M3 / Arctic-embed	Open

One architecture, evolving

Live production

Batch

Training

Start simple — one Nebius VM

Nebius VM · HGX H100 (8× H100) — deployed together

API client
catalog / app

Nemotron Omni – VLM
GPU 0

Nemotron – LLM
GPU 1

FLUX.1-Kontext – image
GPU 2

TRELLIS – 3D
GPU 3

NV-EmbedQA – embed
shares GPU 2

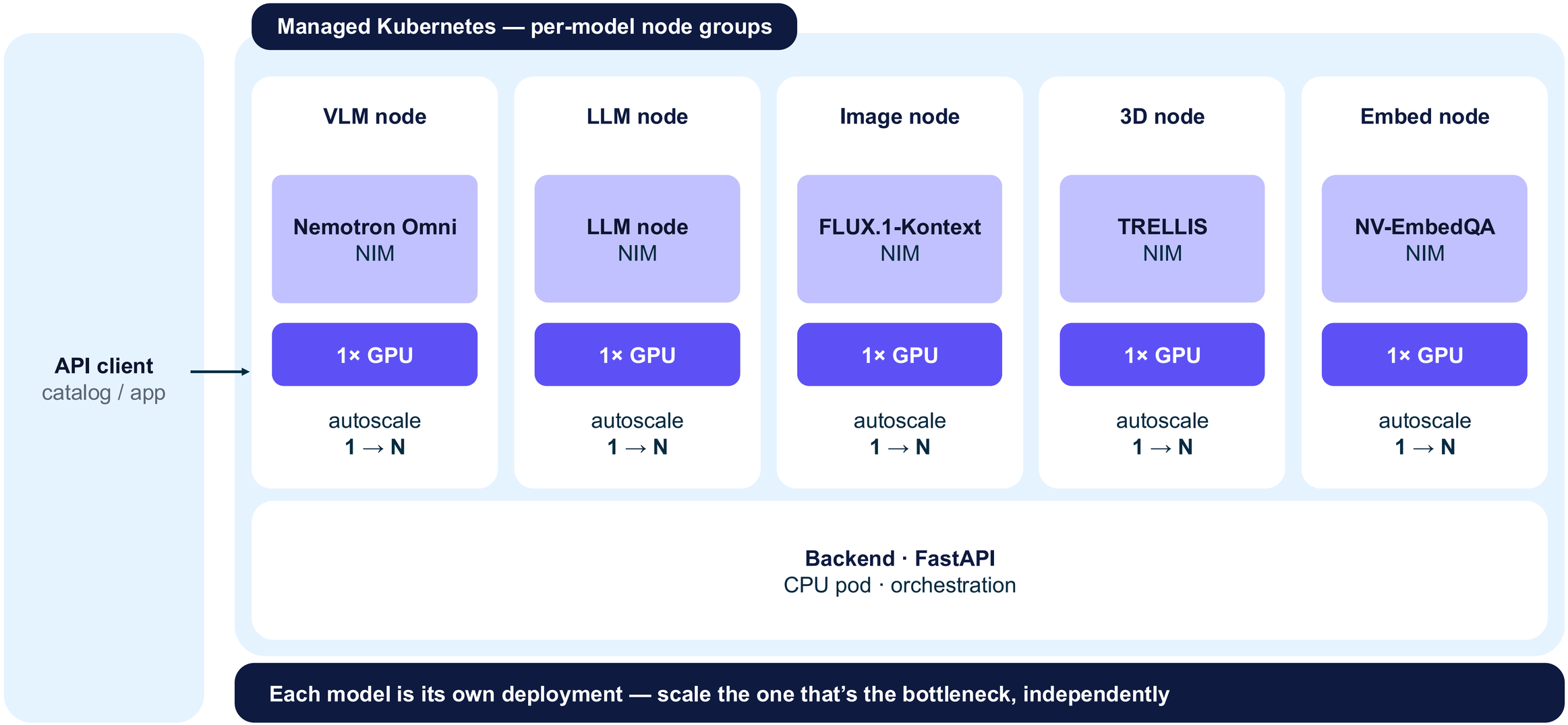
8× H100 – 4 GPUs in use

GPU0	GPU1	GPU2	GPU3
GPU4	GPU5	GPU5	GPU7

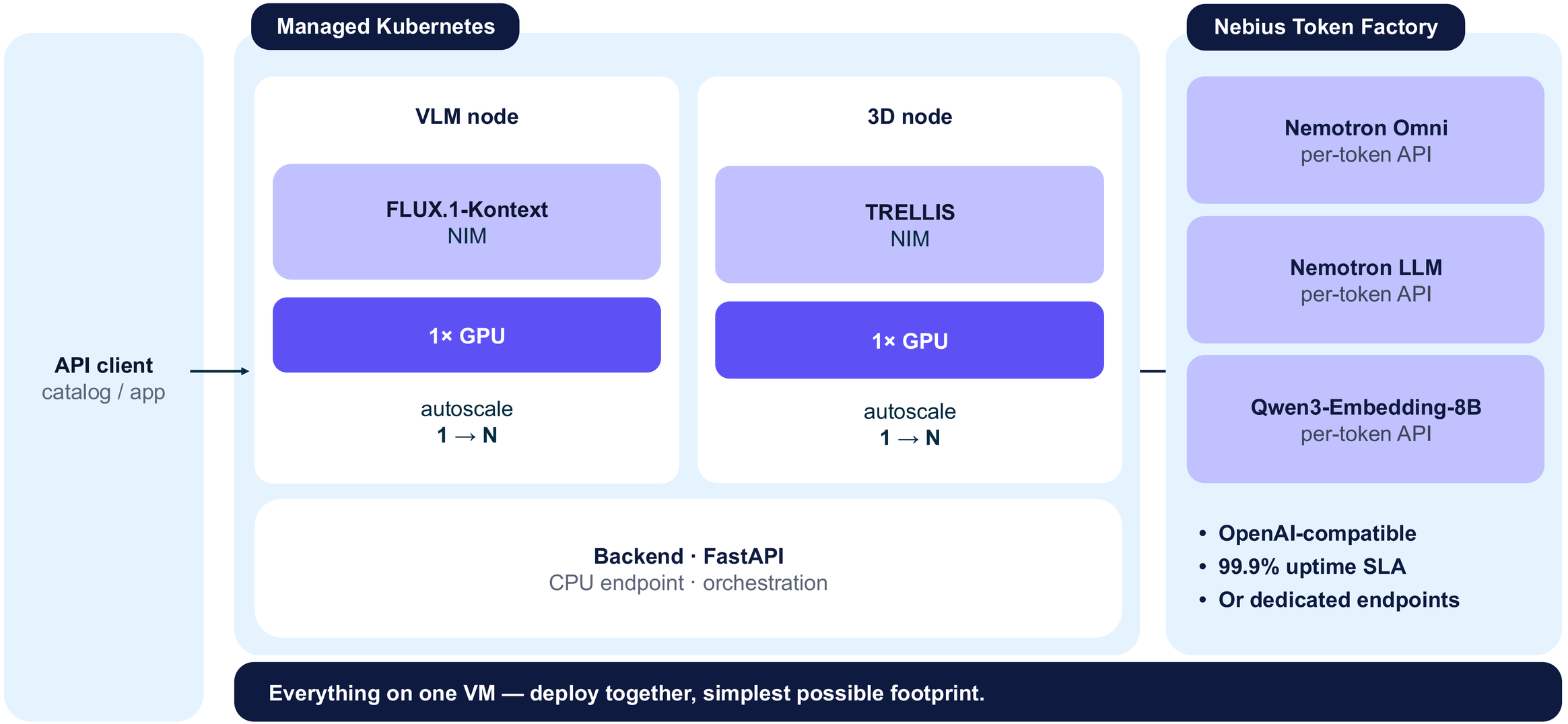
GPUs 4–7
free for other workloads on the same VM

Everything on one VM — deploy together, simplest possible footprint.

Split into Kubernetes — scale each model on its own



Start simple — one Nebius VM



Fully managed — Serverless Endpoints + Token Factory

API client
catalog / app

Nebius Serverless AI — Endpoints

Flux.1-Kontext
managed endpoint

always-on
auto-scales to zero ops

Trellis
managed endpoint

always-on
auto-scales to zero ops

Backend · FastAPI
CPU endpoint · orchestration

Nebius Token Factory

Nemotron Omni
per-token API

Nemotron LLM
per-token API

Qwen3-Embedding-8B
per-token API

- **OpenAI-compatible**
- **99.9% uptime SLA**
- **Or dedicated endpoints**

Image + 3D on Serverless Endpoints; VLM, LLM & embeddings managed on Token Factory.

But do you need it live?

Batch vs live enrichment

Most catalog work isn’t a live request — it’s thousands of SKUs processed once. That changes the best architecture.

Live

Real-time, one SKY at a time

Shopper / editor-facing

- Sub-second, always-on
- Serverless Endpoints or K8s
- On-demand / reserved GPUs

Use when:
PDP, live, preview, single product

Batch

Catalog-scale, offline

- Thousands of SKUs × 10 locales
- Run-to-completion
- Serverless Jobs or Token Factory Batch
- Preemptible / spot GPUs

Use when:
Initial load, re-enrich, new market

Reserved vs on-demand vs preemptible

Nebius GPU lineup

GB300 NVL72 | HGX B300 | HGX B200 | HGX H200 | HGX H100 | RTX PRO 6000

Comitted

Reserved

Lowest unit cost on a commitment — but you pay for idle.

Best for
Steady live serving

Flexible

Oh-demand

Per-second, no commitment, premium rate.

Best for
Bursty traffic · experiments

Cheapest

Preemtible / spot

Cheapest, can be interrupted — checkpoint-safe.

Best for
Batch & offline enrichment

For bath enrichment, preemptible usually wins: the work is checkpoint- and retry-safe, so interruptions cost little and you get the lowest price.

Kubernetes on preemptible GPUs + Token Factory

Batch queue
SKUs x locales

Managed Kubernetes · preemptible

Image node

Flux.1-Kontext
NIM

Spot · 1 x GPU

autoscale
1 → N

3D node

TRELLIS
NIM

Spot · 1 x GPU

autoscale
1 → N

Backend · FastAPI
CPU endpoint · orchestration

Nebius Token Factory

Nemotron Omni
per-token API

Nemotron LLM
per-token API

Qwen3-Embedding-8B
per-token API

- **OpenAI-compatible**
- **99.9% uptime SLA**
- **Or dedicated endpoints**

Image + 3D on spot GPUs; VLM, LLM & embeddings via Token Factory Batch API (~50%).

Serverless AI Jobs + Token Factory (newest)

Batch queue
SKUs x locales

Nebius Serverless AI — Jobs (async)

FLUX.1-Kontext
run-to-completion

async job
no cluster to manage

Trellis
run-to-completion

async job
no cluster to manage

Backend · FastAPI
CPU endpoint · orchestration

Nebius Token Factory

Nemotron Omni
per-token API

Nemotron LLM
per-token API

Qwen3-Embedding-8B
per-token API

- **OpenAI-compatible**
- **99.9% uptime SLA**
- **Or dedicated endpoints**

Image + 3D as managed Jobs; VLM, LLM & embeddings on Token Factory.

Make it yours

Where fine-tuning pays off in retail

Out of the box it’s prompt- & RAG-steered. **Fine-tuning is the production upgrade when prompts aren’t enough.**

VLM

Fine-tune the VLM

On your product taxonomy

- Recognise your categories & attribute schema
- Consistent tagging across millions of SKUs
- Fewer mis-classifications that break search

LLM

LoRA the LLM

On your brand voice

- Tone, terminology & banned claims baked in
- Locale-specific phrasing for your markets
- Less prompt-engineering, lower latency

How you train it →

Both are open-weight with published recipes: LoRA / PEFT or full fine-tune via NeMo, on a Nebius Soperator GPU cluster (next slide).

You may also want to train

Fine-tune on Soperator, push to inference

Managed Soperator — Slurm on Kubernetes (Nebius)

NVIDIA Quantum-2 InfiniBand · 3.2 Tbit/s/host · topology-aware

HGX B300
node 1 · 8×Ultra

HGX B300
node 2 · 8×Ultra

HGX B300
node 3 · 8×Ultra

HGX B300
node 4 · 8×Ultra

NeMo / LoRA / PEFT
fine-tune VLM + LLM

Checkpoints → object storage
resume on preemption

Most powerful GPUs · shared filesystem · fault-tolerant · reserved or preemptible

Push fine-tuned model

Your inference workload

- **VLM**
- **Image**
- **3D**
- **LLM + embed (Token Factory)**

Train on a separate Soperator cluster — only when you need it — then push the improved model into the live or batch pipeline.

Match the Nebius service to the job

Pilot / static

Virtual Machines

All models on one host — simplest start.

Self-managed live

Managed Kubernetes

Per-model node groups, autoscale each.

Live, hands-off

Serverless AI — Endpoints

Always-on managed inference, no cluster. (preview)

Offline batch

Serverless AI — Jobs

Async run-to-completion batch. (preview)

Text & embeddings

Token Factory

Per-token API + dedicated endpoints + Batch API −50%.

Fine-tuning / training

Soperator (Slurm)

Managed Slurm on the most powerful GPUs.

The point

Same open blueprint, same models — Nebius lets you slide from a VM to serverless, batch or training without re-architecting.

Try it yourself

Deploy the blueprint on Nebius in minutes

Run the NVIDIA Retail Catalog Enrichment blueprint as-is on one Nebius VM (8× H100) with Docker Compose.

1 Launch a GPU VM on Nebius

8× H100 · Ubuntu + Docker/CUDA

```
Console → Compute → Create VM → 8× H100 (HGX H100)
```

2 Set keys & log in to NGC

NIMs + FLUX need tokens

```
export NGC_API_KEY=…   HF_TOKEN=…       # EXA_API_KEY optional
echo $NGC_API_KEY | docker login nvcr.io -u '$oauthtoken' --password-stdin
```

3 Clone & launch

Docker Compose brings up all 5 NIMs

```
git clone https://github.com/NVIDIA-AI-Blueprints/Retail-Catalog-Enrichment.git
cd Retail-Catalog-Enrichment && docker compose up -d
```

4 Open the console

Enrich a product live

```
http://<vm-ip>:3000
```

Needs ≥4× H100 — 8× gives headroom. Full steps in docs/DOCKER.md.

Tip: Run it on a **preemptible VM** and stop it whenever you're not using it to save on your budget

Take your $50 promocode to try the solution

Useful links

console.nebius.com

build.nvidia.com

docs.nebius.com

Q&A

Andrew Eliseev
Head of Retail & Commerce AI, Nebius

Antonio Martinez
GenAI Evangelist Product Engineer, NVIDIA

Manuel Gomez
Solution Architect, Nebius

Take your $50 promocode to try the solution

Useful links

console.nebius.com

build.nvidia.com

docs.nebius.com