NVIDIA GTC

Large Language Model Fine-Tuning

Josh Mineroff

Director of Solution Architecture,
Tech Alliances

John Wu

Senior Product Manager,
AI

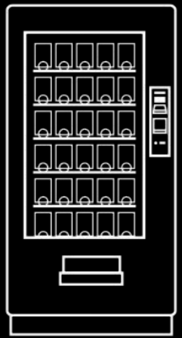
Domino

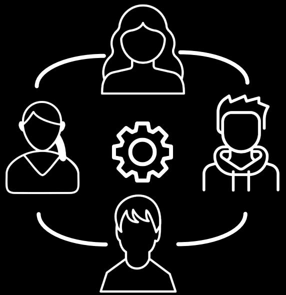
Agenda

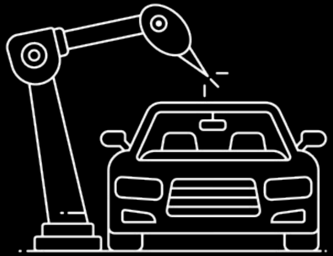
1. Introduction
2. Today’s Gen AI challenge
3. How to customize your LLMs
4. How do Domino & Nvidia help
5. Demo

Domino in 60 seconds

Build and operate AI at scale







On-Demand Infrastructure

Self-service access to compute & secure data

Comprehensive Reproducibility

Collaboration across teams & technologies

AI Factory

Rapid model deployment to production

Model Governance

Responsible AI model monitoring, risk management, & remediation

Today's Challenge

`We all know pretrained LLMs are great, but your use case requires ...`

Business Context

The LLM doesn't know about your business since it wasn't trained on any of your **proprietary** data.

Industry Vocabulary

The LLM doesn't understand unique terminologies and concepts that are used within your industry.

Structured outputs

The LLM doesn't know the specific structure or style of outputs that your application is expecting.

Ways to customize LLMs to your needs

- **Prompt Engineering**: Use carefully structured inputs to guide the outputs.
- **RAG (Retrieval Augmented Generation)**: Adds contextual information to prompts by querying a vector database for related information.
- **Full Fine-Tuning**: Transfer learning approach in which all the parameters are adjusted using task-specific data.
- **Parameter-Efficient Fine-Tuning (PEFT)**: Modifies only a small select amount of parameters for more efficient adaptation.

Different ways to fine-tune using PEFT

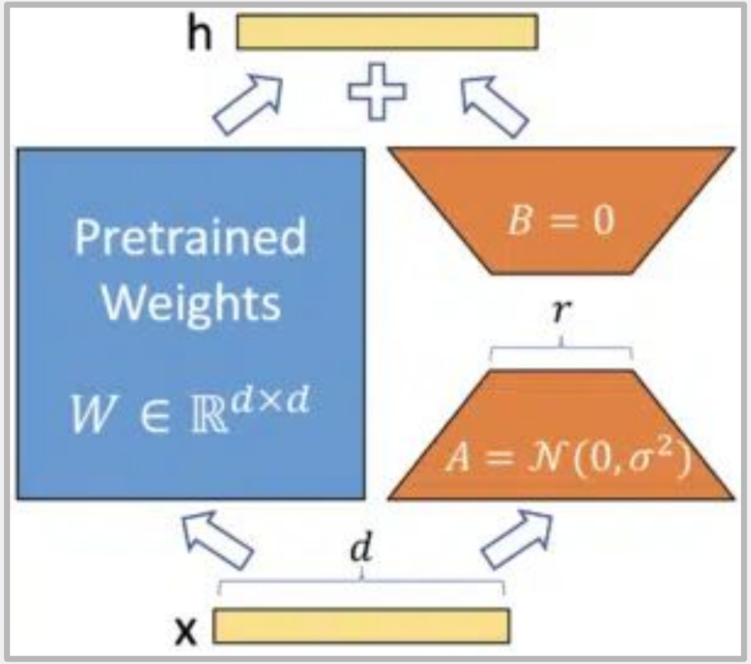
- **Prompt Tuning**: Add task-specific prompt embeddings to the input and parameters are updated independently of the frozen pretrained model.

- **Prefix Tuning**: Similar to prompt tuning, but the embeddings are inserted in all of the model layers.

- **P-Tuning**: A prompt encoder (LSTM model) is used to predict the input embeddings and only weights are updated at each training step.

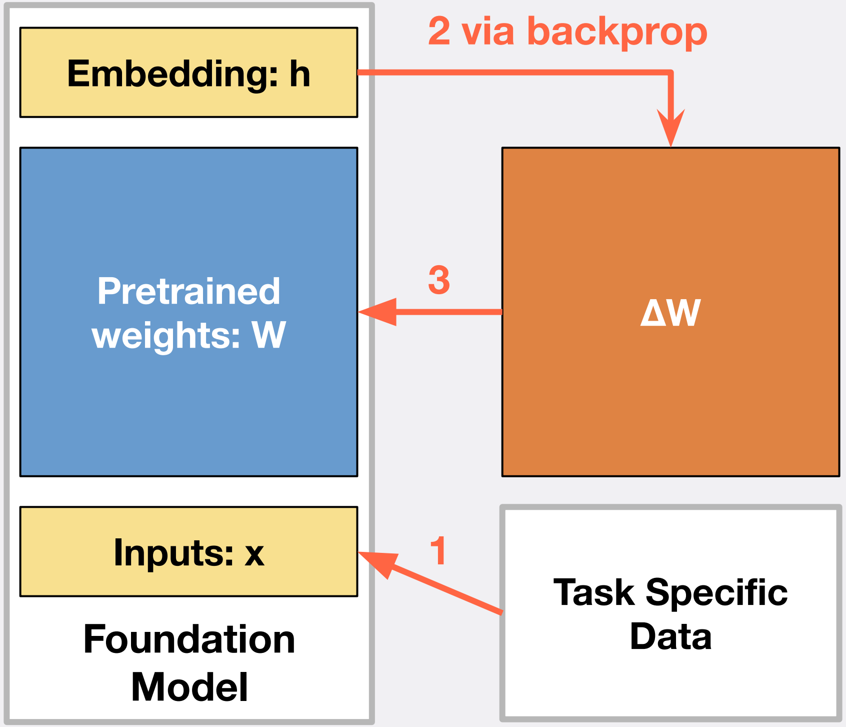
- **LoRA (Low rank adaptation)**: Decomposes a large matrix into two smaller low-rank matrices in the attention layers (drastically reduce number of parameters).

Diving deeper into LoRA

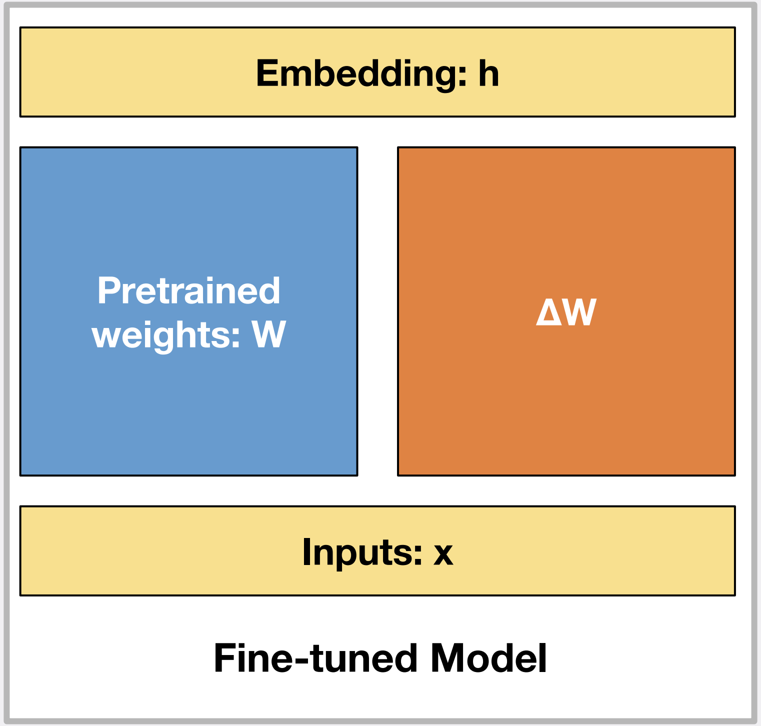

h
Pretrained
Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

Diving deeper into LoRA: Traditional Fine-Tuning

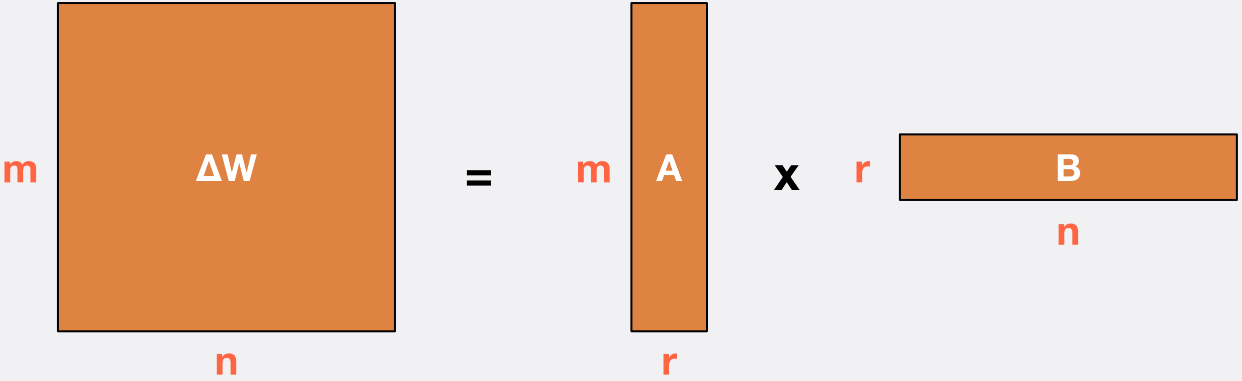


Diving deeper into LoRA

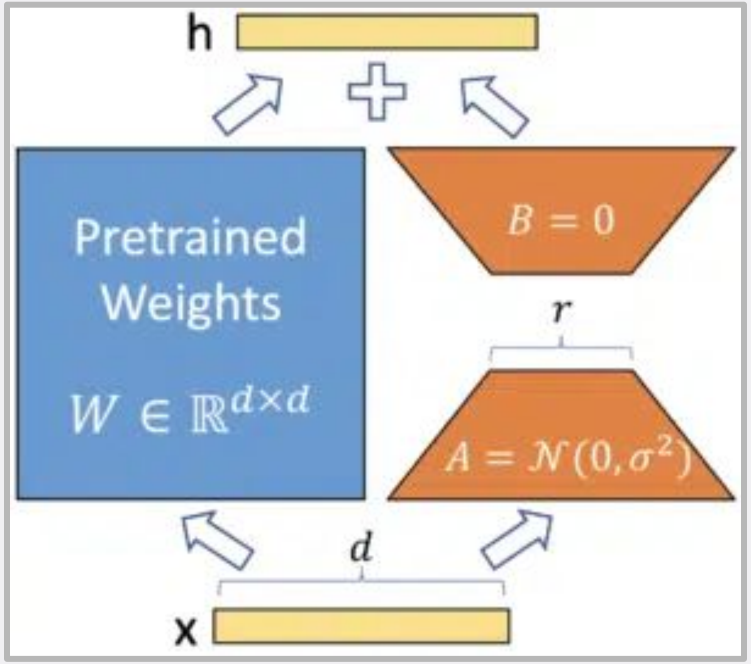


Diving deeper into LoRA: Low-Rank

m x n = m x r * r x n



Diving deeper into LoRA

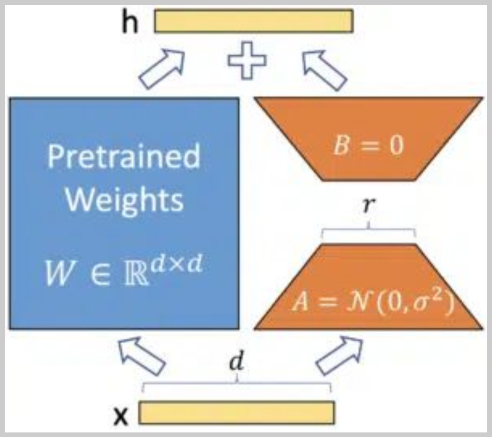


h
Pretrained
Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685 (2021).

Diving deeper into LoRA

1. Let's say you have a 100K x 100K weight matrix = 10B parameters
2. We can create our low-rank adaptor by reparameterizing the original weight into two matrices (A and B) of low rank R.
3. Our new low-rank matrix is then taken to be the product of A and B
4. If r=2, we end up updating (100K x 2) + (100K x 2) = 400K parameters

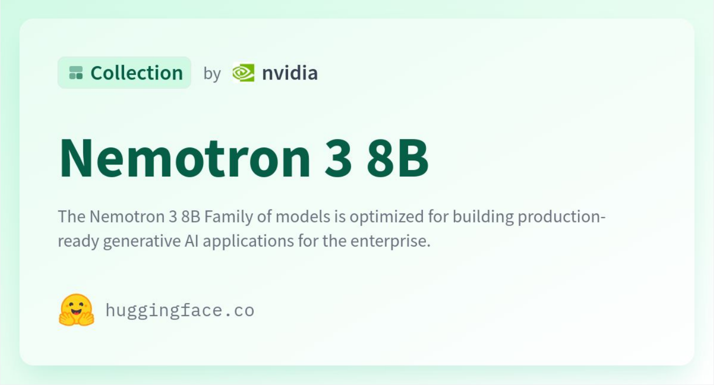

h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

How does Nvidia help?

- NVIDIA NeMo is a Generative AI framework built for researchers and developers working on large language and other types of models.
- Pre-built containers and existing code templates make it easy to apply existing adaptation techniques such as LoRA.
- Pretrained models such as **NVIDIA Nemotron** provide a powerful baseline to start fine-tuning from.

NVIDIA

NEMO



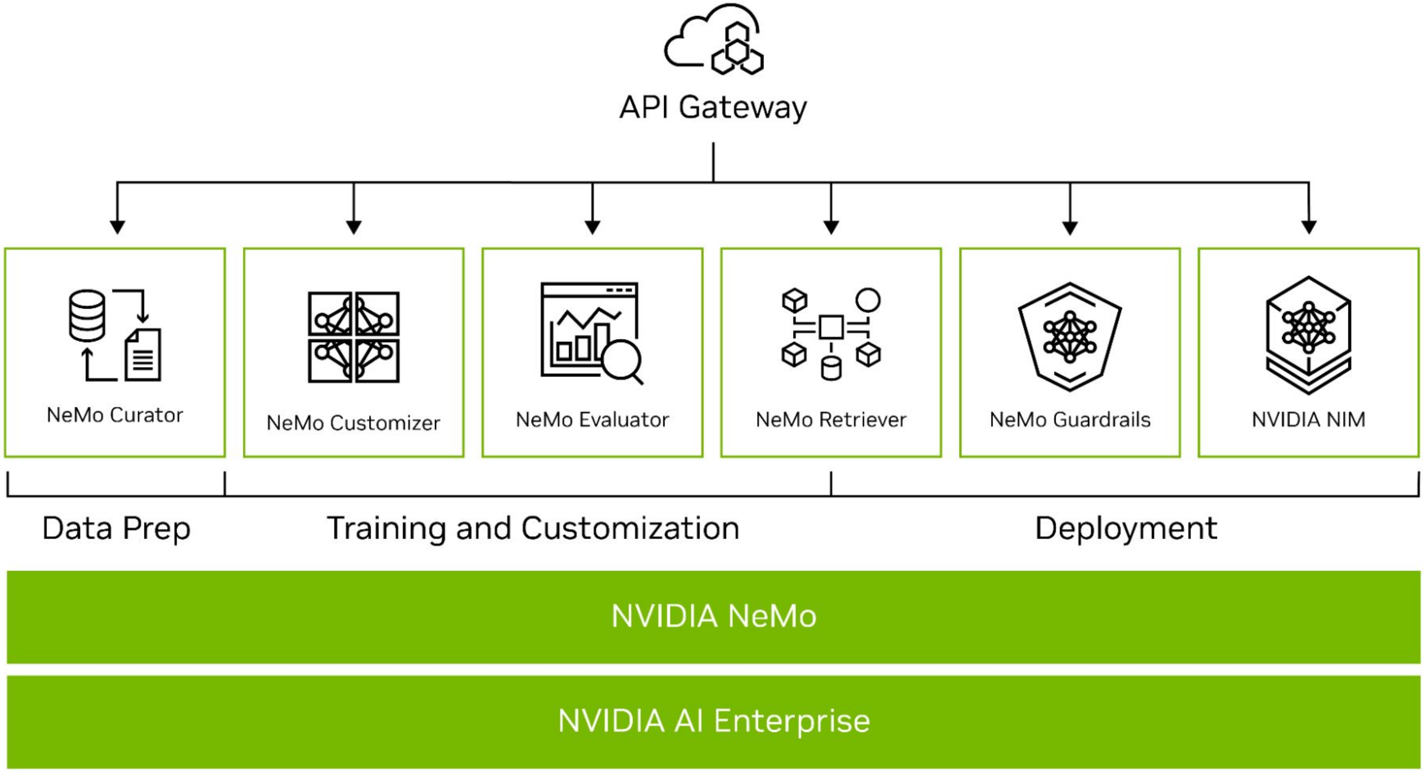
Collection by nvidia

Nemotron 3 8B

The Nemotron 3 8B Family of models is optimized for building production-ready generative AI applications for the enterprise.

huggingface.co

NVIDIA NeMo



API Gateway

NeMo Curator

NeMo Customizer

NeMo Evaluator

NeMo Retriever

NeMo Guardrails

NVIDIA NIM

Data Prep

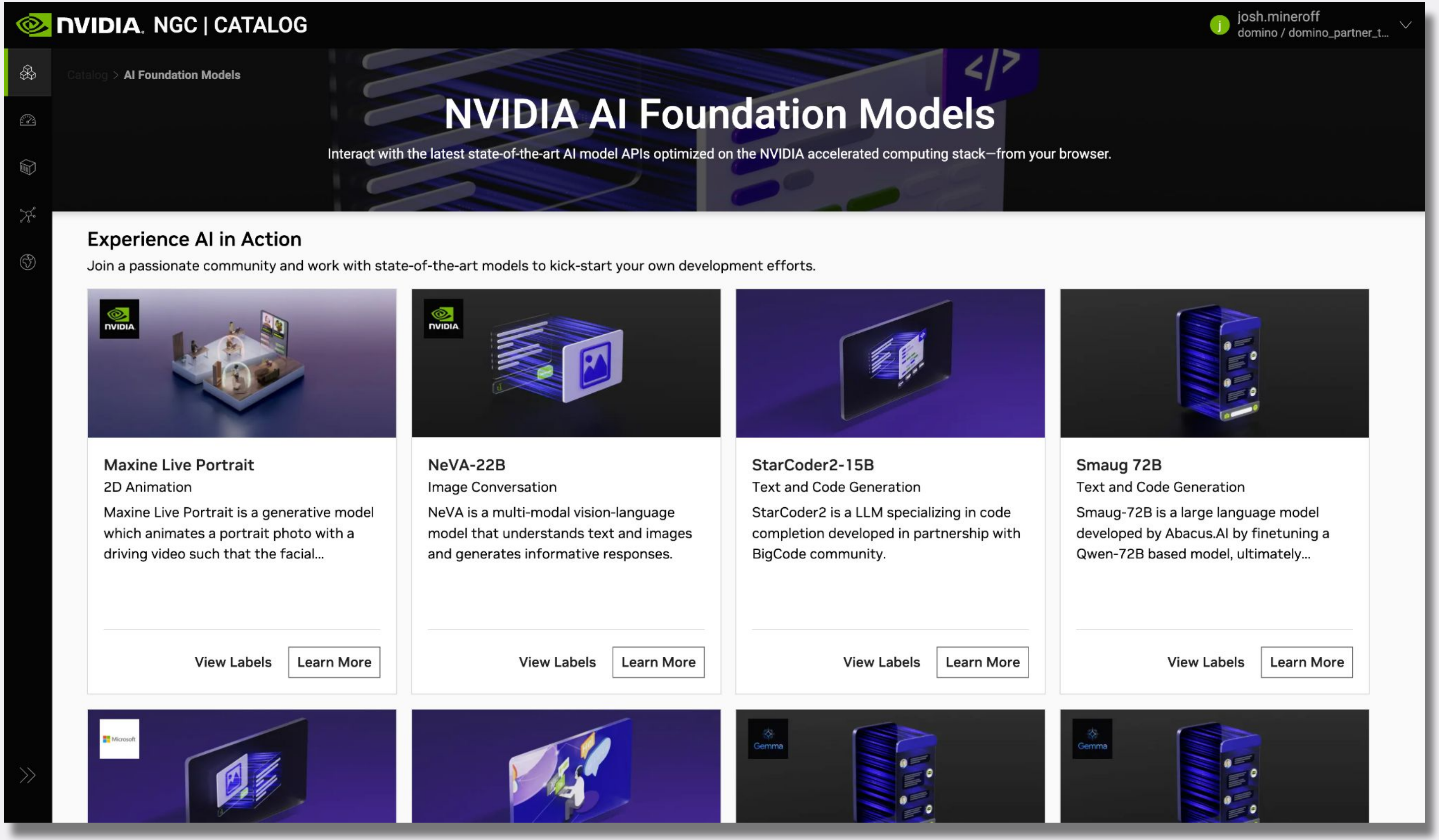
Training and Customization

Deployment

NVIDIA NeMo

NVIDIA AI Enterprise

NVIDIA AI Foundation Models



NVIDIA AI Foundation Models

Interact with the latest state-of-the-art AI model APIs optimized on the NVIDIA accelerated computing stack—from your browser.

Experience AI in Action

Join a passionate community and work with state-of-the-art models to kick-start your own development efforts.

Maxine Live Portrait
2D Animation
Maxine Live Portrait is a generative model which animates a portrait photo with a driving video such that the facial...

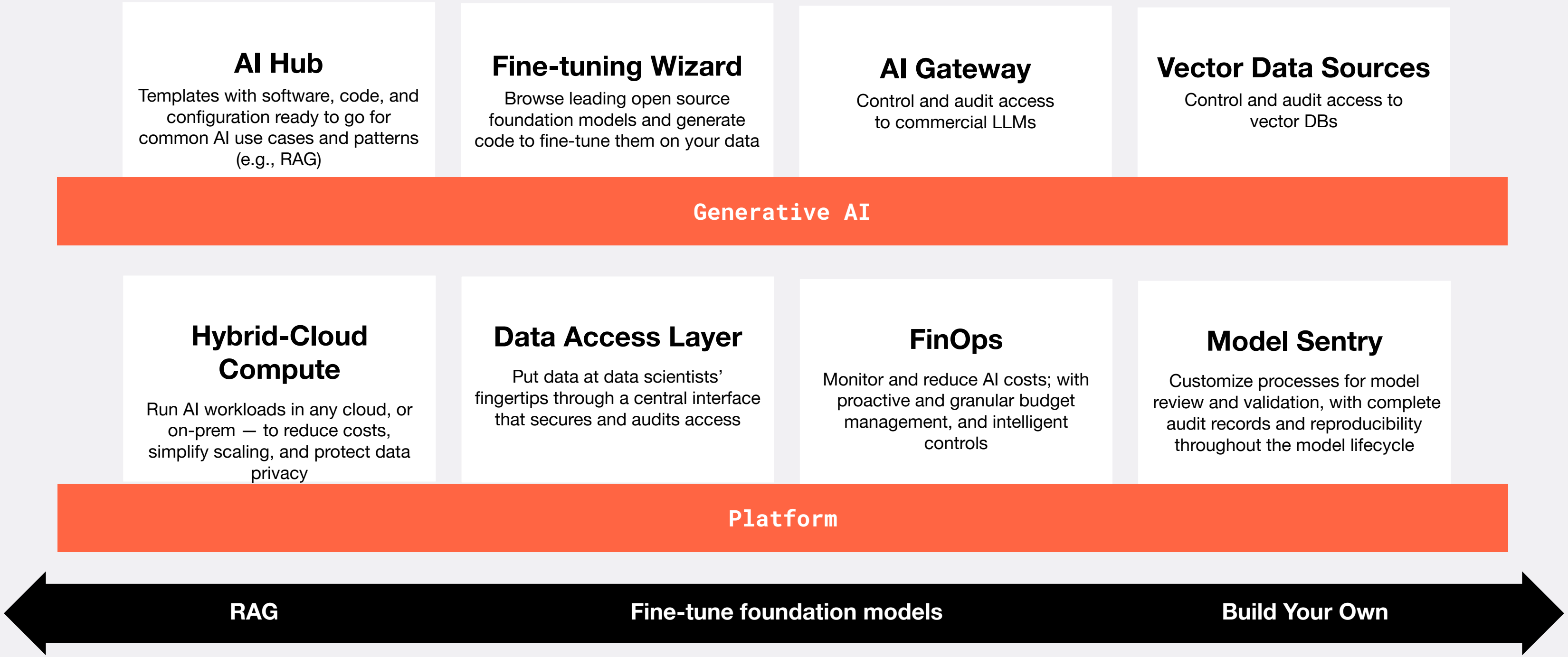
NeVA-22B
Image Conversation
NeVA is a multi-modal vision-language model that understands text and images and generates informative responses.

StarCoder2-15B
Text and Code Generation
StarCoder2 is a LLM specializing in code completion developed in partnership with BigCode community.

Smaug 72B
Text and Code Generation
Smaug-72B is a large language model developed by Abacus.AI by finetuning a Qwen-72B based model, ultimately...

How does Domino help?

Domino



AI Hub

Templates with software, code, and configuration ready to go for common AI use cases and patterns (e.g., RAG)

Fine-tuning Wizard

Browse leading open source foundation models and generate code to fine-tune them on your data

AI Gateway

Control and audit access to commercial LLMs

Vector Data Sources

Control and audit access to vector DBs

Generative AI

Hybrid-Cloud Compute

Run AI workloads in any cloud, or on-prem — to reduce costs, simplify scaling, and protect data privacy

Data Access Layer

Put data at data scientists' fingertips through a central interface that secures and audits access

FinOps

Monitor and reduce AI costs; with proactive and granular budget management, and intelligent controls

Model Sentry

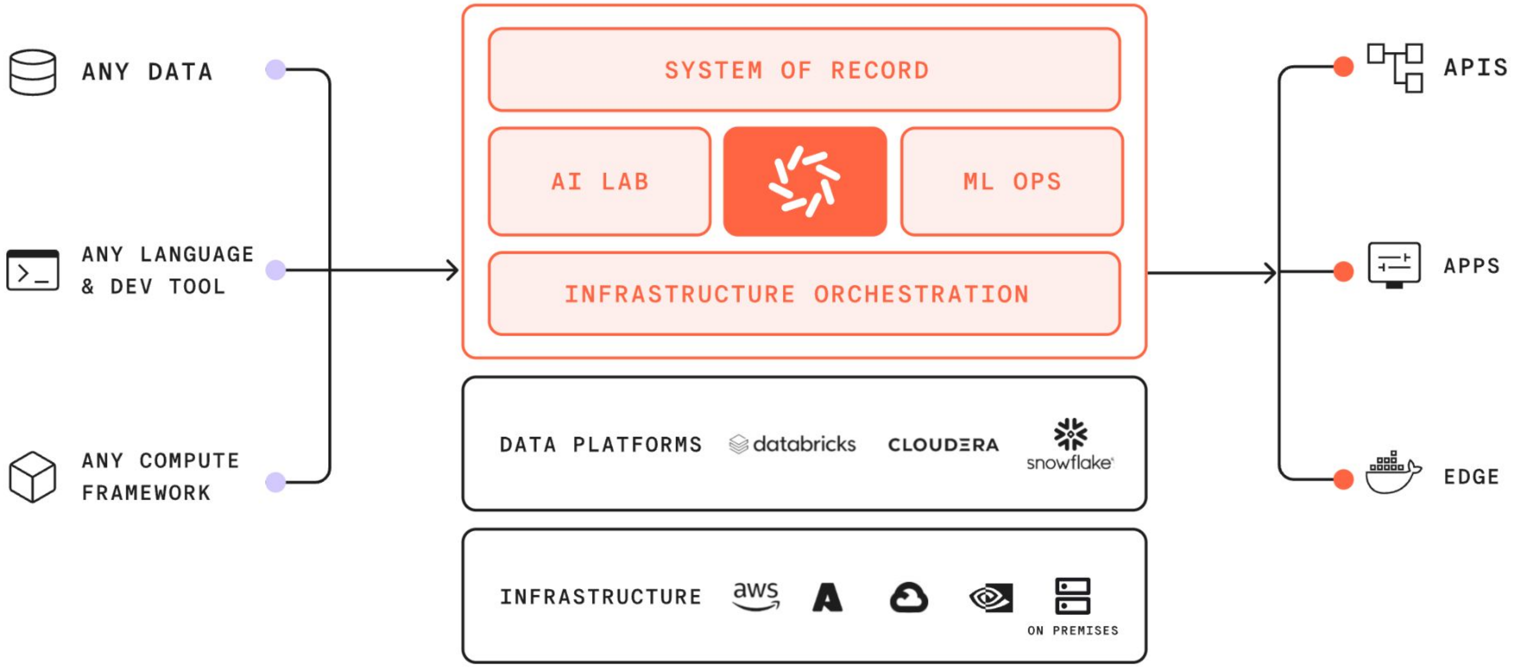
Customize processes for model review and validation, with complete audit records and reproducibility throughout the model lifecycle

Platform

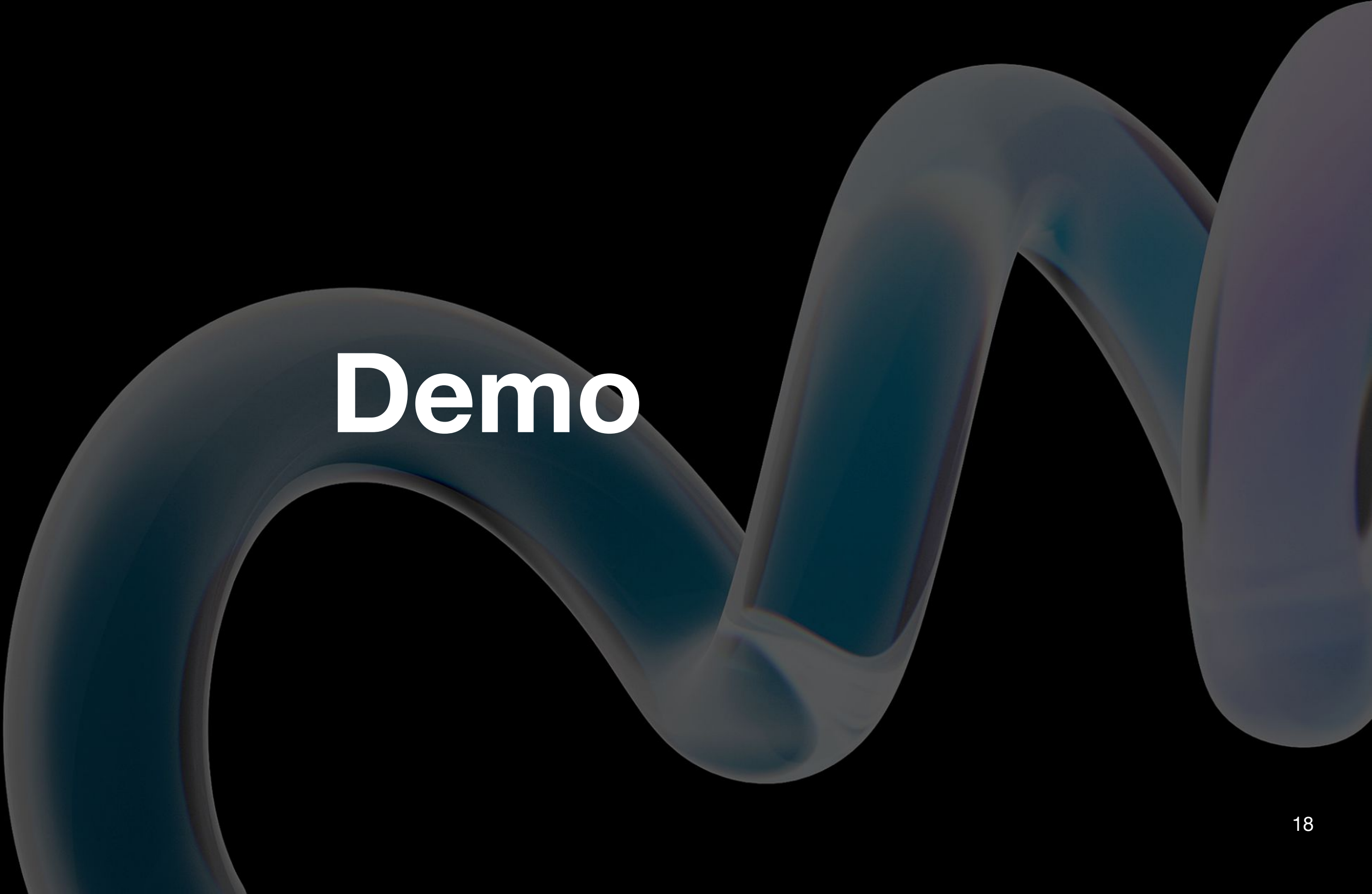
RAG — **Fine-tune foundation models** — **Build Your Own**

Accelerate AI impact with Domino

DATA SCIENCE FREEDOM + ENTERPRISE CONTROL

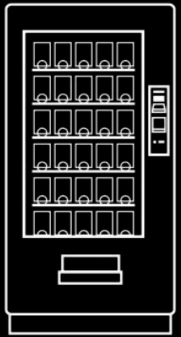


Demo



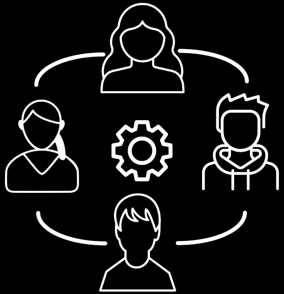
Domino in 60 seconds

Build and operate AI at scale



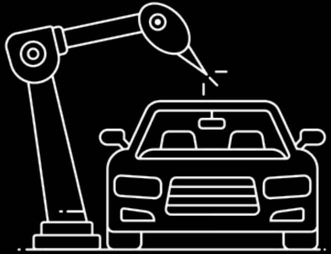
On-Demand Infrastructure

Self-service access to compute & secure data



Comprehensive Reproducibility

Collaboration across teams & technologies



AI Factory

Rapid model deployment to production

Model Governance

Responsible AI model monitoring, risk management, & remediation

Thank you!

1. **LEARN MORE:** domino.ai/NVIDIA
2. **VISIT OUR BOOTH:** #1612 in the AI Center of Excellence Pavilion.
3. **WIN**: NVIDIA Jetson Orin™ Nano Developer Kit!