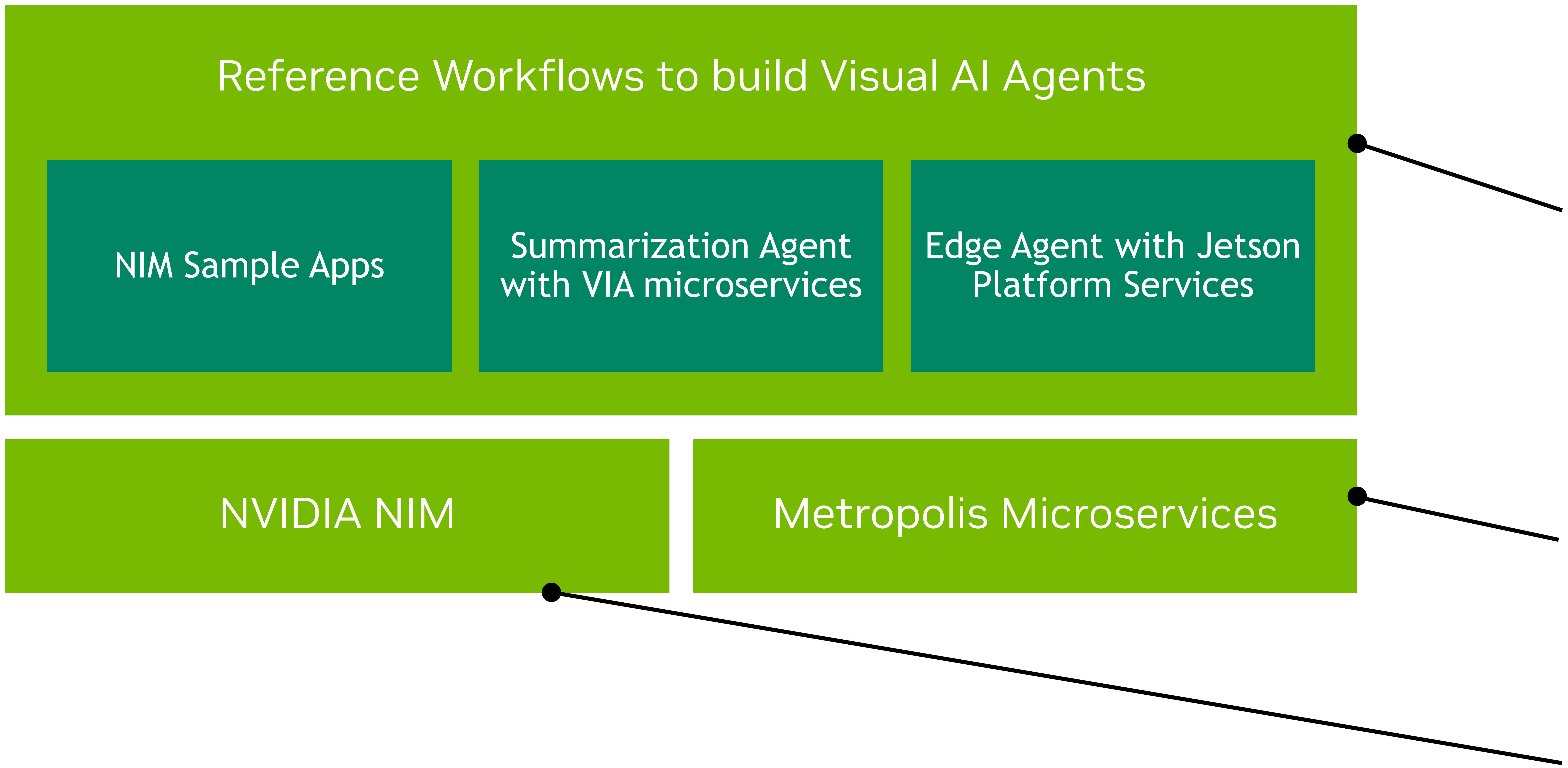
Building Visual AI Agents with Generative AI and NVIDIA NIM

NVIDIA Metropolis

NVIDIA Metropolis Generative AI Stack



Reference Workflows to build Visual AI Agents
NIM Sample Apps
Summarization Agent with VIA microservices
Edge Agent with Jetson Platform Services
NVIDIA NIM
Metropolis Microservices

Recipes and sample apps to make it easy to evaluate Gen AI & VLMs and speed up the development and integration perceptive & interactive AI agents into your business apps.

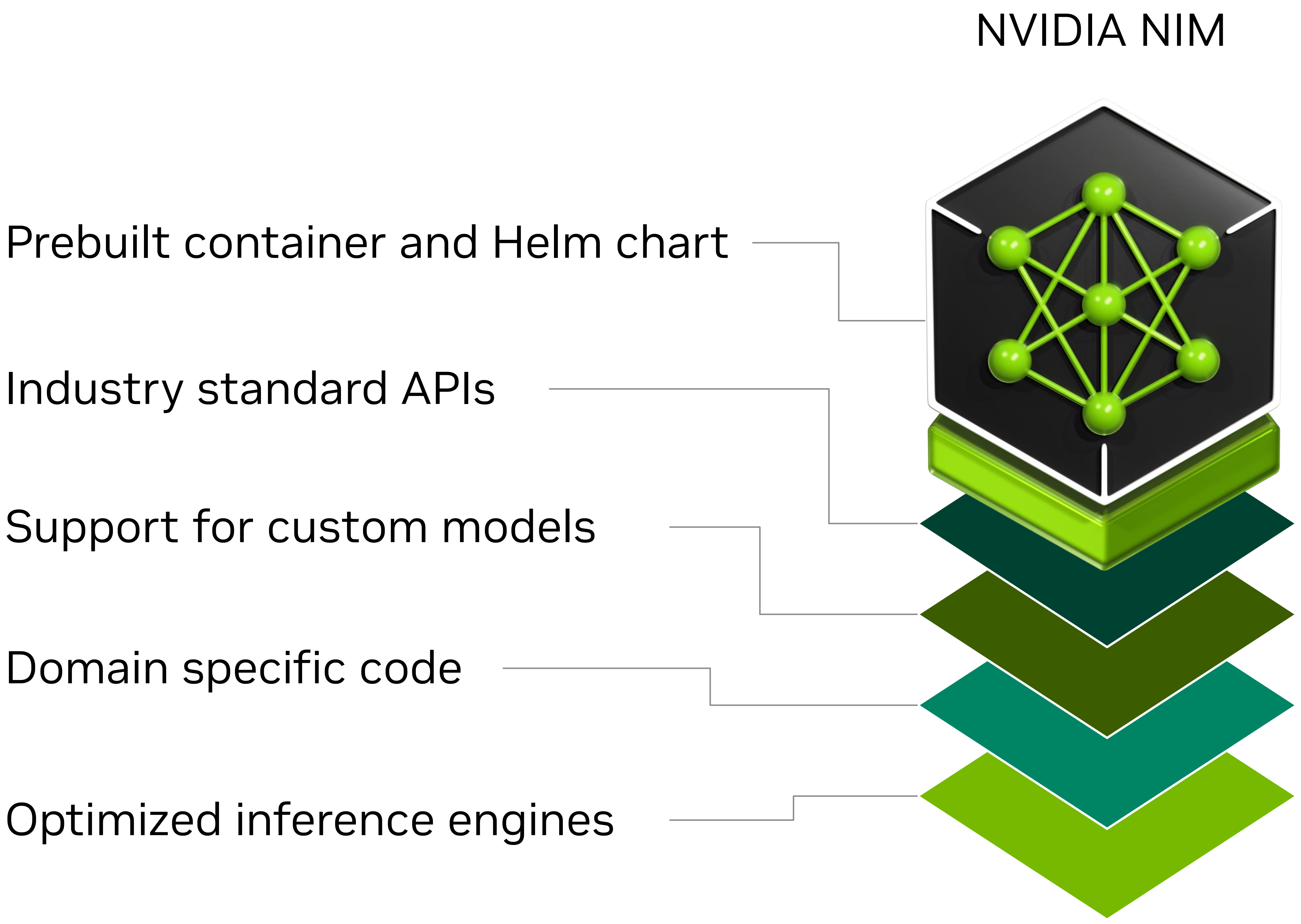
Optimized inference & app microservices for flexible and easy cloud-native development with simple API calls

Powerful and Customizable Vision Language Models (VLM) and specialized computer vision foundation models

9A90AA
9A90BA

NVIDIA NIM - Optimized Inference Microservices

Accelerated runtime for generative AI



NVIDIA NIM

Prebuilt container and Helm chart

Industry standard APIs

Support for custom models

Domain specific code

Optimized inference engines

Deploy anywhere and maintain control of generative AI applications and data

Simplified development of AI applications that can run in enterprise environments

Day 0 support for all publicly available models providing choice across the ecosystem

Improved TCO with best latency and throughput running on accelerated compute

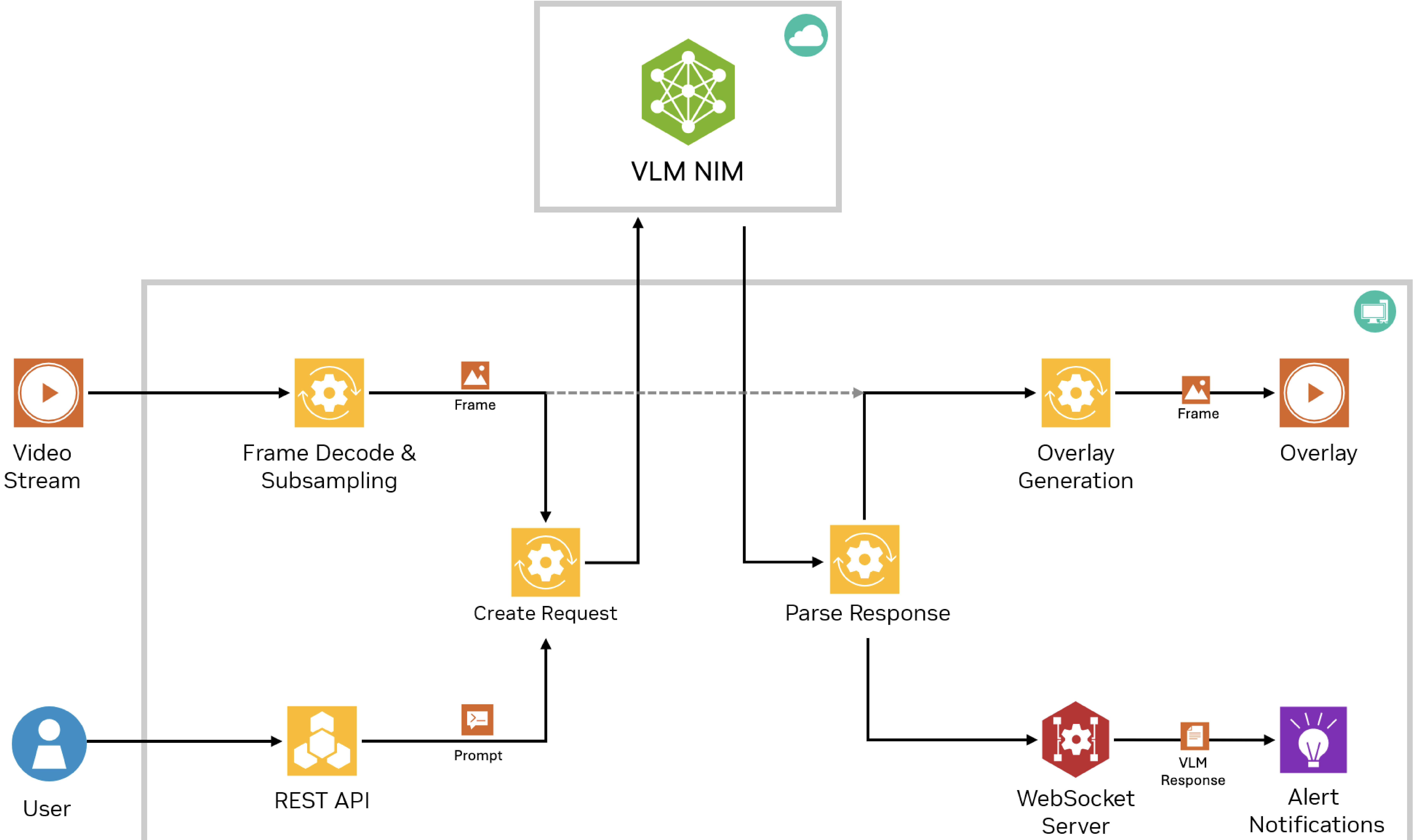
Best accuracy for enterprise by enabling tuning with proprietary data sources

Enterprise software with feature branches, validation and support

Microsoft Azure | aws | Google Cloud | ORACLE | DGX & DGX Cloud | DELL Technologies | Hewlett Packard Enterprise | Lenovo | SUPERMICRO

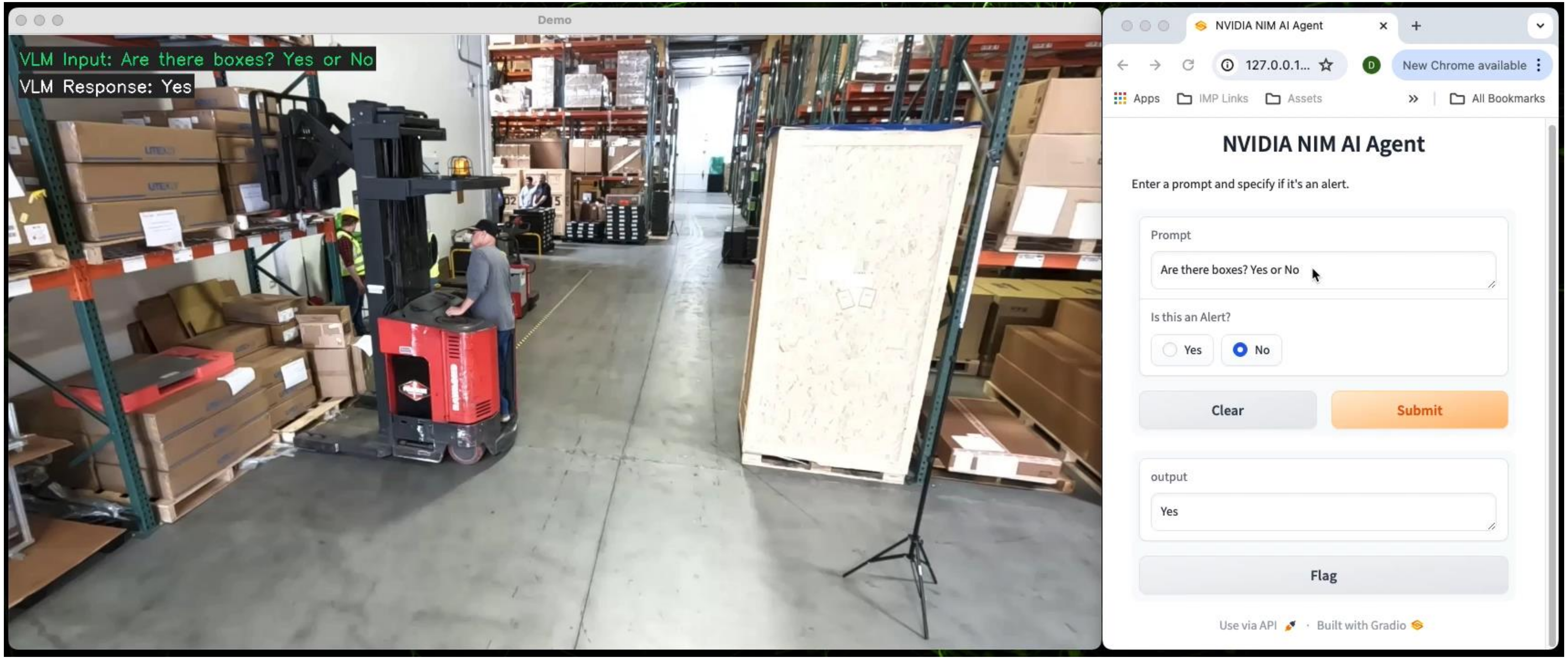
Metropolis Reference Workflows & Sample Apps

Recipes to Build a Wide Range of AI Agents



Evaluate and Experiment with Vision Language Models (VLM)

From simple Jupyter Notebooks to Nearly Production-ready Sample apps



VLM Input: Are there boxes? Yes or No

VLM Response: Yes

NVIDIA NIM AI Agent

Enter a prompt and specify if it's an alert.

Prompt: Are there boxes? Yes or No

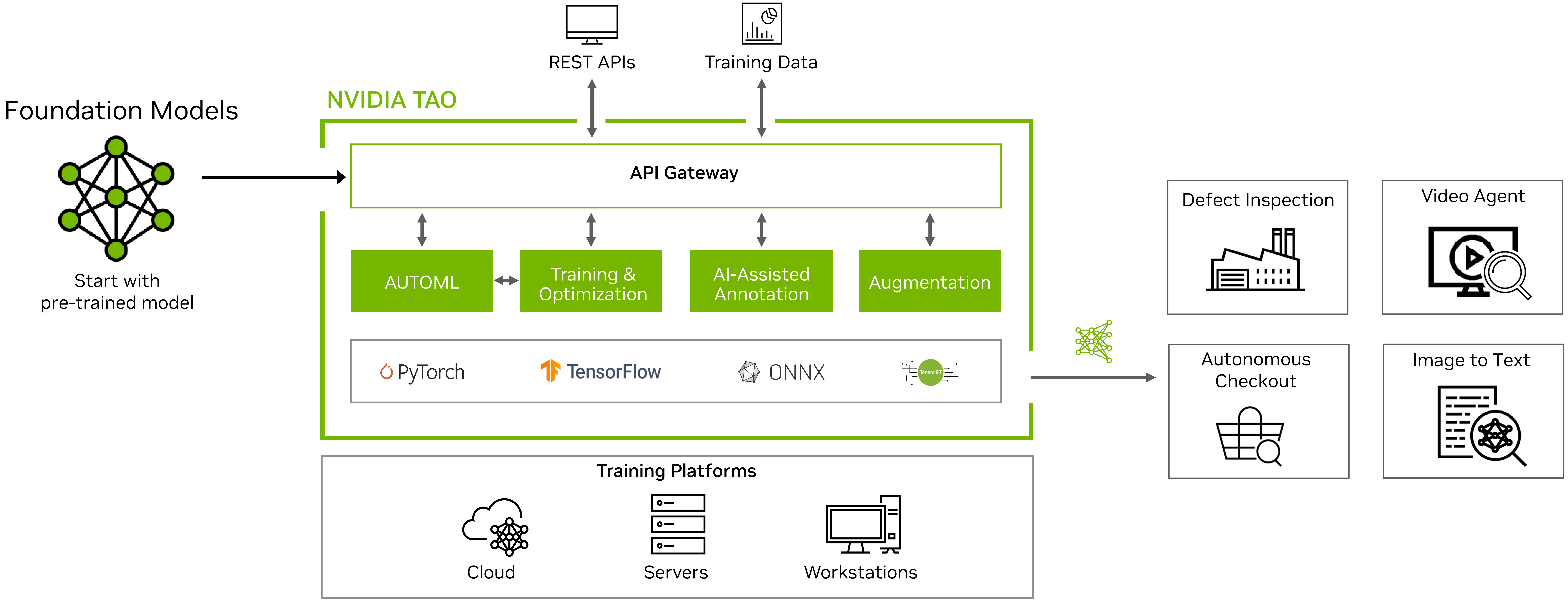
Is this an Alert? ○ Yes ● No

output: Yes

https://github.com/NVIDIA/metropolis-nim-workflows

Get Access to and Customize / Tune Models with NVIDIA TAO

State of the Art Models, Training and Customization for Vision AI



Full fine-Tuning
Update weights of entire model
including the Foundation backbone

Last layer or Head fine-tuning
Freeze the Foundation backbone
and fine-tune the last few layers

In-context Learning
Use visual prompting and model chaining
to improve contextual awareness

NVIDIA®