Open Foundations for Trustworthy Agentic AI in Financial Services

Dr. Jochen Papenbrock

EMEA Head of Financial Technology
E-Mail: jpapenbrock@nvidia.com

AI – The New Industrial Revolution

APPLICATIONS

MODELS

INFRASTRUCTURE

GPUS

ENERGY

CHATBOTS

DIGITAL BIOLOGY

ROBOTAXI

ENTERPRISE AI AGENTS

SCIENCE

ROBOTICS

MANUFACTURING

AI CODER

LLM VLM VLA MMLLM GPT

DM GNN MOE SSM LBM

AI FACTORIES

NVIDIA

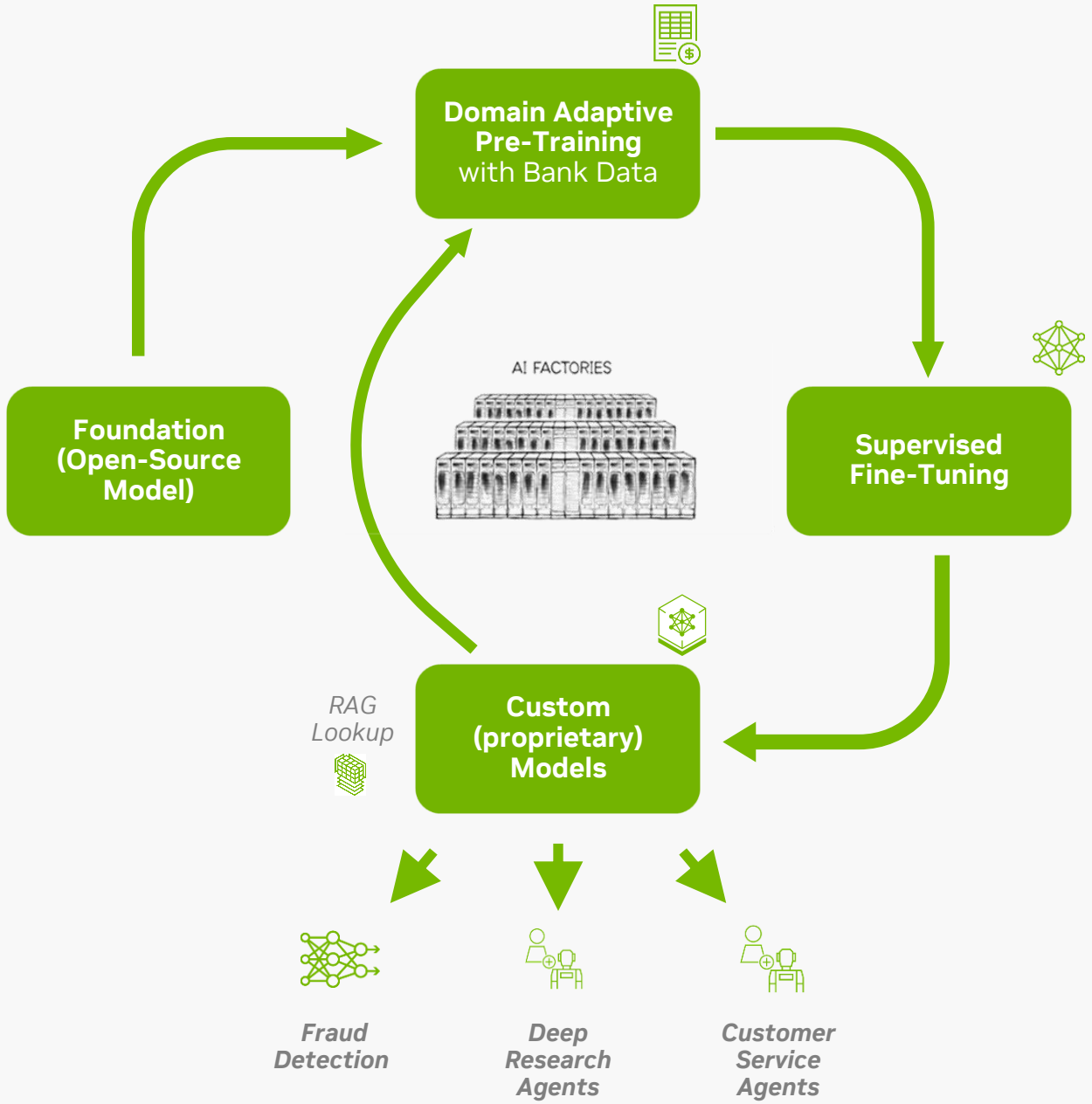
Architecting the AI-Powered Bank

Sovereign Foundations: Leverage open-source models to ensure platform independence, data privacy, and cost efficiency

Proprietary Precision: Transform internal data into a unique competitive moat by post-training and fine-tuning models

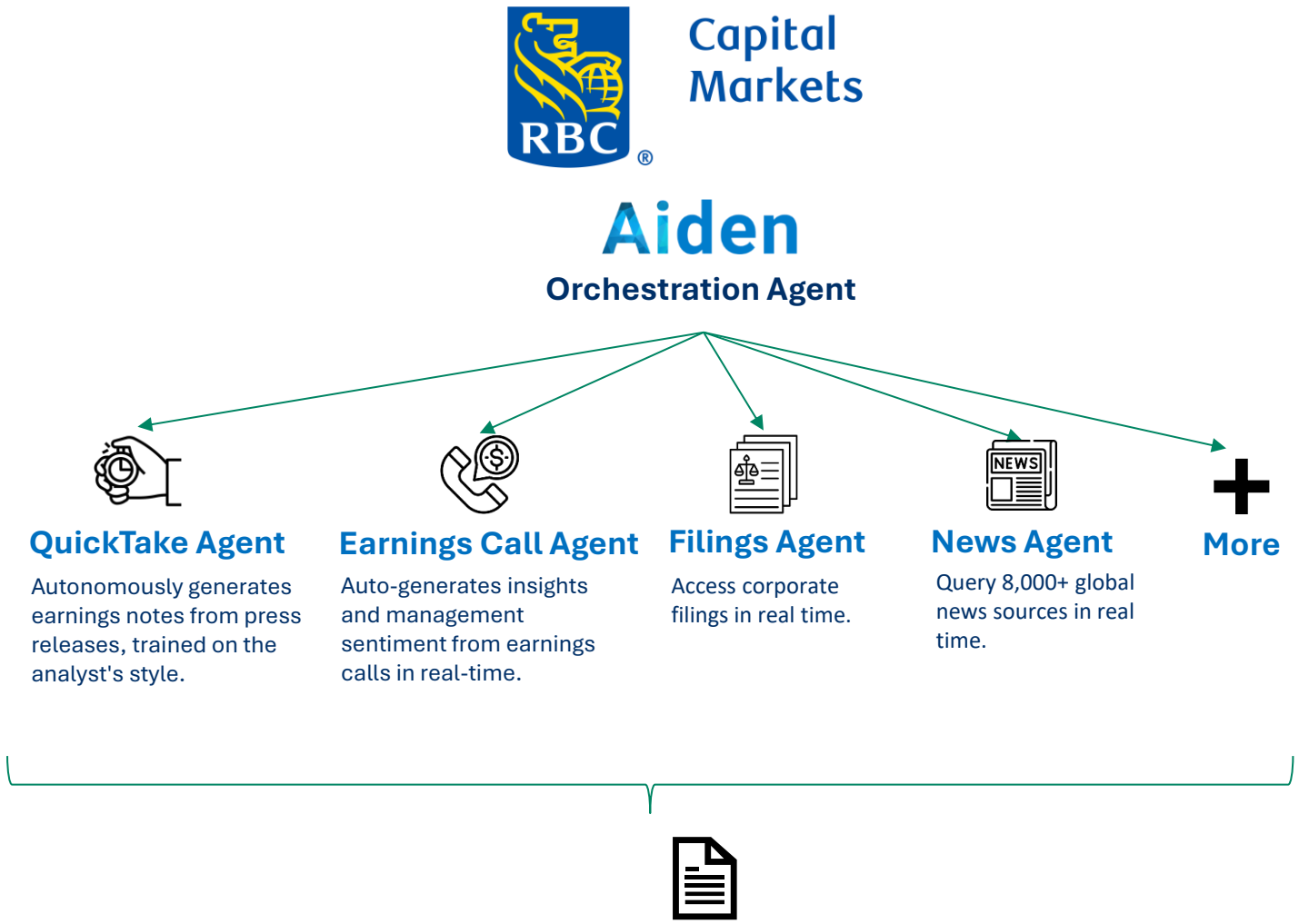
Agentic Execution: Operationalize intelligence via autonomous agents to improve core banking processes and drive efficiency at scale

The Intelligence Flywheel: Capture continuous system learnings to perpetually sharpen decision-making



RBC Capital Markets Agentic AI Solution

- Powered by NVIDIA A-IQ, RBC Capital Markets' Aiden has internal agents perform analysis when companies release SEC filings
- Orchestration oversees numerous other agents that collaborate to provide insight
- Aiden will significantly enhance productivity and efficiency by automating repetitive tasks and generating client insights
- With global research analysts using in deployment today to scale company coverage and deliver deeper insights to clients, RBC Capital Markets will continue to roll out agentic tools across the company



Enabling RBC to deliver deeper insights to clients up to 60% faster.

Blog: https://resources.nvidia.com/en-us-financial-services-industry/rbc-capital-markets

Transaction Foundation Models are the Next Frontier

Transaction Foundation Models leverage Transformers and Graph Neural Networks (GNNs) to convert transactions into embeddings

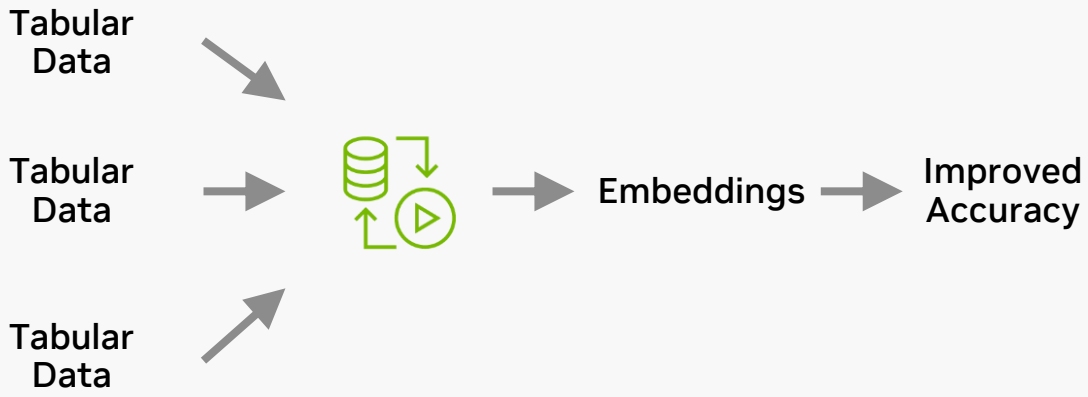
Embeddings learn intricate patterns by codifying customer's persona, preferences, and relationships, providing deeper context and accuracy

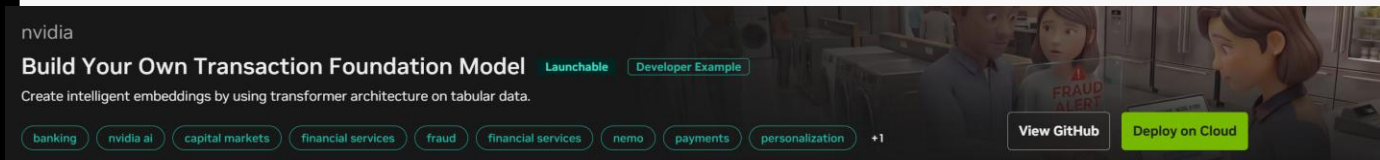
Impacts multiple use cases such as improving fraud detection, card authorization, and offer personalization

Better model accuracy reduces fraud losses, improves offers targeting, and lowers false positives improving customer experience

Financial Services

Revolut Builds a Transaction Foundation Model With NVIDIA Accelerated Computing Platform

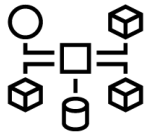




NVIDIA NeMo Microservices & Open-Source Models

Modular, API-first services for rapidly customizing AI agents and building data flywheels

NeMo Data Designer
GENERATE



NeMo Retriever*
LEARN

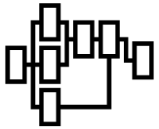
NeMo Safe Synthesizer
SECURE

NeMo Guardrails
GUARDRAIL

NeMo Curator*
PROCESS

NIM
INFERENCE

NeMo Evaluator
EVALUATE



NeMo Customizer
CUSTOMIZE

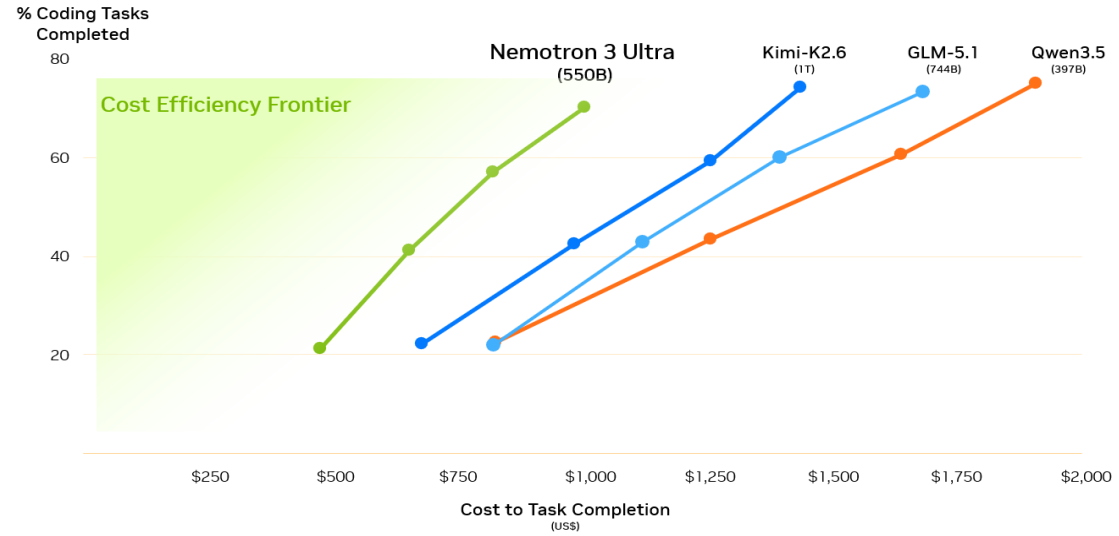
Key Features

NeMo Data Designer: Designing synthetic data from scratch
NeMo Safe Synthesizer: Privacy-safe Data
NeMo Curator: Collect, and preprocess data
NeMo Customizer: LoRA, SFT, DPO, GRPO
NeMo Evaluator: Benchmarking, CI/CD Evals
NeMo Guardrails: Programmable safety enforcement
NeMo Retriever: Open models for information retrieval
NIM: Performant model hosting containers

NVIDIA Nemotron 3 Ultra

Nemotron 3 Ultra

Frontier-reasoning, Orchestration
550B-A55B



SWE-bench Verified, CoreWeave, DeepInfra

Fastest Task Completion
Highest throughput with frontier accuracy completes more tasks

Lower Cost
Highest token efficiency reduced cost per task

Post-Trained for Agent Harness
Tuned across leading agent for multi-turn, multi-tool workflows

Open
Open model, datasets, environments.
Run anywhere – workstation to cloud

NVIDIA

NeMo Agent Toolkit Reduces Agentic System Complexity

Flexible integration, configuration, monitoring, and optimization

Framework Agnostic

Multi-framework support: LangChain, LangGraph, CrewAI, Semantic Kernel, Google ADK

Configuration System

YAML-based configuration for rapid agent deployment and prototyping

Observability Tracing

End-to-end tracing across frameworks and agent invocations

Evaluation System

Customizable evaluation framework for agent workflow components

Profiler

Performance analysis and runtime bottleneck identification

Guardrails

Function-level security controls for agent workflows*

Inference Optimization

NVIDIA Dynamo integration for accelerated performance*

End-to-End Profiling

2 – 8X

latency reduction in real agents, plus insights into token efficiency, bottlenecks, tool usage, and prompt quality

Offline Hyperparameter Tuning

Up to 35%

accuracy gain with 20% fewer tokens by optimizing model settings (temperature, top_p, max_tokens) and prompts

Online Inference Optimization

Up to 4X

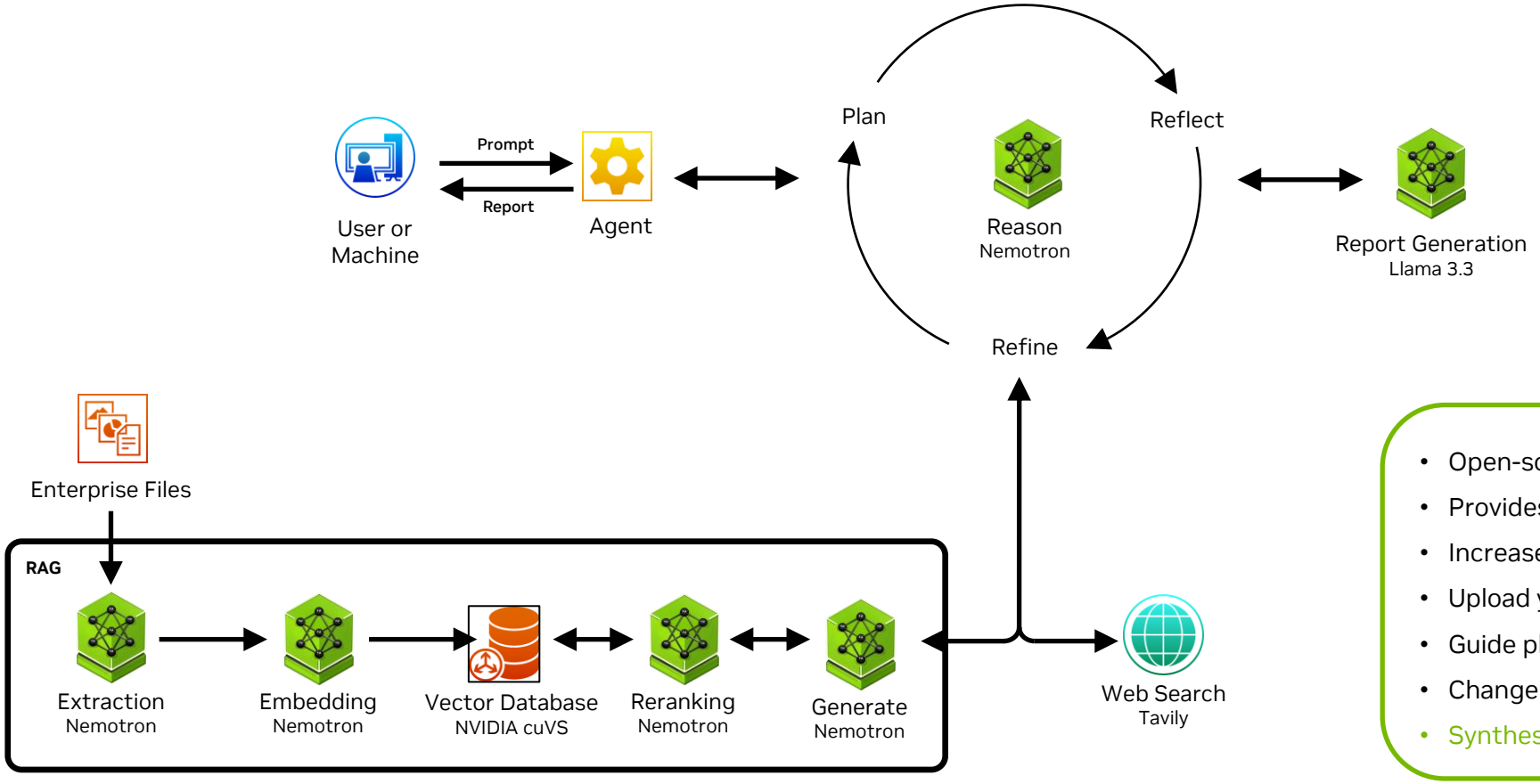
faster time-to-first-token with workload-aware and KV-cache-aware routing for complex agent workflows

GPU Sizing Intelligence

built-in calculator estimates how many GPUs are needed to run each agent workflow at target scale

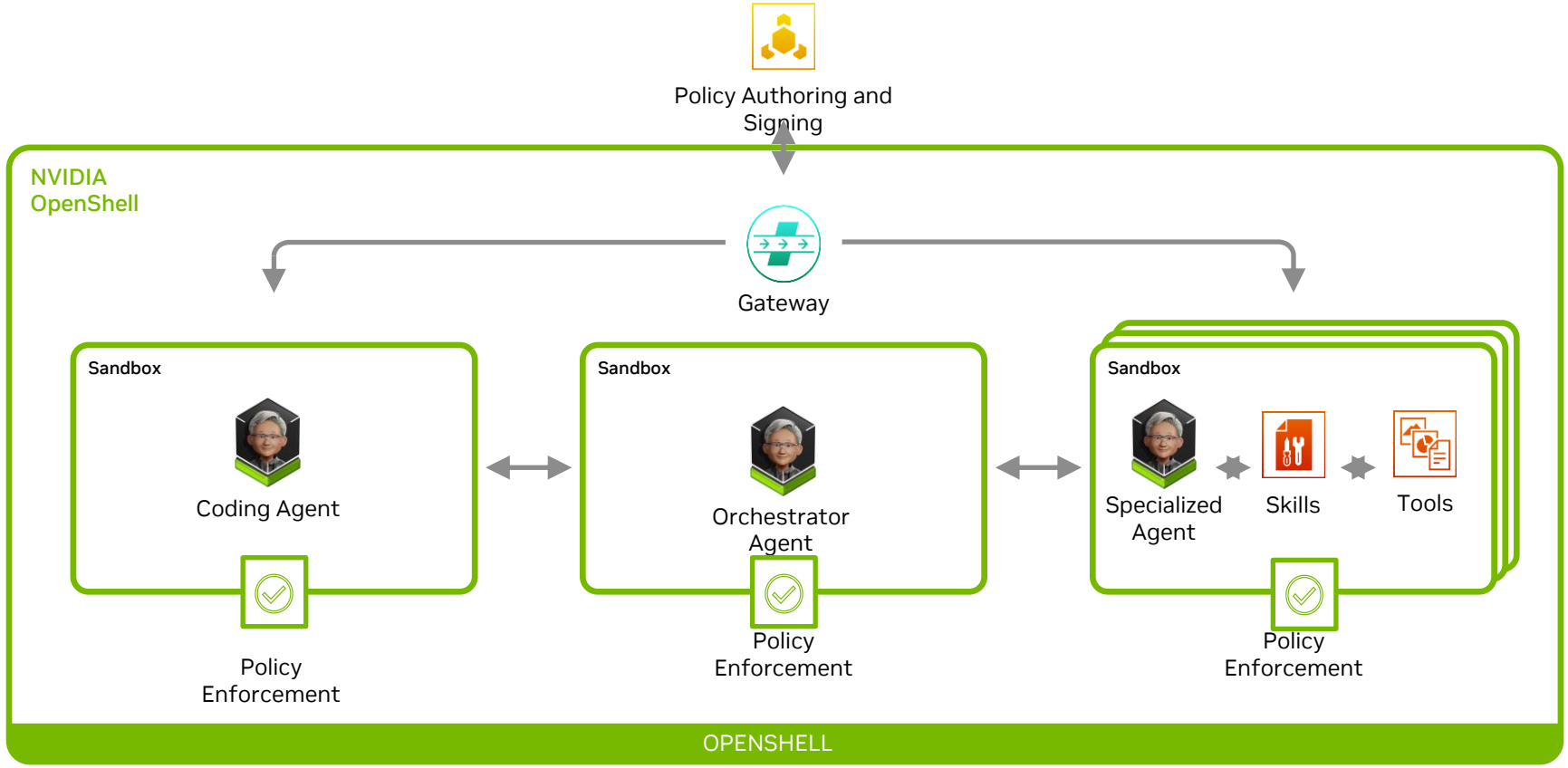
NVIDIA AI-Q Blueprint for Deep Research

Synthesize Hours of Research in Minutes Using Your Own Data



- Open-source blueprint for research & reporting
- Provides observability and transparency
- Increases agent accuracy with reasoning
- Upload your own data sources
- Guide plan and report in real-time
- Change report output dynamically
- Synthesize hours of material in minutes

Securely Govern Any Agent



Secure Runtime for Every Agent

Model-agnostic, Harness-agnostic infrastructure

Governance and Trust

A trust boundary for every autonomous action, at any scale

Accelerated Tooling

Agent native tooling that lets agents run, learn, and self-improve continuously

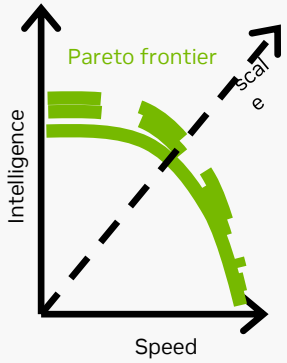
AI Factories Power a New Industrial Revolution

AI Factories: highly purpose-built systems that harness compute, networking, storage, and software

Manufactures intelligence from data, in the form of digital tokens

NVIDIA provides Reference Architecture for AI Factories

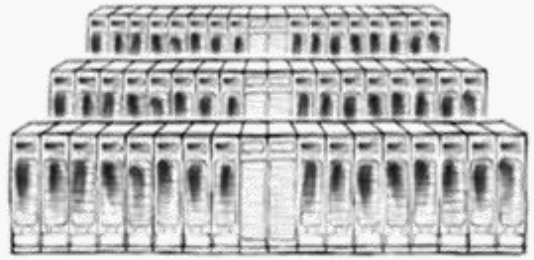
Extend the Pareto Frontier



Agentic AI Banks

Prevent Fraud

AI FACTORIES



NVIDIA

jpapenbrock@nvidia.com