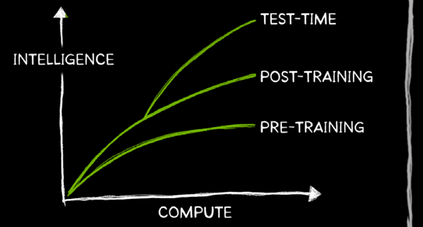

NVIDIA

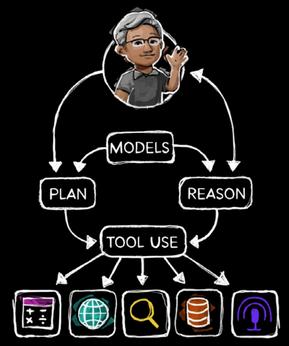
AI Scales Beyond LLMs

"Everything Everywhere All at Once"

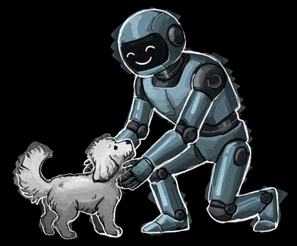


3 SCALING LAWS

COMPUTE IS DATA



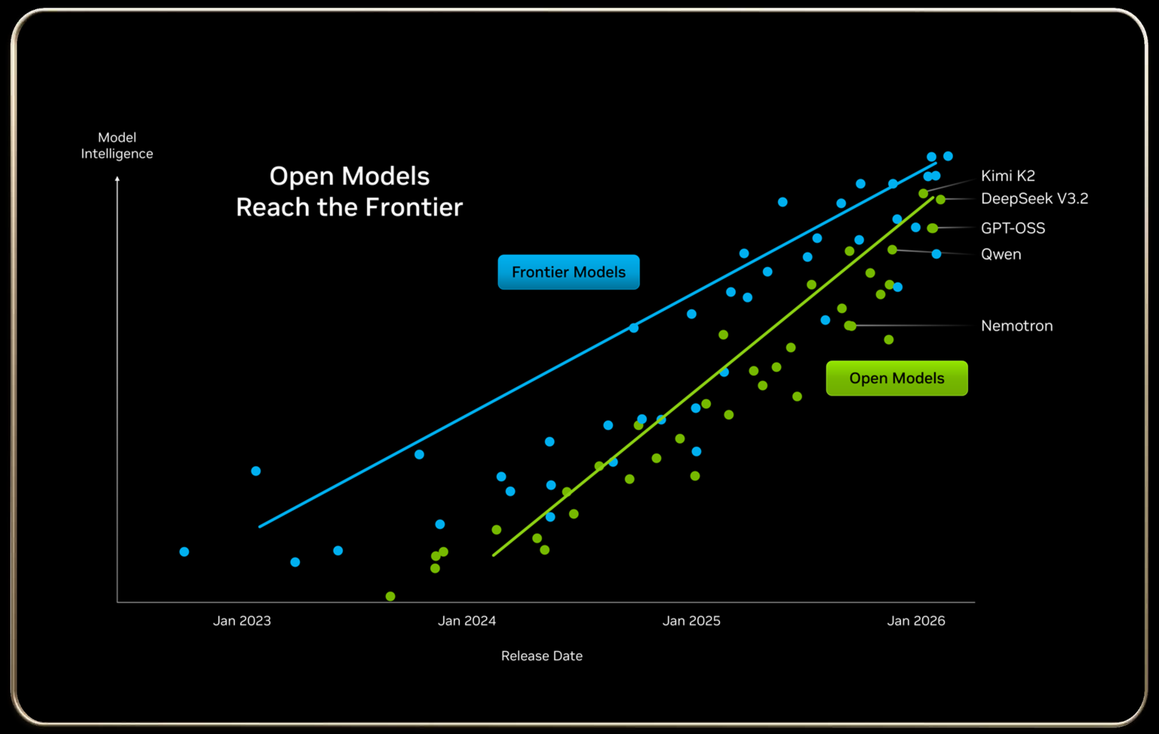
AI BECOMES AGENTIC

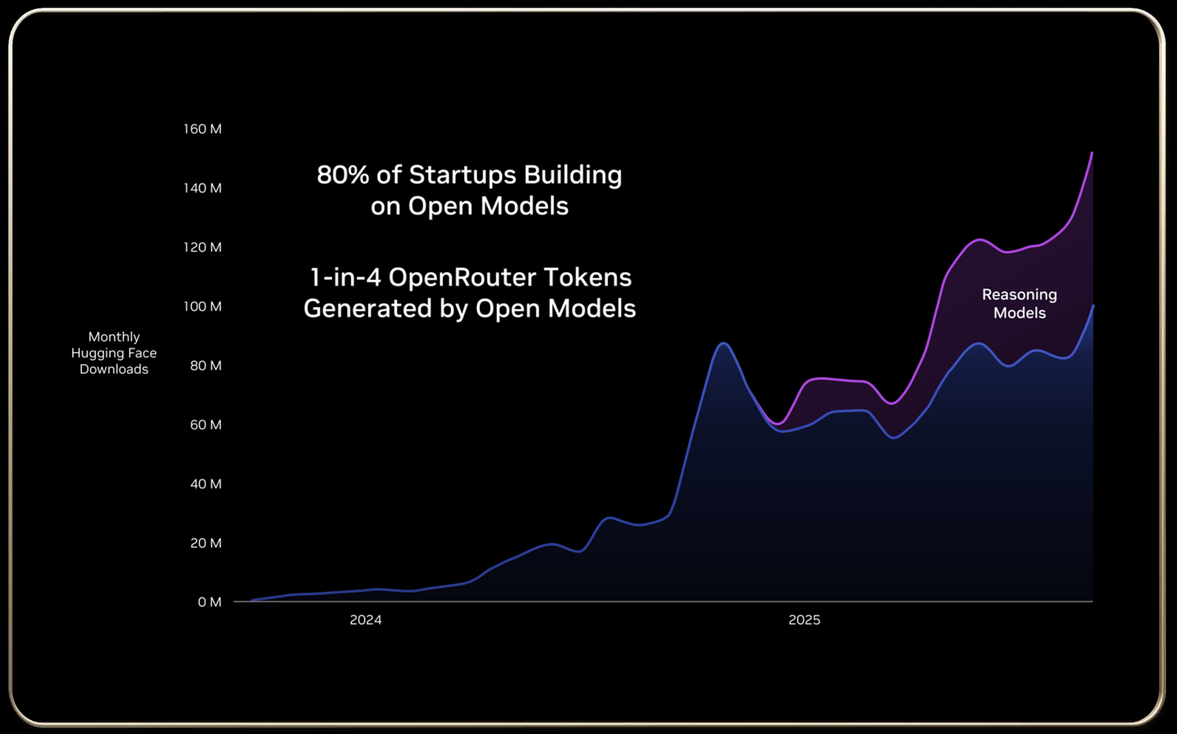


PHYSICAL AI TAKES LEAP

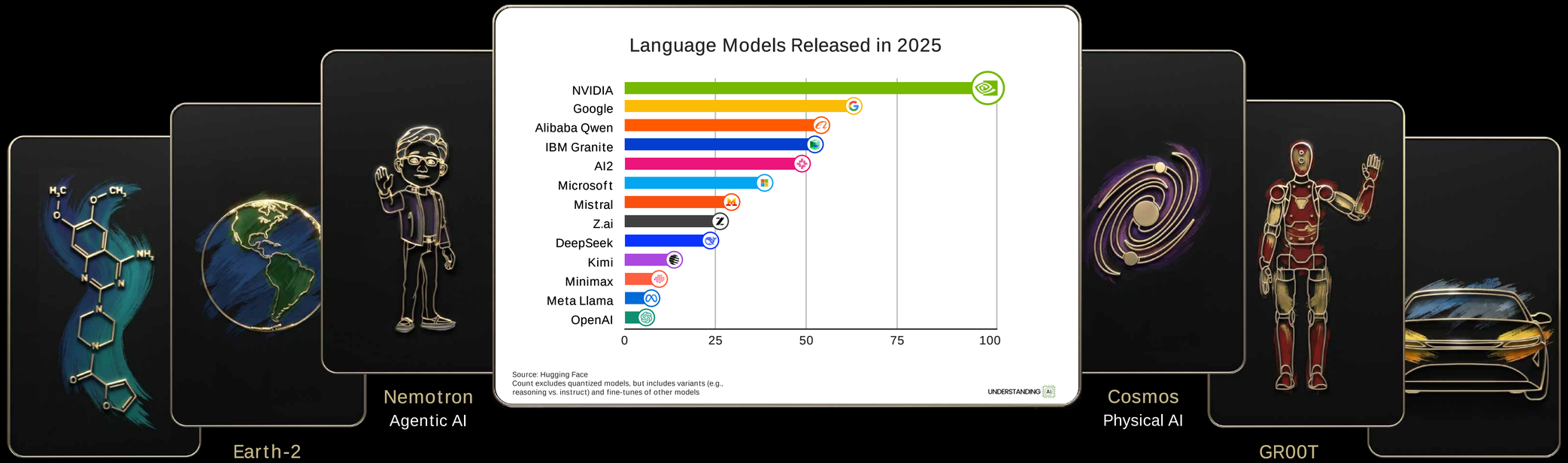
AI LEARNS LAWS OF NATURE

OPEN MODELS REACH FRONTIER


Model Intelligence
Open Models Reach the Frontier
Frontier Models
Open Models
Kimi K2
DeepSeek V3.2
GPT-OSS
Qwen
Nemotron
Jan 2023
Jan 2024
Jan 2025
Jan 2026
Release Date


80% of Startups Building on Open Models
1-in-4 OpenRouter Tokens Generated by Open Models
Monthly Hugging Face Downloads
160 M
140 M
120 M
100 M
80 M
60 M
40 M
20 M
0 M
Reasoning Models
2024
2025

NVIDIA Leads Open Model Ecosystem



Clara
Biomedical AI

Earth-2
AI-Physics

Nemotron
Agentic AI

Cosmos
Physical AI

GR00T
Robotics

Alpamayo
Autonomous Vehicles

Generate and Process Data

Model Building

Train and Customize

Evaluate

Research

Guardrail & Deploy

NVIDIA Tops Leaderboards

Artificial Analysis
Intelligence Index

Artificial Analysis
Intelligence per Token

#	User (Team)	Algorithm	Created	Mean Position
1st	sina.amirrajab (IJM_PM)	Report2CT	11 Aug. 2025	1.7
1st	zohaibsalahuddin00 (IJM_PM)	CT Gen with Classifier Guidance	29 Aug. 2025	1.7
3rd	daniele.molino (ArCo Lab)	Rep2CT No Rag	15 July 2025	3
4th	RayIW (CTFlow)	CTFlow	30 Aug. 2025	4.7
5th	daiweicheng4536 (always_wrong)	Medsyn v2 448	31 Aug. 2025	5.7
6th	pmath2012	LoRAD: Low-Resource Adaptive Diffusion via Pretrained Latent Representations	11 Aug. 2025	6.7
7th	vlm3dchallenge (CT-RATE Team)	GenerateCT (Scores will be updated)	6 July 2025	7.7
8th	ctagarow	LtmVAE decoder	2 Sept. 2025	8
8th	bhuhm (Potato Rebellion)	baseline	25 July 2025	8
10th	Stefan (MIC-DKFZ)	ctgpn2	22 Aug. 2025	8.3

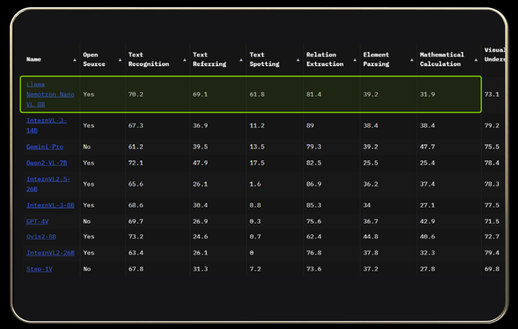
VLM3D

model	Average WER	RTFx	License	AMI	Earnin
nvidia/canary-qwen-2.5b	5.63	418.28	Open	10.19	10.45
ibm-granite/granite-speech-3.3-8b	5.74	146.42	Open	8.98	9.42
ibm-granite/granite-speech-3.3-2b	6	270.57	Open	8.9	10.25
microsoft/Phi-4-multimodal-instruct	6.02	151.1	Open	11.09	10.16
nvidia/parakeet-tdt_0.6b-v2	6.05	3386.02	Open	11.16	11.15
aquavoice/avalon-v1-en	6.24	NA	Proprietary	11.58	11.37
nvidia/parakeet-tdt_0.6b-v3	6.32	3332.74	Open	11.39	11.19
nvidia/canary-1b-flash	6.35	1045.75	Open	13.11	12.77
krystal/stt-2.6b-en	6.4	88.37	Open	12.17	10.99
nvidia/canary-1b	6.5	235.34	Open	13.9	12.19
nyrahealth/CrisperWhisper	6.67	84.05	Open	8.71	12.89
elevenlabs/scribe_v1	6.88	NA	Proprietary	14.43	12.14

OpenASR

Model Name	IntPhys 2 (%)	MVPBench (%)	CasualVQA (%)	Model Type
Human	92.44	92.9	84.78	Human
Cosmos-Reason1-7B	59.88	41.31	48.17	Open
V-JEPA 2	56.4	44.5	44.89	Open
GPT-4o	53.19	32.5	50.95	Closed
Qwen2.5-VL	49.12	36.7	49.05	Open
Gemini 2.5 Flash	56.1	-	61.66	Closed
Perception Language Model (PLM)	-	39.7	50.86	Open
InternVL2.5	-	39.9	47.54	Open
Gemini 1.5 Pro	52.1	29.6	-	Closed
LLaVA-OneVision	-	20.7	45.27	Open
test	0	-	25.47	Open
BAM-MMAI (12e8_70B)	-	-	67.59	Closed

Physical Reasoning
MVPBench, IntPHys, CasualVQA



Name	Open Source	Text Recognition	Text Referring	Text Spotting	Relation Extraction	Element Parsing	Mathematical Calculation	Visual Underst
Llama Nemotron Nano VL 8B	Yes	70.2	69.1	61.8	83.4	39.2	31.9	73.1
InternVL 3-14B	Yes	67.3	36.9	11.2	89	38.4	38.4	79.2
Gemini Pro	No	61.2	39.5	13.5	79.3	39.2	47.7	75.5
Qwen2 VL 7B	Yes	72.1	47.9	17.5	82.5	25.5	25.4	78.4
InternVL2.5-26B	Yes	65.6	26.1	1.6	86.9	36.2	27.4	78.3
InternVL 3-8B	Yes	68.6	30.4	8.8	85.3	34	27.1	77.5
GPT 4V	No	69.7	26.9	8.3	79.6	36.7	42.9	71.5
Ovis2-8B	Yes	73.2	24.6	0.7	62.4	44.8	40.6	72.7
InternVL2-26B	Yes	63.4	26.1	0	76.8	37.8	32.3	79.4
Step-1V	No	67.8	31.3	7.2	73.6	37.2	27.8	69.8

OCRBench

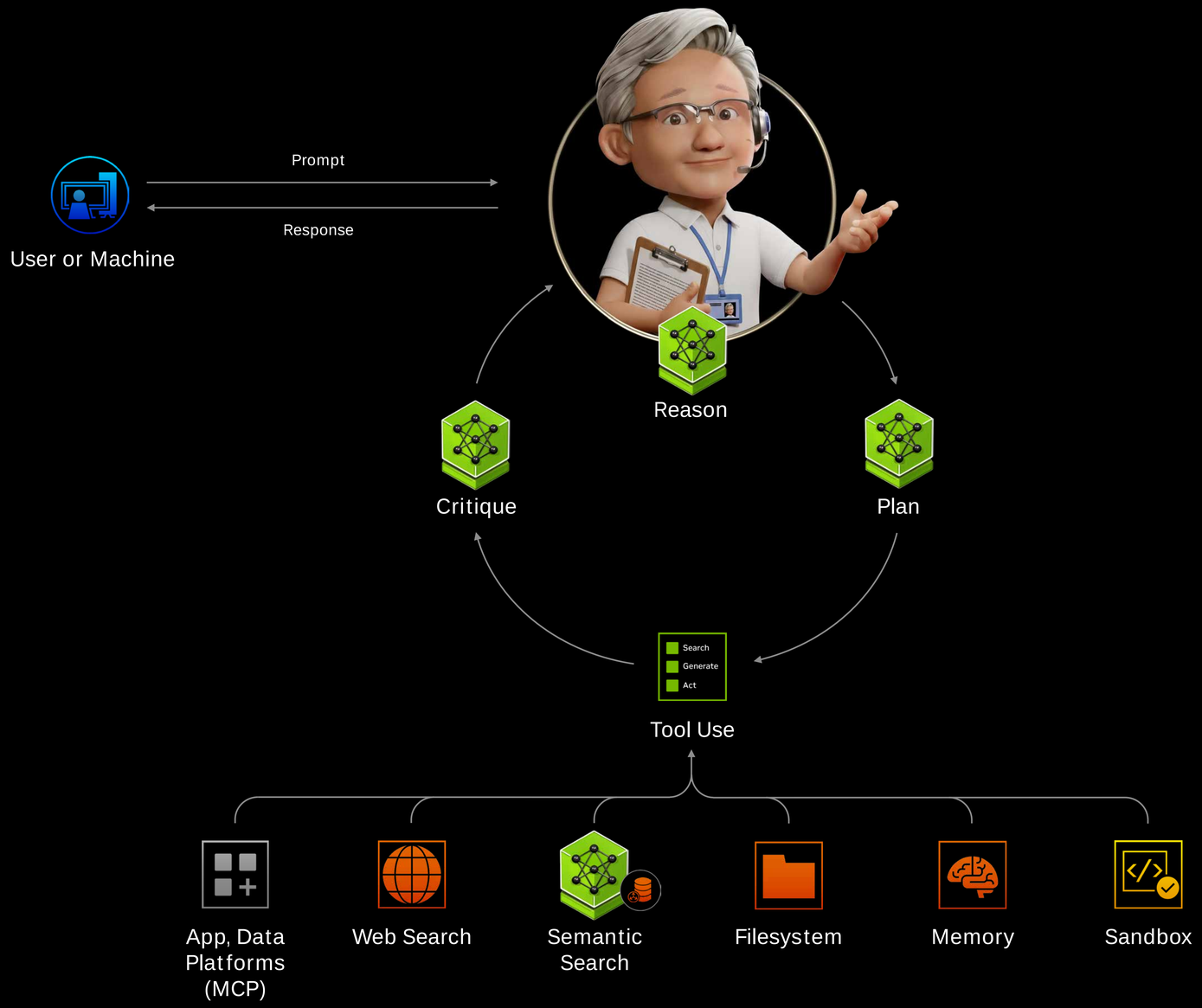
Model	Zero-shot	Memory Usag…	Number of Para…	Embedding Dime…	Max Tokens	Mean (T
KaLM-Embedding-Gemma3-12B-2511	73%	44884	11.8	3840	32768	72.32
Llama-embed-nemotron-8b	99%	28629	7.5	4096	32768	69.46
Qwen3-Embedding-8B	99%	14433	7.6	4096	32768	70.58
gemini-embedding-001	99%			3072	2048	68.37
Qwen3-Embedding-4B	99%	7671	4.0	2560	32768	69.45
Octen-Embedding-8B	99%	14433	7.6	4096	32768	67.85
Seed1.6-embedding-1215	89%			2048	32768	70.26
Qwen3-Embedding-0.6B	99%	1136	0.596	1024	32768	64.34
gte-Qwen2-7B-instruct	NA	29040	7.6	3584	32768	62.51
Linq-Embed-Mistral	99%	13563	7.1	4096	32768	61.47
multilingual-e5-large-instruct	99%	1068	0.560	1024	514	63.22
embeddinggemma-300m	99%	1155	0.308	768	2048	61.15

Retrieval
ViDoRe, MTEB, MMTEB

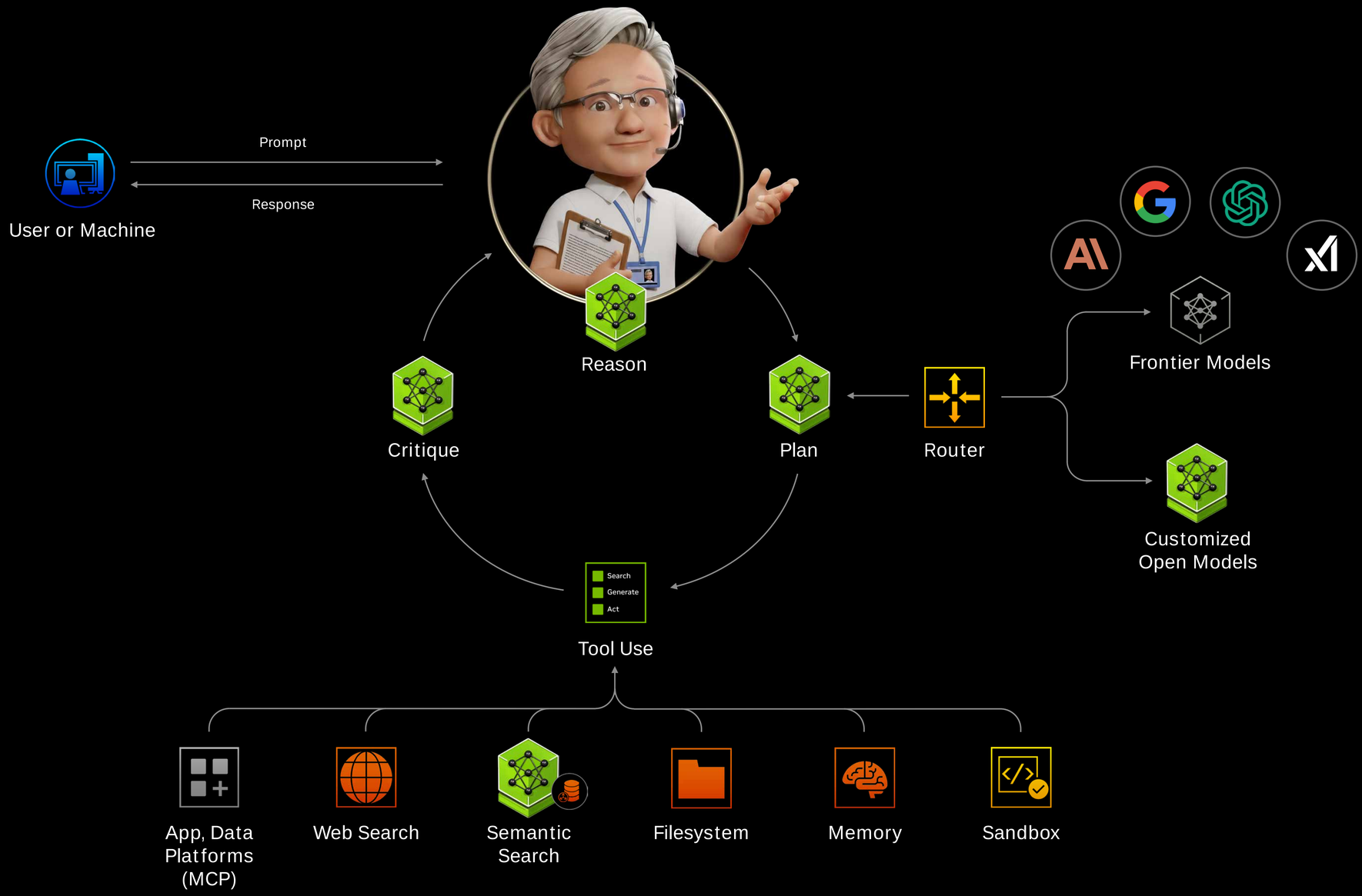
Model	Overall	Domain
Source	82.6	87.1
Vec-3	82.1	86.7
Cosmos-Predict2.5-14B	81.0	83.8
Cosmos-Predict2.5-2B	81.0	84.0
Wan2.2-I2V-A14B	80.6	84.1
Wan2.2-TI2V-5B	80.4	83.4
Cosmos-Predict2-14B-Video2World	80.0	84.3
Wan2.1-I2V-14B-720P	79.8	82.7
Cosmos-Predict2-2B-Video2World	79.6	83.9
MAGI-1-24B	78.5	80.5
CogVideoX1.5-5B-I2V	78.3	80.1

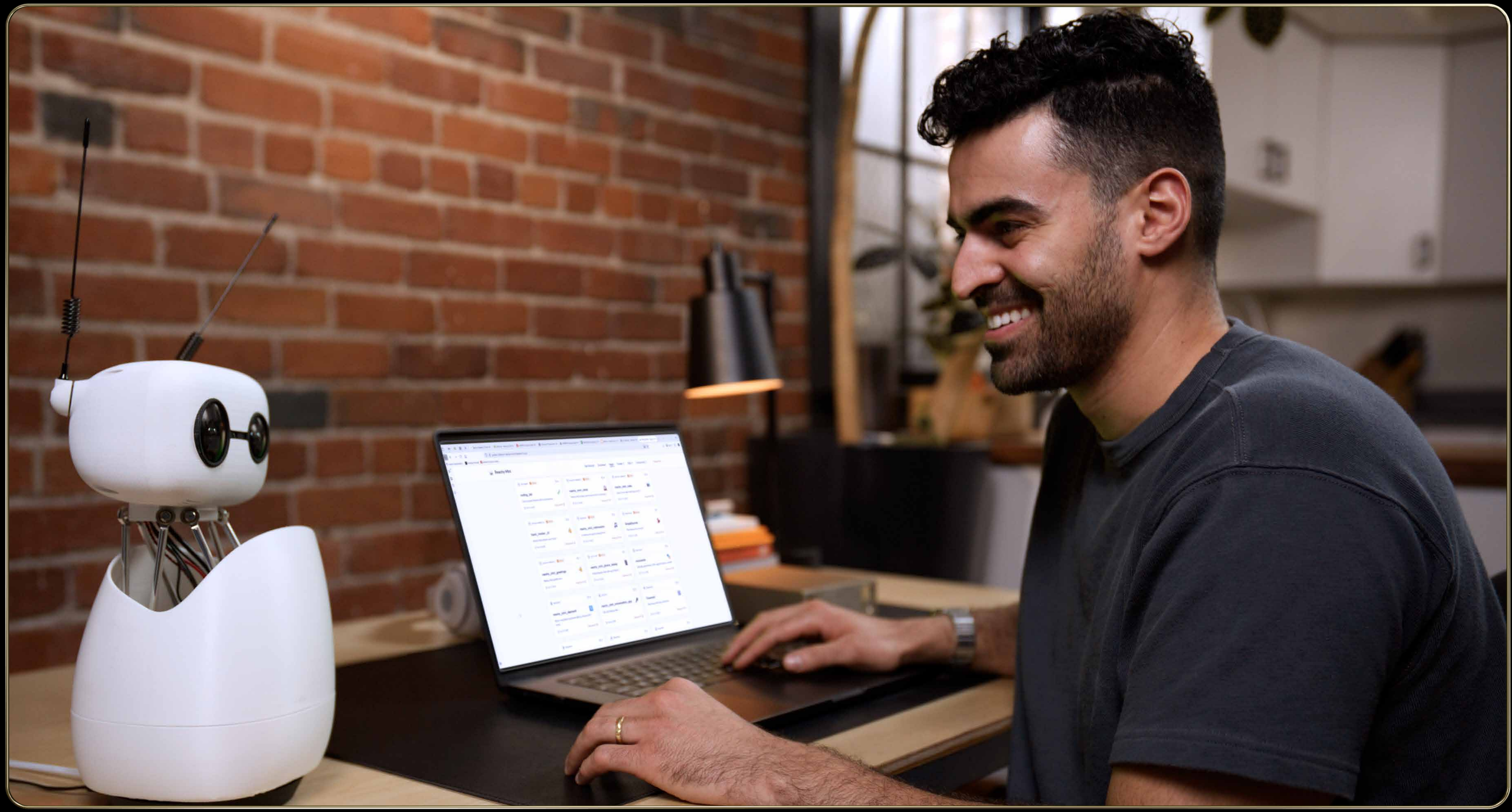
Physical AI
PAIBench

Agents Are Multi-Model, Multi-Cloud, and Hybrid-Cloud



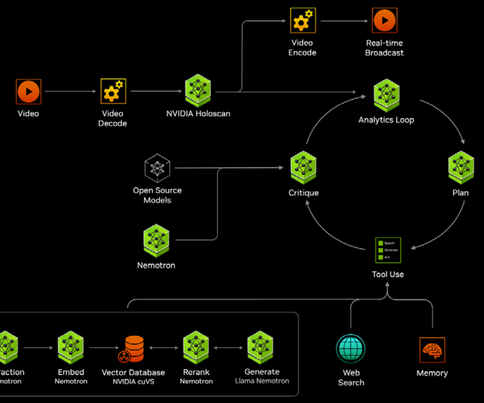
Agents Are Multi-Model, Multi-Cloud, and Hybrid-Cloud





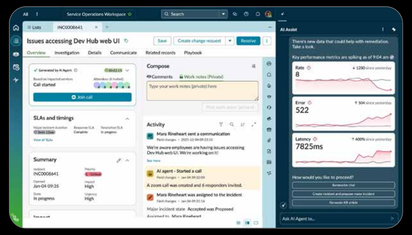
Enterprise Platforms Adopt NVIDIA AI

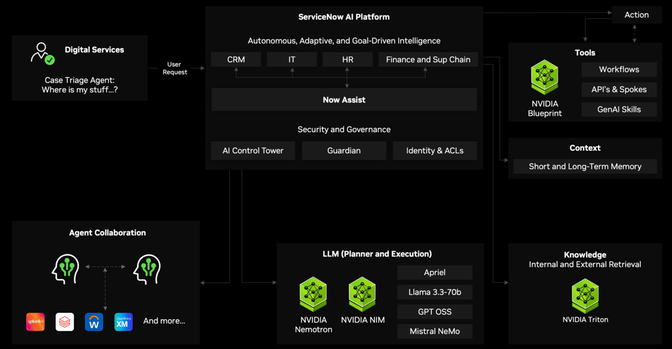
Edge Video Analytics



Palantir

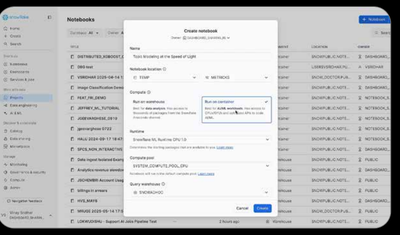
IT and HR Workflow Automation

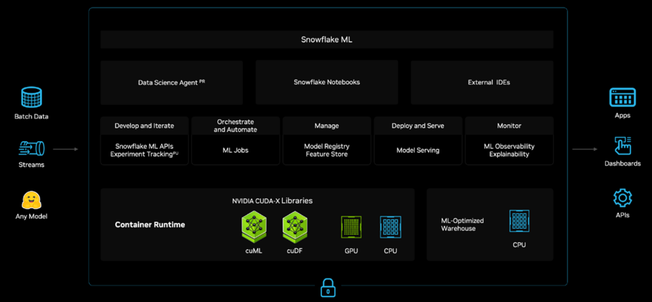




servicenow

Accelerated Data Science and ML

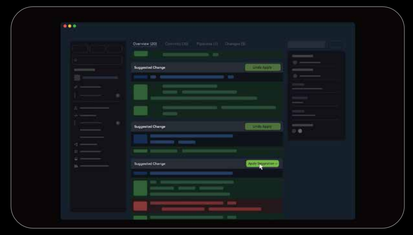


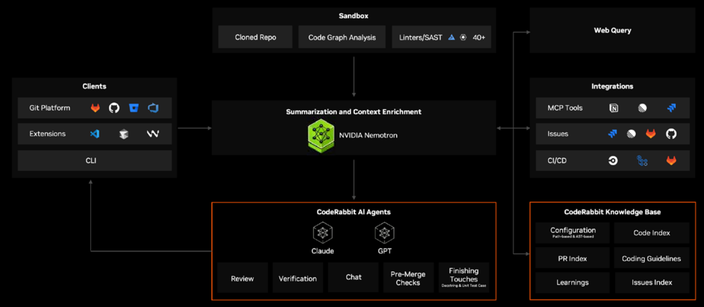


snowflake

Enterprise Platforms Adopt NVIDIA AI

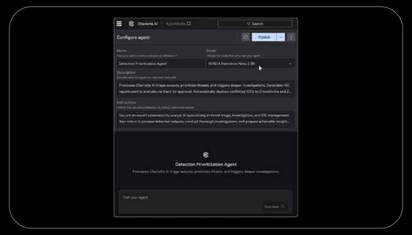
Code Review

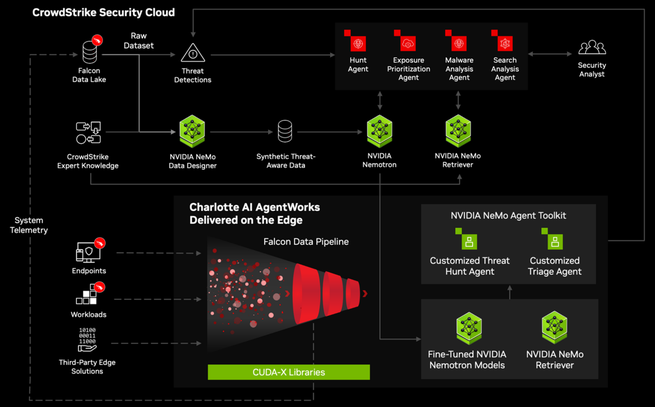




CodeRabbit

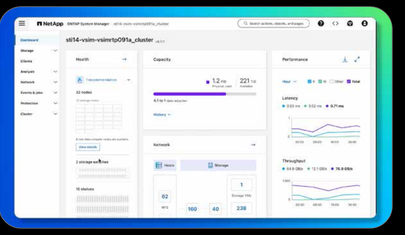
Real-Time Security Analyst

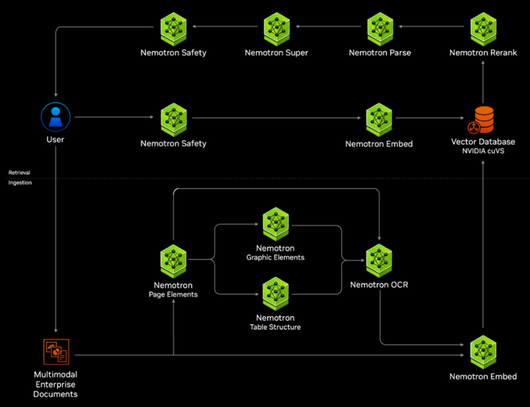




CROWDSTRIKE

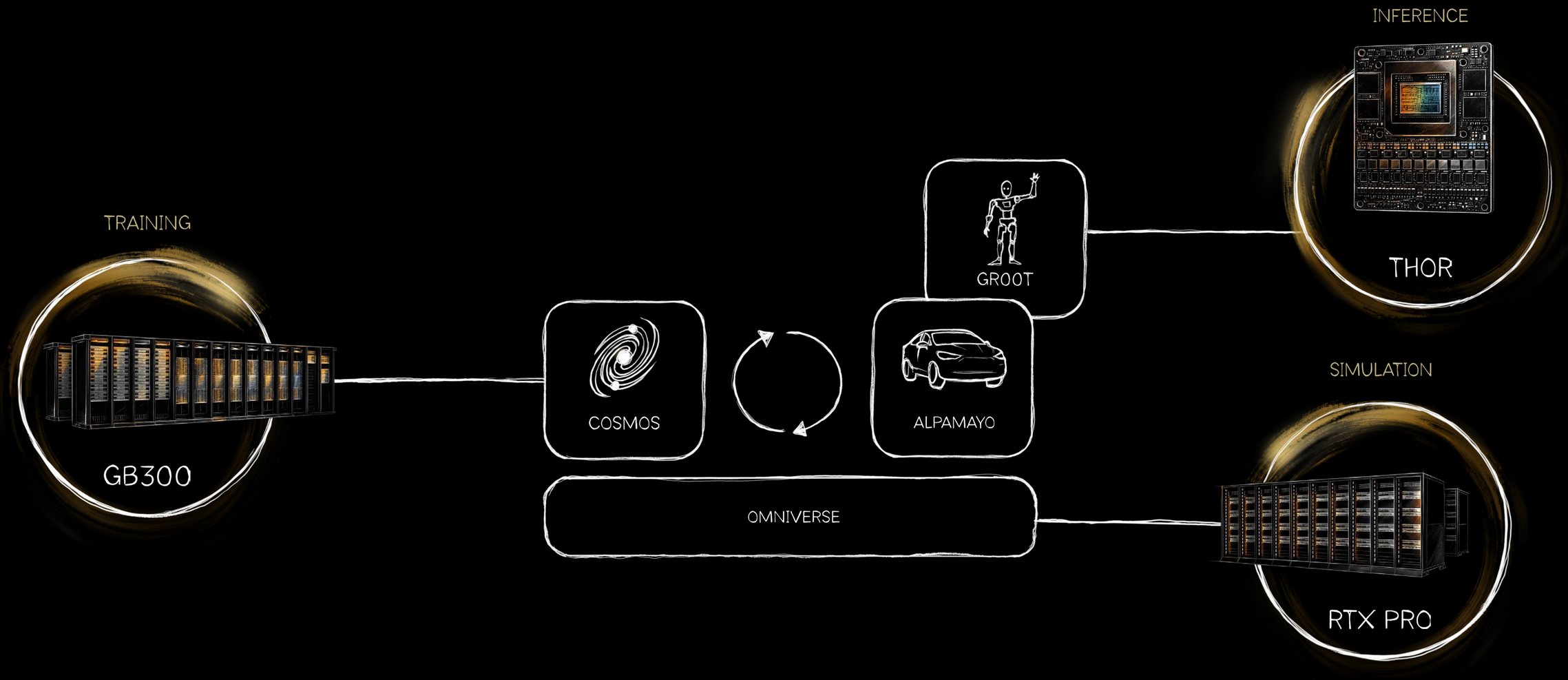
Retrieval Augmented Generation



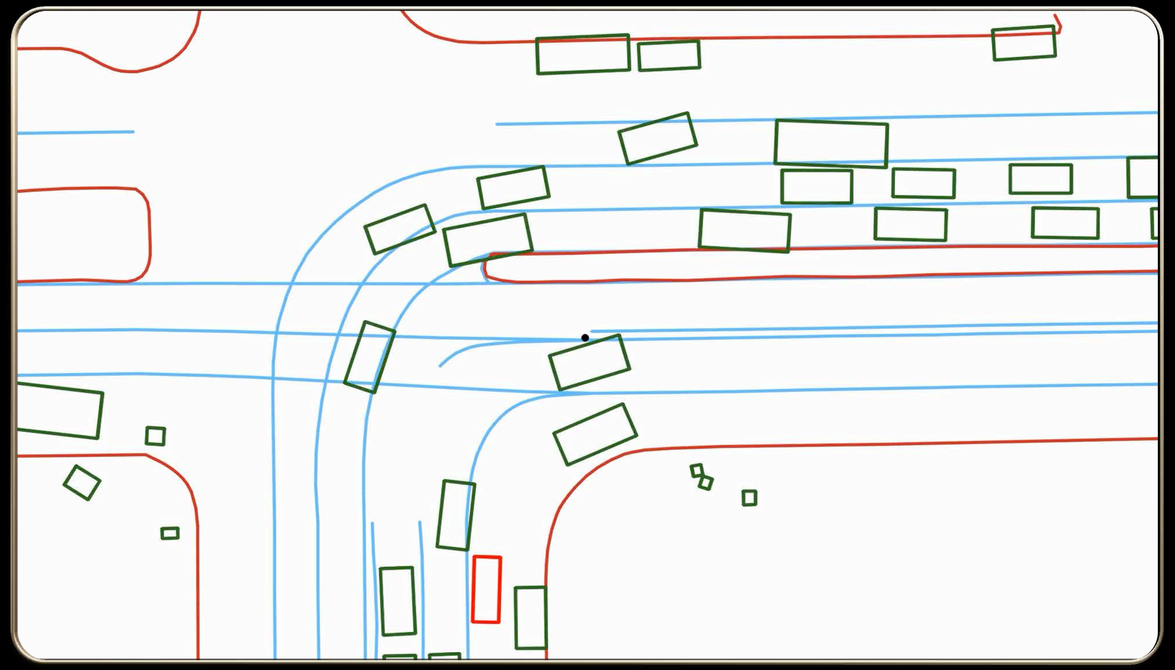


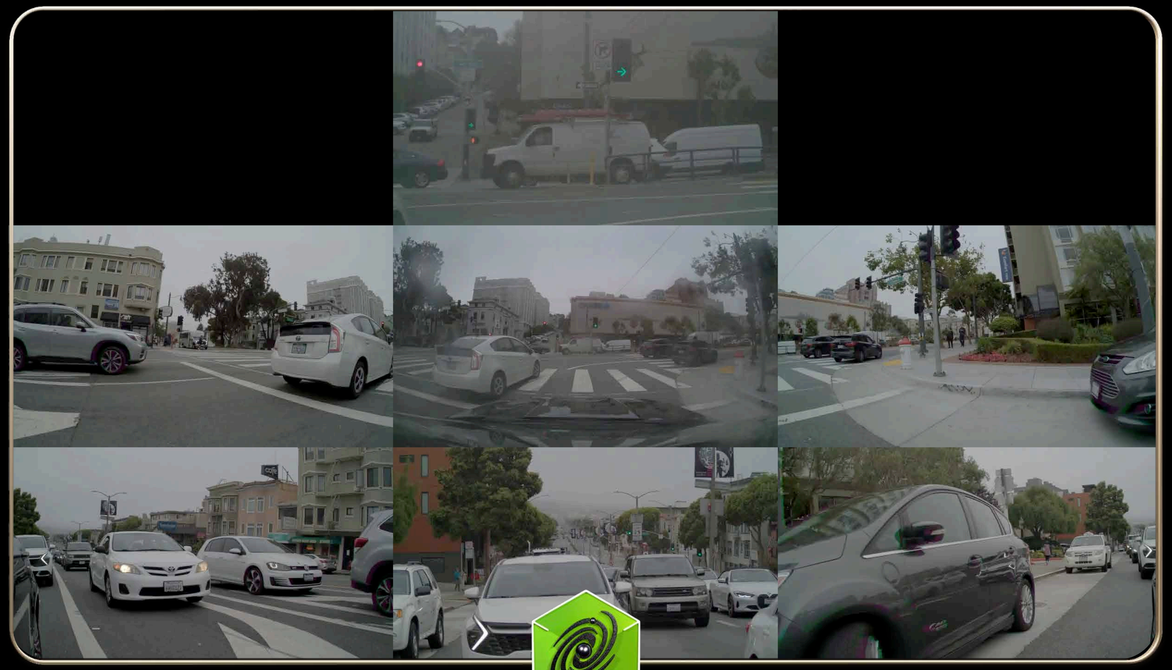
NetApp

NVIDIA Full-Stack Physical AI Platform



Compute is Data





Cosmos

Announcing NVIDIA Alpamayo

Open Reasoning VLA for Autonomous Vehicles

Announcing NVIDIA Alpamayo

Open Reasoning VLA for Autonomous Vehicles

Alpamayo

“Parked cars mean the ego needs to be cautious of sudden movements”

“Ball rolling could be a hazard-maybe a child or pet following it”

“Slow down and prepare to stop if the ball or pedestrians pose a risk”


09:19am
750 ft
Fillmore Street
Chestnut Street
102.1 FM | KRBQ ...
Tap Here
to Connect

NVIDIA Ships Full-Stack AV on 2025 Mercedes Benz CLA

Alpamayo

Policy and Safety Evaluator

Classical AV Stack

Halos Safety OS

Global L4 and Robotaxi Ecosystem Building on NVIDIA

AV Software

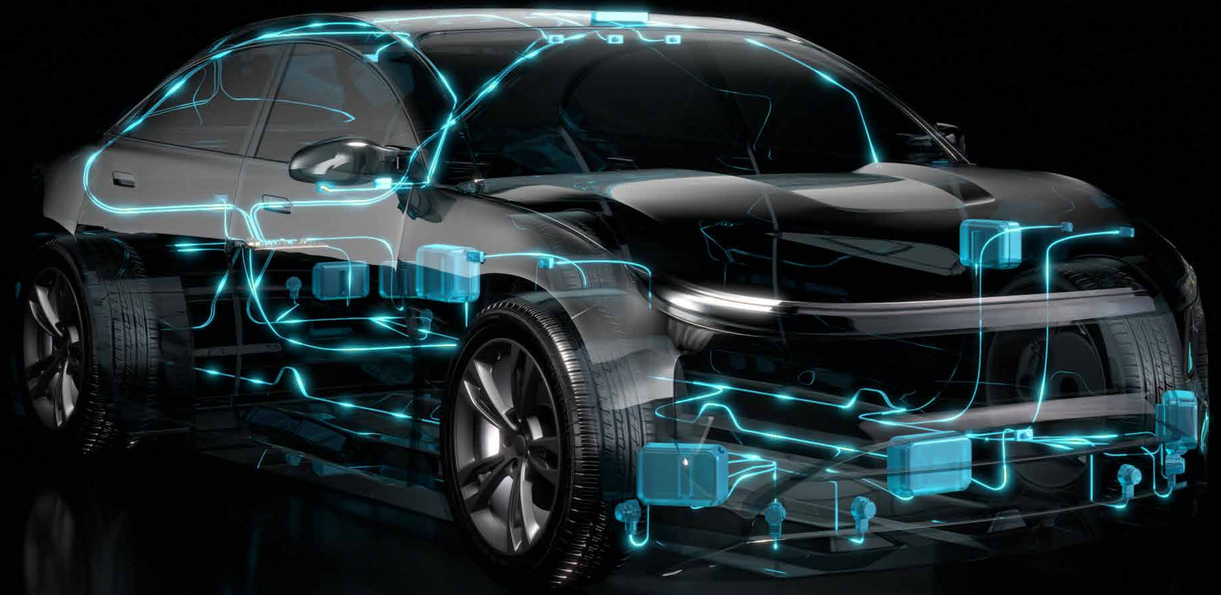
Aurora · DEEPROUTE.AI · momenta · nuro · Plus · pony.ai · waabi · WAYVE · WeRide · ZYT

OEMs and Mobility Platforms

BYD · DiDi · GEELY · GWM · IM MOTORS · LUCID · SAIC · STELLANTIS · xiaomi · Uber

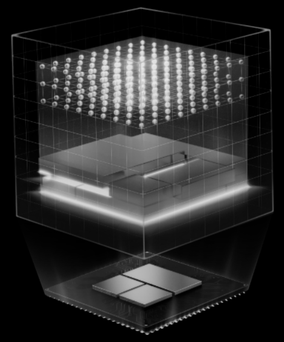
Tier 1 and Hardware Partners

AEVA · Astemo · AUMOVIO · BOSCH · DENSO · DESAY · FOXCONN · HESAI · Lenovo · MAGNA · OMNIVISION · Quanta Computer · SONY · ZF GROUP




CAT
745
CAT
320

cadence® | NVIDIA



Agentic AI

Optimization AI

Digital	Custom	Verification	PCB/Package	System
Cadence Cerebrus®	Virtuoso® Studio	Verisium™	Allegro® X AI	Optimality™

Silicon Agent

Systems Agent

Cadence JedAI Platform

Principled Simulation & Optimization

Electronic Design Automation

Digital & Signoff	Custom	Verification
Innovus™	Virtuoso	Xcelium™
Voltus™	Spectre® X	Jasper®
Celsius™	…	…
…		

System Design Automation

Package & PCB	System	Data Center
Allegro	Fidelity®	Cadence Reality™
Allegro APD	Clarity™	…
…	Beta CAE	
	…	

Sciences

Molecular
Orion
…

Accelerated Computing

Hardware Emulation	Software and System	AI Supercomputer
Palladium®	Protium™	Millennium™

X86 CPU	Arm® CPU	FPGA
GPU	DPU	Cadence IP

Knowledge Sources

cadence | 3rd Party

LLM

Vendor

Customer

RAG

Vendor

Customer

Compute

cadence | OnCloud

Customer Farm

3rd Party CSP

This slide contains forward-looking statements regarding Cadence's business or products. Actual results may differ materially from the information presented here.

SYNOPSYS® | NVIDIA

Electronic Design Automation

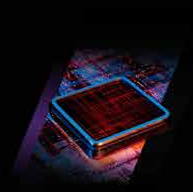
Design	Manufacturing	Verification
PrimeLib	Proteus	PrimeSim
Redhawk-SC	S-Litho	VCS
Totem SC	S-Device	
	Quantum ATK	

Computer Aided Engineering

Structures	Fluids	Electronics, Electromagnetics	Optics
Mechanical	Fluent	HFSS	SPEOS
LS-Dyna	FreeFlow	Icepak	Zemax
Discovery	Rocky	Maxwell	Lumerical FDTD

Synopsys AgentEngineer™ Technology

Physical AI






Semiconductor DT | Industrial DT | Medical DT | Aerospace DT | Robotics DT | Automotive DT

NVIDIA Omniverse

Targeting Availability in Every Cloud and OEM

aws | CoreWeave | DELL Technologies | Google Cloud | Hewlett Packard Enterprise | Lenovo | Microsoft Azure | Oracle Cloud Infrastructure | Supermicro

Legend

In production

SIEMENS | NVIDIA

Industries

Chip Design

- Fuse
- Solido
- Calibre
- Tessent
- Questa One
- Veloce
- Aprisa
- Catapult
- Xpedition
- HyperLynx

Design / Engineering

- Teamcenter Digital Reality Viewer
- Designcenter
- Simcenter
- Digital Twin

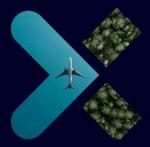
Operations

- Industrial AI Suite
- Industrial Edge
- Totally Integrated Automation (TIA)
- Industrial Cybersecurity Services
- Opcenter

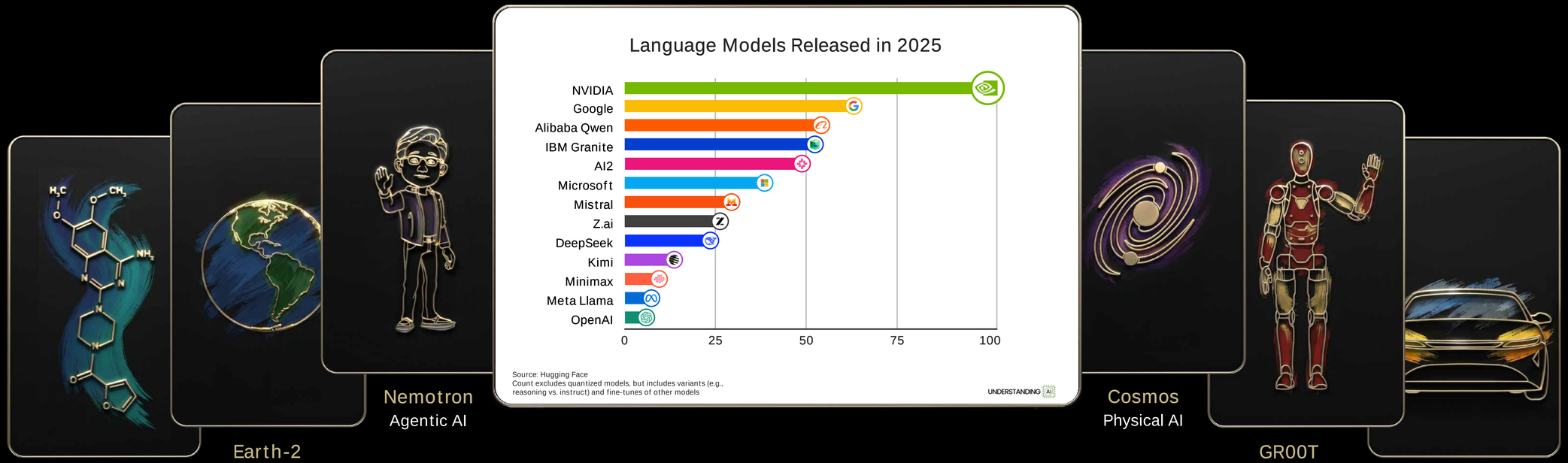
Infrastructure

- BuildingX
- Gridscale X
- Electrification X

Powered by Siemens Xcelerator



NVIDIA Leads Open Model Ecosystem



Clara
Biomedical AI

Earth-2
AI-Physics

Nemotron
Agentic AI

Cosmos
Physical AI

GR00T
Robotics

Alpamayo
Autonomous Vehicles

Generate and Process Data

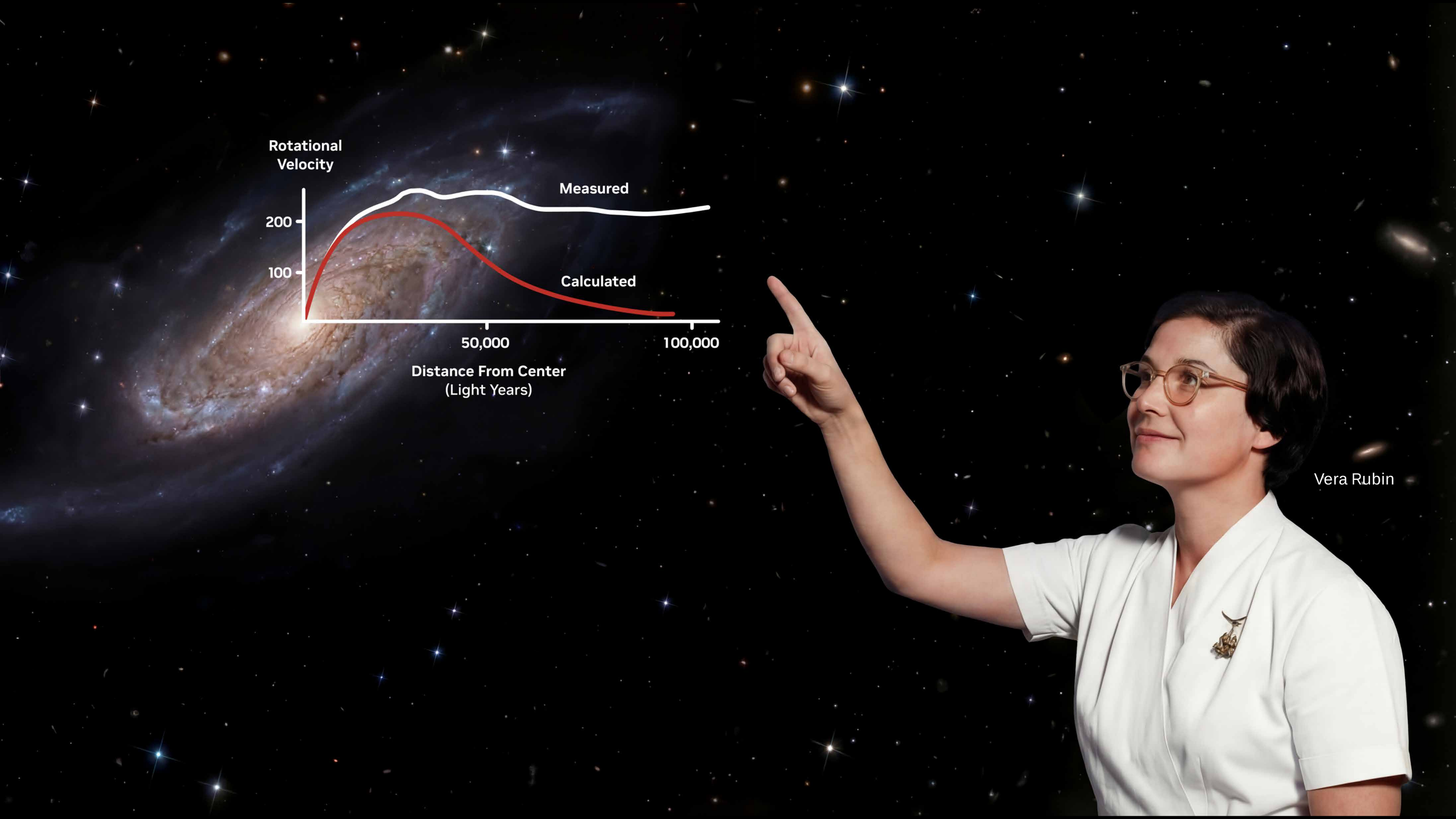
Model Building

Train and Customize

Evaluate

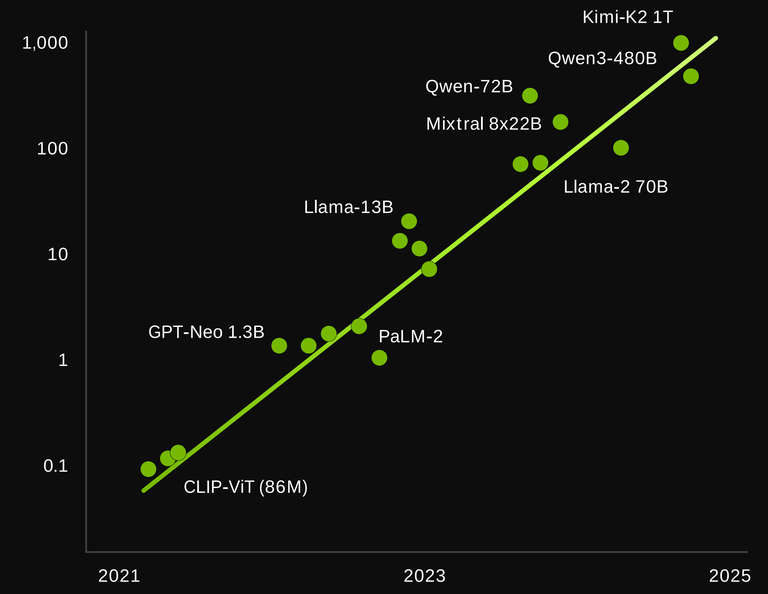
Research

Guardrail & Deploy

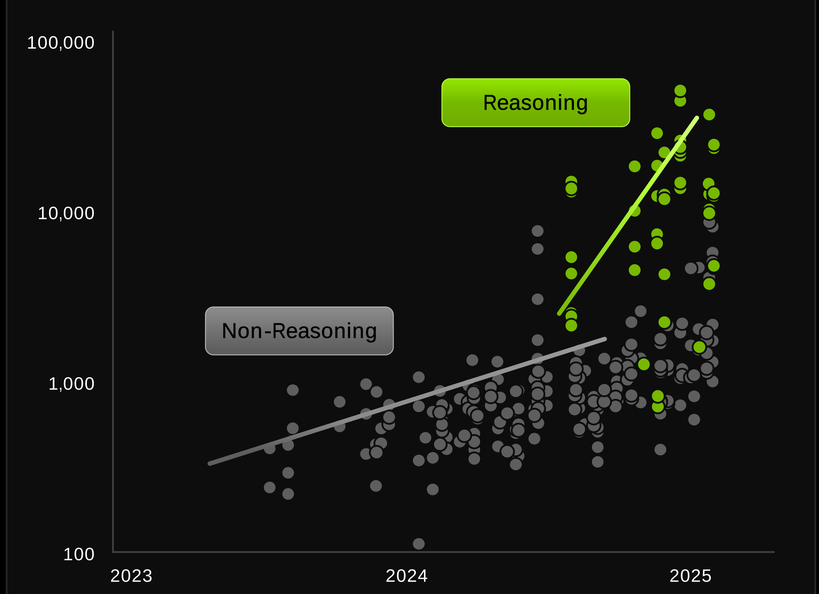

Rotational Velocity
Measured
200
100
Calculated
50,000
100,000
Distance From Center (Light Years)
Vera Rubin

Insane Demand for AI Computing

Model Size Growing
10X Parameters Per Year



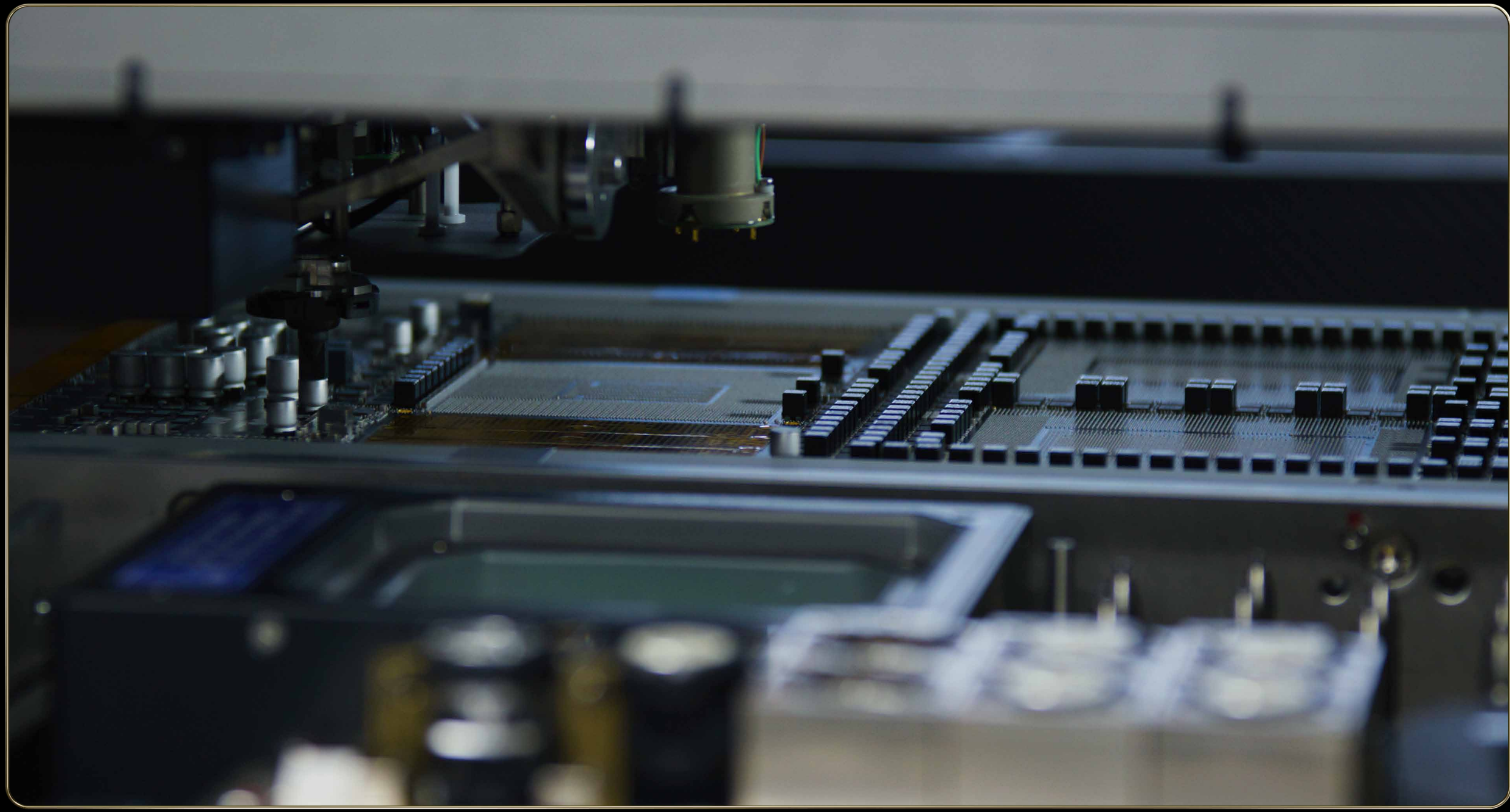
Test-Time Scaling “Thinking”
5X Tokens Per Year



Token Cost
10X Cheaper Per Year

Source : EPOCH AI

Source : Artificial Analysis



NVIDIA Vera CPU

88 NVIDIA Custom Olympus Cores

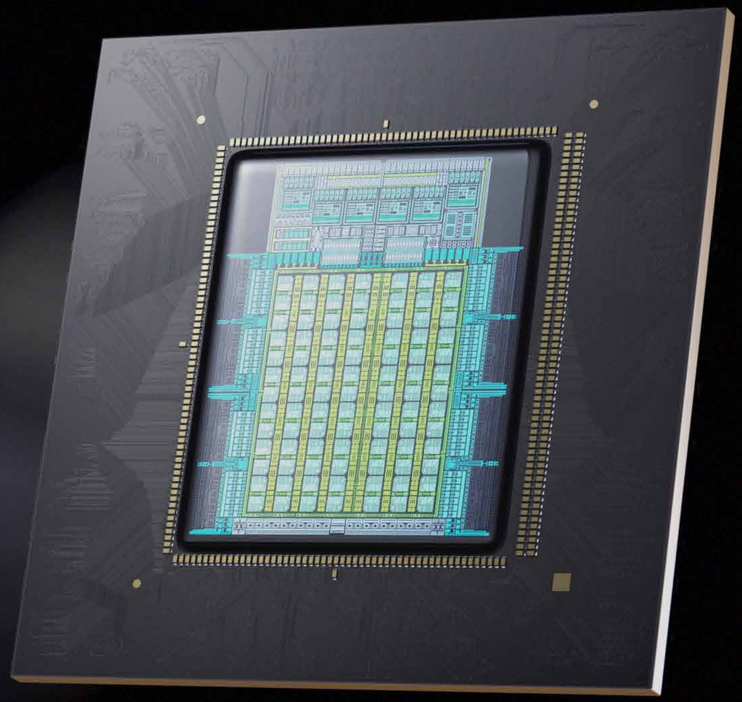
176 Threads with NVIDIA Spatial Multi-Threading

1.8 TB/s NVLink-C2C

1.5 TB System Memory (3X Grace)

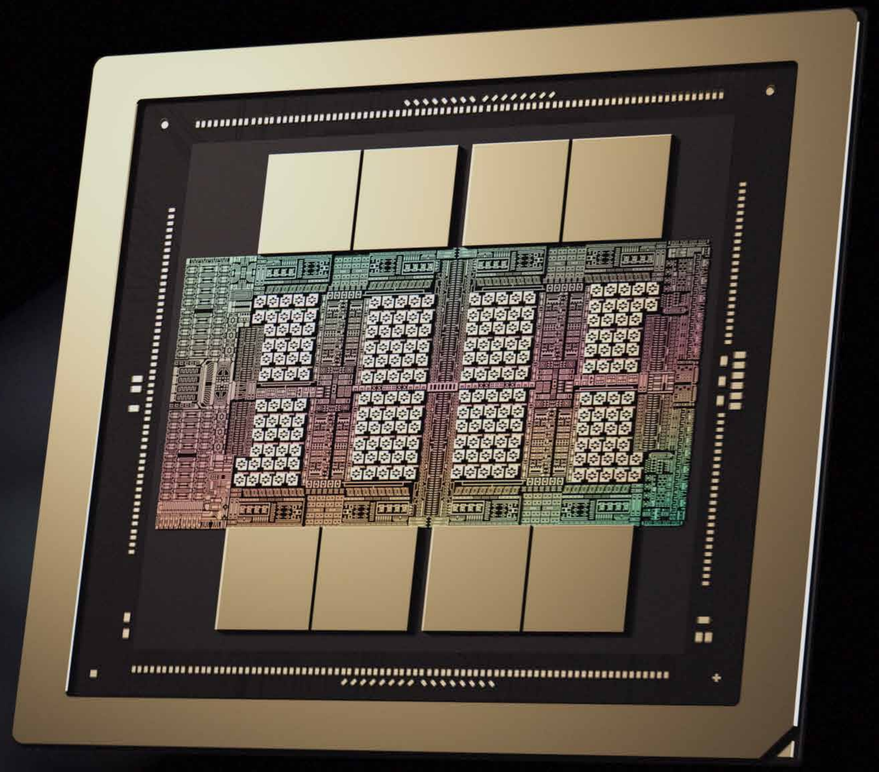
1.2 TB/s LPDDR5X

227 Billion Transistors



NVIDIA Rubin GPU

NVFP4 Inference	50 PFLOPS	5X Blackwell
NVFP4 Training	35 PFLOPS	3.5X
HBM4 Bandwidth	22 TB/s	2.8X
NVLink Bandwidth per GPU	3.6 TB/s	2X
Transistors	336 Billion	1.6X



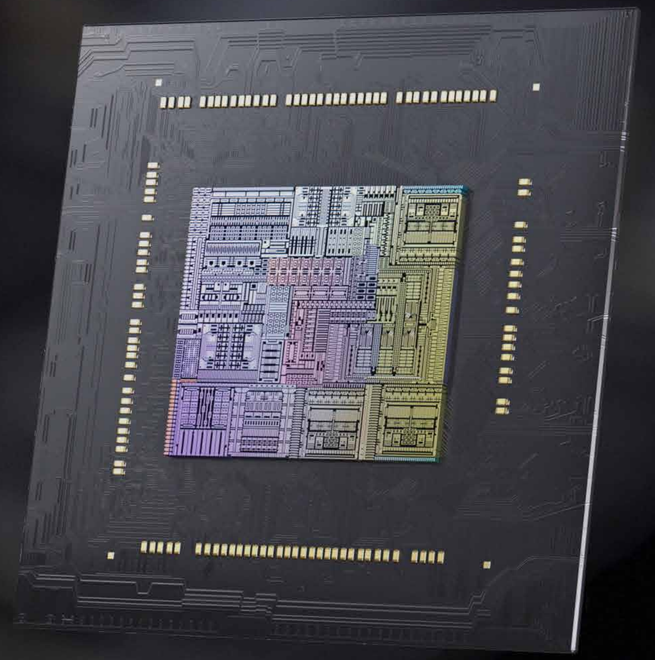
NVIDIA ConnectX-9
Spectrum-X SuperNIC

800 Gb/s Ethernet with 200G PAM4 SerDes

Programmable RDMA and Data Path Accelerator

State-of-the-Art Security
Line-Speed Encryption | Security Enclave and Attestation
CNSA and FIPS Certified

23 Billion Transistors



NVIDIA BlueField-4

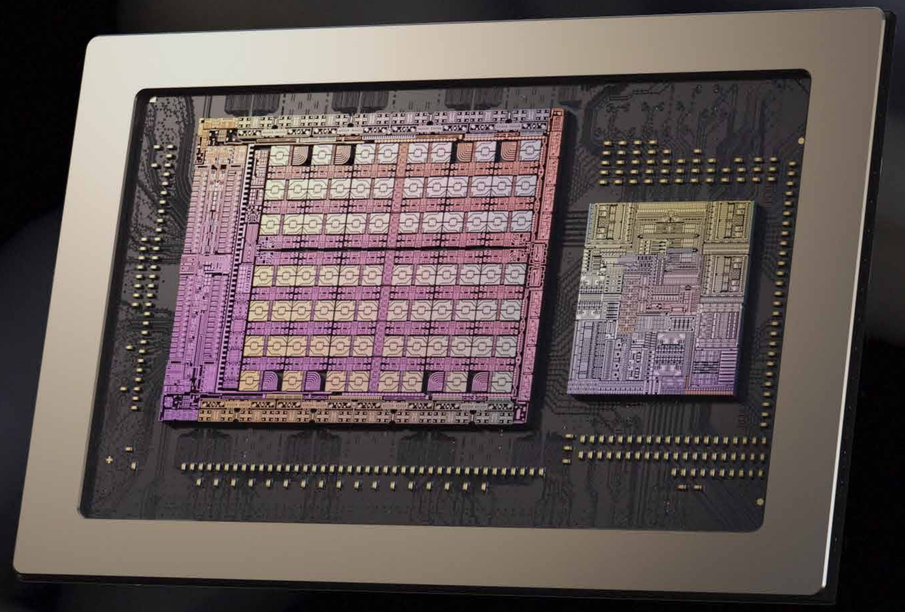
800G Gb/s DPU for SmartNIC and Storage Processor

64 Core Grace CPU with ConnectX-9

2X Networking
6X Compute
3X Memory BW

vs BF3

126 Billion Transistors



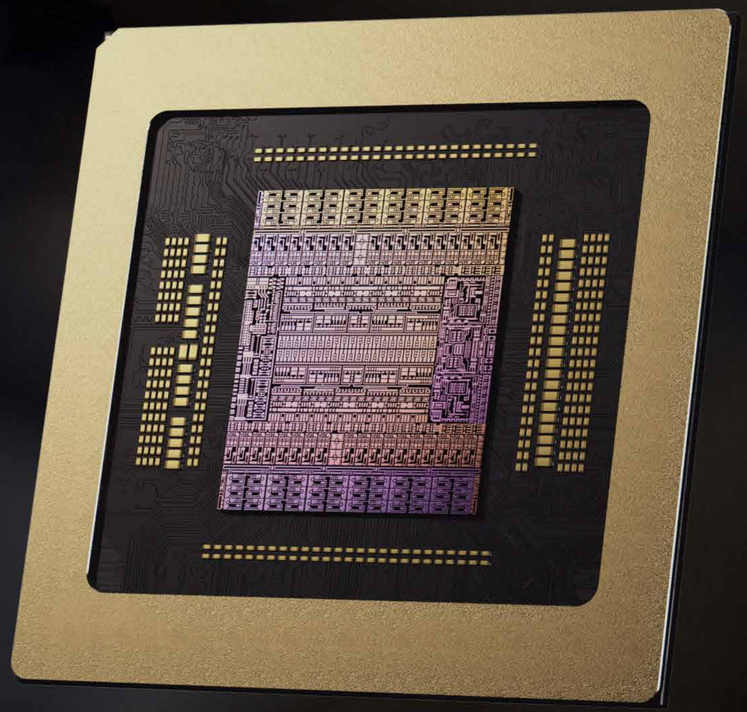
NVIDIA NVLink 6 Switch

Scale-Up Fabric with 3.6 TB/s Per-GPU All-to-All BW

400G SerDes

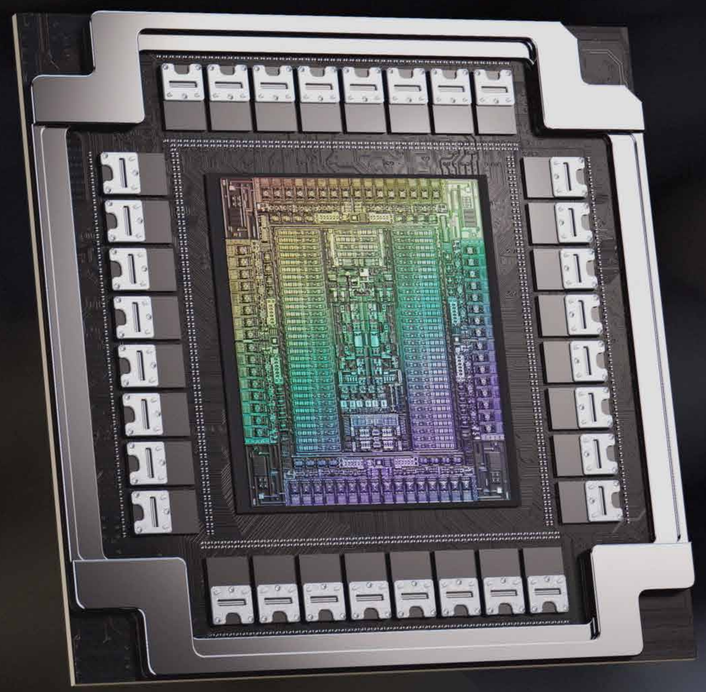
In-Network SHARP Collectives

108 Billion Transistors



NVIDIA Vera Rubin NVL72

NVFP4 Inference	3.6 EFLOPS	5X Blackwell
NVFP4 Training	2.5 EFLOPS	3.5X
LPDDR5X Capacity	54 TB	3X
HBM Capacity	20.7 TB	1.5X
HBM4 Bandwidth	1.6 PB/s	2.8X
Scale-Up Bandwidth	260 TB/s	2X
Transistors	220 Trillion	1.7X



NVIDIA Spectrum-X Ethernet Co-Packaged Optics

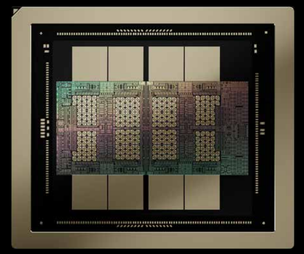
102.4 Tb/s Scale-Out Switch Infrastructure

Co-Packaged 200G Silicon Photonics

128 Ports of 800 Gb/s

512 Ports of 200 Gb/s

352 Billion Transistors



HBM

Memory

Rack SSD

Network SSD

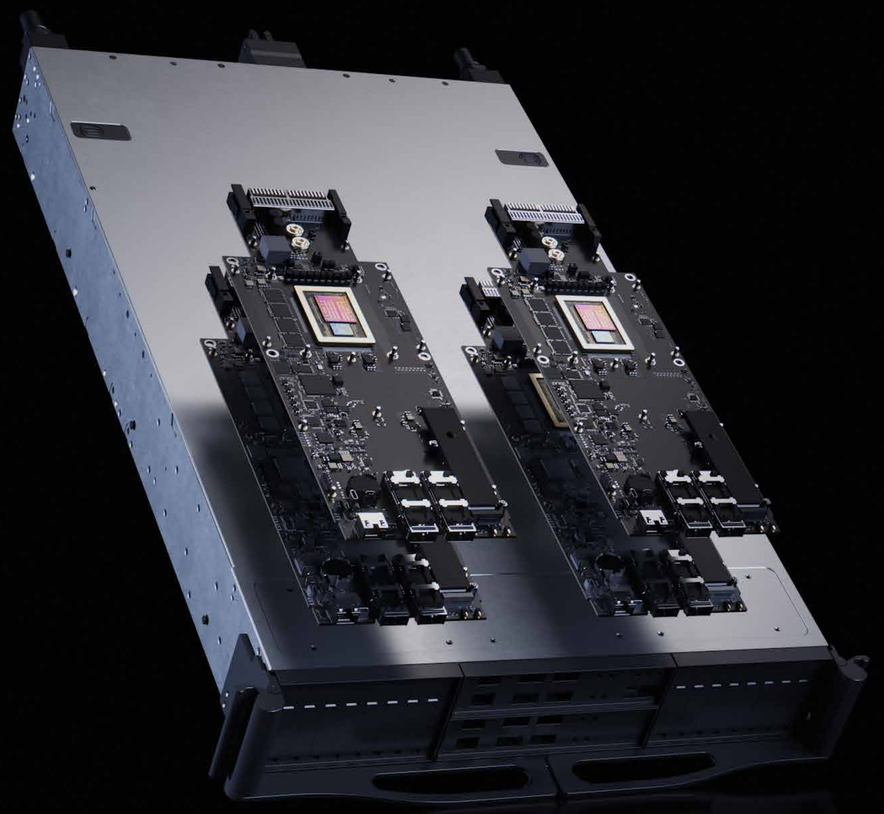
Context is the New Bottleneck – Storage Must be Rearchitected

Model Size Growing

Context Length Growing

Multi-Turn Conversations - Context Accumulates

More Concurrent Users and Sessions



NVIDIA Context Memory Storage Platform

A New POD-Level Context Memory Platform

BlueField-4 Data and Storage Processor

Spectrum-X Ethernet RDMA Infrastructure

5X Higher Tokens per Second

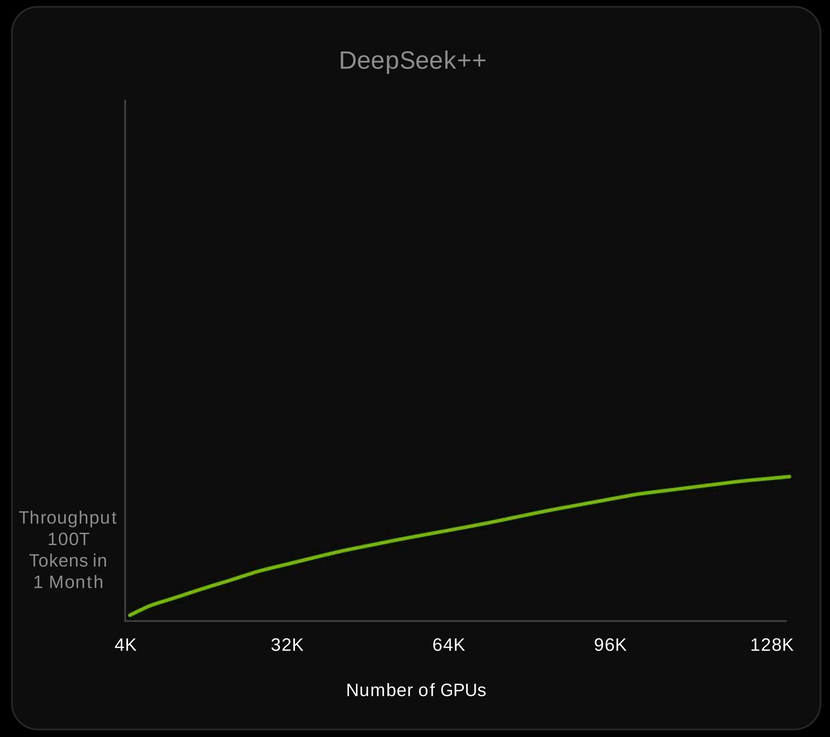
5X Higher Power Efficiency

AIC CLOUDIAN ddn DELL Technologies HPE Hitachi Vantara IBM NUTANIX PURESTORAGE SUPERMICRO VAST WEKA

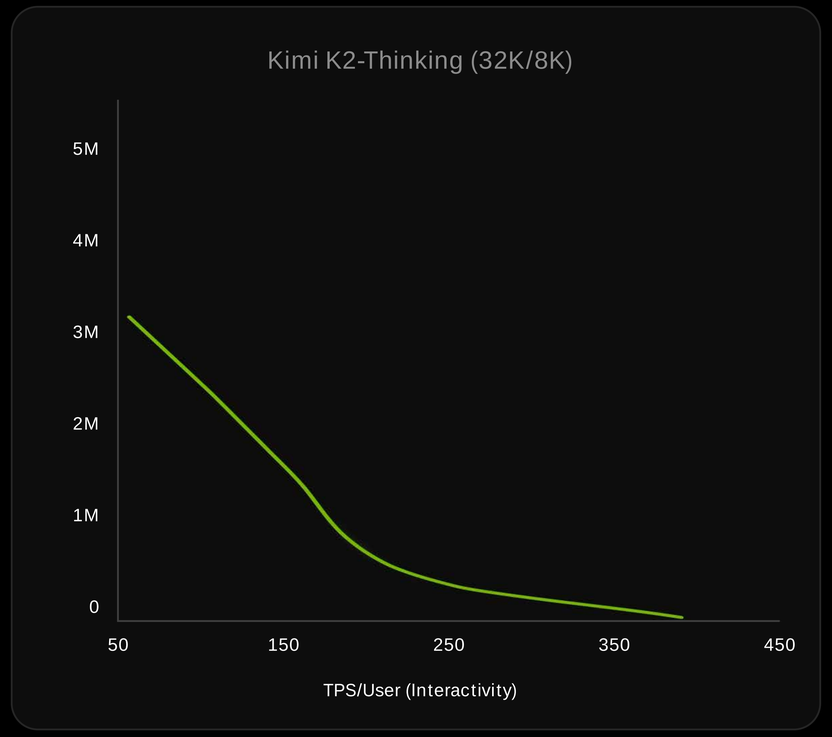
Vera Rubin
Six New Chips – One Giant Leap to the Next Frontier

Six New Chips – One Giant Leap to the Next Frontier

Time to Train
1/4th Fewer GPUs



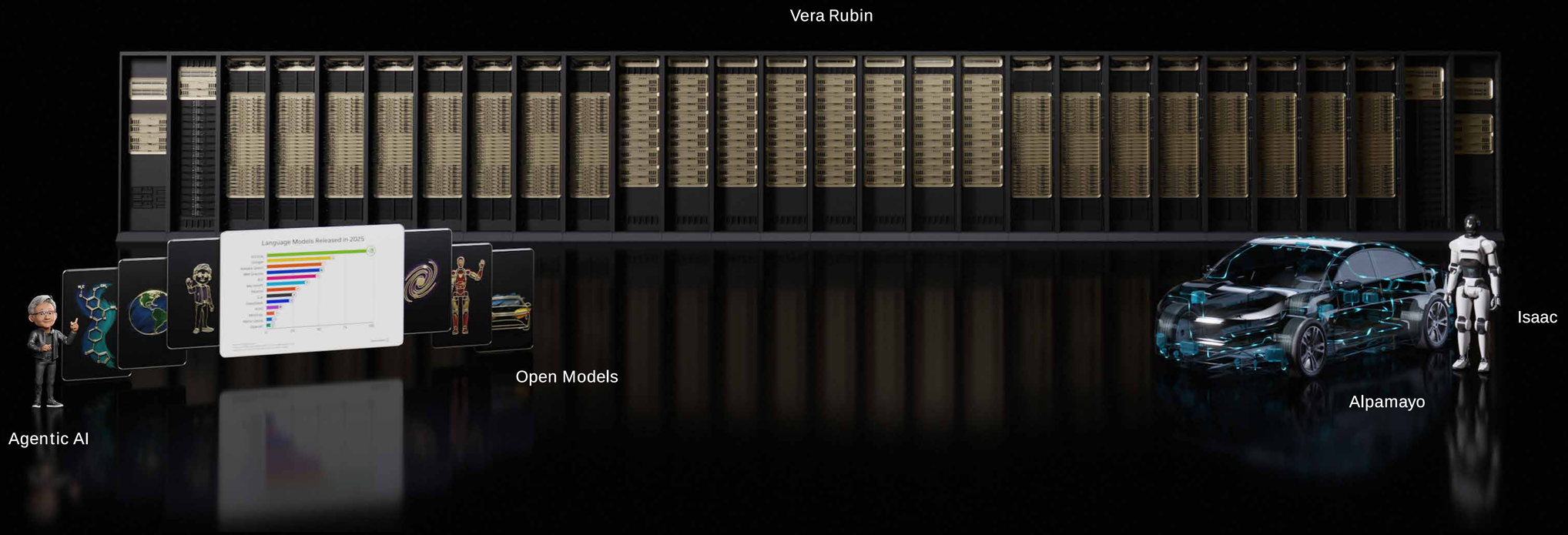
Factory Throughput
Up to 10X More Tokens



Token Cost
1/10th Lower Cost

Blackwell NVL72 Rubin NVL72

NVIDIA
One Platform for Every AI




NVIDIA