AI Agent Infrastructure & Post Coding-Agent Cybersecurity

Robert Blumofe, CTO

April 17, 2026

AI Agent Infrastructure

Frontier Model Myth

Myth

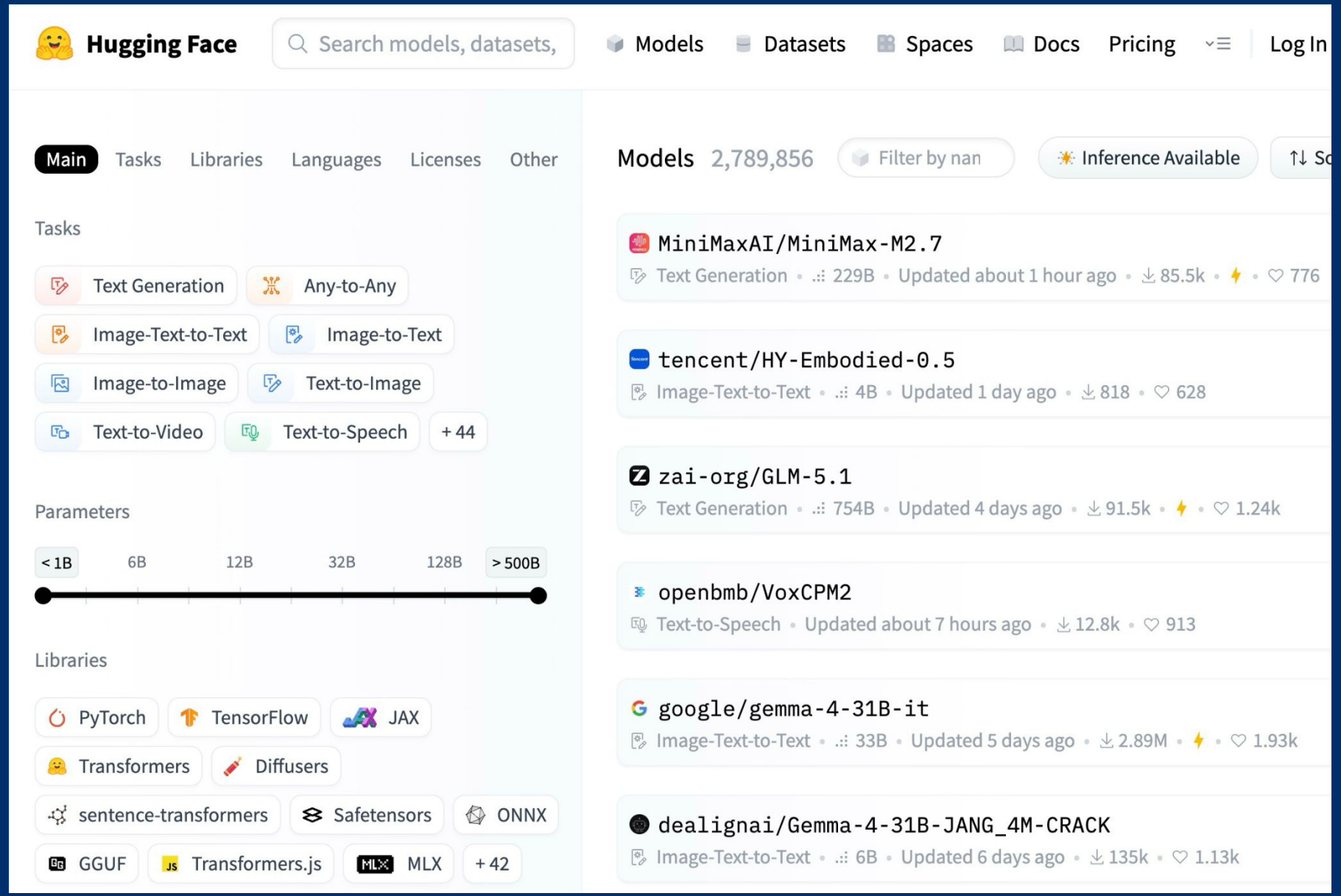
The most capable and useful AI models are the frontier models from a small handful of AI companies such as Anthropic and OpenAI.

Reality

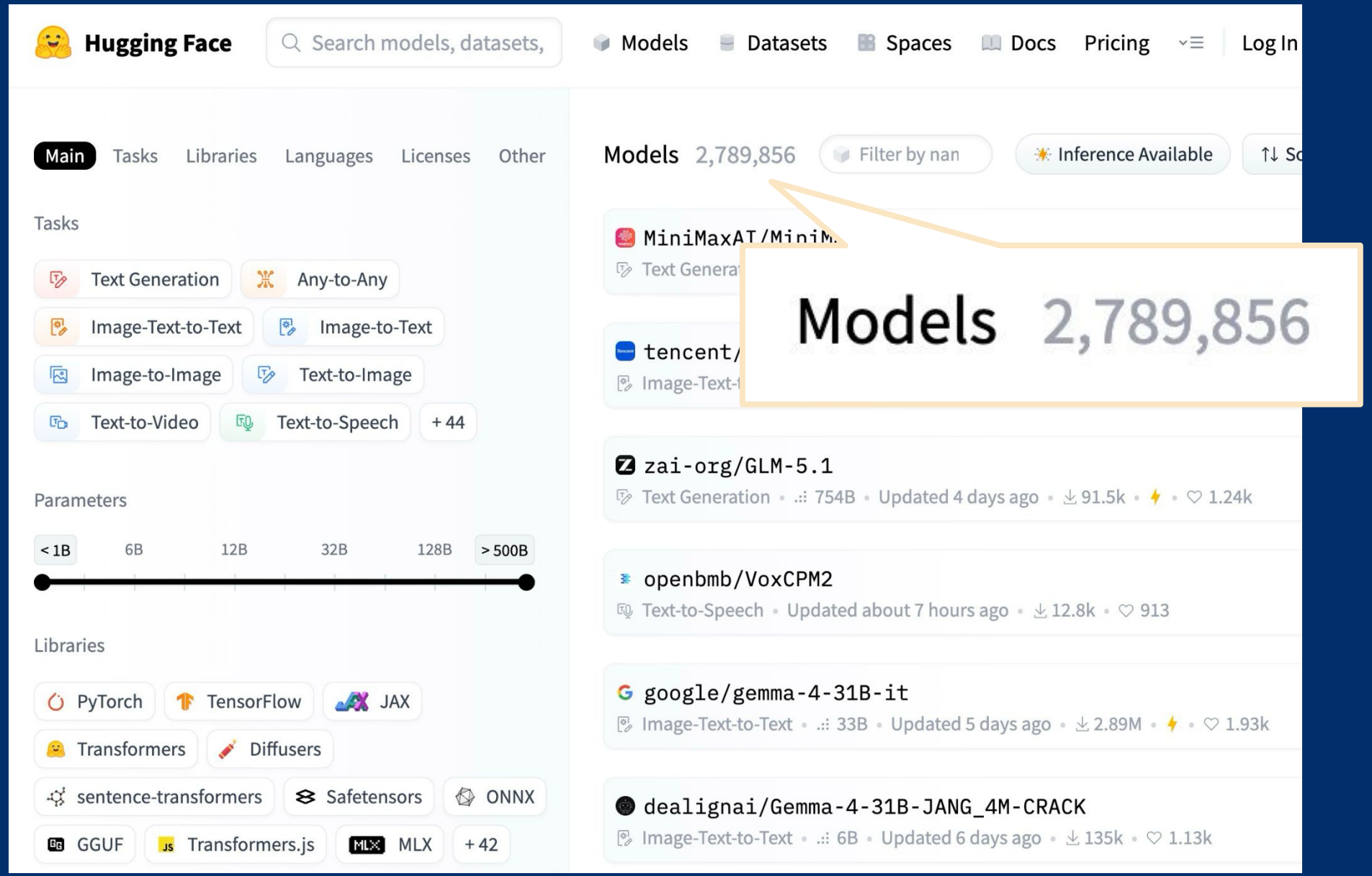
There is a myriad of AI models specialized for specific use cases.

- Specialized models outperform frontier models on their specific use cases.
- Specialized models are orders of magnitude smaller than frontier models and run at a small fraction of the cost.
- These models can be further specialized by the enterprise.
- These models can be downloaded and run on any cloud.

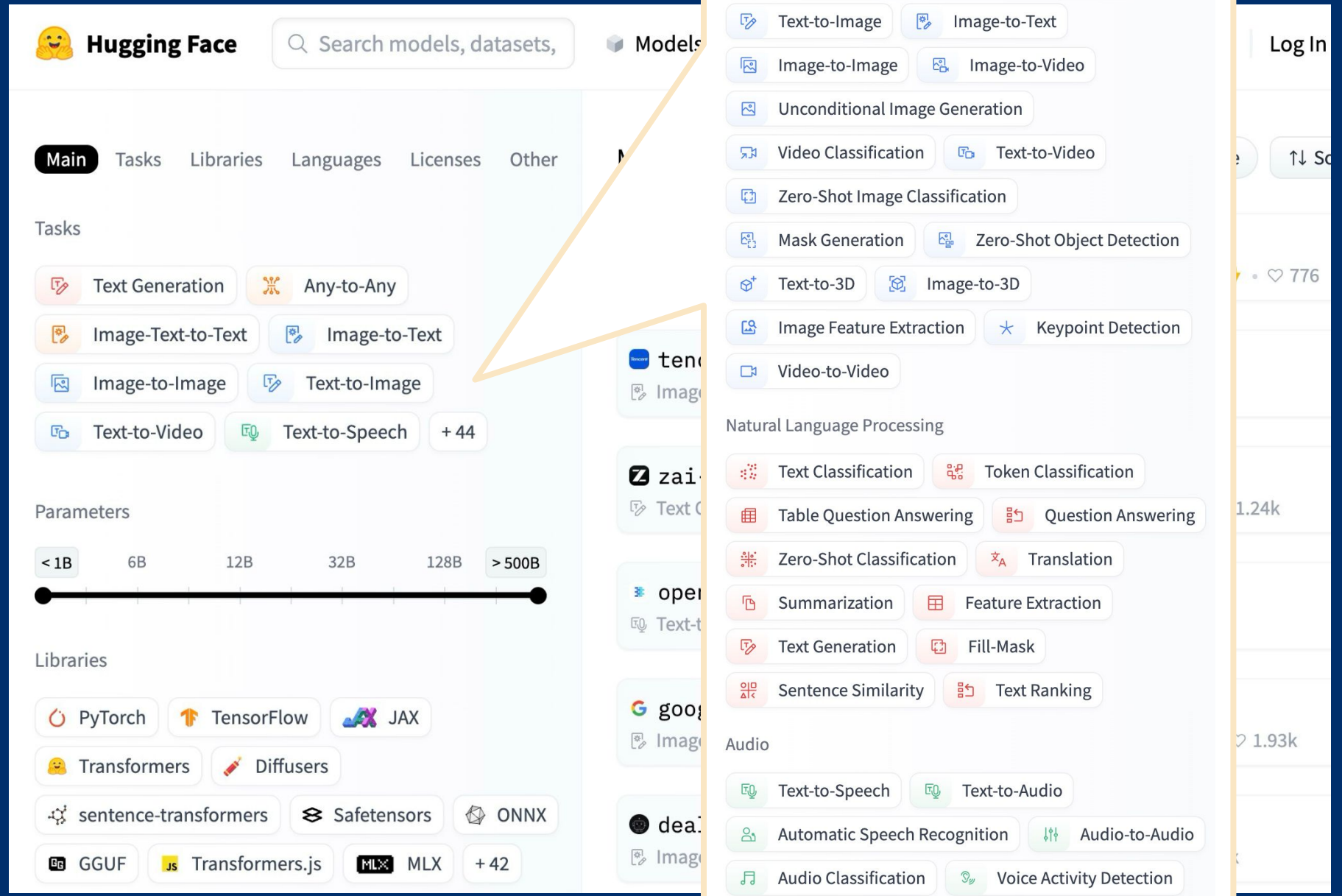
Hugging Face Models



Hugging Face Models



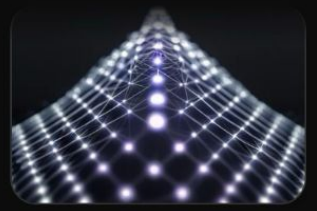
Hugging Face Models



NVIDIA Models

NVIDIA Explore Models Blueprints GPUs Docs Search ? Login

Filter by text | Sort By: Most Recent | Publisher | Use Case | NIM Type



deepseek-ai
deepseek-v3.1
DeepSeek V3.1 Instruct is a hybrid AI model with fast...
reasoning chat +2

nvidia
nvidia-nemotron-nano...
High-efficiency LLM with hybrid Transformer-Mamba design,...
mamba +8

stabilityai
stable-diffusion-3.5-la...
Stable Diffusion 3.5 is a popular text-to-image generation model
image generation +2

black-forest-labs
FLUX.1-Kontext-dev
FLUX.1 Kontext is a multimodal model that enables in-context...
image generation +3

nvidia
cosmos-reason1-7b
Reasoning vision language model (VLM) for physical AI an...
video understanding +9

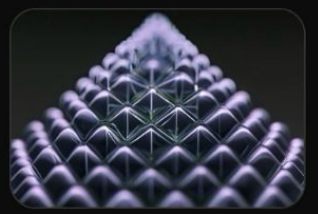
nvidia
nemoretriever-ocr-v1
Powerful OCR model for fast, accurate real-world image text...
+6

openai
gpt-oss-20b
Smaller Mixture of Experts (MoE) text-only LLM for efficie...
text-to-text chat +3

openai
gpt-oss-120b
Mixture o... reasoning LLM (text-only)...
gpt-oss-120b
text-to-text chat +3

nvidia
parakeet-tdt-0.6b-v2
Accurate and optimized English transcriptions with punctuatio...
asr english +4

nvidia
llama-3.3-nemotron-s...
High efficiency model with leading accuracy for reasoning,...
chat math +4



opengpt-x
teuken-7b-instruct-co...
Multilingual 7B LLM, instruction-tuned on all 24 EU languages f...
sovereign ai +5

sarvamai
sarvam-m
Multilingual, hybrid-reasoning model optimized for Indian...
coding +7

Agent Architecture Myth

Myth

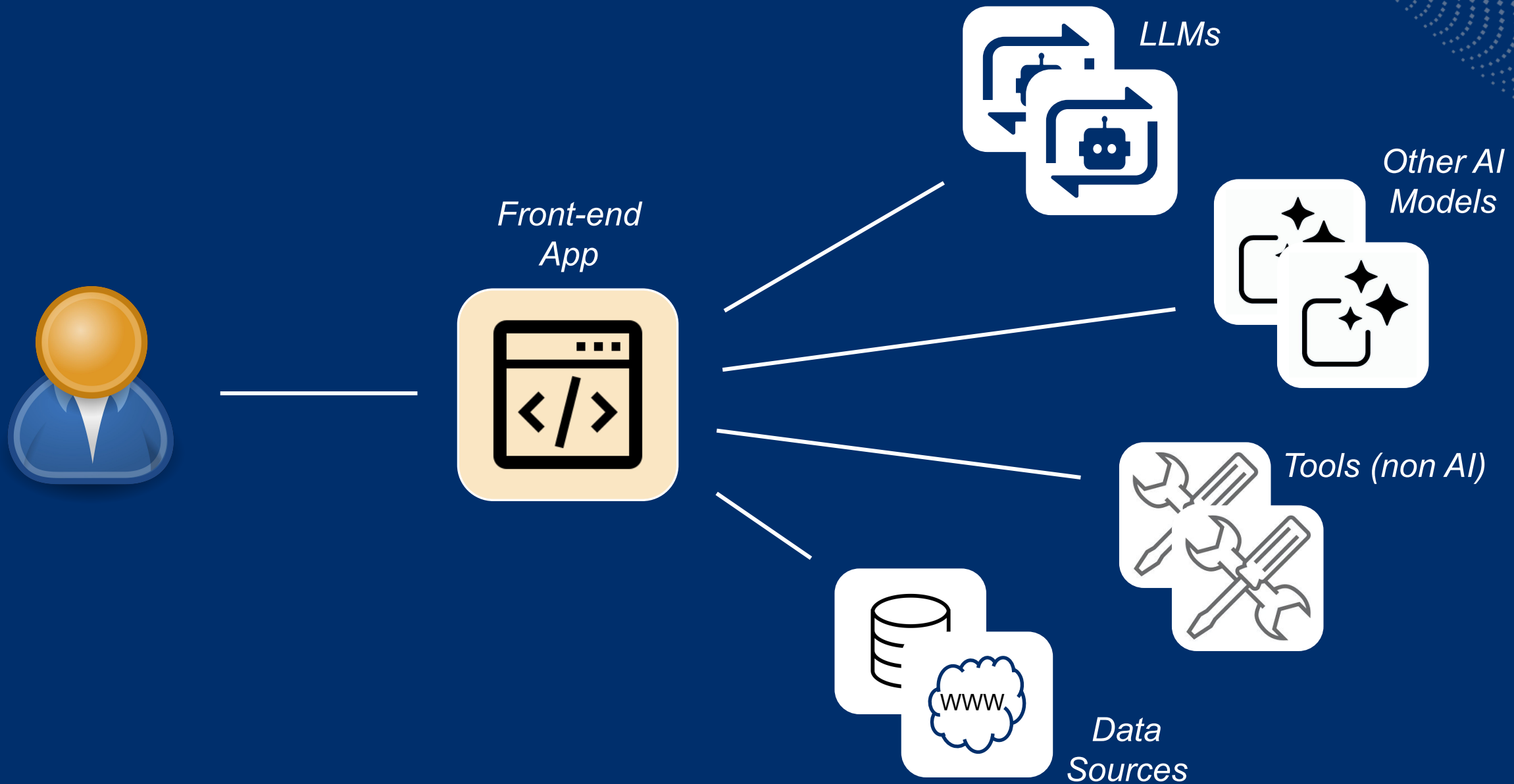
An agent is a leading-edge LLM with “reasoning” capabilities.

Reality

An agent is an application with many components, only one of which is an LLM.

- Agents use tools – local or 3rd-party – for things like math, code execution, path calculation, scheduling, and communication.
- Agents use persistent storage for memory and context management.
- Agents use vector stores for access to proprietary information.
- Agents use Web retrieval and APIs to access public information and services.

Agent Architecture



What is OpenClaw?

The architecture on the previous slide, prebuilt and downloadable.

Agent Infrastructure Myth

Myth

Agents will run on centralized, dense GPU clusters.

Reality

Agents need GPUs and CPUs, storage, connectivity, and CDN.

- The AI models themselves run on GPUs, but the tools require CPU.
- Storage is needed for memory, context management, and vector databases.
- Connectivity is needed to communicate with users, invoke 3rd-party tools, access remote storage, and retrieve Web content.
- CDN and distributed infrastructure is needed for low-latency, high-bandwidth communication between all of these components.

NEW!

Inference Cloud

A hardened, globally distributed cloud built for the AI Agent era, bringing GPU and CPU powered compute closer to users, data, services, and other agents.

Fast | Scalable | Affordable

NVIDIA

RTX 6000 Pro Blackwell

Post Coding-Agent Cybersecurity

Vulnerabilities Myth

Myth

New AI capabilities to find and fix vulnerabilities mean that software will be (largely) vulnerability free.

Reality

Software systems will always have vulnerabilities.

- Some software is difficult or impossible to fix: legacy software, IoT software, unmaintained software.
- New AI models will find new vulnerabilities.
- Fixing vulnerabilities is more than just finding and authoring fixes, so even with AI automation, it takes time: there's testing and rollout (remember Cloudstrike?).
- Many vulnerabilities are not related to coding flaws.

The “AI Vulnerability Storm”: Building a “Mythos-ready” Security Program

Expedited Strategy Briefing

By the CSA CISO Community, SANS, [un]prompted,
the OWASP Gen AI Security Project, and the wider community.

From the Report:

What do we believe will happen next?

- The capabilities seen in Mythos will quickly become more widely available, dramatically increasing the number and frequency of complex, novel attacks organizations will face.

Web Application Firewall (WAF) Myth

Myth

New AI capabilities to find and fix vulnerabilities mean that the need for a WAF is largely mitigated.

Reality

A WAF is more important than ever.

- The attackers also have tools to find vulnerabilities, and those tools can also author exploits.
- With zero-day exploits, the WAF is the only defense.
- The time available to fix vulnerabilities is reduced.
- Without a WAF, attackers can find vulnerabilities even in proprietary software (with no access to source code).

From the Report:

What to do now to deal with the current risk spike?

- Focus on the basics and harden your environment further. Segmentation, egress filtering, multifactor authentication, and defense-in-depth/breadth all increase the difficulty for attackers.

Commercial Cybersecurity Myth

Myth

New AI capabilities to write software mean that the need for commercial cybersecurity systems is largely mitigated.

Reality

Cybersecurity systems are more than just software.

- Need access to large amounts of data to train models to create effective detection and mitigation rules.
- Need domain knowledge and experience to create and fine-tune these models.
- Need distributed infrastructure to run the software at scale and without bottlenecks.
- Need domain knowledge and experience to make it all work reliably.

How Do You Replicate This?

37T bot attacks

2PB of DDoS attack traffic

310B app & API attacks

23T L7 DDoS

1000
TBPS Capacity

4,350+
Edge PoPs

130+
Countries

700
Cities

5+
24/7 SOCC

THE AKAMAI DATA UNIVERSE

1,700+ TB
of data analyzed daily

Depth
Customer-specific traffic patterns

Breadth
Cross-customer threat intelligence

Thank You