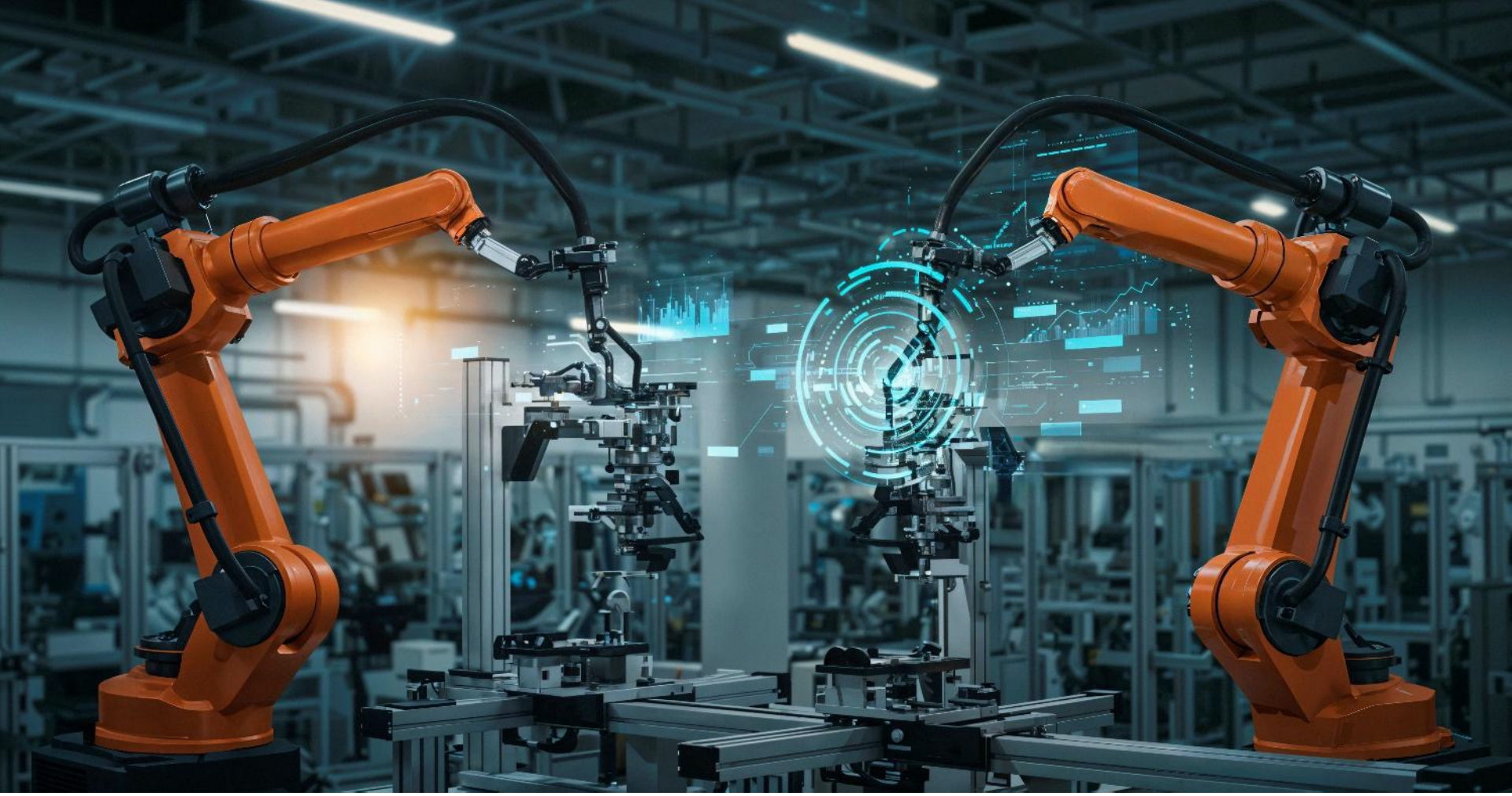


AI Factories: The New Operating Model for Intelligence

2026

ALLIED ADVISERS

Table of Contents

ALLIED ADVISERS

Executive Summary

Executive Summary

ALLIED ADVISERS

As inference emerges as the dominant AI workload, data centers are evolving into AI factories, reshaping infrastructure, economics, and enterprise strategy for the next phase of AI at scale

From Data Centers to AI Factories: Rise of AI Factories

AI factories evolved from general-purpose data centers into purpose-built systems for large-scale AI production, converting data, compute, and energy into AI outputs and decision-making

What Is Driving This Shift

Driven by the need for:

High Compute Power	Rise of Reasoning Models	Need for Full Stack Control	Rising Demand for Real-time Inference

These demands exceed traditional data center capabilities, with inference as the primary driver

Inference as the Primary Driver

- Inference share in AI Infra* spend is rising from **33% (2023)** to **55% (2026)**, becoming a key AI workload**
- It pushes AI factories to optimize storage, network, and resource allocation

Inference Workloads Are Shaping the AI Economics

- Inference shifts AI economics from one-time training costs to continuous, usage-based expenses, making **cost per token** a key business metric
- This pressures AI factories to lower token costs, as profitability depends on efficient intelligence production

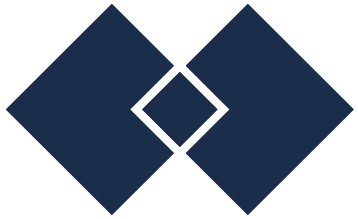
Implications of AI Factories and Inference Across:

Infrastructure:
Driving demand for specialized chips, custom silicon, low-latency design, and GPU-aware orchestration

Enterprises:
Pushing providers to expand offerings, prompting users to invest in custom chips, and creating space for inference-first players to grow

What's Next

- **70%** of large enterprise leaders plan to scale AI factories by 2028, nearly doubling the current rate
- AI factories will become the operating system of every enterprise, where intelligence is mass-produced, continuously deployed, and measured through decisions, not reports



Introduction to AI Factories

Transformative Journey: From Data Centers to AI Factories

ALLIED ADVISERS

AI factories are redefining data centers from general-purpose compute environments into purpose-built systems for developing, deploying, and continuously improving AI at scale

Evolution of Data Centers Toward the AI Factory Model

Pre-2020: Traditional Data Centers

- Built for general-purpose computing on CPU-centric architectures
- As AI emerged, GPUs* were added but remained an add-on to existing stack

Not built for AI at scale, which requires far greater parallel compute, power density, and data movement

2020-22: AI-Ready Data Centers

- Rise of large AI models pushed data centers toward:
  - **Denser GPU/TPU clusters**
  - **Faster networking**
  - **Advanced cooling**

From 2022: Emergence of AI Factory Model

2022 — NVIDIA introduced "**AI factories**," redefining data centers as intelligence-producing systems

2023-24 — AI infrastructure players began scaling AI factories, shifting from AI-ready stacks to unified systems for training and inference

2025 onwards — AI factories started scaling globally as the standard model for next-generation AI infrastructure

> *"Data centers are no longer just compute hubs, but "AI factories" – per the new terminology."*
>
> ***- Jensen Huang, CEO,*** NVIDIA.[1]

Key Takeaway

AI factories representing the next evolution of data centers, extending rather than replacing them through purpose-built, full-stack AI systems

1. RCR Wireless

*Graphics Processing Units (handling many tasks at once)

AI Factories: Concept and How They Operate

ALLIED ADVISERS

AI factories are integrated operating models that enable continuous AI production, supporting real-time workloads, faster time-to-value, and optimized operations

- An **AI factory** is a vertically integrated operating model that transforms data centers from storage and compute facilities into end-to-end AI production systems
- It turns data, compute, memory, and energy into AI models, inference services, and automated decisions at scale

Core Characteristics of AI Factories

1 Inference-first Operating Model

Prioritizes real-time decision-making, shifting from training-centric operations to production-grade, always-on inference systems

2 Integrated Stack

Combines hardware, data pipelines, models, and infrastructure into a unified platform for AI workloads

3 Hardware and Software Tuning

Optimizes GPU clusters, high-throughput storage, and scheduling frameworks for real-time inference performance

Why AI Factories Matter

Provides infrastructure to process large-scale AI workloads, accelerate time-to-value with automated pipelines, and optimize operations with AI

AI Factory: Production Lifecycle

Input
(Data, Compute, Tools, Base Models)

→ **Build & Customize**
(Design, train, fine-tune, test)

→ **Deploy & Run**
(Run in production environments)

→ **Collect Production Feedback**
(Logs, user signals, outcomes, metrics)

→ **Improve Continuously**
(Retrain, validate, and redeploy)

Key Drivers Accelerating the Shift to AI Factories

ALLIED ADVISERS

The shift to AI factories is gaining momentum as demand for inference, reasoning workloads, and compute power exceeds the capabilities of traditional data centers

Key Growth Drivers	Supporting Facts
Massive Compute Demands — Modern AI models demand far **higher rack power density** and cooling than traditional data centers can handle	Recent AI factories run at **120–142 kW**[1] per rack (power density), vs. 5–10 kW (up to ~30 kW for high-density) in traditional data centers
Rise of Reasoning Models — Reasoning models solve problems step by step, requiring **more compute** and **memory** than standard AI, making the AI factory stack crucial for efficiency	NVIDIA's GB300 NVL72^ can deliver up to **50x**[2] **higher output for reasoning workloads in AI factories** NVIDIA. **50x**
Need for Full-Stack Control — AI systems **require sync** across hardware, software, and models, pushing data centers to operate as integrated AI factories	NVIDIA's **five-layer AI framework**[2] (energy, chips, infrastructure, models, applications), built as an architecture for AI factories NVIDIA.
Inference* Surge — Rising inference demand from real-time AI applications requires infrastructure that delivers **continuous and efficient inference** at scale	By 2030, inference will drive **~30–40%**[3] of **total data center demand**, becoming the dominant AI workload, boosting the shift to AI factories McKinsey & Company **~30–40%**

Key Takeaway

While multiple factors are driving this transition, inference is emerging as the primary driver, reshaping AI infrastructure design and scale requirements

1. LinkedIn
2. Nvidia
3. Mckinsey

^ A rack-scale AI factory system for reasoning and inference

*Inference is the stage where trained AI models generate live responses, predictions, or decisions, making it the core workload behind real-time AI applications

Inference: Primary Driver of AI Factories

Inference: The Core Driver Reshaping AI Factories

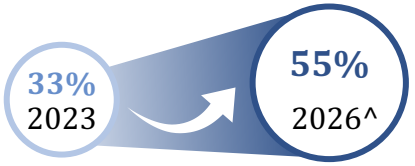
ALLIED ADVISERS

As inference becomes the dominant AI workload, AI factories are being redesigned to meet the growing need for real-time processing and its business impact

> *"Inference inflection has arrived and the world is racing to build AI factories."*
>
> 
>
> *–Jensen Huang, CEO, Nvidia[1]*

- As AI factories enable real-time decision-making, inference demand is rising, making it central to workload efficiency in AI factories
- This shift is resulting in changing infrastructure requirements for AI factories

Inference Share in AI Infrastructure Spending[2]



Driving next-gen AI factories

How Inference Is Reshaping AI Factory Infrastructure Needs

	What Inference Demands	Impact on AI Factory
Design Priorities	Inference requires systems handling low latency*, high throughput**, and variable demand	Forcing AI factories to rethink **storage, compute,** and **network design**
Cost Pressures Driving Resource Optimization	Inference runs continuously in production, so its lifecycle costs often surpass training costs • Accounts for **80–90%**[3] of the lifetime cost of production AI system	Pushing AI factories to optimize **compute, workload orchestration, and resource allocation**
New Optimization Priorities for Inference-Heavy Workloads	Unlike training, inference runs constantly in AI factories and must be optimized for speed, efficiency, and cost per request	Shifting priorities toward **specialized chips, memory, and edge deployment**

Takeaway **Rising inference workloads are transforming AI factory design, cost structures, and operating models, while also reshaping AI economics within AI factories**

1. Reuters
2. Byteiota
3. Introl

^ 2026YTD
*Latency is the time delay between a request and the response
**Higher throughput means the ability to process more requests or tokens in a given amount of time.

New Economics of AI

The Shift to Inference Is Redefining AI Economics

ALLIED ADVISERS

Inference-led workloads are redefining AI factory economics, shifting from traditional cost models to cost per token and revenue per token

AI Factories Are Moving to a New Economic Model

- AI factories are shifting from upfront, one-time training costs to continuous inference expenses that scale with usage

Inference drives **60–90%**[1] of enterprise AI compute spend, reshaping financial models

- As tokens become the core unit of value in AI systems, **cost per token*** is emerging as a key metric, redefining how performance and costs are measured in AI factories

> *"AI is no longer about models or chips, but about monetizing inference, where tokens become the core unit of value, and data centers evolve into revenue-generating factories."*
>
> ***- Jensen Huang, CEO, Nvidia***[2]

To scale profitably, AI factories need to lower the cost per token, as it directly affects performance and financial outcomes

How AI Factories Are Lowering Cost per Token

This is being enabled by advances in:

Better Hardware

New systems like NVIDIA Blackwell produce up to **35x**[3] more AI output at a lower cost than older systems

~30%[4] **annual** ⇩ in AI inference hardware costs

Smarter Serving Systems

AI inference platforms (baseten, deepinfra, Fireworks AI, and together.ai) are cutting token costs by reducing compute and infrastructure overhead

More Efficient Models

Open-source models (LLaMA, Mistral AI_, and Qwen, deepseek) are closing the performance gap with proprietary models

- Reducing computation and hardware needed per token

1. Introl
2. LinkedIn
3. Nvidia
4. Stanford AI Index 2025

*A token is the smallest unit of data an AI model processes, used to measure output, cost, and performance in an AI factory

Why Power Efficiency Matters in AI Factories

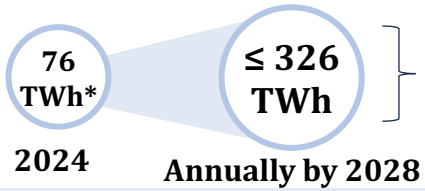
ALLIED ADVISERS

As cost per token becomes a central business metric, AI factories are increasingly focused on optimizing performance per watt to address the growing power demands of inference workloads

Power Constraints in AI Factories

- Rising token demand at scale requires more efficient energy use to support continuous inference
- As these workloads grow, power is becoming a key constraint on capacity deployed in AI factories

AI Inference Power Consumption[1]



Highlighting the growing need for efficient power use in AI factories

Performance per Watt**: The Critical Metric for AI Factory Infrastructure Efficiency

- Driven by the need for efficient power use, **performance per watt** is emerging as the key metric for how efficiently power is converted into valuable AI output
- Optimizing it is essential to align AI factory performance with economic outcomes

> *In the AI era, power is the key constraint, and every AI factory operates within a hard limit, making **performance per watt**—the rate at which power is converted into revenue generating intelligence — the defining metric for modern AI infrastructure.*
>
> ***- Jensen Huang, CEO, Nvidia[2]***

From Power Efficiency to AI Factory Economics

Higher performance per watt → **More tokens per watt** → **Lower energy cost per token** → **Higher economic output per AI factory**

Takeaway

- **As AI economics shift toward cost per token, power is emerging as a key infrastructure constraint for AI factories, making performance per watt critical to economic outcomes**
- **Along with rising inference demand, this is reshaping broader infrastructure and enterprise decisions**

1. Tensormesh
2. Nvidia

*Terawatt Hour (TWh)
**Performance per watt refers to how much AI output a system can generate for each unit of energy consumed.

Rise of AI Factories and Inference: Key Implications

Implications for AI Infrastructure

ALLIED ADVISERS

As AI factories scale to meet inference demand, the effects are clear: infrastructure needs are shifting toward low-latency design, GPU-aware orchestration, and custom silicon, moving beyond general-purpose compute

Rising Focus on GPU-Aware Infrastructure

Infrastructure focus is shifting from adding raw compute to maximizing GPU utilization, with scaling AI factories

AI GPU Orchestration Platforms (2025–2030)[1]

Expected to grow at a **25.9% CAGR,** adding **$6.6B** by 2030

NVIDIA's Multi-Instance GPU (MIG)[2] splits a single GPU into multiple instances, enabling more efficient utilization across concurrent AI workloads

NVIDIA

Expansion of the AI Hardware Supplier Base

The rise of AI factories is expanding the AI hardware ecosystem*, encouraging new players to enter the market with specialized chips

Startups like **Graphcore** and **Cerebras** are entering the stack with specialized AI chips, challenging established vendors

GRAPHCORE cerebras

Shift Toward Custom AI Chips

Rising costs and delays in 3rd-party chips are pushing hyperscalers to develop custom AI chips, making silicon optimization key to AI infrastructure strategy

- Google: Built its 7th-generation custom AI chips
- amazon Project Rainier[3]: Deployed **500K+** Trainium 2 chips across US data centers, with plan to **double** that count
- BROADCOM and OpenAI: Co-developing a **$10B**[4] custom AI accelerator

Rise of AI factories and inference is driving fundamental changes in infrastructure needs, while extending its impact beyond data centers to how enterprises operate, compete, and invest in AI

1. Technavio
2. Nvidia
3. AIcerts
4. Reuters

*AI Hardware Ecosystem (vendors, startups, custom chips, and AI infrastructure)

Implications for Enterprises (1/2)

ALLIED ADVISERS

Enterprises across infrastructure providers and platform users are recalibrating AI strategies around token-based economics, investing in tools and infrastructure to scale AI production efficiently

Key Implications for:

AI infrastructure Providers

Scaling to Meet Enterprise Demand

- Growing demand from cloud and enterprise customers for AI apps is driving vendors to expand AI factory platforms
- Expanding end-to-end solutions across data, models, and inference

DELL Technologies

Expanded AI Factory with **NVIDIA**[1], adding AI services and enterprise-scale infrastructure, adopted by 4K+ customers

cognizant

Launched AI Factory[2], powered by **Dell** and **NVIDIA**, to scale the AI lifecycle securely and efficiently

AI Platform Users

Rising Inference Spending

Rising inference compute demand is making AI workloads more costly and less predictable

OpenAI spent **$8.7B**[3] on inference via Azure in the first 3Q's of 2025, which is **~2x** the **$3.7B in 2024**

Token-Based Pricing for AI Workloads

Enterprises are adopting token-based pricing to manage compute variability and scale AI deployments efficiently

Rafay Systems[4] supports this shift with token-based access via its Token Factory, enabling scalable AI models and services

RAFAY

> *With NVIDIA, we are enabling AI factories to treat token production as a measurable business metric.*
>
> ***- Kunal Anand, CPO, F5 (Technology Company)***[5]

1. IT Pro
2. Cognizant
3. The Register
4. Rafay
5. F5

Implications for Enterprises (2/2)

ALLIED ADVISERS

Rising inference demand is further driving incumbents to strengthen control over the inference stack, while inference-first providers continue to capture share

AI infrastructure Providers

Strategic Partnerships to Control the Inference Stack

NVIDIA

- Made a **$20B**[1] big licensing deal (2026) with **Groq,** an inference-focused company, and hired key engineers to own the full-stack inference pipeline
- Partnered with **F5**[2] to improve inference-stack control, enhancing token throughput, GPU utilization, and latency

AI Platform Users

Building In-house Inference Capabilities

- aws, Google, and Meta are launching custom AI chips (**Amazon Inferentia/ Trainium, Google TPUs, Meta MTIA**)
  - These chips are engineered for large-scale AI workloads, including inference-heavy production

Emerging Players in Inference Infrastructure

Specialized inference platforms are emerging as a distinct layer in AI factory infrastructure, enabling faster, cost-efficient inference

Players like cerebras and Fireworks AI are already capitalizing, Fireworks AI, for e.g., grew traffic by **~100x**[3] in just 6 months

There's investment by the bigger companies to strengthen their inference portfolio through products and by buying engineering talent. But there are so many chip startups that are going to emerge, and they are going to be significant"
- Matt Kimball, VP and Principal Analyst of Data Center, Moor Insights & Strategy[4]

Enterprises are realigning as incumbents expand their factory platforms, users capitalize on inference-driven, token-based economics, and inference-first players gain share, raising a key question: what does the future of AI factories look like?

1. LinkedIn
2. F5
3. Webmaster
4. Data Center Knowledge (Futurum Group)

The Next Phase of AI Factories

AI Factories: Future Trends and Adoption

ALLIED ADVISERS

AI factory adoption is set to rise sharply, with inference expected to dominate AI operations and create opportunities for enterprises to innovate, invest, and redefine infrastructure

Adoption of AI Factories Will Rise

70%[1*] of large enterprises leaders plan to scale AI factory and edge AI deployments by 2028

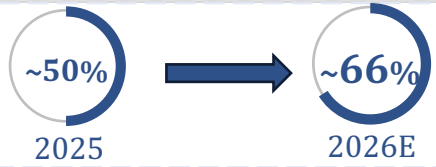
~2x the current adoption rate

This expansion will bring compute closer to data sources, reducing reliance on centralized cloud

Key Trends Shaping the Next Phase of AI Factories

Inference Continues to Dominate AI Factory Workloads

Inference is projected to consume **~66%**[2] of all AI compute by 2026, up from 50% in 2025



Rising AI Infrastructure Investment

To meet growing adoption, major tech and industrial firms are set to invest in AI factory stacks and its related components

Nvidia's next gen AI chips are being designed for large-scale AI and data center workloads

Projected to exceed **$1 trillion**[3] **in** revenue by 2027

Siemens[4] has committed to invest **$165M+** to expand its US manufacturing capacity, boosting production of electrical stack for AI factories and large-scale data centers

SIEMENS

Every company will have two factories... one for what they build, and one for the AI.

- Jensen Huang, CEO[5], NVIDIA

Looking ahead, centralized AI factories will mass-produce models and agents powering autonomous factories, smart cities, and enterprise operations. Winners will be those who operationalize this intelligence across their business, turning data into a continuous stream of decisions

1. Deloitte
2. Introl
3. Yahoo Finance
4. Siemens
5. LinkedIn

*Survey conducted by Deloitte; 515 US leaders across five industries from enterprises with more than $500M in annual revenue in December 2025

About Allied Advisers

Allied Advisers: Investment Banking for Technology Companies and Investors

Allied Advisers is a global technology-focused boutique advisory firm focused on investment banking for entrepreneurs and investors. The Silicon Valley-based firm, with a presence in Los Angeles, Israel, and India, serves entrepreneurs and investors of technology growth companies globally on strategic advisory including M&A and capital raises. Allied Advisers bankers have completed technology transactions globally for clients with Fortune 50 buyers and top tier Private Equity firms.