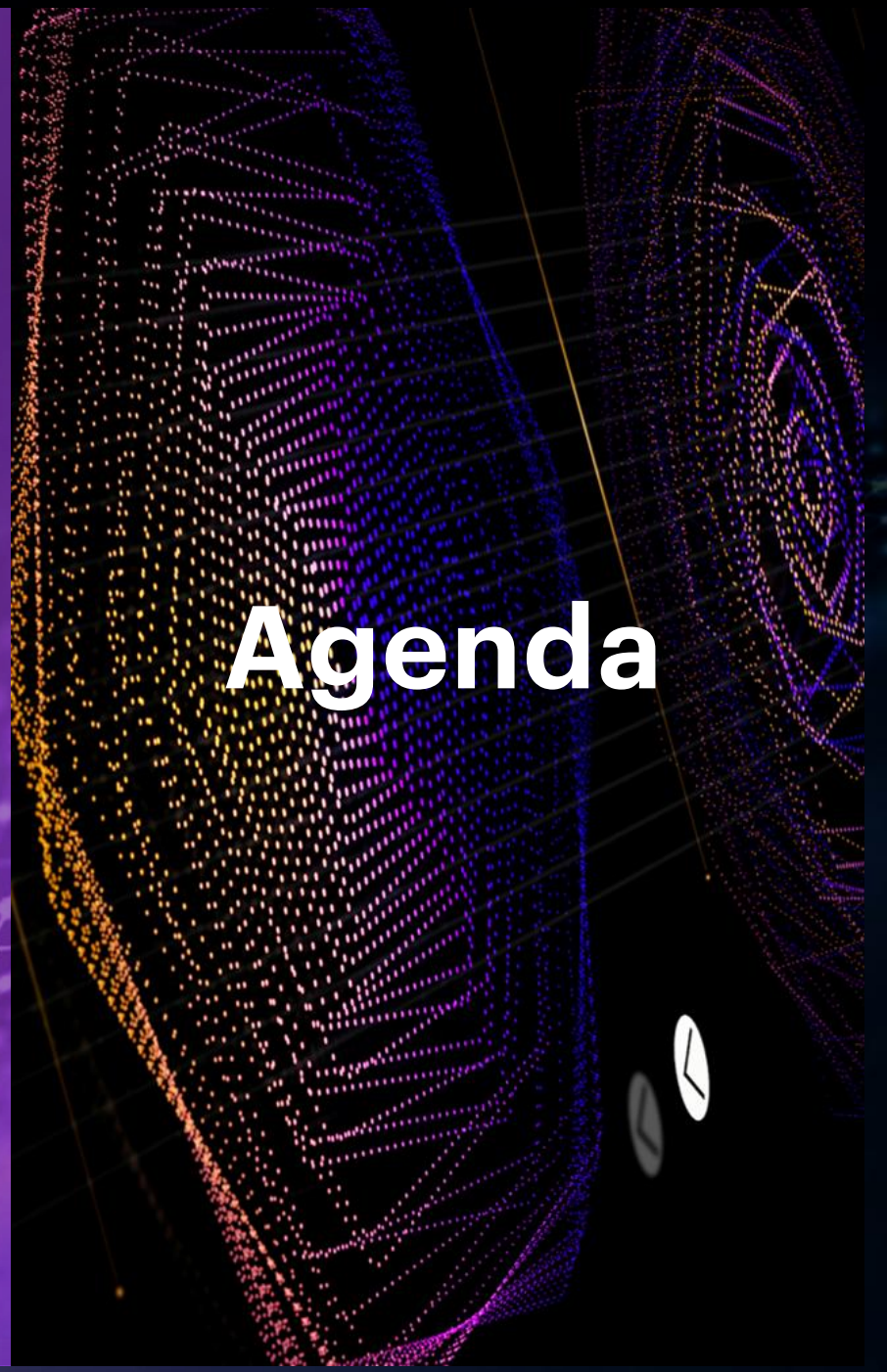
Global Trends in Scaling Agentic AI

East meets West

Malcolm Hsiao
Accenture Managing Director
AI & Data Lead
May 2026

Global AI Trends 2026

East meets West



01 **The AI Leap**

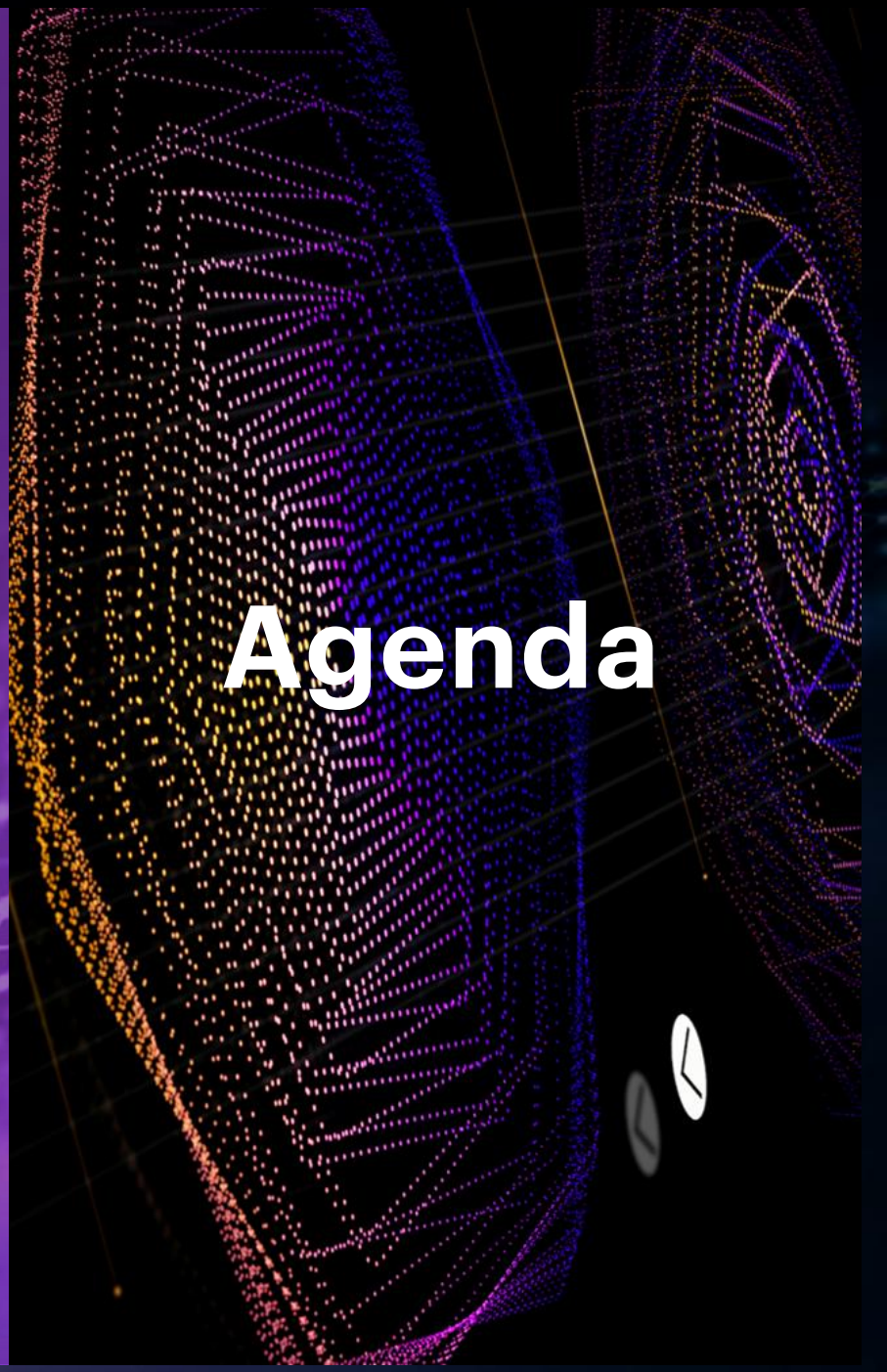
02 **How LLMs Are Getting Smarter**

03 **The Rise of Democratized Agentic AI**

04 **Scaling Agentic AI**

Global AI Trends 2026

East meets West



01 **The AI Leap**

02 **How LLMs Are Getting Smarter**

03 **The Rise of Democratized Agentic AI**

04 **Scaling Agentic AI**

The AI Revolution

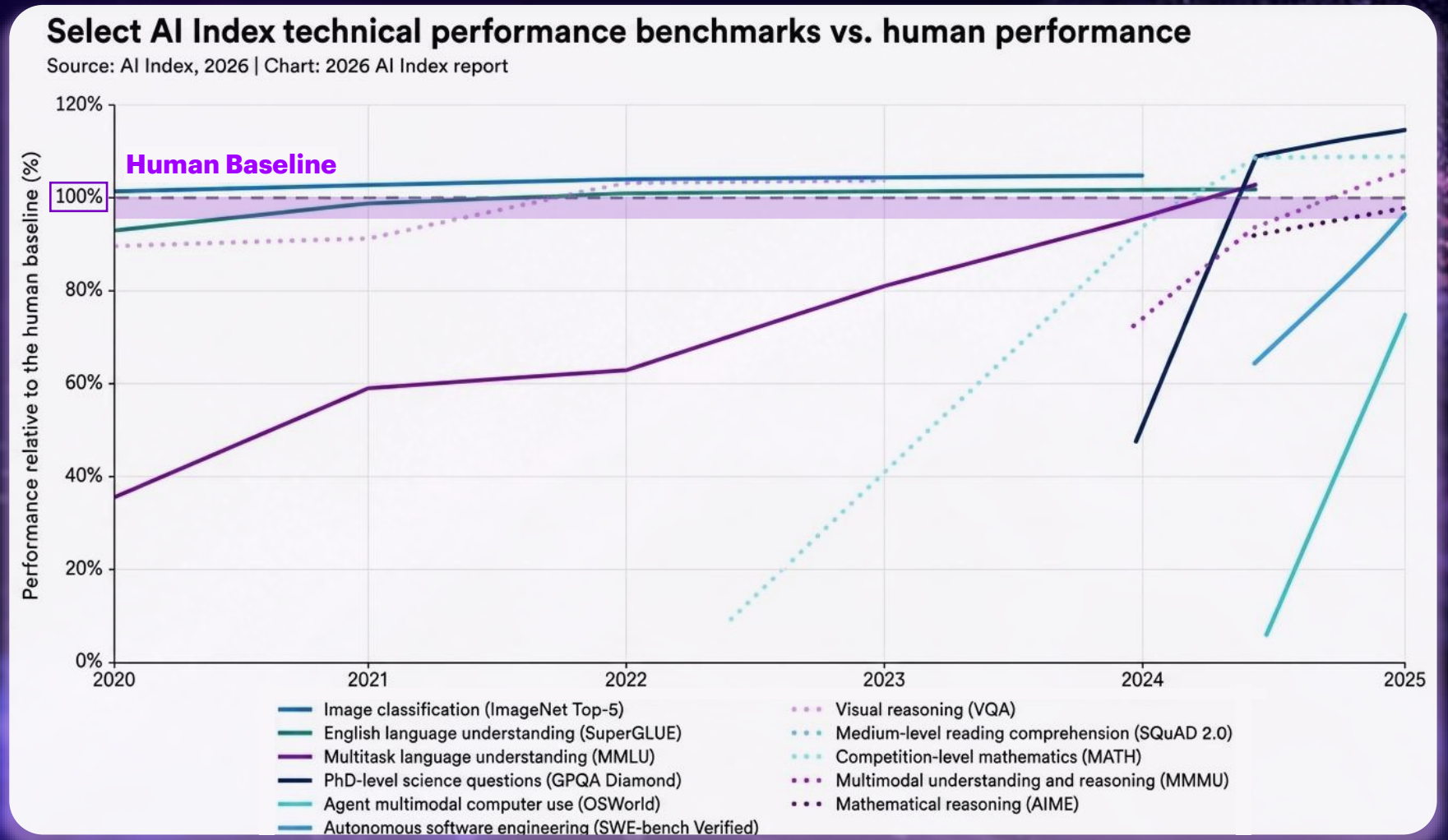
An Era of Unprecedented Change

Imagine a technology that gains **30 IQ-points annually**, **reduces inference costs** by **10-20x every year**, continuously evolves into **new modalities**, and reshapes entire ecosystems in mere months.

This isn't science fiction—it's today's AI reality.

The AI Revolution

An Era of Unprecedented Change

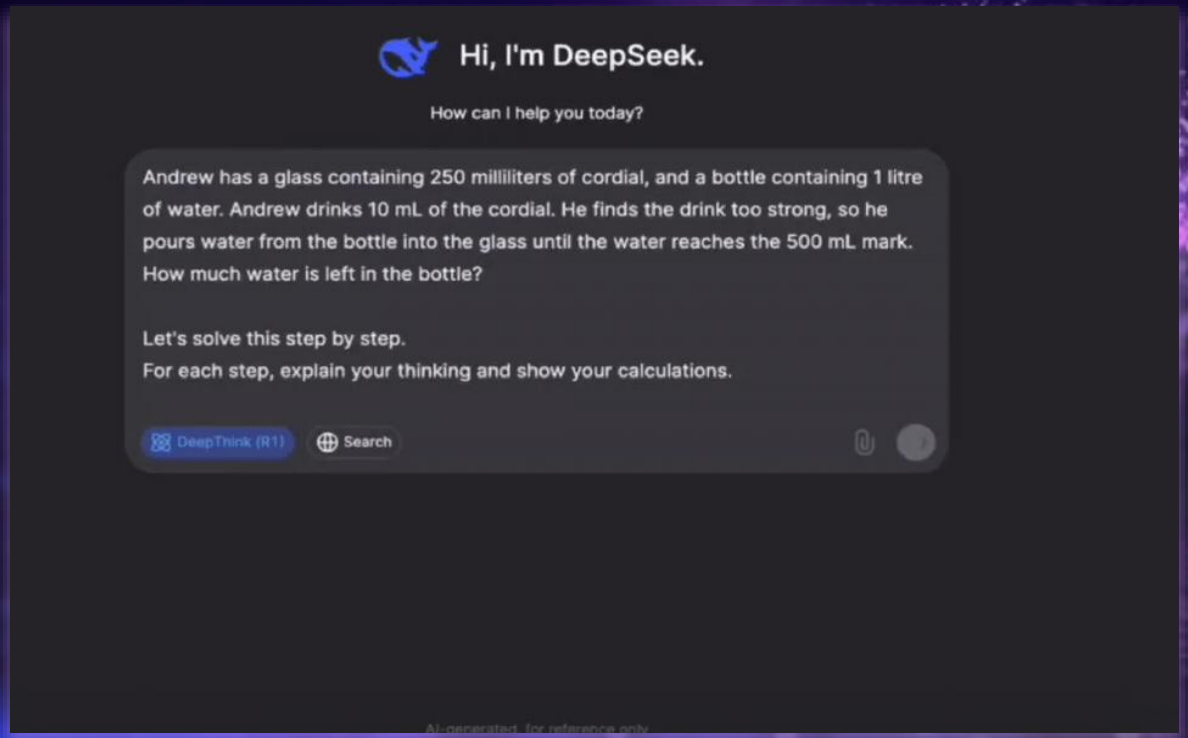


Source: Stanford Institute for Human-Centered AI (2026)

Reasoning Capability

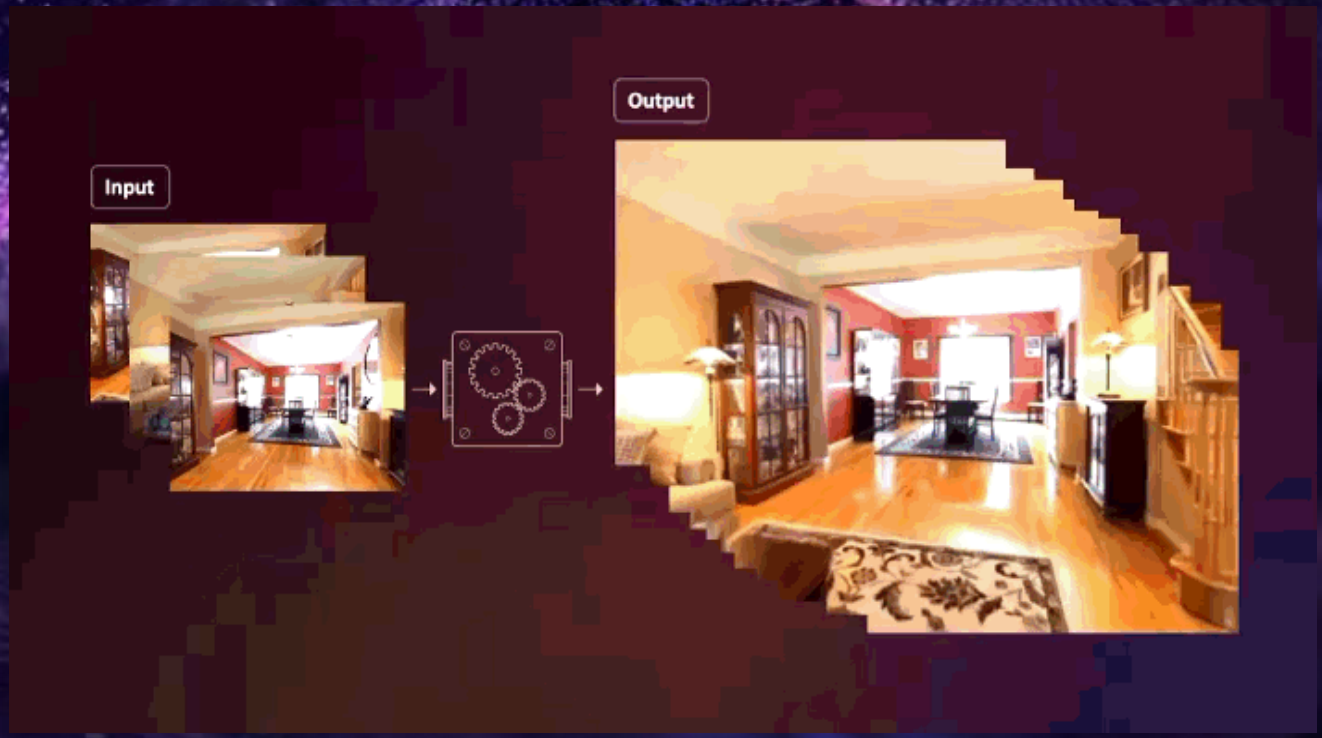
DeepSeek V3
Jan 2025

Reasoning - Chain of Thought



Stanford –Fei-Fei Li
2026

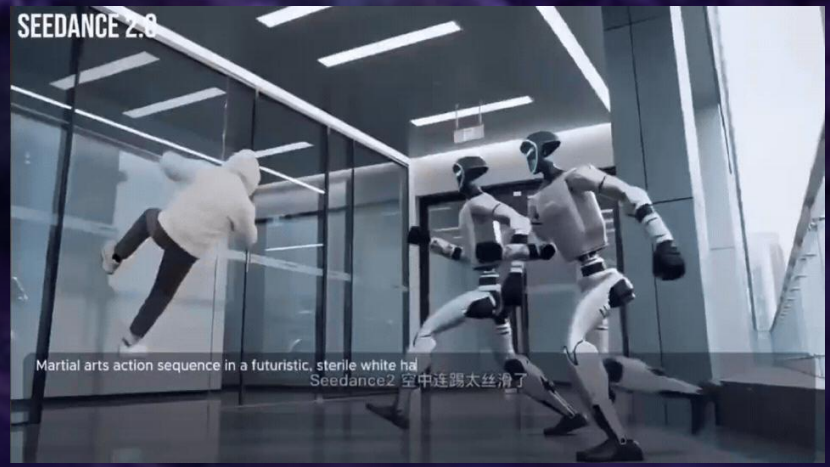
Spatial Intelligence – AI Understands the Real World



Video Generation

Sora Turbo Jan 2025

SeedDance 2.0 Mar 2026



Physical AI – Humanoid Robots

Unitree Robotics
Feb 2025

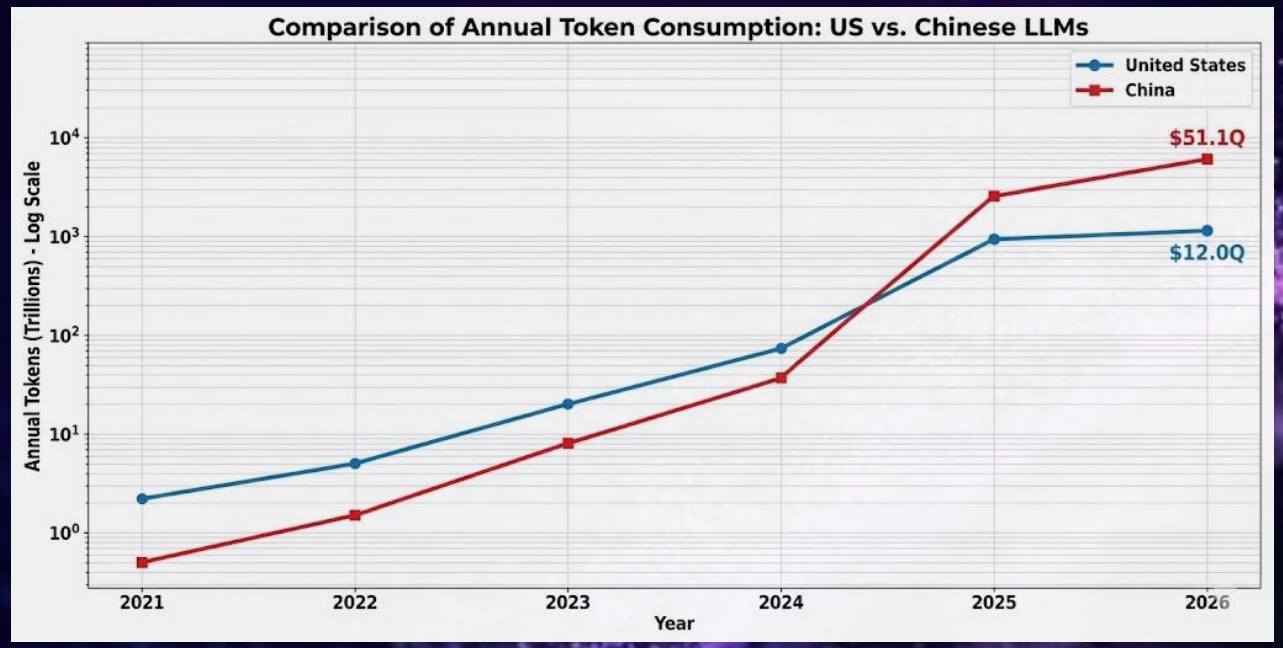
Performing Dongbei Dance

Unitree Robotics
Feb 2026

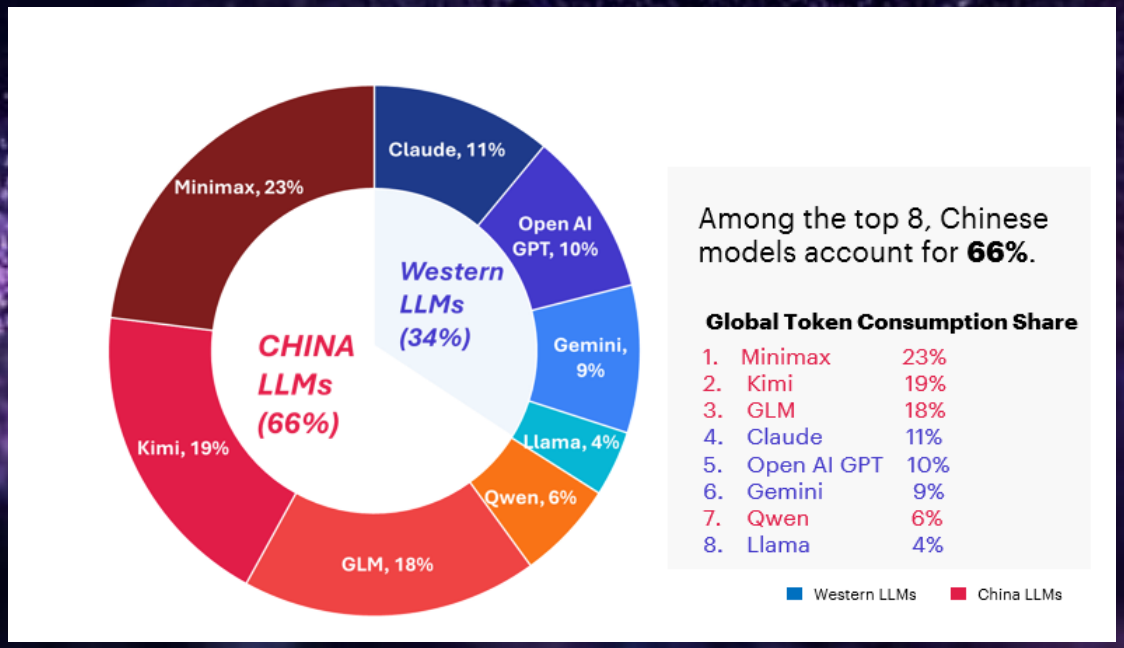
Performing Martial Arts

Token Consumption Comparison

US vs. Chinese Annual Token Consumption (2025-2026)



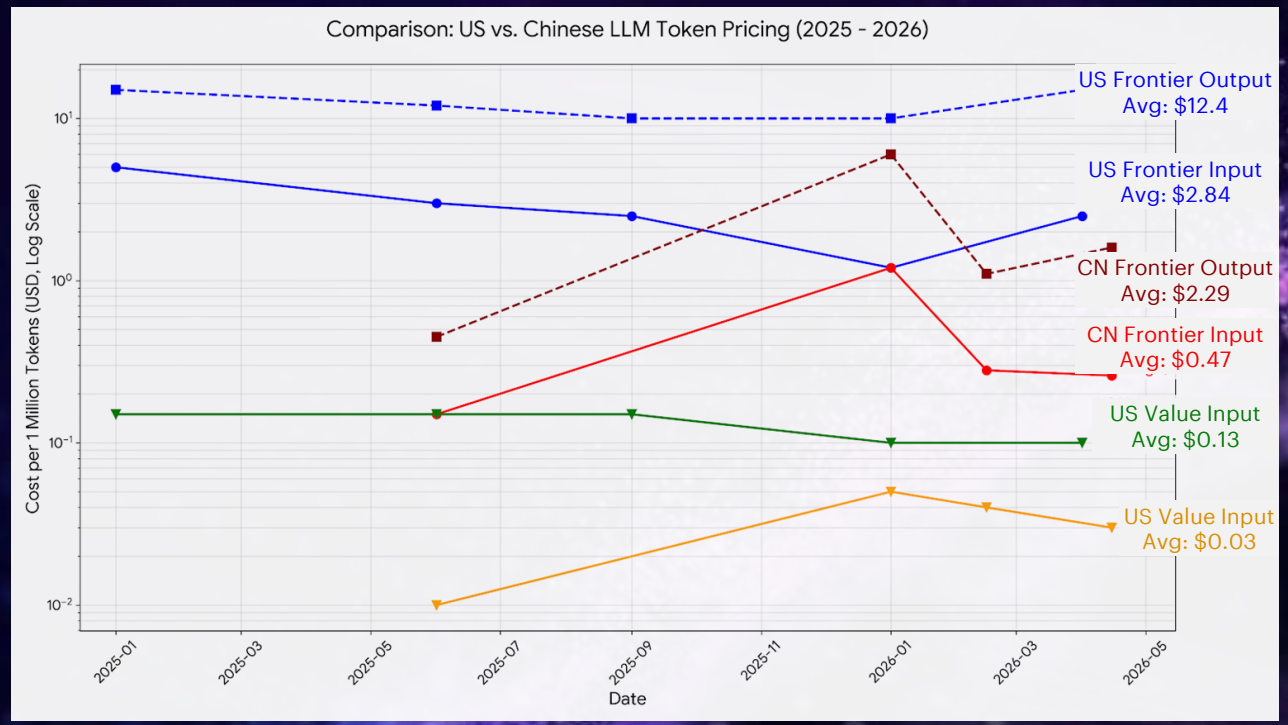
Global LLM Token Consumption Share (May 2026)



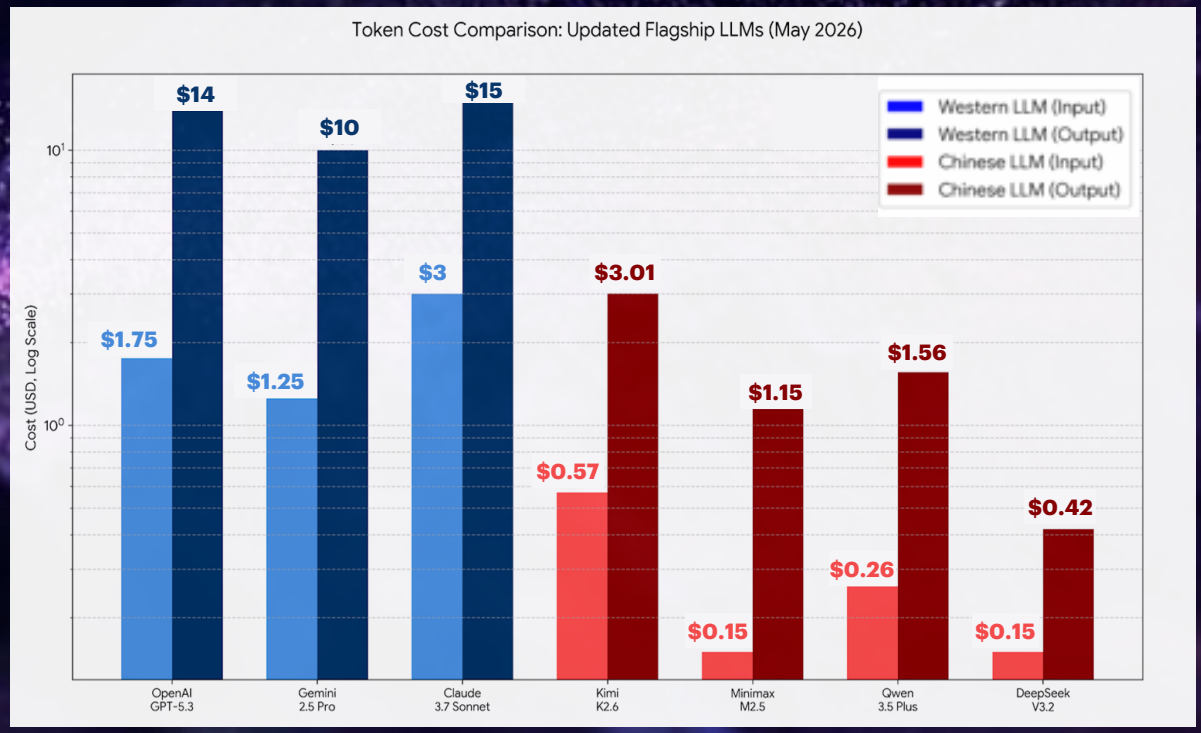
Source:
China National Data Administration - Announcements regarding national daily token consumption benchmarks (March 2026)

Token Pricing Comparison

US vs. Chinese LLM Token Pricing (2025-2026)



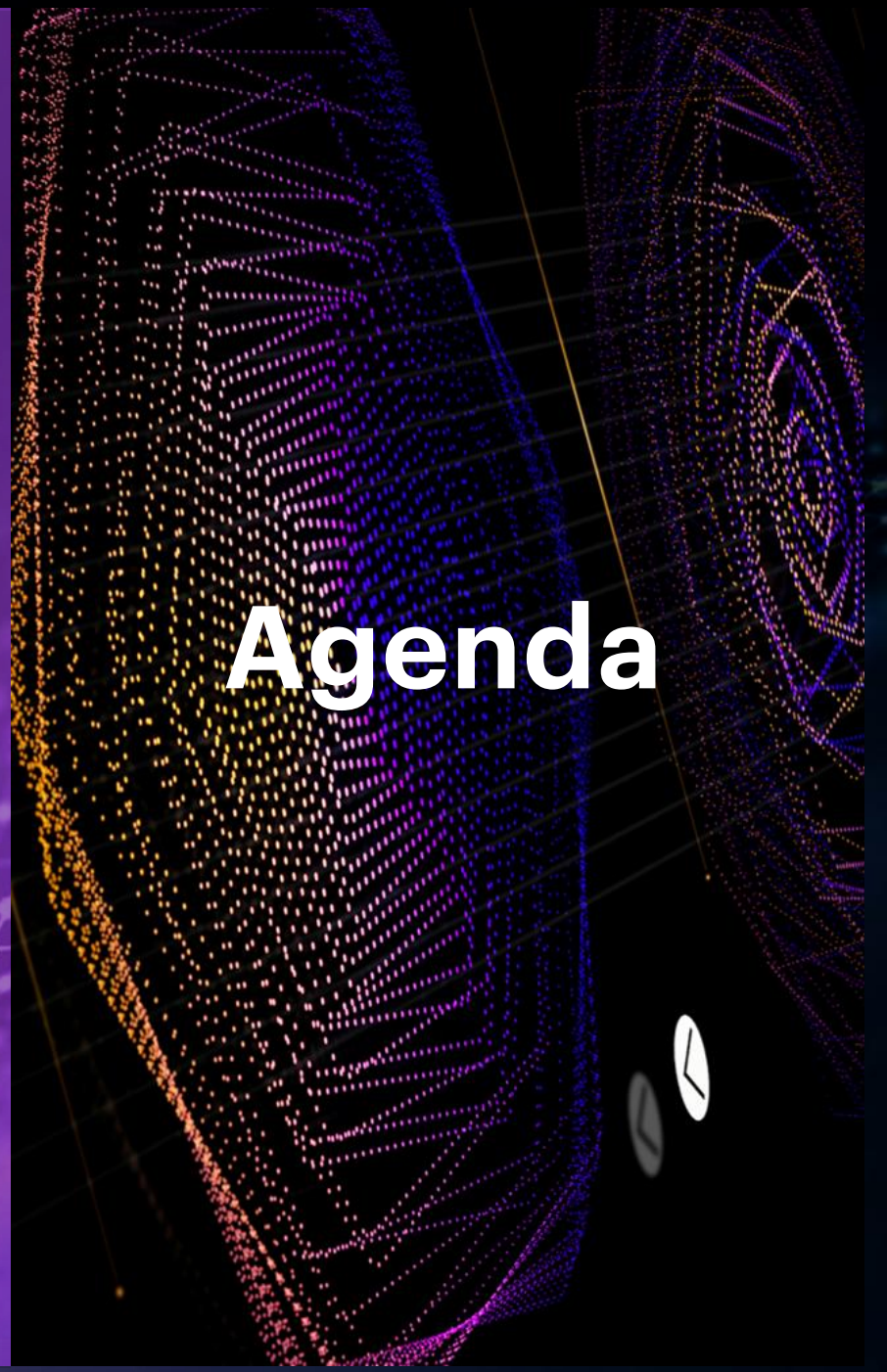
US vs. Chinese Token Pricing by LLM (May 2026)



Source: China National Data Administration: Announcements regarding national daily token consumption benchmarks (March 2026)

Global AI Trends 2026

East meets West



Agenda

01 The AI Leap

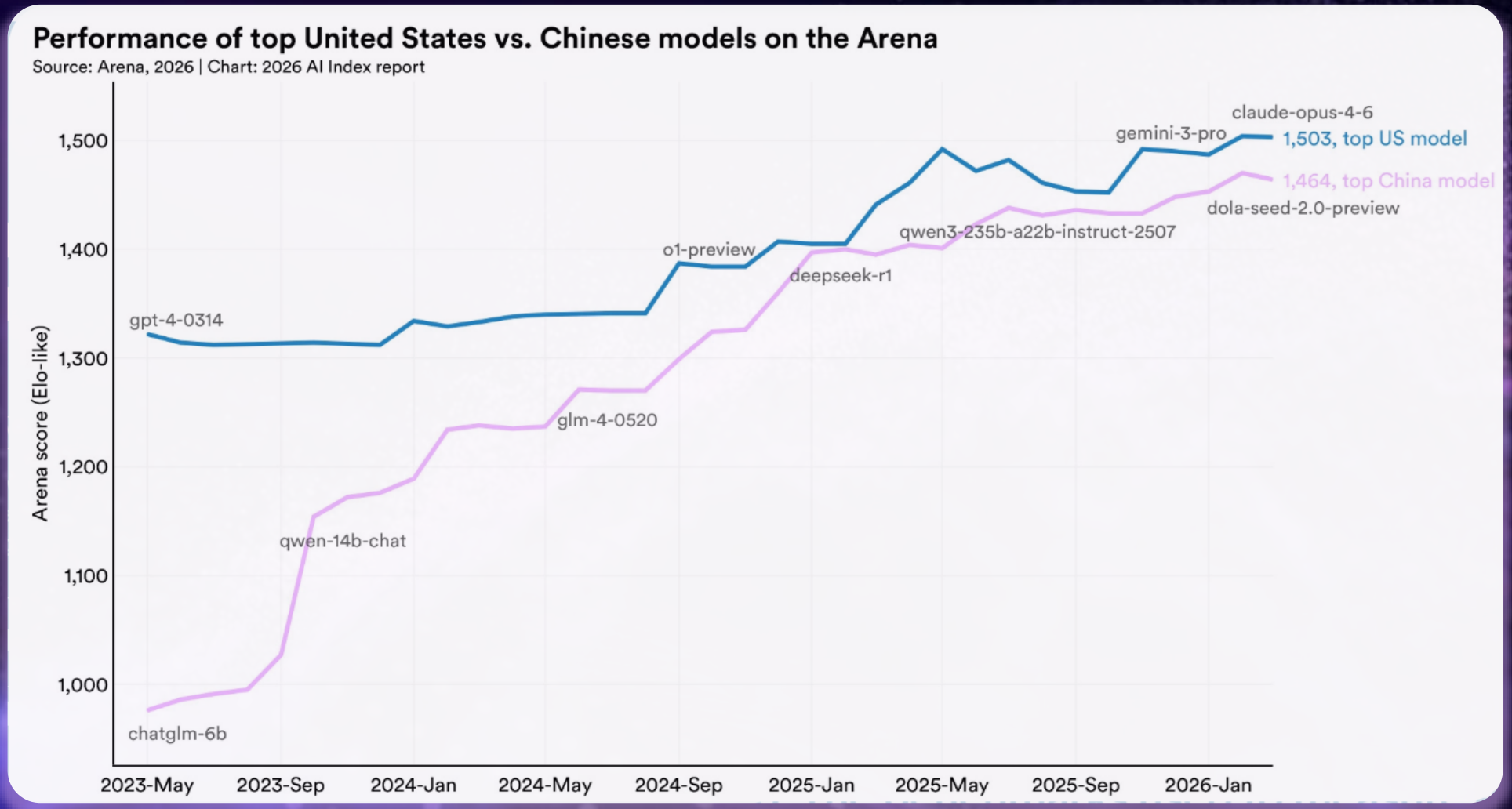
02 How LLMs Are Getting Smarter

03 The Rise of Democratized Agentic AI

04 Scaling Agentic AI

>

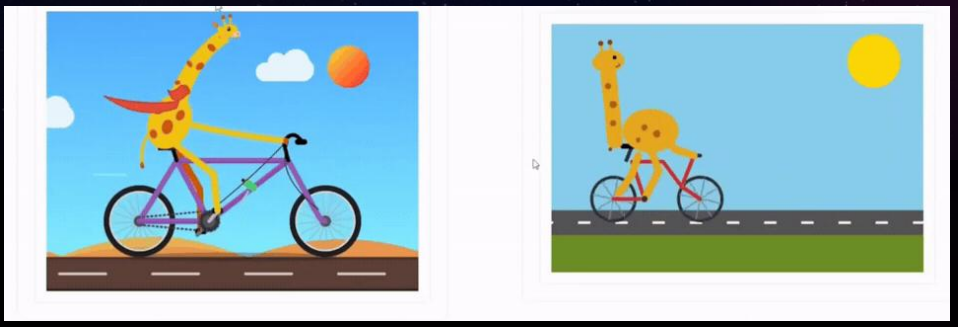
Model Comparison



Source: Stanford Institute for Human-Centered AI (2026)

Gemini 3.1

- **Release:** February 2026
- **Variants:** Gemini 3.1 Pro, Flash, Flash-Lite
- **Context:** 1M input tokens (~2700 pages), 64K output tokens (~175 pages)
- **Native multimodal:** Yes
- **Open source:** No
- **Parameters:** Not disclosed
- **Knowledge cutoff:** January 2025



Gemini 3.1 Pro **Gemini 3 Pro**

Training Methods

- **End-to-End Multimodal Pre-training**
- **Long-Context Specialised Training**
- **Progressive Reasoning Curriculum Training**
- **Reinforcement Learning Human Feedback**
- **Adversarial Red Teaming**

Technical Breakthroughs

Reasoning Power: Dominates benchmarks like ARC-AGI-2 (Gemini 3.1 scoring 77.1%, doubling Gemini 3 Pro's 31.1%). This "AI IQ test" uses never-before-seen logic puzzles, proving 3.1 Pro's ability to derive new rules rather than just recall facts.

Architecture Shift: Architecture Shift: Moves beyond "probability-based text prediction" to true logical reasoning—it breaks complex problems into smaller steps and solves them methodically.

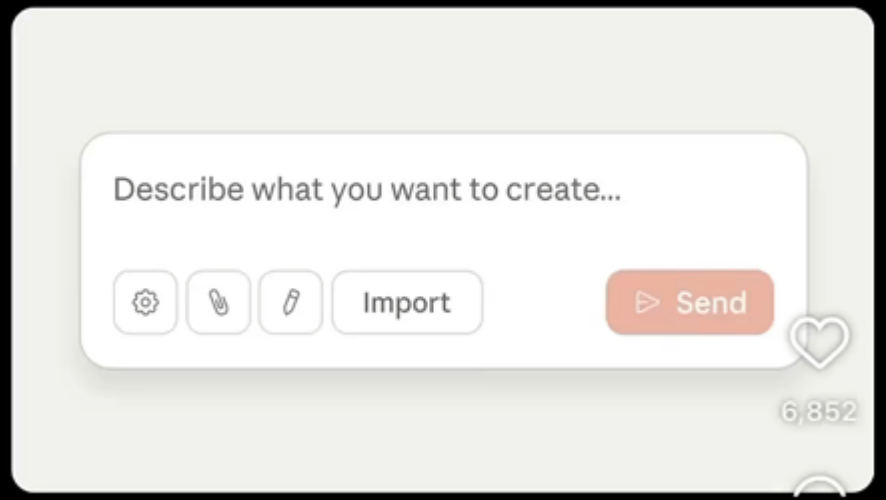
Multimodal & Coding: Multimodal & Coding: Excels in text, image, video, and audio integration. It writes SVG animations (pure code, infinitely scalable without quality loss) and 3D interactive programs (e.g., a flock of starlings with gesture-controlled formations and real-time adaptive music).

Claude Opus 4.7

- **Release:** April 2026
- **Variants:** None
- **Context:** 1M input tokens (~2700 pages), 128K output tokens (~350 pages)
- **Native multimodal:** Yes
- **Open source:** No
- **Parameters:** Not disclosed
- **Knowledge cutoff:** Dec 2025

Claude Opus 4.7

Claude Opus 4.6



Training Methods

Constitutional AI + RLHF + Red-Teaming for safety and low hallucinations

Adaptive Thinking RL — Effort-Calibrated Reasoning Training

Agentic RL in Long-Horizon Harnesses with Interleaved Thinking

Technical Breakthroughs

Adaptive Thinking: lets Claude dynamically allocate thinking token budgets based on the complexity of each request — rather than applying a fixed reasoning depth to every problem. No previous model could achieve this natively.

Sustained Long-Horizon Autonomous Reasoning: handles complex, long-running tasks with rigor and consistency, pays precise attention to instructions. The self-verification capability is technically significant — the model catches its own logical faults mid-execution rather than reporting back with errors.

High-Resolution Vision: the first Claude model with high-resolution image support — maximum image resolution increased to 2576px / 3.75MP, up from the previous limit of 1568px / 1.15MP. It enables major improvements in multimodal understanding, from reading chemical structures to interpreting complex technical diagrams

Claude Mythos

Leaked in late March 2026. Anthropic states it did not explicitly train Mythos Preview to specialize in software exploitation — these capabilities are a downstream consequence of general improvements in AI reasoning and software engineering capabilities. Preview available to 40 enterprise customers only

deepseek v4

- **Release:** April 2026
- **Variants:** V4-Pro and V4 Flash
- **Context:** 1M input tokens (~2700 pages), 384K output tokens (~1000 pages)

- **Native multimodal:** Partial
- **Open source:** Yes
- **Parameters:** 1.6T
- **Knowledge cutoff:** 2025

Training Methods

Hybrid Attention Architecture (CSA + HCA)

Manifold-Constrained Hyper-Connections (mHC)

Muon Optimizer

Technical Breakthroughs

mHC Training Stability and Hybrid Compressed Sparse Attention: Standard transformers scale quadratically with sequence length — making 1M token contexts prohibitively expensive. DeepSeek V4 introduces a hybrid Compressed Sparse Attention (CSA) and Heavily Compressed Attention (HCA) system that reduces inference FLOPs by up to 73%.

On-Policy Distillation — Replacing Reinforcement Learning: Rather than training a single model with RL rewards (as V3R1 did), V4 builds specialist domain experts independently then distils their combined knowledge into the unified V4 model. This produces strong and consistent cross-domain performance while keeping training costs manageable.

Trained Natively on Huawei GPU: V4's entire inference stack runs natively on Huawei's CANN (Compute Architecture for Neural Networks) platform, with day-one support across the full Ascend 950, A2, and A3 supernode lineup.

HUAWEI

DeepSeek V4 = Built on Huawei → Trained on Huawei → Launched on Huawei → Then ported to Nvidia.

Seedance 2

ByteDance Seed

- **Release:** February 2026
- **Variants:** Seed 2.0 Base LLM, Seedream 5.0 image generation, Seedance 2.0 video generation
- **Context:** 272K input tokens (~400 pages), 131K output tokens (~200 pages)
- **Native multimodal:** Yes
- **Open source:** No
- **Parameters:** Not disclosed
- **Knowledge cutoff:** December 2025

Hollywood Shock

- Production cost collapsed
- Cinematic AI at scale
- Democratized blockbuster power
- Existential threat to studios & unions

Technical Breakthroughs

Dual-Branch Diffusion Transformer:
Separates spatial quality (like Seedream) and temporal motion (video logic). Understands 3D space & physics, not just 2D pixels.

True Multi-shot Cinematic Storytelling:
One prompt = full sequence: wide shots → close-ups → tracking → cuts.
Character consistency across angles, lighting, scenes. Built-in "director logic": composition, rhythm, pacing

Native Audio-Video Sync (Industry First):
Generates video + dialogue + SFX + BGM in one pass. Perfect lip-sync for many languages. No separate audio post-production needed

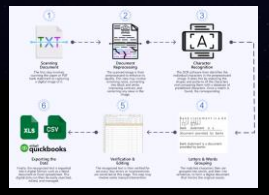
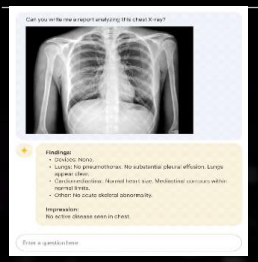
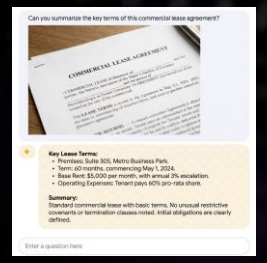
Early March 2026
Hollywood coalition formally files collective lawsuit

Mid March 2026
ByteDance halts global public rollout of Seedance 2.0 overseas; closes open API access for international users

Late March 2026
ByteDance rolls out strict content filters. Maintains limited service only in mainland China with restricted content rules

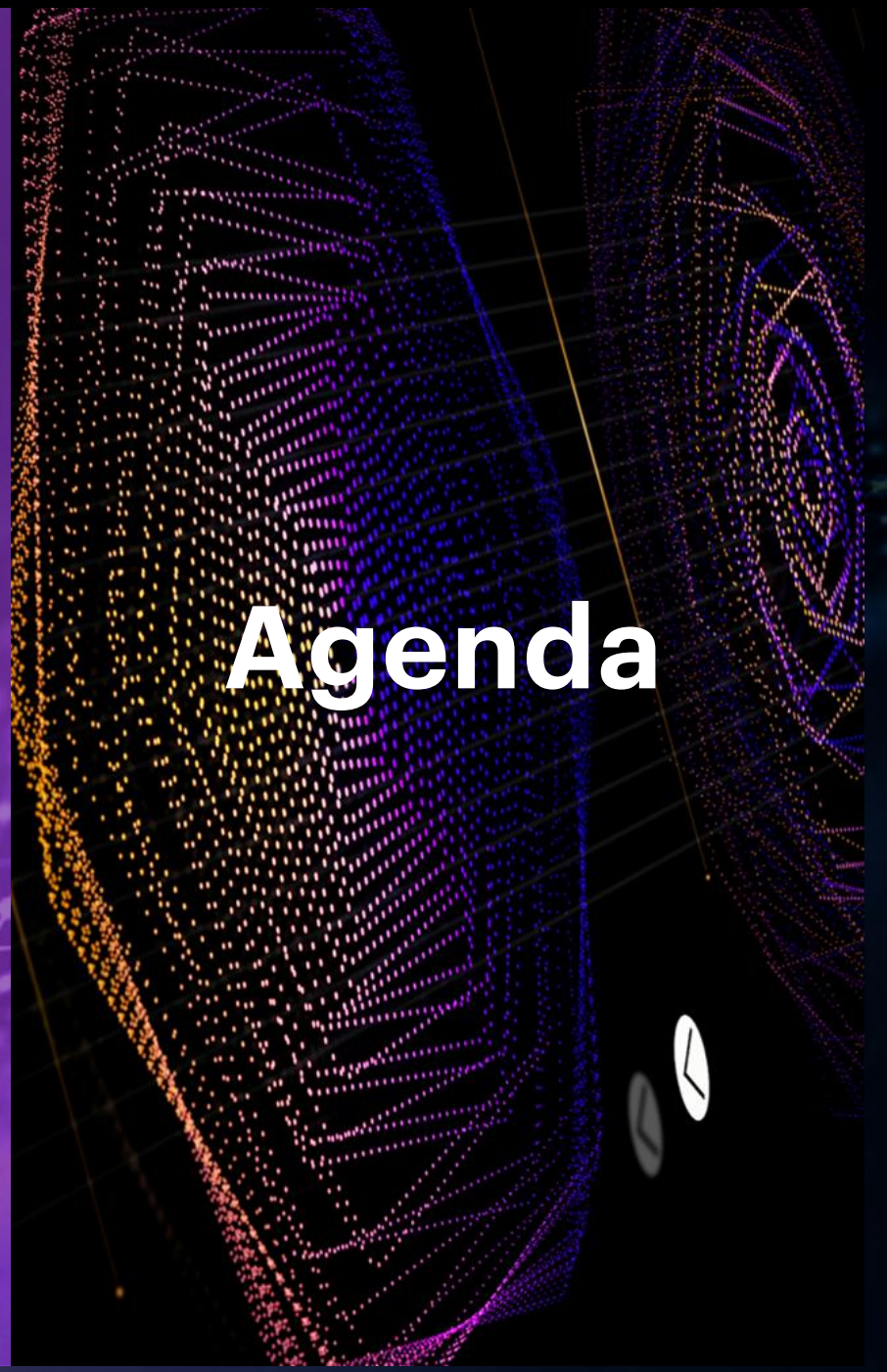
Task-specific LLMs emerging

The next wave of AI value will come from specialist models that understand industry language, documents, risks, and decisions.

Main Domain-Specific LLM	Example Models	Business Use Cases
 **OCR / Document AI**	• LayoutLMv3, • Donut, • PaddleOCR-VL	• Invoices, contracts, KYC forms, statements, insurance claims, tables, scanned PDFs
 **Medical / Healthcare**	• Med-PaLM, • Med-PaLM 2, • MedLM, • BioGPT	• Medical Q&A, clinical summarization, patient note drafting, healthcare research support
 **Finance**	• BloombergGPT, • FinGPT, • FinBERT / financial LLM variants	• Market intelligence, financial sentiment, research summarization, credit analysis, robo-advisory, risk monitoring
 **Legal / Compliance**	• LegalBERT, • Legal-domain LLMs, • Contract review models	• Contract review, clause extraction, regulatory monitoring, policy interpretation

Global AI Trends 2026

East meets West



01 **The AI Leap**

02 **How LLMs Are Getting Smarter**

03 **The Rise of Democratized Agentic AI**

04 **Scaling Agentic AI**

>

The Rise of Democratised Agentic AI

OPEN-SOURCE AGENT

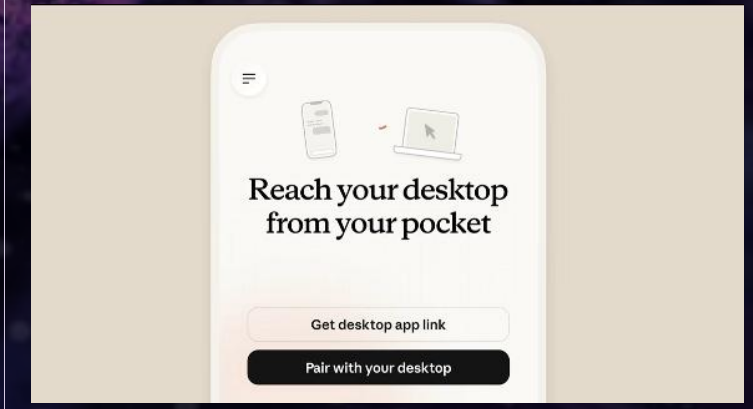
Open Claw

Open-source autonomous agent with full system access. Powerful, but high attack surface — needs sandboxing and a human in the loop.

DESKTOP AGENT

Claude Cowork

Sandboxed desktop agent from Anthropic. Enterprise-safe with limited system access — friendlier for non-technical users.



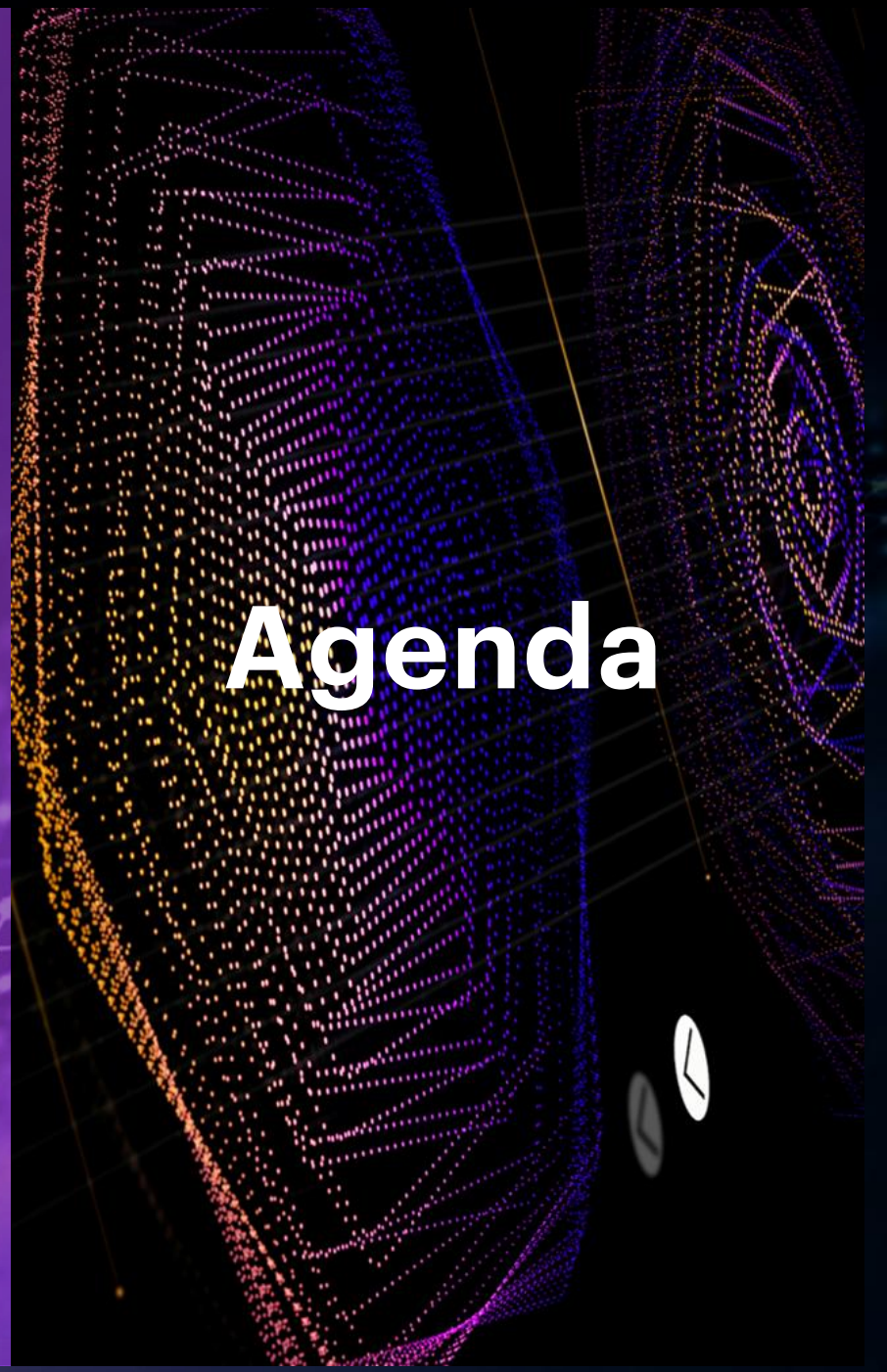
WORKFLOW ENGINE

n8n

Workflow automation engine with predefined flows. Not truly autonomous, but easy to govern and familiar to IT teams from day one.

Global AI Trends 2026

East meets West



Agenda

01 The AI Leap

02 How LLMs Are Getting Smarter

03 The Rise of Democratized Agentic AI

04 **Scaling Agentic AI**

Why Are Enterprises Failing at Scaling AI?

82% of organizations are not seeing **scaled value creation** from Gen AI initiatives.

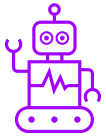
Technology is Moving Exponentially Faster Than Organizations Can Absorb

Agentic capabilities are advancing rapidly, but enterprise operating models, skills, and governance evolve too slowly to keep pace—creating a persistent execution gap.

Key Challenges

Lack of a Secure, Governed Agentic & AI Platform Foundation
Most organizations experiment with agents in isolation, without a trusted, enterprise-grade control plane for security, policy enforcement, lifecycle management, and observability.

Fragmented Data, Knowledge, and Context Limit Agent Effectiveness
Agents lack access to unified, semantic, and institutional knowledge—resulting in brittle reasoning, shallow automation, and limited decision confidence.



Point Solutions Instead of End-to-End Agentic Orchestration
Agents are deployed as task-level tools rather than orchestrated across full business journeys, preventing compound value creation and cross-functional impact.

Lack of Enterprise Capability to Continuously Monitor, Learn, and Improve Agents
Enterprises lack the operational capability to monitor agent behavior, learn from outcomes, and continuously improve agents—blocking trust and scale.

Successful AI Reinvention Requires Focus on Key Imperatives: Value, Work, Workforce, Workbench

Powered by a robust AI & Data Engine – the Intelligent Digital Brain

Lead with

Value

Start with

Art of Reinvention

Reimagine Work

Re-engineer processes around outcomes

Enable data-driven decisions

Reshape the Workforce

Lead in new ways

Build a future-ready organization

Access and create talent

Deliver continuous change

Redesign the Workbench

AI-powered tools

Connected Platform

Seamless user experience

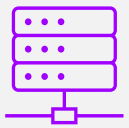
*Rewire the **Digital Core** to create an*

Intelligent Digital Brain

Adaptive learning | **Curated enterprise ontologies** | **Specialized models**

Intelligent Digital Brain (The Five Layer Cake) Creates Value in Several Ways

It unlocks trapped data and institutional knowledge to create **actionable specialized intelligence, drive agentic automation and transform systems into dynamic, evolving ones.**



Unlock Trapped Data

Unlock trapped multi-modal data in different silos and turn them into shared insights

Codify Institutional Knowledge

Codify all institutional learning – explicit, implicit and tacit – into institutional knowledge

Create Specialized Intelligence

Turn general intelligence – external and/or internal into specialized intelligence embedded into functions

Drive Autonomous Workflow

Drive multi-systems, trusted multi agent automation and orchestration

Continuously Learn & Adapt

Convert static, rule-based systems into dynamic, intelligence-based systems that adaptively and continuously learn

Intelligent Digital Brain is Made Up of 5 Capability Building Blocks

The Brain's full capabilities emerge when each of these components functions in a **coordinated and unified way**

1. Data (Modern Data Foundation)

Base Layer for Structured & Unstructured Data

- Data as a Product and Unlocking Trapped Data
- Seamless data sharing via secure connectors/APIs, model context protocol or zero copy integration

3. Model (Specialized Industry Models)

Cognition for Thinking & Reasoning

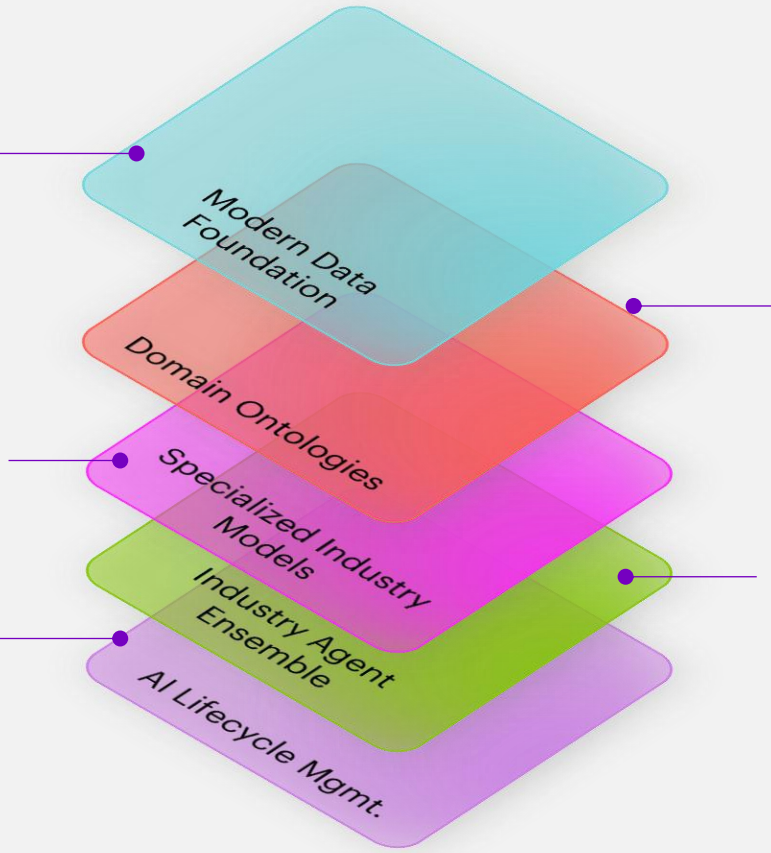
- Model Recipe for specialized model customization
- Adaptive learning
- Experiential knowledge acquisition

5. Arch & Governance (AI Lifecycle Mgmt.)

Structure and Control for Scaled AI

- Lifecycle governance and asset registries
- Accurate and explainable outputs
- Secure data and privacy controls

The Five Layer Cake



2. Knowledge (Domain Ontologies)

Contextual Data Understanding

- Convert data into specialized and domain contextual knowledge
- Codify institutional learning to scale expertise
- Implemented via semantic layers & knowledge representation graphs

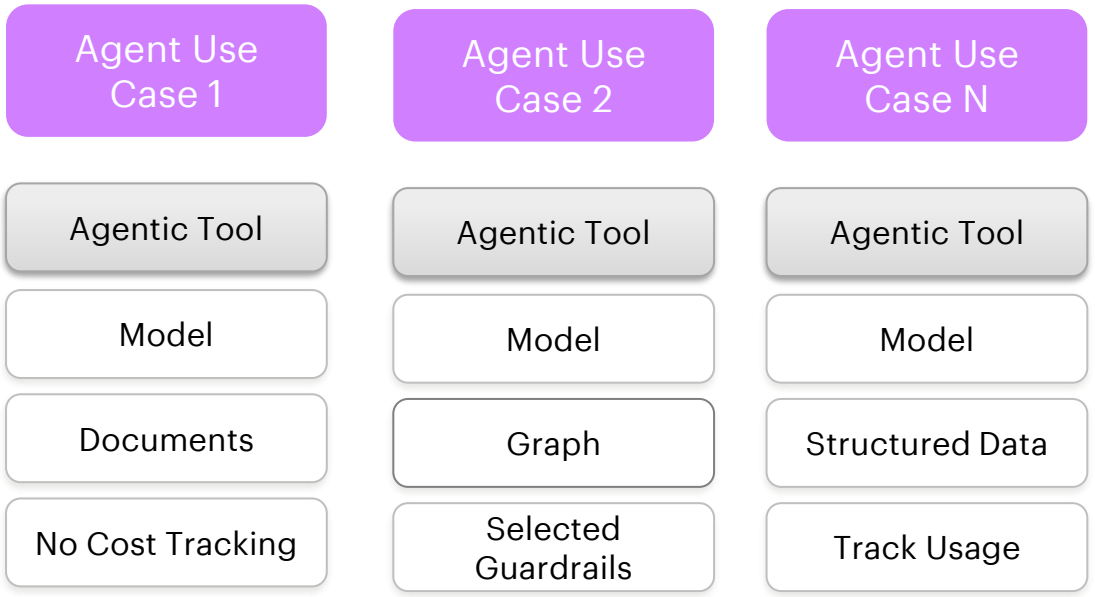
4. Agent (Industry Agent Ensemble)

Tasks & Autonomous Decisions

- Multi-agent collaboration and orchestration
- Agent certification & evaluation
- Industry specific agents

Enabling a “Platform based approach” to scale Agentic AI rapidly, responsibly & securely

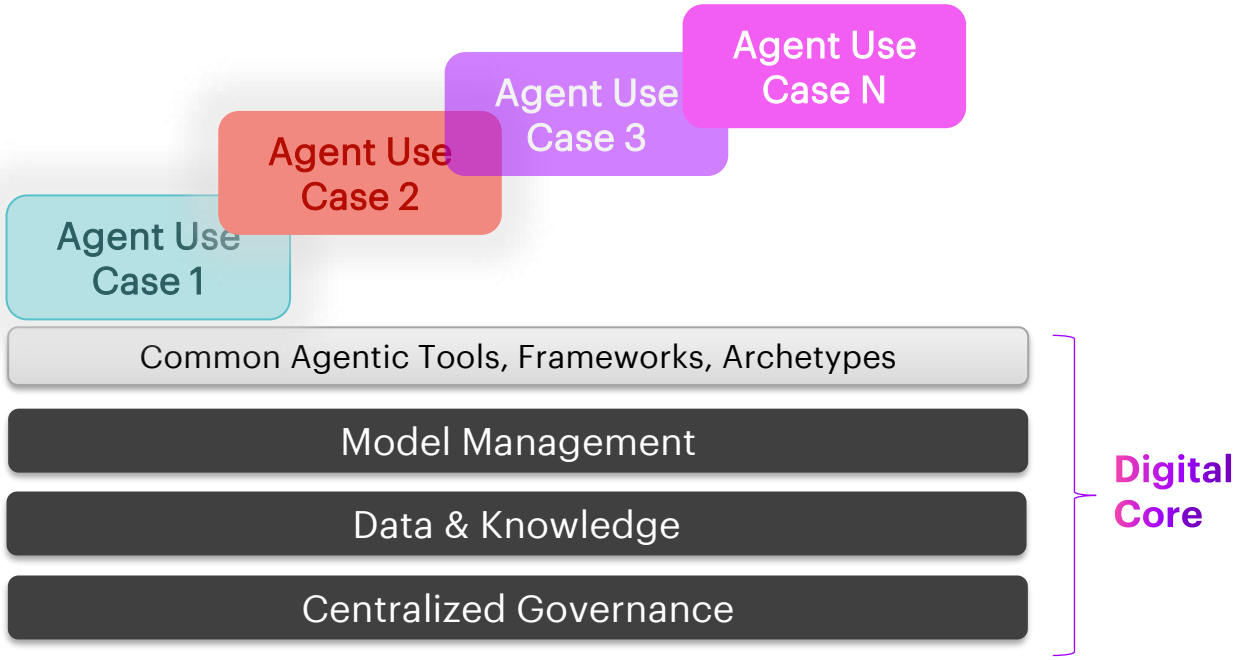
Without Digital Core
“Building One by One” in fragmented fashion



- Lack of Scalability
- Duplication of Effort
- Inconsistency
- Higher Costs

- Slow Time to Market
- Operational Burden
- Technical Debt

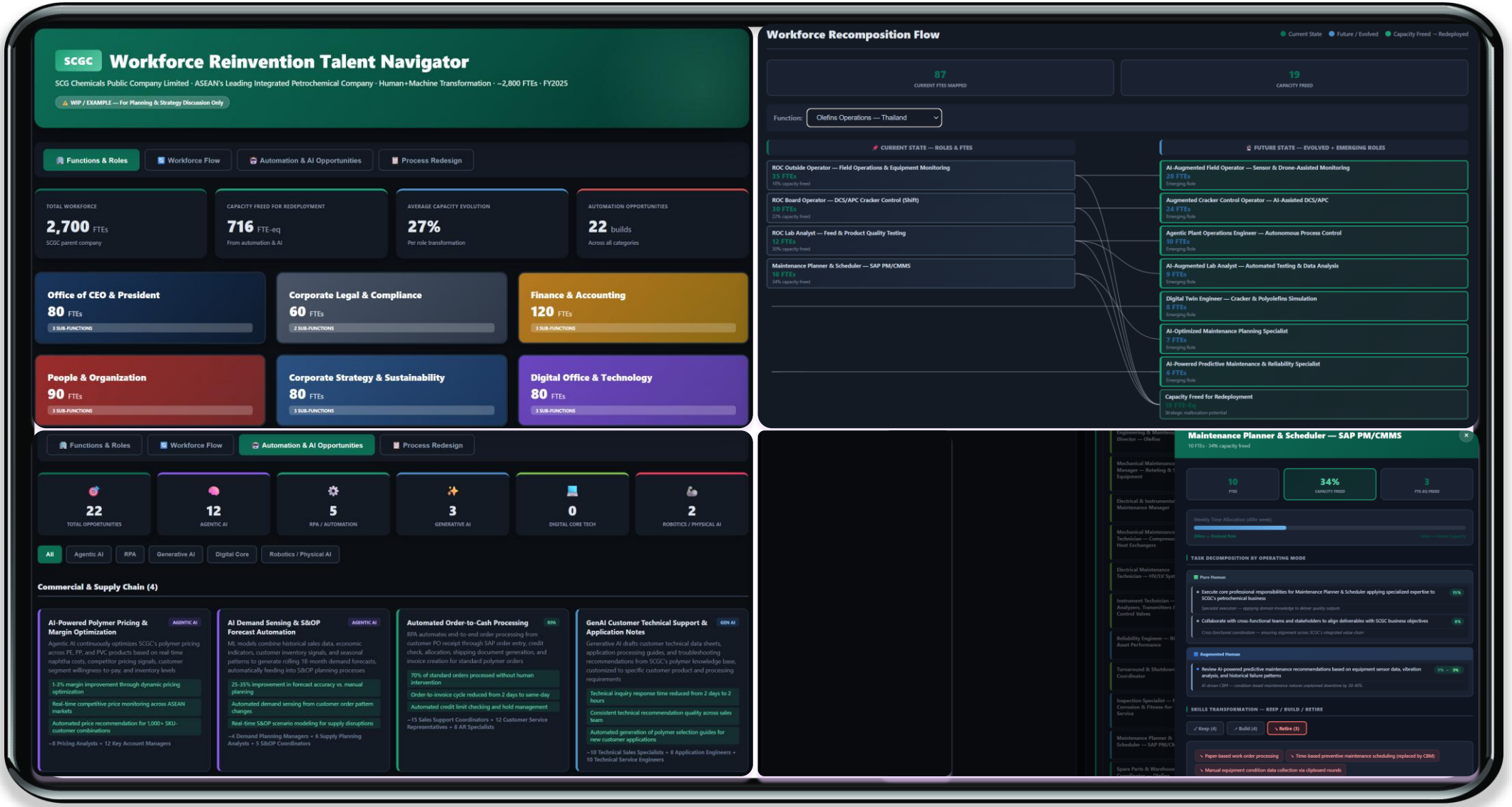
With Digital Core
“Platform Based Approach” in integrated fashion



- Scalability
- Efficiency and speed
- Cost optimized (higher upfront)
- Consistency and quality

- Improved Collaboration
- Better Governance and compliance
- Fast innovation
- Robustness

AI Reinvention Demo



Call for Action
Individuals - New Paradigms of Work

- **Believe in AI**
- From users of AI to **Co-creators & Orchestrators**
- Developing **composite skills** that blend domain expertise, tech fluency and judgment
- **Learning in Real time** — embedded in the flow of work, not bolted on

Call for Action

Leaders - Leadership Qualities in the Age of AI

- Become a **coach in AI**
- Model visible, responsible and **empowering use of AI**
- Don’t just ask what AI can do — ask **how your people can grow with it**
- Look beyond how AI affects specific tasks and roles. Start to scale AI by **redesigning processes across the organizations — breaking silos**

Interested in exploring how these solutions can support your business?

Connect with IT One team to learn more.

IT1.BSA@accenture.com

SCG Chemicals

K. Piyachat Dhanaraks
piyachat.dhanaraks@itone.co.th

SCG Packaging

K. Kitti Boonyakitnotai
kitti.boonyakitnotai@itone.co.th

SCG CBM

K. Sunion Panichattra
sunion.panichattra@itone.co.th

SCG Corporate

K. Soraphot Markathub
soraphot.markathub@accenture.com