

INDUSTRY DEEP DIVE

# The AI Industry

Value Chain · Key Players · Applications & Outlook

---

Core thesis: value now pools in the compute and cloud layers, shifting over 3–5 years toward apps with data, distribution & switching costs.

Vincent Wang

Data as of mid-2026 | Compiled July 2026

## Application Layer

Coding · Office · Consumer · Vertical

## Model Layer

Closed-source duo · Open-source wave

## Cloud & Infrastructure

Hyperscalers · Neoclouds · Power

## Compute Hardware

GPUs · Advanced packaging · HBM

# Three Core Judgments



01

**Profits concentrate in compute & cloud**

**\$725B**

Combined 2026 capex, top four hyperscalers

NVIDIA booked \$75.2B of data center revenue in a single quarter (+92%) at ~75% gross margin; model-layer revenue is exploding but deeply unprofitable.



02

**Token prices fall ~10x every year**

**\$20 → \$0.40**

GPT-4-class, per 1M tokens (late 2022 → early 2026)

Great for application growth, brutal for model margins; agent-driven inference volume means “volume over price” still entrenches compute.



03

**Top risk: capex–ROI divergence**

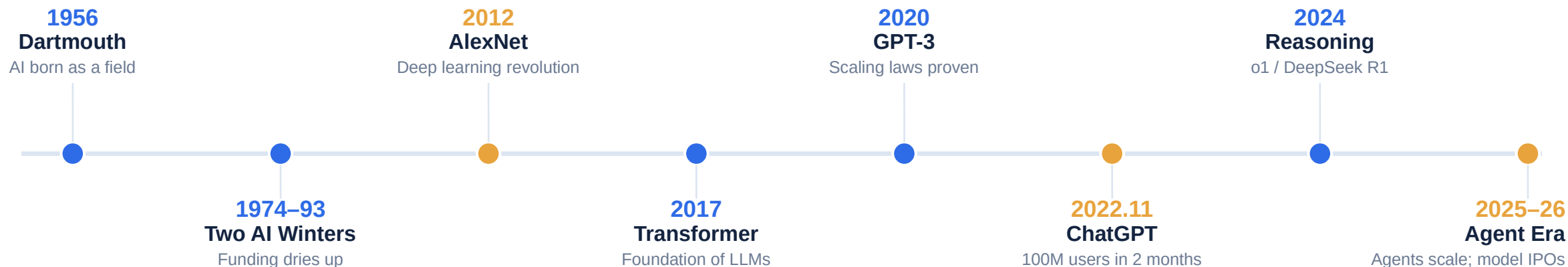
**5%**

MIT: only 5% of enterprise pilots deliver \$M-scale returns

The NVIDIA–OpenAI–Oracle–neocloud circular-financing loop builds systemic risk — yet \$1.4T+ of cloud backlog says demand is real, for now.

**Base case:** not a 2000-style pure bubble — but the capex–revenue gap is real systemic risk; the outcome hinges on enterprise ROI and credit conditions.

# Seventy Years, Three Waves: Symbolic AI to the Agent Era



## Symbolic AI & Classic ML

1956 – 2012

## Deep Learning Era

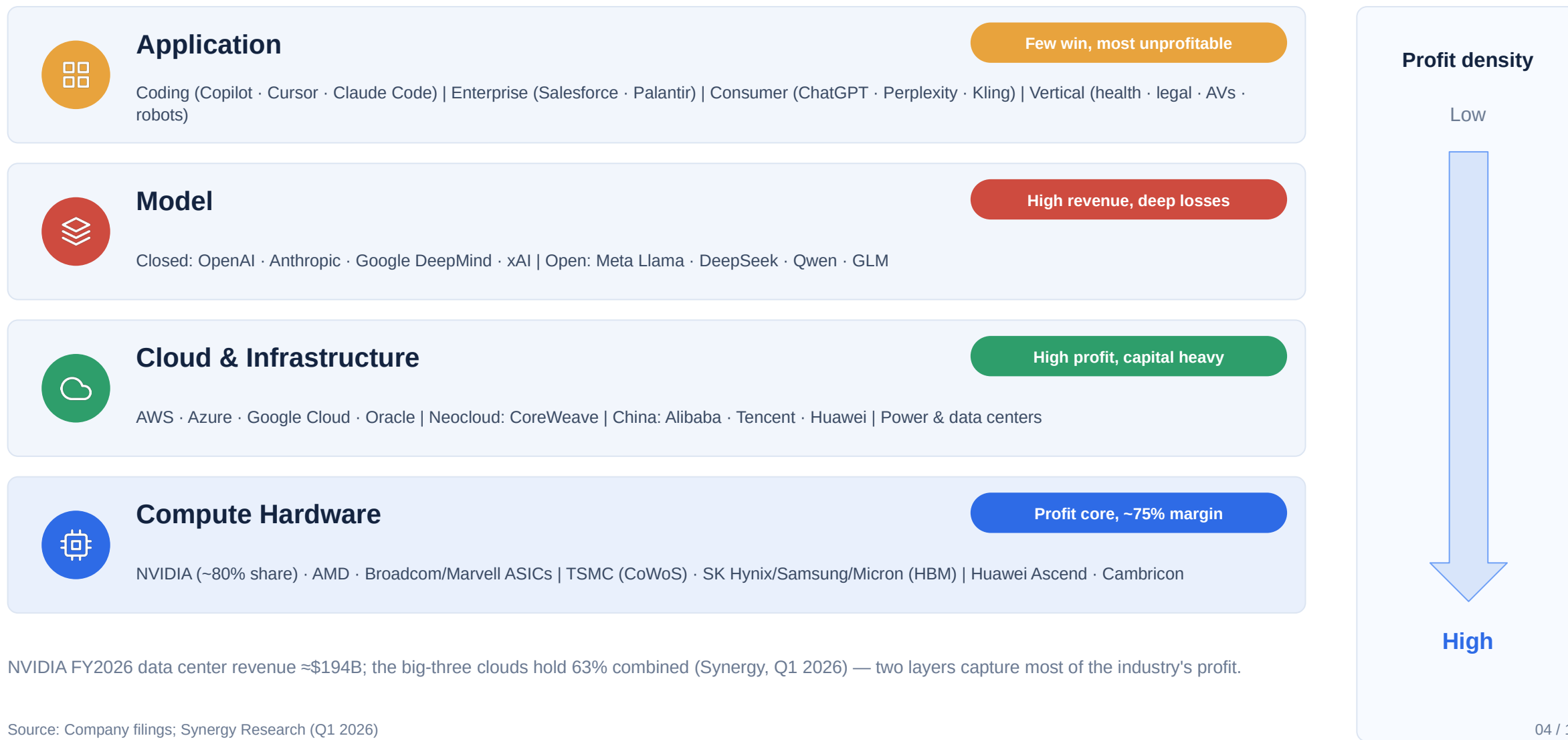
2012 – 2022

## Generative AI & Agents

2022 – Present

\* Spacing is illustrative, not a linear time scale

# The Four-Layer Value Chain: Profits Pool at the Bottom



# NVIDIA: The Industry's Most Certain Value Capturer

**\$81.6B**

FY2027 Q1 total revenue (+85% YoY)

**\$75.2B**

Quarterly data center revenue (+92%;  
networking \$14.8B)

**≈75%**

Gross margin — pricing power made  
visible

**≈80%**

AI accelerator share; hyperscalers ≈50%  
of DC revenue



## The Moat

- CUDA: two decades of developer and framework lock-in
- NVLink & networking: system-level, not chip-level, competition
- Priority claim on CoWoS & HBM (~60% of TSMC CoWoS locked)



## Risk Register

- China DC revenue at zero (H2O export ban; \$4.5B write-down)
- Circular-financing exposure: up to \$100B pledged to OpenAI
- Custom ASICs: 2026 shipment growth 44.6% vs 16.1% for GPUs
- Valuation priced for flawless execution, quarter after quarter

## From Gaming GPUs to the Bedrock of AI

- CUDA launched 2006 — a 20-year software-ecosystem moat
- Pivoted to AI after AlexNet (2012); biggest GenAI winner since 2023
- Rubin platform ramps H2 2026; fresh \$80B buyback authorized

# The 2026 Flip: Anthropic Overtakes on Revenue & Valuation

## Anthropic

Run rate (mid-2026)

≈\$47B

Valuation (May 2026, as reported) ≈\$965B

**Revenue mix** Enterprise / API ≈80%; Claude Code at ≈\$2.5B run rate

**Unit economics** Revenue per MAU ≈\$192; paid conversion ≈13%

**Customers** 8 of the Fortune 10 run critical workloads on Claude

**Milestones** May 2026: \$65B Series H; confidential IPO filed Jun 1

VS

## OpenAI

Run rate (mid-2026)

≈\$25B

Valuation (Mar 2026, as reported) ≈\$852B

**Revenue mix** ChatGPT ≈\$17B + API ≈\$6.5B + Sora / licensing ≈\$1.5B

**User scale** ≈900M weekly, 1B+ monthly actives — king of consumer reach

**Profitability** Q1 2026 operating margin ≈ −122% (\$1.22 lost per \$1 of revenue)

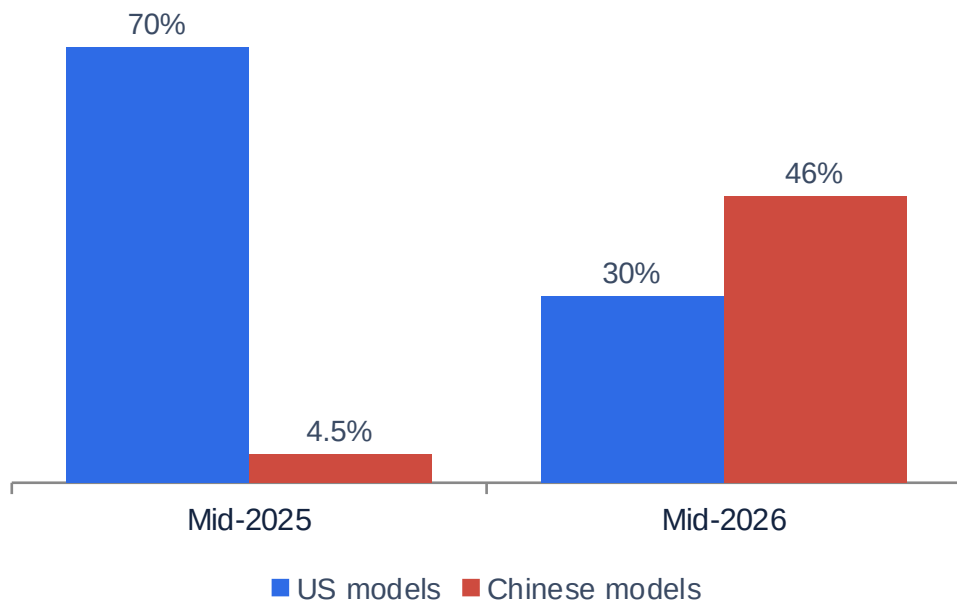
**Milestones** Confidential IPO filed Jun 8, 2026; targeting \$1T+ valuation

*Both are deeply unprofitable, and both filed confidential IPOs in June 2026 — the ultimate test of the business model moves to public markets.*

Source: CNBC (May 28, 2026); Sacra; The Information — private-company figures as reported, unaudited

# China's Route: Extreme Efficiency + Open Source at Scale

OpenRouter enterprise token share (%)



\* 46% is a single-week peak; US share fell 70% → 30% in a year



## DeepSeek

R1 matched o1-class at ≈\$5.6M training cost (MoE 671B / 37B active); Jan 27, 2025: NVIDIA –16.9% in a day, \$589B erased



## Alibaba Qwen

King of open source: 170k+ derivative models; #1 domestic enterprise token share at 32.1% (H2 2025)



## ByteDance Doubao

Wins on distribution: 21.3% domestic share; peak 63.3B tokens per minute on New Year's Eve



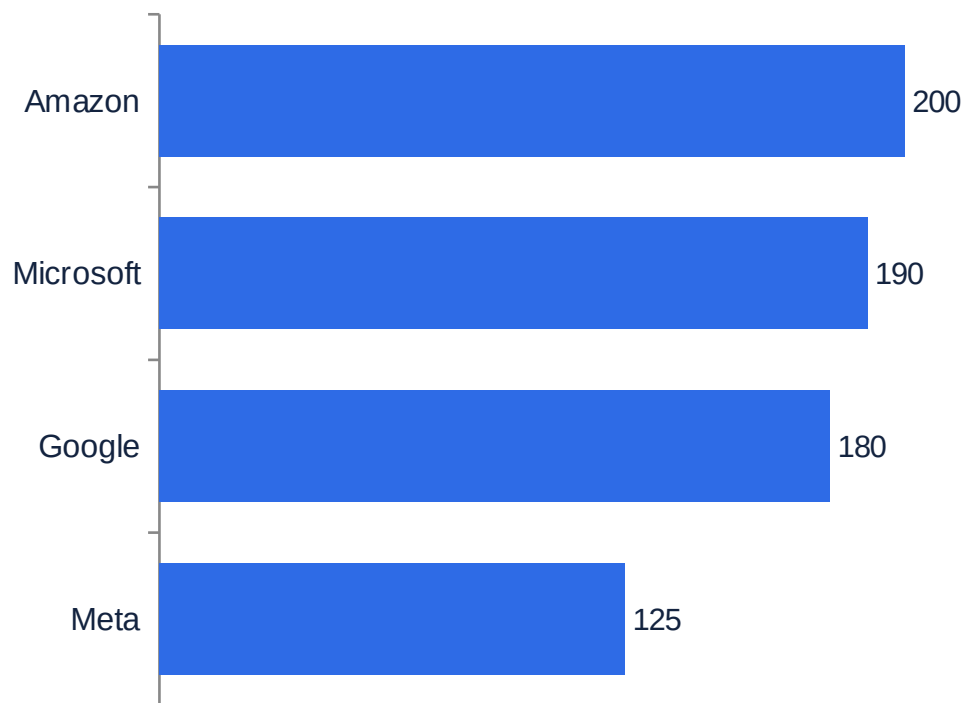
## Kimi / Zhipu

Kimi: fastest-ever decacorn (\$10–12B); Zhipu's GLM-5 open-sourced, coding demand rationed

**So what:** open source plus rock-bottom prices commoditize the base layer, dragging down global token prices.

# The 2026 Capex Arms Race: ≈\$725B Across the Big Four

## 2026 capex guidance (\$B)



\* Google / Meta shown at guidance midpoints (\$175–185B / \$115–135B)

## Order Books (RPO / Backlog) — Demand Visibility

|              |               |                                            |
|--------------|---------------|--------------------------------------------|
| Oracle       | <b>\$638B</b> | OpenAI ≈54% (\$300B five-year contract)    |
| Microsoft    | <b>\$627B</b> | Commercial RPO, +99% YoY                   |
| Google Cloud | <b>\$460B</b> | Roughly doubled QoQ; “compute-constrained” |
| AWS          | <b>\$364B</b> | Annualized run rate \$150B                 |



## Apple: The Edge-AI Watchpoint

Apple Intelligence lags; whether AI phones and PCs spark an upgrade supercycle is the key test of the edge story. Meanwhile Microsoft AI runs at ≈\$37B annualized; Copilot 20M+ seats (3–4% penetration).

# The Bottleneck Migrates: Chips → Packaging, Memory, Power



## Advanced Packaging: CoWoS

- TSMC at 125k wafers/month (+79%) — and still sold out
- NVIDIA locks ~60% of capacity; 2nm fully booked
- Packaging binds tighter than wafer supply itself



## HBM: High-Bandwidth Memory

- All three makers sold out for 2026; \$54.6B market (+58%)
- Share: SK Hynix ≈53% · Samsung ≈35% · Micron ≈11%
- HBM4 in mass production since Feb 2026; rising BOM share



## Power: The Final Constraint

- Data center demand: 460 TWh (2024) → ≈1,300 TWh (2035E)
- PJM interconnection queue >2,600 GW; 5+ year waits
- Power has replaced chips as the #1 limit on growth

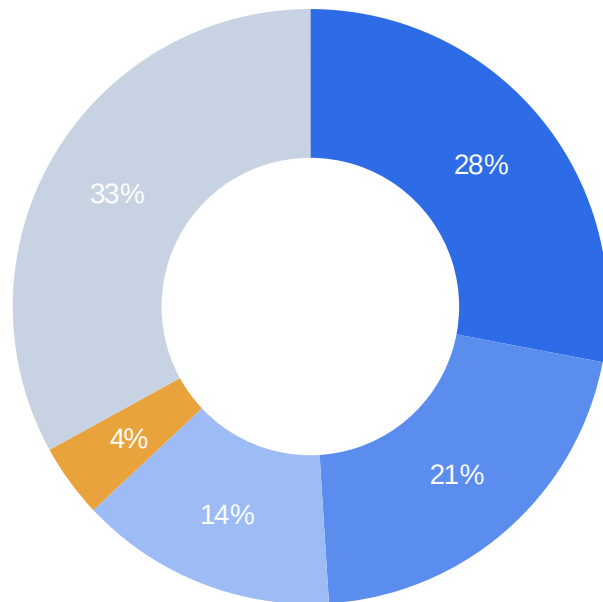


## Nuclear Returns (incl. SMRs)

- Big-four hyperscalers have signed ≈10 GW of nuclear
- Microsoft: 20-yr PPA restarts Three Mile Island (835 MW, 2028)
- Amazon (X-energy) · Google (Kairos) · Meta (1–4 GW RFP)

# Cloud: A Stable Oligopoly, Plus AI-Fueled Neoclouds

Global cloud infrastructure share (Q1 2026, Synergy)



■ AWS ■ Azure ■ Google Cloud  
■ Oracle ■ Others

Big three combined: 63%



## CoreWeave — the neocloud template

2025 revenue \$5.1B (+170%); Q1 2026 backlog \$99.4B; Microsoft ≈67% of revenue — hypergrowth and hyperconcentration in one company.



## Oracle — extreme AI order concentration

OCI +93% (FY26 Q4); RPO of \$638B with OpenAI ≈54% (\$300B over five years) — single-customer risk made explicit.



## China's clouds: local silicon + open source

Alibaba, Tencent and Huawei pair Ascend-class domestic compute with Qwen/GLM open models — a value-plus-sovereignty play.

# The Profit-Pool Map: Deflating Tokens, Expensive Compute

Profit king

## Compute hardware

NVIDIA gross margin ≈75%; HBM and CoWoS sold out across the board

Cash cow

## Cloud & infrastructure

High margin but capital heavy; \$1.4T+ of combined backlog

Burn for share

## Model layer

Revenue exploding, losses deeper (OpenAI op. margin ≈ -122%)

Divergent

## Application layer

Margins eaten by API bills: Cursor at 30–50% gross; \$0.40–0.70 of each \$1 goes to models

## Token Price Deflation

**\$20 → \$0.40**

GPT-4-class, per 1M tokens (late 2022 → early 2026)

- ≈10x cheaper per year; ≈200x incl. efficiency (Epoch, a16z)
- Inference is ≈2/3 of AI compute; agents use 5–50x tokens
- Net effect: volume beats price — deflation entrenches compute

**Base case:** profits migrate gradually to applications over 3–5 years — but only apps with proprietary data, distribution and switching costs capture them; compute stays the surest winner.

# Porter's Five Forces Across the Four Layers



Highest walls

## Compute hardware

- Near-monopoly: CUDA plus priority capacity claims
- Substitution rising: cloud ASICs growing 44.6%
- Buyer power grows as clouds build own silicon



Stable oligopoly

## Cloud & infrastructure

- Extreme capital walls; big three at 63% share
- Weak vs NVIDIA upstream, strong downstream
- Neoclouds wedge in as AI-only specialists



Fiercest fight

## Model layer

- Open-source substitutes plus price wars; thinnest moats
- Rising capability keeps swallowing thin wrappers
- Forced downstream into apps and agents for margin



Barbell split

## Application layer

- Low entry walls: wrappers are easy to copy
- Distribution, data and switching costs create local monopolies
- Value accrues to deep workflow integrators

# Adoption Is Layered: Coding Leads, Enterprise ROI Lags



## Enterprise Reality: Broad, Not Deep

88%

of organizations now use AI routinely (McKinsey, Nov 2025)

≈7%

have fully scaled it; only ≈6% qualify as high performers

5%

of pilots deliver \$M-scale returns (MIT NANDA) — despite \$30–40B already invested

**Failures trace to data and integration, not models:** vendor-led deployments succeed ≈67% of the time; internal builds only ≈1/3.

# Five Storylines for the Next 3–5 Years

01



## Agents

From chat to autonomous task execution; 5–50x tokens per task creates a compute-demand flywheel — the cycle's master narrative.

02



## Multimodal & video

Sora · Kling · Gemini Omni; capability is maturing fast, monetization is still early.

03



## Embodied AI & robotics

\$34B+ raised in 2025; Figure at \$39B, Unitree filing a Shanghai IPO, PI shipping  $\pi$  models — dense pilots, still no “ChatGPT moment.”

04



## AI for Science

The AlphaFold path extends into drug discovery and materials — long cycles, high walls, deeply tied to compute.

05



## Edge AI & devices

Whether AI phones and PCs trigger an upgrade supercycle decides the AI story for Apple, Qualcomm and consumer electronics.

# Who Wins, Who Bleeds: The Impact Matrix

**Four tests:** workflow automatability · proprietary-data moat · distribution edge · customer switching costs



## Beneficiaries

- Compute & power chain: NVIDIA · SK Hynix · Vertiv · nuclear/SMRs
- Clouds: AWS · Azure · GCP · Oracle · neoclouds
- Data-plus-distribution apps: Microsoft · Palantir · Anthropic's enterprise base

*Common thread: supply bottlenecks, or irreplaceable customer assets*



## In Transition

- Adobe & legacy software: AI speed vs pricing-power defense
- Seat-based SaaS: forced toward outcome-based pricing
- IT services: from labor stacking to AI delivery, margins squeezed

*Key variable: disrupt yourself faster than you are disrupted*



## Under Pressure

- Labor-heavy IT services, outsourcing, content mills
- Chegg: ~99% from its 2021 peak; 636 laid off across two 2025 rounds
- Stack Overflow: monthly questions down >90% from peak

*Common thread: workflows replaced outright, no data flywheel*

**The litmus test: augmented by AI (data flywheel, owned distribution) — or replaced by it (pure execution, bypassed by direct answers).**

# Five Questions That Must Be Answered Head-On



## 1. Bubble vs. early innings

- Bull: \$1.4T+ cloud backlog; corporate cash flows 3x 1999 levels
- Bear: circular financing (NVIDIA → OpenAI up to \$100B; Microsoft 27%; OpenAI → AMD 10%) + MIT's 5%
- Call: not a 2000-style bubble, but the gap is systemic risk



## 2. Power & capacity limits

- PJM queue >2,600 GW with 5+ year waits
- Power has replaced chips as the binding constraint
- Nuclear / SMRs won't deliver until 2028 or later



## 3. US–China chip controls

- NVIDIA's China DC revenue already at zero
- Ascend 910C targets 600k units; state buying excludes NVIDIA
- China scales out on domestic parts, HBM-constrained



## 4. Open source vs. closed

- Chinese open models hit a 46% weekly peak of enterprise tokens (4.5% a year earlier)
- The base layer is commoditized; price wars are the norm
- Closed labs push value into apps, enterprise trust and agents



## 5. Regulation

- EU: GPAI rules live since Aug 2025; high-risk items pushed to 2027–28
- US: Dec 2025 executive order asserts federal primacy over state laws
- China: content labeling in force; 796 GenAI services registered

# 2026–2030: Bull / Base / Bear



## Bull

Agents deliver

- Agents scale; enterprise ROI turns positive and measurable
- Capex is absorbed by real revenue; AI spend heads toward trillions
- Profits migrate healthily to apps; new platform giants emerge

**Implication: applications and full-stack platforms gain most**



## Base

Growth downshifts

- Demand persists at a slower clip; profits stay in compute & cloud
- Leveraged neoclouds and me-too startups get washed out
- Regulation tightens mildly; open and closed coexist

**Implication: shovel-sellers and data/distribution apps hold up**



## Bear

The gap bites

- Credit tightens; circular financing unwinds, capex slams the brakes
- At least one major infrastructure player blows up (the Fink warning)
- Valuations reset hard — but demand doesn't go to zero, unlike 2000

**Implication: leveraged infrastructure is most exposed**

**The switch signal:** if 2027 hyperscaler capex keeps accelerating (consensus sees \$1T+), bull/base holds; a collective guide-down flips the odds to bear.

# Where Value Settles — and What to Track

## Near term · 0–12 months

The most certain beneficiaries are the shovel-sellers: compute (NVIDIA / ASIC chain), HBM, power & cooling — the highest-visibility links.

## Mid term · 1–3 years

App-layer value accrues to proprietary data, distribution and switching costs; generic wrappers and answer-engine roadkill are most exposed.

## Long term · 3–5 years

Watch ROI: if profits truly migrate, value shifts toward applications; if the capex–revenue gap persists, leveraged infrastructure is most exposed.

**Core conclusion:** value settles first and thickest in compute and cloud; application-layer winners will be decided by data, distribution and switching costs.



## High-Frequency Tracking List

- Big-four quarterly capex and 2027 guidance (past \$1T?)
- NVIDIA data center revenue (and any China revival)
- HBM / CoWoS capacity and utilization
- Model API price curve (speed of token deflation)
- Frontier benchmark progress (SWE-Bench and peers)
- Enterprise AI spend & ROI surveys; backlog conversion

**What would change the call:** a collective capex guide-down = risk realized; a jump in measurable enterprise ROI = migration accelerating; API prices below self-hosted breakeven = closed-source squeeze deepens.