EXECUTIVE SUMMARY · THE AI ECONOMY IN NUMBERS

The State of the AI Economy

Key figures from Exponential View's inaugural report — the first bottom-up, deduplicated measure of AI demand, spend, infrastructure and tokens · June 25, 2026

GENERATIVE AI ECONOMY REVENUE — DEDUPLICATED, $BN / YEAR

Jan 2023 → Jun 2026 · global ex-China

1 REVENUE & DEMAND

$175B
Annualized revenue run rate

$110B
Banked trailing 12-month revenues

1.6x
Run-rate vs trailing gap — the spread is the growth rate

$22.7T
Stock-market value underpinned by AI

REAL EXTERNAL CUSTOMERS · EVERY $ DEDUPLICATED

Q1 2026 quarterly revenue	$25B
Quarterly depreciation to cover	$21B

2 UNPRECEDENTED SPEED

3x
Faster than any prior IT wave

90x
Faster to each new $1B of revenue

TIME TO ADD $1B OF CUMULATIVE REVENUE

2023	180 days
Today	< 2 days

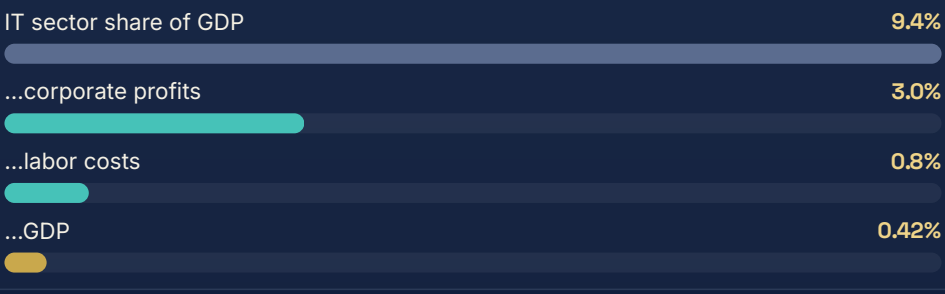
Growth has held at **35% QoQ** across every adoption phase: chatbots → subscriptions → agentic coding.

3 SHARE OF THE ECONOMY

3x
AI revenue / GDP vs Q1 2025 (0.13%)

10x
vs Q1 2024 (0.04%) — still small, still early

GLOBAL AI REVENUE (EX-CHINA) RELATIVE TO US...



4 CAPEX — THE BIGGEST BUILDOUT IN TECH

$2.0T
Cumulative CapEx through 2026E

$535B
Above pre-AI trend

$111B
2026E depreciation

HYPERSCALER CONTRACT BACKLOG (RPO)

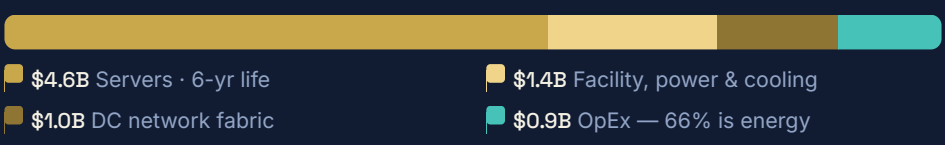
Q4 2025: quarterly AI revenues **first exceeded depreciation**
Q1 2026 headroom: **19%** infra-only · **32%** all GenAI

5 ECONOMICS OF 1 GW OF AI CAPACITY

$7.9B/yr
Cost to own & operate 1 GW (480,000 GPUs)

75–85Q
Quadrillion tokens produced per GW per year

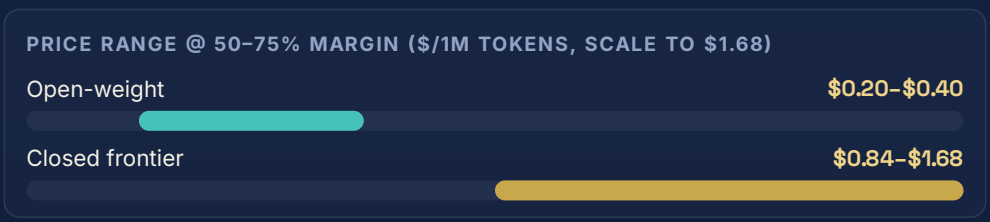
WHERE THE $7.9B GOES



Capital costs are **89%** of the total; OpEx just 11%. Data center economics set the hurdle for token pricing.

6 TOKEN PRICING PER 1M TOKENS

	Open-weight (Kimi K2.5)	Closed licensing (illustrative)
Cost for inference provider	$0.10	$0.42
Price @ 50–75% gross margin	$0.20–$0.40	$0.84–$1.68
Customer value for 25% ROI	$0.25–$0.50	$1.05–$2.10



Blended market price fell **$17 → $2** per 1M tokens since 2023 while capability rose (Epoch Index 112 → 158). Divide $7.9B/yr by token output — that's the pricing floor.

POWER

+9 TWh/mo

Annual growth in US electricity generation since 2024

vs **±0** during 2008–2024 and +6 TWh/mo during 1950–2008. AI is reigniting a moribund power sector.

COMPUTE SUPERCYCLE

66% → 80%

Global compute CAGR — AI bent a 50-year trend upward

The largest data centers grew **50x** in four years.

NVIDIA & THE GRID

$31B → $95B

Nvidia supply commitments in one year

Expected extra US grid load by 2030: **24 → 166 GW** (~7x since 2022); data centers = ~55% of growth.

BUILD-COST MIX

40% → 60%

Chips' share of DC build cost, 2021 → 2026E

Memory is the biggest mover: **2% → ~18%**. Each dollar buys more silicon, less concrete.

TOKENS

30Q / month

Global inference token volume (quadrillions, incl. China)

Growing **14x year-on-year**, propelled by agentic workloads.

DEMAND ELASTICITY

1.2 – 1.8

Price elasticity of token demand

Every 10% price cut → **12–18% more tokens** — total spend still rises as prices fall.

CHAT → AGENTS

12 → 36

Tokens processed per output token — reasoning & agents multiply consumption

Total tokens **+39x** vs output tokens +30x (Jan 2025 → Apr 2026).

THE STACK — **WHERE THE VALUE IS CAPTURED**

The AI Economy Value Chain — Singular Link Portfolio Map

The stack turns capital and energy into cognitive work: chips convert energy to tokens, hosting converts CapEx to token OpEx, models convert tokens to intelligence, apps convert intelligence to customer value. Gold nodes mark Singular Link's current & pipeline positions.

01 Power · 02 Chips · 03 Compute · 04 Network · 05 Memory · 06 Inference · 07 Models · 08 Evals · 09 Agents · 10 Apps

01 Power

EXAMPLES

Utilities, gas turbines, SMR nuclear, fusion

SINGULAR LINK — CURRENT & PIPELINE

Type One Energy (stellarator fusion)

02 Chips

EXAMPLES

Nvidia · AMD · TSMC

SINGULAR LINK — CURRENT & PIPELINE

Cerebras (wafer-scale, Nasdaq: CBRS) · **Cortical Labs** (biological compute)

03 Compute

EXAMPLES

CoreWeave · Nebius · AWS · GCP

SINGULAR LINK — CURRENT & PIPELINE

Fluidstack (AI GPU cloud) · **Starcloud** (orbital data centers) · **FluidCloud** (multicloud infra software)

04 Networking

EXAMPLES

Arista · Broadcom · Cisco

SINGULAR LINK — CURRENT & PIPELINE

Ayar Labs (optical I/O) · **Lightmatter** (silicon photonics)

05 Memory

EXAMPLES

Micron · SK Hynix · Samsung

SINGULAR LINK — CURRENT & PIPELINE

— open

06 Inference / Hosting

EXAMPLES

RunPod · Together AI · Fireworks

SINGULAR LINK — CURRENT & PIPELINE

— served via Fluidstack GPU capacity

07 Foundation Models

EXAMPLES

Anthropic · OpenAI · Google DeepMind

SINGULAR LINK — CURRENT & PIPELINE

Reka AI (multimodal frontier) · **Hologen AI** (Large Medicine Models)

08 Evals

EXAMPLES

Scale AI · Epoch AI

SINGULAR LINK — CURRENT & PIPELINE

— open

09 Agents

EXAMPLES

OpenAI · Anthropic

SINGULAR LINK — CURRENT & PIPELINE

Lucid Capital (autonomous AI trading agents)

10 Apps & Human Interface

EXAMPLES

Cursor · Harvey · OpenRouter

SINGULAR LINK — CURRENT & PIPELINE

NovaLexi (AI IP management, KSA) · **Synchron** (endovascular BCI) · **Evolve Neuro** (neurotech)

Stack dynamics per the report: revenue is concentrated but the mix is shifting toward apps and models. Labs hold pricing power only at the frontier — last year's frontier commoditizes quickly into open weights. Among OpenRouter users, the Google + OpenAI + Anthropic token share fell from **72% to 33%**. H100 rentals recovered from $1.70 (overbuild fears) to **$2.40/hr** on surging inference demand.

KEY CONCLUSIONS

- ✓ Demand is **real, big and fast** — more revenue-validated than any prior platform shift, driven by external paying customers.
- ✓ The biggest buildout in tech history is **paying back (for now)**: revenues first cleared quarterly depreciation in Q4 2025, again in Q1 2026.
- ✓ Coverage remains thin — depreciation still absorbs **68–81%** of GenAI revenue before power, labor and financing.
- ✓ The marginal AI-infra dollar is increasingly **externally financed**, moving risk outside the hyperscalers.
- ✓ Value is moving **up the stack** toward apps and models; chat → agents is multiplying token use.
- ✓ Tokens are AI's billing metric but **not yet its unit of value** — quality-adjusted output tokens come closest.

BIGGEST WINNERS THIS CYCLE

1. Compute providers & GPU clouds
2. Memory & high-bandwidth memory (2% → ~18% of build cost)
3. AI chips (inference-optimized)
4. Networking & silicon photonics
5. Inference & efficiency software
6. RL / post-training & data engines
7. AI applications (enterprise & dev) — gaining share of stack revenue

BIGGEST RISKS

- × The depreciation base rises as committed CapEx enters service — growth, utilization and pricing must compound or headroom compresses.
- × Open-weight models commoditize last year's frontier, squeezing lab margins and generic "wrapper" pricing power.
- × Debt-heavy external financing (especially neoclouds) moves repayment risk to third parties.
- × Physical constraints: grid connection queues, power and land don't scale exponentially.
- × Falling token prices may not move enough volume to earn a return on $2T of CapEx.

FOUR SCENARIOS FOR WHAT'S NEXT

- **Open-weight models catch up:** tokens get cheaper to consume and serve; labs lose licensing fees.
- **General-purpose models absorb the app layer:** labs take on integration work; apps defend with proprietary data and domain workflows.
- **Reduced compute needs:** smaller models and efficient compute grow the addressable market; some hosting moves to the edge.
- **All three combined:** a perfect storm of efficiency — every layer compresses and **consumers capture the surplus**.

This is not a bubble — it is a rebuild of the world's computing infrastructure, and demand is more revenue-validated than any prior platform shift.

The investment case comes down to whether falling prices can move enough token volume to earn a return on CapEx.

Source: **Exponential View — The State of the AI Economy** · June 25, 2026 · Azeem Azhar, William Gildea, Hannah Petrovic PhD, Nathan Warren & Marija Gavrilov · intelligence.exponentialview.co · Figures: global ex-China, deduplicated (tokens incl. China)