The Structural Disequilibrium of AI Economics

The Consolidation Phase

3 Pilgrim LLC
Independent Research

Abstract

Artificial intelligence economics is entering a structural inflection point. Current architectures integrate two incompatible regimes: a capital-intensive training model and a low-margin inference model. This vertical integration creates reflexive instability—training demands hyperscale investment while inference cannot amortize those costs under intermittent, low-utilization workloads. Physical constraints compound the imbalance: datacenters operate near steady-state power draw even when GPUs idle, converting capital into heat rather than productive work. The result is negative-return scaling: larger clusters yield diminishing economic efficiency despite rising energy and depreciation costs. This paper argues that equilibrium requires bifurcation—training infrastructure will consolidate into utility-like providers, while inference matures into a service economy optimized for throughput and latency. The consolidation phase will mark normalization, not decline: a shift from narrative-driven expansion to rational pricing and structural specialization.

Keywords: AI Economics, Structural Disequilibrium, Training vs Inference, Capital Intensity, Utilization Efficiency, Negative-Return Scaling, Reflexivity, FOMO Economics, Bifurcation, Compute Economics, Energy Elasticity, AI Infrastructure, Consolidation Phase, Economic Equilibrium, Marginal Productivity

Correspondence:
https://3pilgrim.com/contact

Recommended Citation:

3 Pilgrim LLC (2025). The Structural Disequilibrium of AI Economics. Working Paper. Available at: https://3pilgrim.com/

NOTICE:
This work is intended for conceptual and educational purposes only. All examples are illustrative and based on publicly available data or stylized assumptions. No statements herein should be interpreted as predictions or financial advice regarding any specific company or entity. This document may contain errors or inaccuracies; readers should verify all information independently.

1. Introduction — The End of Indefinite Scaling

For more than a decade, the organizing principle of artificial-intelligence research has been simple: larger models perform better. The empirical “scaling laws” identified by Kaplan et al. (2020) provided a quantitative basis for that intuition, linking model loss inversely to total compute through a power-law relation. This logic—more data, more parameters, more GPUs—became the guiding strategy for both research and investment.

But scaling cannot proceed indefinitely. Physical constraints at the transistor, thermal, and synchronization levels now cap feasible cluster sizes; economic constraints cap the willingness to finance them. The theoretical frontier has flattened into a practical ceiling. What remains is not a race toward new discovery but a process of rationalization: the search for balance between cost, capacity, and sustained economic return.

This paper begins from that inflection point. It treats the AI sector not as a mystery of emergent intelligence but as an industrial system subject to the same cyclical laws that have governed every prior technology boom.

2. The Dual Economies of AI

Modern AI operates through two fundamentally different economic logics.

Dimension	Training Regime	Inference Regime
Economic Type	Capital-intensive R&D	Demand-driven service
Cost Behavior	High fixed, low variable	Low fixed, high variable
Revenue Source	Internal investment or compute resale	End-user consumption (APIs, subscriptions)
Frequency / Scale	Infrequent, project-based	Continuous, mass-market
Business Analogy	Semiconductor fabrication	Cloud or SaaS distribution
Optimization Goal	Minimize model loss	Maximize throughput per watt / dollar
Efficient Market Form	Few firms, high barriers	Many firms, price competition

The two regimes are orthogonal. Training behaves like a heavy-industry process—few customers, high sunk cost, long depreciation cycles. Inference behaves like a consumer service—many customers, elastic demand, thin margins. Integrating them under one balance sheet creates structural tension: the capital allocation and risk profile appropriate to training are incompatible with the throughput and pricing dynamics of inference.

3. The Physics of Idle Capital

Empirically, this tension manifests as persistent under-utilization. Once a model is trained, the computational load required for inference drops by roughly an order of magnitude. Public data from hyperscale deployments suggest sustained utilization between 5 and 15 percent of installed GPU capacity during inference periods. The remaining hardware, along with its cooling and power infrastructure, remains energized but idle.

This phenomenon is not a transient inefficiency; it is a physical consequence of the asymmetry between burst-like training cycles and intermittent inference demand. Power and thermal systems at this scale are engineered for stability, not elasticity. Variable cooling, staged power delivery, and selective shutdown are economically and mechanically prohibitive. The datacenter therefore operates near steady-state draw regardless of computational load, converting capital investment into heat rather than productive work.

The result is idle capital: billions of dollars in depreciating equipment consuming energy without proportional output. In macroeconomic terms, the industry has created a stock of fixed assets whose marginal productivity is declining even as total investment accelerates—a classical symptom of structural oversupply.

4. The Inevitability of Bifurcation

The resolution to structural oversupply is specialization. History provides precedent: the semiconductor industry of the 1980s faced the same disequilibrium. Integrated device manufacturers combined design and fabrication under one roof; costs ballooned; margins collapsed. Equilibrium was restored only when the market bifurcated—design firms went fabless, and dedicated foundries (TSMC, UMC) absorbed the capital burden as independent suppliers.

The same logic applies to artificial intelligence.

> **Proposition 1.**
> When two orthogonal economic regimes—a high-fixed-cost, low-frequency training regime and a low-margin, high-frequency inference regime—are vertically integrated, the composite system becomes reflexively unstable. Equilibrium is restored only through separation into distinct functional markets.

Training infrastructure behaves like semiconductor fabrication: capital-intensive, cyclical, and technology-driven. Inference behaves like software or consumer electronics: demand-driven, commoditized, and margin-sensitive. Forcing both onto a single corporate balance sheet obliges investors to fund two incompatible cost functions.

Bifurcation transforms this conflict into efficiency:

- Training entities focus on process yield, depreciation, and throughput—selling compute capacity to any model vendor, regardless of brand.
- Inference entities purchase that capacity as an operating expense, optimizing user experience, latency, and application value rather than hardware scale.

In this arrangement, specialization replaces reflexivity. Capacity is priced by utility, not by expectation.

5. Reflexivity and the FOMO Incentive

The persistence of disequilibrium is not irrational; it is reflexive.
In the present structure, each new expansion raises corporate valuation, which in turn justifies further expansion. Capital begets scale; scale begets narrative; narrative begets capital. The system sustains itself through FOMO economics—the fear of missing the next discovery.

Cross-subsidization by trillion-dollar conglomerates further masks inefficiency. Losses incurred by AI divisions are offset by profits elsewhere, permitting continued growth despite declining marginal productivity. The market interprets this persistence as validation rather than distortion.

This dynamic is self-reinforcing until an external constraint—economic, physical, or regulatory—interrupts it. At that point, the cycle resolves not through collapse but through specialization: the emergence of independent training utilities and cost-efficient inference providers. Every prior technology boom has ended this way—from mainframes to semiconductors to cloud computing.

6. Toward the Consolidation Phase

The disequilibrium described here marks the close of AI's expansionary era. The sector's next frontier is not scale but equilibrium. The economics of artificial intelligence are converging toward the same structural resolution that every capital-intensive industry eventually reaches: consolidation, specialization, and rational pricing.

Training and inference will separate. The former will stabilize as a utility industry governed by throughput and depreciation; the latter will mature into a service economy governed by efficiency and differentiation. The transition will not signal decline—it will mark normalization.

The frontier of artificial intelligence has moved.
It no longer lies in building larger models, but in aligning physical capacity, economic structure, and real demand.
What comes next is not another leap in size, but the consolidation phase.

Appendix A: Utilization and Power Calculations

A.1 Framework and Definitions

To understand the physical and economic imbalance between training (capital expenditure) and inference (revenue generation), we first define a minimal formal structure linking compute utilization, energy consumption, and throughput. The core objective is not precision forecasting but bounded reasoning: to show that small changes in utilization or architectural balance yield orders-of-magnitude differences in efficiency and viability.

Let:

- N_{GPUN} = number of deployed GPUs
- P_{GPU} = average power draw per GPU (Watts)
- U = sustained utilization (fraction of peak load, e.g., 0.10 = 10%)
- PUE = power usage effectiveness (ratio of total site energy to IT energy)
- H_{year} = annual hours of operation = 8,760
- E_{site} = annual site energy use (kWh or TWh)

Then the site energy can be approximated as:

$$E_{site} = N_{GPU} \times P_{GPU} \times U \times PUE \times H_{year}$$

This gives a baseline energy envelope for a given configuration and utilization regime.
Because inference workloads are bursty and demand-driven, U for inference typically ranges from 5–15%, whereas training workloads sustain 70–90% during epochs that last weeks or months. The delta between these modes is the economic source of inefficiency: inference fleets idle far more often than training clusters, yet remain electrically and thermally active.

A.2 Case 1: xAI Colossus Cluster (Memphis, 2025)

Inputs

N_{GPU} = 200,000 $(H100/H200 - class)$

P_{GPU} = 700 W (liquid-cooled SXM average)

PUE = 1.4 (fleet-level, accounting for cooling and networking)

U = 0.10 (inference baseline)

H_{year} = 8,760 h

Computation

E_{site} = $200{,}000 \times 700 \times 0.10 \times 1.4 \times 8{,}760 = 1.717 \times 10^{10}\ kWh = 17.17\ TWh/yr$

That is, at 10% sustained utilization, the Colossus cluster would consume roughly 17 TWh annually, equivalent to the electricity use of ~1.5 million U.S. homes. At 80% training utilization, the same cluster would rise to 137 TWh, illustrating the enormous energy spread driven purely by workload composition.

A.3 Case 2: ChatGPT Fleet (OpenAI, 2025)

This example uses the **query-driven approach**, linking FLOPs per query to physical compute.

Given parameters:

- $Q = 2.5 \times 10^9$ queries/day
- $\tau = 50$ tokens/query
- $P_m = 1.76 \times 10^{12}$ parameters (GPT-4o scale)
- $f_{op} = 2$ FLOPs/param/token
- $F_{GPU} = 60 \times 10^{12}$ FLOPs/s per GPU
- $P_{GPU} = 400\ W$ (inference load average)
- $PUE = 1.25$

FLOPs per query:

$$F_{query} = \tau \times P_m \times f_{op} = 50 \times 1.76 \times 10^{12} \times 2 = 1.76 \times 10^{14}$$

Total daily FLOPs:

$$F_{day} = Q \times F_{query} = 2.5 \times 10^9 \times 1.76 \times 10^{14} = 4.4 \times 10^{23}$$

GPU-seconds per day:

$$GPU_{sec/day} = F_{day}/F_{GPU} = \frac{4.4\times10^{23}}{60\times10^{12}} = 7.33 \times 10^9$$

Concurrent GPUs:

$$NGPU, active = \frac{GPU_{sec/day}}{86{,}400} = 84{,}800$$

IT Power:

$$P_{IT} = N_{GPU,active} \times P_{GPU} = 84{,}800 \times 400 = 33.9\ MW$$

Site Power (with PUE):

$$P_{site} = 33.9 \times 1.25 = 42.4\ MW$$

Annual Energy:

$$E_{site} = 42.4 \times \frac{8{,}760}{10^6} = 372\ GWh/yr = 0.372\ TWh/yr$$

This aligns closely with top-down financial proxies (based on $/query cost), supporting a utilization range between **10–15%**, consistent with smaller benchmark clusters.

A.4 Sensitivity Analysis

Let U vary from 5% to 15%. Holding $N_{GPU}, P_{GPU}\ and\ PUE$ constant:

$$E_{site}(U) \propto U$$

Thus, for Colossus:

Utilization (U)	Annual Energy (TWh)	Relative Efficiency
5%	8.6	0.5× baseline
10%	17.2	1.0× baseline
15%	25.8	1.5× baseline
80% (training)	137.3	8× baseline

This proportionality exposes the pathological energy elasticity of large-scale inference.
Even modest overbuilding multiplies fixed energy costs—because cooling, pumping, and base infrastructure cannot scale linearly down. Idle GPUs still radiate heat and draw power for basic cycling and synchronization

A.5 Linking Physical and Economic Disequilibrium

From these calculations, several principles emerge:

1. **Structural Inefficiency:**
   Inference operates in the 5–15% utilization band, while fixed site infrastructure (power, cooling, real estate) remains ~80–90% active. This creates a large negative leverage between energy cost and economic output.
2. **Capital Misallocation:**
   Training and inference are orthogonal economic functions:
   - Training = low frequency, high CAPEX, high specialization.
   - Inference = high frequency, low margin, scale-dependent.Merging them under one corporate balance sheet forces capital to chase both high-end specialization and mass-service elasticity—an inherently unstable model.
3. **Energy Irreversibility:**
   Because steady-state thermodynamics dominate data center design, energy use cannot

drop linearly with demand. Thus, even during off-peak hours, power draw remains near constant, invalidating "elastic compute" economics at hyperscale.

4. **Analogy to Fabrication Economics**:
   This mirrors the semiconductor industry before bifurcation: fabs (capital-intensive, slow-cycle) were decoupled from design firms (fast-cycle, innovation-driven). A similar split between training infrastructure providers and inference operators is the natural equilibrium outcome.

A.6 Conclusion

The utilization and power calculations illustrate the boundary condition: training drives peak demand, inference sustains minimal utilization, and the two together form a structurally loss-making loop under a single-entity model. This disequilibrium explains why AI infrastructure appears to defy standard return-on-capital expectations—because the real physical utilization lags the financial narratives of growth. Only architectural and corporate separation can restore balance.

Appendix B — Stylized Economic Equilibrium

The following model is not intended as a predictive financial analysis, but as a conceptual framework illustrating the structural imbalance between cost and utilization in large-scale AI systems.
It abstracts the operating dynamics of an AI enterprise into a minimal algebraic form sufficient to expose the disequilibrium mechanism.

B.1. Framework

Let:

U = average utilization rate (fraction of installed GPU capacity actively used)

R_q = average revenue per query (USD/query)

Q = total number of queries processed per period

D = depreciation of hardware and infrastructure per period

C_t = training cost amortized over the model's useful life

O = operating expense per period (labor, energy, networking)

Then:

$$\Pi = (U \cdot R_q \cdot Q) - (D + Ct + O)$$

where Π is the net operating profit per period.

B.2. Derivatives and Sensitivities

Differentiating with respect to key parameters yields:

$$\frac{\partial \Pi}{\partial U} = R_q Q > 0$$

$$\frac{\partial \Pi}{\partial C_t} = -1$$

$$\frac{\partial \Pi}{\partial R_q} = UQ$$

Interpretation:

Profitability scales linearly with utilization; however, U is empirically bounded at ~10–20%.

Training cost C_t is fixed and periodic, introducing a negative offset that cannot be amortized effectively when utilization is low.

Revenue per query R_q is constrained by market competition and user willingness to pay; its elasticity is minimal.

Thus, under realistic conditions:

$$\frac{\partial \Pi}{\partial U} \gg \partial \Pi / \partial R_q$$

meaning utilization dominates all other drivers of profitability, yet is the one variable least amenable to improvement under current architectures.

B.3. Structural Implication

If U remains below the breakeven threshold:

$$U *= \frac{D + C_t + O}{R_{qQ}}$$

then the enterprise operates at structural loss regardless of incremental revenue growth. Empirically, with $U \approx 0.1$ and R_q on the order of $(\$10^{\{-3\}}\ to\ \$10^{\{-4\}})$ per query, U^* is unattainable without drastic cost reduction or architectural change.

Because training cost (C_t) and depreciation (D) scale superlinearly with model size, and inference revenue ($R_q Q$) scales sublinearly with user adoption, the equilibrium cannot close through volume alone.
This produces a negative-return scaling regime—larger models yield lower return on capital even if total usage rises.

B.4. Economic Interpretation

The algebraic form reveals the fundamental disequilibrium:

Training costs behave as front-loaded capital expenditures with long payback periods.

Inference revenues behave as low-margin, volume-dependent cash flows with diminishing marginal yield.

Utilization, the only lever connecting the two, is physically capped by the intermittency of inference workloads and the inflexibility of datacenter infrastructure.

In such a system, no combination of scale or price can reconcile cost and revenue. Equilibrium requires structural change — either by reducing fixed costs (outsourced training) or increasing utilization through architectural specialization (dedicated inference providers).

B.5. Comparative Note

This pattern mirrors the structural transformation of semiconductor economics in the early 1990s. Integrated device manufacturers reached negative marginal returns at low fab utilization, triggering the rise of fabless design firms and independent foundries.
The same algebra applied: capital intensity outpaced throughput, forcing bifurcation between high-fixed-cost producers and low-fixed-cost consumers of fabrication services.

B.6. Summary Statement

When U is low and fixed costs dominate, the system's steady state is negative return.The only path to equilibrium is structural specialization, not scale

---

7. Appendix C: Query-Level Cost Modeling (Financial Sensitivity)

C.1 Purpose

To complement the physical-utilization analysis of Appendix A, Appendix C translates compute performance and energy consumption into financial efficiency metrics. It evaluates the cost dynamics of AI workloads — showing that the marginal cost per query remains high and non-scalable under low utilization.

C.2 Framework

We define:

- C_{GPUh}: cost per GPU-hour (e.g., \$1.50 – \$2.00 on cloud H100 instances)
- Q: queries/day
- c_q: cost per query (\$/query)
- H_{GPU}: total GPU-hours/day
- U: utilization fraction
- N_{GPU}: number of deployed GPUs

Core relationships:

$$H_{GPU} = N_{GPU} \times 24 \times U$$

$$c_q = \frac{H_{GPU} \times C_{GPUh}}{Q}$$

C.3 Worked Example 1 — OpenAI (ChatGPT, 2025)

(Based on Method B in Appendix A)

$Q = 2.5 \times 10^9$ queries/day

$$C_{GPUh} = \$1.50$$

$$N_{GPU} = 85{,}000$$

$$U = 0.10$$

$$H_{GPU} = 85{,}000 \times 24 \times 0.10 = 204{,}000$$

$$c_q = \frac{204{,}000 \times 1.50}{2.5 \times 10^9} = 1.22 \times 10^{-4} = \$0.00012$$

That is, \$0.0001 per query just in GPU time — before overhead, networking, or profit margin. At 15% utilization, c_q falls only to \$0.00008, confirming the weak scaling with demand.

C.4 Worked Example 2 — xAI (Colossus, 2025)

(Using same parameters as Appendix A)
If Colossus runs 200,000 GPUs at 10% utilization, even without knowing Q we can invert to find the break-even query volume for $0.001/query:

$$Q = \frac{H_{GPU} \times C_{GPUh}}{c_q} = \frac{200{,}000 \times 24 \times 0.10 \times 1.50}{0.001} = 7.2 \times 10^9 \; queries/day$$

Thus, Colossus must process ≈ 7 billion daily queries to achieve a $0.001 marginal cost per query — several times OpenAI's 2025 load — simply to break even on GPU time alone.

C.5 Sensitivity Table

Utilization	Cost/query ($, GPU-only)	Notes
5%	0.00024	Grossly underutilized; majority of cost sunk in idle power
10%	0.00012	Baseline in main text
15%	0.00008	Upper realistic bound
50%	0.000024	Approaches physical limits; requires near-constant demand

C.6 Interpretation

1. Economic Compression:
   Cost per query scales inversely with utilization, not with hardware size. Adding GPUs without proportional load increases unit cost.
2. Revenue–Cost Mismatch:
   Inference demand (millions of lightweight queries) cannot absorb CAPEX and OPEX designed for training-scale power budgets.
3. Path to Equilibrium:
   Economic efficiency requires bifurcation of business models:
   - Training clusters as high-CAPEX, low-frequency service providers (fab-like).
   - Inference operators as high-volume, thin-margin distributors.