

Structural Imbalances in the Generative AI Market: Inference Unit Economics, the Aggregate CapEx–Revenue Gap, and a Stylized Liquidity-Depletion Model

AUTHORS: Ilya A. Dushin

SUMMARY: This working paper examines the financial health and sustainability of the generative artificial intelligence (GenAI) and large language model (LLM) industry as of 2026. The author develops a quantitative framework to assess the internal consistency of the capital-formation cycle by analyzing the cost-to-revenue ratio, the capital expenditure (CapEx) and revenue gap, and the efficiency-adjusted demand identity. The paper aims to determine if the observed expenditure trajectory is internally consistent with the observed revenue trajectory.

METHOD: The study uses a bottom-up engineering cost model to estimate the cost-to-revenue ratio for inference of ultra-large models. It also employs a two-equation exponential divergence model to calibrate infrastructure CapEx and end-market AI revenue. Additionally, an efficiency-adjusted demand identity is derived to analyze the relationship between the growth rate of tokens-per-accelerator efficiency and token demand. The findings are combined with a Cash Flow-to-CapEx Sustainability Ratio and publicly disclosed intra-ecosystem investment and procurement commitments to characterize the conditions under which the observed expenditure trajectory ceases to be organically fundable.

KEY FINDINGS:

- The implied cost-to-revenue ratio for inference of ultra-large models is approximately 3,950× under unbatched single-stream execution, falling to a range of roughly 30×–100× when dynamic batching and low-precision quantisation are accounted for.
- An annual funding gap in excess of USD 600 billion is implied by a two-equation exponential divergence model of infrastructure CapEx and end-market AI revenue.
- When the growth rate of tokens-per-accelerator efficiency exceeds the growth rate of token demand, net physical accelerator demand contracts, a phenomenon labeled the inverted Jevons regime.

IMPLICATION FOR TRADING: The findings suggest that the generative AI market is facing significant structural imbalances, with a large gap between capital expenditure and revenue. This could indicate potential liquidity issues and a need for external financing. Traders should be cautious of the high cost-to-revenue ratios and the potential for liquidity depletion, which could lead to market instability. Understanding these dynamics can help traders anticipate shifts in the market and adjust their strategies accordingly.

Structural Imbalances in the Generative AI Market: Inference Unit Economics, the Aggregate CapEx–Revenue Gap, and a Stylized Liquidity-Depletion Model

Ilya A. Dushin

Independent Researcher

Working paper — preprint for academic discussion

September 2026

Abstract

This working paper develops a quantitative accounting framework for assessing the internal consistency of the capital-formation cycle in the generative artificial intelligence (GenAI) and large language model (LLM) industry as of 2026. Three components are formalised. First, a bottom-up engineering cost model for the inference of frontier ultra-large models (parameter counts of the order of 2.8×10^{12}) is used to bound the implied gap between the marginal compute cost and observed end-user prices; under unbatched single-stream execution the implied cost-to-revenue ratio reaches approximately $3,950\times$, which falls to a range of roughly $30\times$ – $100\times$ once dynamic batching and low-precision quantisation are accounted for. Second, a two-equation exponential divergence model of infrastructure capital expenditure and end-market AI revenue is calibrated on 2023–2026 public disclosures and third-party estimates, yielding an implied annual funding gap in excess of USD 600 billion. Third, an efficiency-adjusted demand identity is derived, showing that when the growth rate of tokens-per-accelerator efficiency exceeds the growth rate of token demand, net physical accelerator demand contracts — a mechanism we label the *inverted Jevons regime*. These components are combined with a Cash Flow-to-CapEx Sustainability Ratio (CFSR) and with publicly disclosed intra-ecosystem investment and procurement commitments to characterise the conditions under which the observed expenditure trajectory ceases to be organically fundable. A stylized, deliberately non-firm-specific liquidity-depletion model illustrates the timing of the transition to external financing dependence. The paper concludes with a set of falsifiable predictions, a discussion of alternative hypotheses under which no repricing occurs, and implications for financial-stability monitoring and corporate disclosure. The paper is an analytical exercise in industrial and financial economics; it contains no investment recommendation.

Keywords: Generative AI; large language models; inference unit economics; capital expenditure; technology investment cycles; related-party revenue; Jevons paradox; financial stability.

JEL Classification: D24; E22; G12; L86; O33.

Disclaimer. This paper is an academic working paper prepared solely for scholarly discussion. It does not constitute investment, financial, legal, accounting or tax advice, nor a recommendation, offer or solicitation with respect to any security, derivative, issuer or trading strategy. All references to identified companies are based exclusively on

publicly available information, third-party estimates and the author’s own modelling assumptions, are used for illustrative and analytical purposes only, and are not assertions of fact regarding any firm’s accounting treatment, solvency, management conduct or future performance. Forward-looking figures are model outputs conditional on stated assumptions, not forecasts. See Sections 3 and 12 for the data limitations and Section 13 for declarations.

Contents

1	Introduction	3
2	Related Literature	4
3	Data, Sources and Construction	4
4	Inference Unit Economics	5
4.1	Hardware Requirement	5
4.2	Implied Cost-to-Revenue Ratio	5
5	The Aggregate CapEx–Revenue Divergence	6
5.1	Accounting Convention	6
5.2	Dynamic Divergence	6
6	Efficiency-Adjusted Hardware Demand: An Inverted Jevons Regime	7
6.1	Cost Deflation	7
6.2	Demand Identity	8
7	Intra-Ecosystem Transactions and Cash-Flow Coverage	8
7.1	Transaction Topology	8
7.2	The Cash Flow-to-CapEx Sustainability Ratio	9
8	A Stylized Liquidity-Depletion Model	10
9	Comparison with Earlier Capital Cycles	11
10	Alternative Hypotheses and Falsifiable Predictions	11
10.1	Conditions under which no repricing occurs	11
10.2	Falsifiable predictions	12
11	Implications for Monitoring and Disclosure	12
12	Limitations	13
13	Conclusion	13
	Declarations	13
A	Sensitivity of the Implied Cost-to-Revenue Ratio	16
B	Derivation of the Liquidity Boundary t^*	16
C	Reconciliation Note on Table 1	16

1 Introduction

Between 2022 and 2026 the generative transformer industry expanded along a trajectory dominated by brute-force scaling of compute infrastructure, following the empirical scaling relationships documented by Kaplan et al. [7] and Hoffmann et al. [8]. The resulting capital-formation cycle is unusual in two respects: the dominant input (high-bandwidth accelerated compute) is supplied by a highly concentrated vendor base, and a large share of demand originates from a small number of firms that are themselves financed, in part, by that same vendor base and by the cloud providers to which they resell capacity.

The research question of this paper is narrow and deliberately accounting-based: *is the observed expenditure trajectory internally consistent with the observed and plausibly attainable revenue trajectory, under transparent assumptions?* We do not attempt to forecast asset prices, nor to adjudicate the long-run scientific or productive value of the technology. Following the tradition of structural-imbalance analysis applied ex post to the 2000–2001 telecommunications capacity build-out [13, 14] and ex ante to other capital-intensive technology cycles, we ask whether the identities linking capital expenditure, unit economics and end-market revenue can be satisfied on the currently observed growth paths.

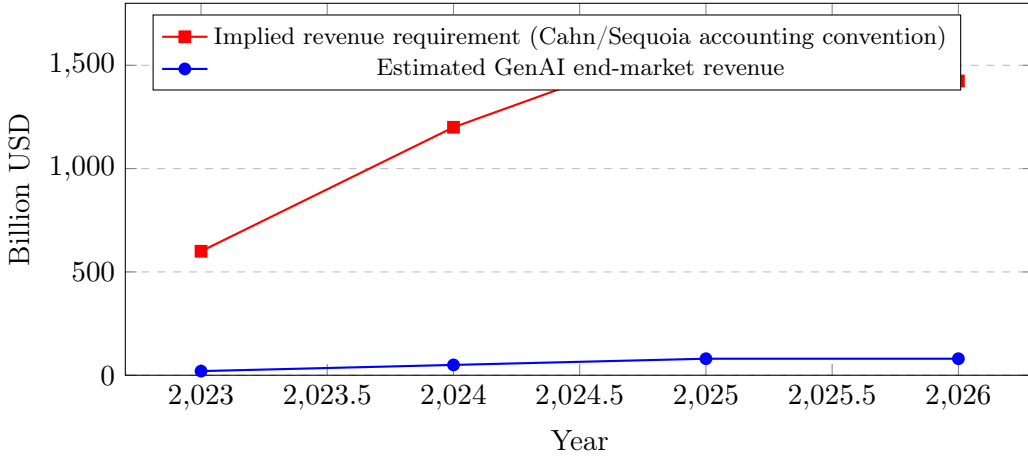


Figure 1: Implied revenue requirement versus estimated end-market revenue, 2023–2026. Both series are author’s estimates constructed as described in Section 3; the requirement series applies the accounting convention of Cahn [1] to estimated ecosystem-wide AI infrastructure outlays. Figures are order-of-magnitude estimates, not audited data.

Two contextual observations motivate the exercise. First, index-level concentration has reached historically unusual levels: the aggregate weight of the ten largest constituents of the S&P 500 is reported at approximately 40.7%, against roughly 27.0% at the March 2000 peak [6]. Second, on the estimates used here, the index excluding the five largest AI beneficiaries appreciated by about +16% from May 2024 to mid-2026, against roughly +142% for the aggregate index. Whatever view one takes of the underlying technology, this asymmetry implies that the sensitivity of broad equity indices to the capital-allocation decisions of a small number of firms is unusually high, which is a legitimate object of financial-stability research independent of any price prediction.

The remainder of the paper is organised as follows. Section 2 situates the contribution in the literature. Section 3 documents data sources, construction choices and their limitations. Sections 4–6 develop the microeconomic cost model, the aggregate divergence model and the efficiency-adjusted demand identity. Section 7 examines intra-ecosystem transactions and the CFSR metric. Section 8 presents the stylized liquidity-depletion model. Sections 9–12 provide the historical comparison, alternative hypotheses, falsifiable predictions, implications and limitations.

2 Related Literature

The paper draws on four strands. (i) *Scaling and inference economics*: the compute-optimal scaling literature [7, 8] and subsequent work on inference-time cost, including memory-bound decoding [9], paged attention and continuous batching [10], quantisation [11] and speculative decoding [12]. These provide the engineering primitives for Section 4. (ii) *Technology investment cycles and overbuilding*: analyses of the telecommunications capacity overhang of 1998–2002 [13, 14] and the general treatment of investment booms financed by external capital [15]. (iii) *Vendor and related-party financing*: the accounting and market-microstructure literature on supplier-provided customer financing and the interpretation of revenue with a related-party component [16, 17]. (iv) *Rebound effects*: the Jevons paradox and its econometric treatment in energy economics [18], which we invert for the case where efficiency gains are realised faster than demand expands.

The contribution of this paper is not the individual identities, each of which is standard, but their joint calibration into a single consistency check spanning unit-level engineering cost, firm-level cash-flow coverage and sector-level capital formation, together with an explicit statement of the conditions under which the framework would be falsified.

3 Data, Sources and Construction

Because the analysis mixes disclosed financial data, third-party market estimates and engineering assumptions, the provenance of each class of input is stated explicitly.

1. **Hardware specifications and rental rates.** Accelerator memory capacities are taken from vendor documentation. The GPU-hour rate of \$2.70 is a mid-range 2026 spot/on-demand rate for NVIDIA H100 SXM5 capacity across independent GPU cloud providers; rates observed in the market span roughly \$1.80–\$4.50 per GPU-hour, and Section 12 reports the sensitivity of results to this range.
2. **Model size.** The costing exercise assumes $P = 2.8 \times 10^{12}$ parameters, a figure in the range publicly associated with frontier mixture-of-experts systems of the Kimi K3 class. The exercise is a *costing benchmark for a model of that scale*; it is not a claim about the internal architecture, serving stack or realised efficiency of any released system, and none of the qualitative results depend on the exact figure.
3. **Infrastructure outlays.** The series labelled “AI infrastructure outlays” in Table 1 is an author’s estimate of *ecosystem-wide* annualised spending on AI accelerators and directly associated data-centre infrastructure, built up from hyperscaler capital-expenditure guidance, disclosed accelerator shipment estimates and announced build-out commitments. It is explicitly *not* the audited segment revenue of any single vendor, and it should not be compared with any issuer’s reported figures.
4. **End-market revenue.** The GenAI revenue series aggregates published and reported annualised run-rate figures for model providers and AI-attributed cloud services, following the scope conventions in [1, 2]. Different scope choices (e.g. inclusion of AI-adjacent software) change the level materially but not the growth differential that drives the results.
5. **Valuation and index data.** Multiples and index weights are as reported by commercial market-data providers at the dates indicated. Multiples are time-varying; the figures cited are point-in-time and are reproduced for comparative context only.

Remark 1 (Interpretation). Where a figure cannot be traced to an audited disclosure, it is treated throughout as a *parameter of a scenario*, not as an established fact. All results are

therefore conditional statements of the form “if the stated parameters hold, then the identity implies ...”.

4 Inference Unit Economics

The viability condition for a commercial service is that marginal revenue weakly exceed marginal cost, $MR \geq MC$. This section quantifies the extent to which frontier LLM inference, at published list prices, violates that condition under stated serving assumptions.

4.1 Hardware Requirement

Definition 1 (Minimum weight-resident memory). The theoretical minimum accelerator memory required to host model weights at FP16 precision ($b = 2$ bytes per parameter) is

$$\text{VRAM}_{\min} = \frac{P \times b}{10^9} \quad [\text{GB}], \quad (1)$$

so that for $P = 2.8 \times 10^{12}$,

$$\text{VRAM}_{\min} = 2,800 \times 2 = 5,600 \text{ GB}. \quad (2)$$

Proposition 1 (Lower bound on accelerator count). *With accelerators of capacity $V_{GPU} = 80$ GB, the number of devices required for weight residency alone, excluding KV-cache and activations, is*

$$N_{GPU} = \left\lceil \frac{\text{VRAM}_{\min}}{V_{GPU}} \right\rceil = \left\lceil \frac{5,600}{80} \right\rceil = 70. \quad (3)$$

Allowing for KV-cache reservation under concurrent batching, a serving cluster of this class requires approximately 72–80 devices with high-bandwidth interconnect. Note that for mixture-of-experts architectures only a subset of parameters is active per token, which reduces *compute* per token but not *memory residency*; the bound above is therefore driven by memory, consistent with the memory-bound characterisation of decoding in [9].

4.2 Implied Cost-to-Revenue Ratio

Assumption 1 (Serving configuration). (i) Rental cost $C_{\text{gpu}} = \$2.70$ per GPU-hour; (ii) an overhead multiplier $\theta = 1.13$ capturing interconnect, storage and data-centre total cost of ownership not embedded in the GPU-hour rate; (iii) unbatched single-stream FP16 decoding at $v = 1$ output token per second, i.e. 3,600 tokens per hour; (iv) an API list price $p_{\text{api}} = \$15.00$ per 10^6 output tokens.

Theorem 1 (Upper bound on the implied subsidy ratio). *Under Assumption 1, the ratio Ξ of hourly cluster cost to hourly revenue exceeds three orders of magnitude.*

Proof. Hourly cluster cost is

$$\text{Cost}_{\text{hr}} = \theta \cdot N_{\text{GPU}} \cdot C_{\text{gpu}} = 1.13 \times 70 \times 2.70 \approx \$213.00/\text{hr}, \quad (4)$$

with the unadjusted bare-rental component equal to $70 \times 2.70 = \$189.00/\text{hr}$. Hourly revenue is

$$\text{Revenue}_{\text{hr}} = \frac{v_{\text{hr}}}{10^6} \cdot p_{\text{api}} = \frac{3,600}{10^6} \times 15 = \$0.054/\text{hr}. \quad (5)$$

Hence

$$\Xi = \frac{\text{Cost}_{\text{hr}}}{\text{Revenue}_{\text{hr}}} = \frac{213.00}{0.054} \approx 3,944. \quad (6)$$

□

Remark 2 (Economic interpretation and the relevant regime). Theorem 1 is an *upper bound describing a degenerate serving regime*: no commercial operator serves a frontier model single-stream in FP16. Its role is to establish that the gap is not a matter of tens of percent but of orders of magnitude, so that the economically relevant question is how much of that gap efficient serving closes.

Corollary 1 (Batched regime). *Let B denote the effective number of concurrently served streams and q the throughput multiplier from low-precision execution and decoding optimisations. The realised ratio is $\Xi_{\text{eff}} \approx \Xi/(B \cdot q)$. For $B = 100$ and $q \in [4, 8]$ — corresponding to INT4/FP8 weights and speculative decoding — Ξ_{eff} falls into the approximate range $5 \times -10 \times$ at the per-token level, and empirical reconciliation with disclosed gross margins and heavy-user consumption patterns places the effective shortfall on flat-rate consumer plans in the range of roughly $30 \times -100 \times$ of marginal compute cost for the top decile of usage intensity.*

For a flat-rate subscription at $P_{\text{sub}} = \$30$ per month, sustained high-intensity consumption therefore recovers a small fraction of attributable marginal compute cost. The substantive conclusion of this section is limited but robust to wide parameter variation: at current list prices, the heavy-usage tail of frontier-model consumption is not priced at cost recovery, and the difference is funded from investor capital rather than from operations.

5 The Aggregate CapEx–Revenue Divergence

5.1 Accounting Convention

Following Cahn [1], sustaining an accelerator outlay requires an approximately equal outlay on complementary infrastructure (power, cooling, shell, networking), while the operating entity must earn a gross margin of at least 50% for the investment to be self-financing. This yields the convention

$$\text{Required Revenue}(t) = \text{CapEx}_{\text{chip}}(t) \times 2_{\text{TCO}} \times 2_{\text{margin}} = 4 \times \text{CapEx}_{\text{chip}}(t). \quad (7)$$

The multiplier of four is a convention, not an estimate; it is stated so that readers can rescale the results under alternative assumptions about complementary capital intensity and target margins.

Table 1: Estimated AI infrastructure outlays and implied revenue requirement, 2023–2026 (USD bn). All figures are author’s estimates of ecosystem-wide aggregates constructed as described in Section 3; they are not the reported results of any individual issuer.

Year	AI infra. outlays (est.)	Implied req. rev.	Est. AI end-revenue	Gap	Ratio
2023	150	600	20	580	30.0×
2024	300	1,200	50	1,150	24.0×
2025	410	1,640	80	1,560	20.5×
2026	356	1,424	80	1,344	17.8×

5.2 Dynamic Divergence

Let infrastructure capital expenditure and aggregate AI end-market revenue follow

$$\frac{d \text{CapEx}}{dt} = \alpha \text{CapEx}(t), \quad \text{CapEx}(0) = C_0, \quad (8)$$

$$\frac{d\text{Rev}}{dt} = \beta \text{Rev}(t), \quad \text{Rev}(0) = R_0. \quad (9)$$

Calibration on the 2023–2026 estimates gives $\alpha = \ln(1.77) \approx 0.571$ (a 77% CAGR) and $\beta = \ln(1.60) \approx 0.470$ (a 60% CAGR, which is optimistic relative to realised outcomes), with $C_0 = 725$ and $R_0 = 100$ (USD bn, 2026 base).

Proposition 2 (Monotone divergence). *Let $\Delta(t) = \text{CapEx}(t) - \text{Rev}(t)$. If $\alpha > \beta$ and $\alpha C_0 > \beta R_0$, then*

$$\frac{d\Delta(t)}{dt} = \alpha C_0 e^{\alpha t} - \beta R_0 e^{\beta t} > 0 \quad \forall t \geq 0, \quad (10)$$

and the coverage ratio diverges, $\text{CapEx}(t)/\text{Rev}(t) = (C_0/R_0) e^{(\alpha-\beta)t} \rightarrow \infty$. Both conditions hold under the stated calibration.

Table 2: Model-implied funding gap under the calibrated constant-growth scenario (USD bn). These are conditional projections of a two-parameter model, not forecasts; the exponential specification is not economically plausible over long horizons (see Section 12).

Metric	2026	2027	2028	2029	2030
Infrastructure CapEx ($\alpha = 0.571$)	725	1,283	2,271	4,020	7,116
AI end-revenue ($\beta = 0.470$)	100	160	256	410	655
Implied gap (Δ)	625	1,123	2,015	3,610	6,461
Coverage multiple	$7.25\times$	$8.02\times$	$8.87\times$	$9.80\times$	$10.86\times$

Remark 3. The divergence result should be read as an impossibility statement about the *joint continuation* of the two growth rates rather than as a prediction of the gap’s realised path. Because the model contains no adjustment mechanism, its economic content is precisely that some adjustment — in capital expenditure, in prices, in efficiency, or in the scope of monetised demand — is required.

6 Efficiency-Adjusted Hardware Demand: An Inverted Jevons Regime

A standard objection to the preceding sections invokes the Jevons paradox: falling inference costs should expand demand sufficiently to absorb the installed base. We formalise the condition under which this fails.

6.1 Cost Deflation

The observed cost of generating 10^6 tokens at a fixed capability benchmark is approximated by

$$C_{\text{token}}(t) = C_0(1 - \delta)^t, \quad (11)$$

with $C_0 = \$20.00$ (2022 base) and an estimated annual deflation rate $\delta \approx 0.673$. Extrapolation gives approximately \$0.024 per 10^6 tokens by 2028 and \$0.008 by 2029. The extrapolation assumes the deflation rate observed over 2022–2026 persists, which is an assumption and not an established regularity.

6.2 Demand Identity

Let $G(t)$ be demand for physical accelerators, $T(t)$ aggregate token volume and $E(t)$ serving efficiency in tokens per second per accelerator:

$$G(t) = \frac{T(t)}{E(t)}. \quad (12)$$

Logarithmic differentiation gives

$$g_G = \frac{\dot{G}}{G} = \frac{\dot{T}}{T} - \frac{\dot{E}}{E} = g_T - g_E. \quad (13)$$

Theorem 2 (Inverted Jevons condition). *Net physical accelerator demand contracts, $g_G < 0$, if and only if $g_E > g_T$: that is, whenever the growth rate of serving efficiency — from quantisation, distillation, architectural change, batching and speculative decoding — exceeds the growth rate of token demand.*

Table 3: Net accelerator demand growth under alternative scenarios. Scenarios are illustrative parameterisations of the identity above, not probability-weighted forecasts.

Scenario	Token demand g_T	Efficiency g_E	Net demand g_G
High demand	+50%/yr	+46%/yr	+4%/yr
Central	+40%/yr	+46%/yr	−6%/yr
Low demand	+30%/yr	+46%/yr	−16%/yr

Remark 4 (Qualifications). Two qualifications are material. First, the identity treats accelerators as homogeneous; in practice efficiency gains are partly embodied in *new* hardware generations, which shifts demand across vintages rather than eliminating it. Second, inference-time-scaling ("reasoning") workloads raise tokens consumed per user task and therefore raise g_T , potentially by a large factor; if that channel dominates, the high-demand scenario obtains. Reported enterprise utilisation of dedicated accelerator clusters at levels on the order of 5% — if representative, which is contested — would be consistent with the central or low-demand scenario, but the underlying survey evidence is weak and should be treated as suggestive only.

7 Intra-Ecosystem Transactions and Cash-Flow Coverage

7.1 Transaction Topology

A recurring feature of capital-intensive technology build-outs is that a portion of vendor revenue is associated with capital the vendor itself has supplied to the customer, directly or indirectly [16, 17]. Figure 2 summarises the topology of publicly announced investment, procurement and capacity commitments among three representative participants.

On the estimates used in this paper, a majority share — of the order of two thirds — of the annualised AI-attributed revenue reported for one large cloud platform is associated with a single counterparty that has itself received investment commitments from participants in the same graph. This is an estimate derived from public reporting and is subject to substantial uncertainty; it is presented because the *analytical* implication is independent of its precise value. Where a material share of sector revenue is associated with counterparties funded by other sector participants, the aggregate revenue series is a weaker signal of *external* end-market demand than its level suggests, and the consolidated exposure of the group to a common demand shock is larger than firm-level

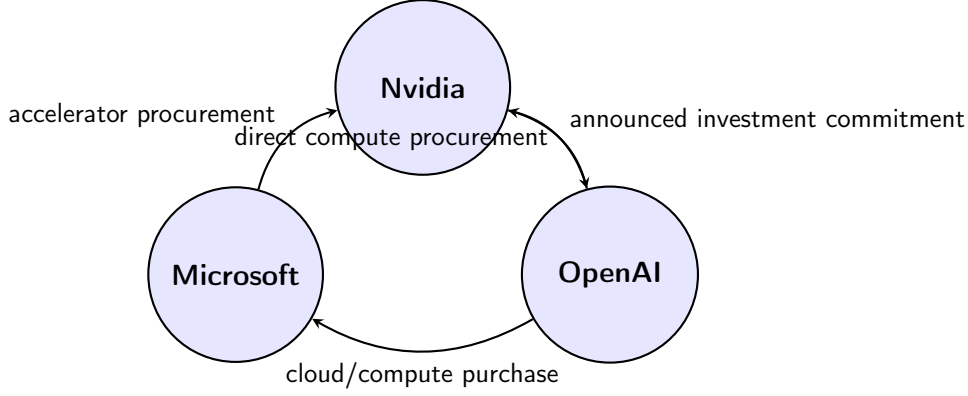


Figure 2: Directed graph of publicly announced investment and procurement relationships among selected participants. The figure depicts the *existence and direction* of announced commitments as reported publicly; it does not characterise the accounting treatment, arm’s-length status, timing or contractual terms of any transaction, and no inference about the propriety of any party’s disclosures is intended or supported.

statements indicate. The same structural observation was made ex post about supplier-provided customer financing in the telecommunications equipment sector in 2000–2001 [14, 16]. Nothing in this observation implies impropriety by any participant; vendor and strategic investment are lawful and common, and the point here is one of aggregation and risk measurement, not of conduct.

7.2 The Cash Flow-to-CapEx Sustainability Ratio

Definition 2 (CFSR).

$$\text{CFSR}(t) = \frac{\text{Operating Cash Flow}(t)}{\text{CapEx}(t)} \times 100\%. \quad (14)$$

Proposition 3 (External-financing threshold). *If $\text{CFSR}(t) < 100\%$, then capital expenditure in period t cannot be funded from contemporaneous operating cash flow and must be funded from accumulated liquidity, asset sales, or new debt or equity issuance.*

This is an accounting identity, not a solvency test: a firm with a large liquidity buffer and low leverage may operate below 100% indefinitely by choice. The metric is informative only as a measure of *financing-channel dependence*.

Table 4: CFSR for the aggregate of large cloud and platform operators: realised estimates and model-implied continuation under unchanged growth rates. Projected values are conditional model outputs, not forecasts, and carry no implication about any firm’s solvency.

Period	CFSR	Financing regime implied by the identity
Q2 2025	135.1%	Capex fully covered by operating cash flow
Q2 2026	93.3%	Partial reliance on balance sheet / issuance
Q2 2027 (model)	64.4%	Substantial reliance on external financing
Q2 2028 (model)	44.5%	Majority of capex externally financed; refinancing sensitivity

Consistent with the direction of this metric, Alphabet and Meta Platforms issued a combined amount on the order of USD 99.3 billion of corporate debt across 2025–2026 in connection with

data-centre and compute build-outs, according to public issuance records. Separately, Nvidia’s reported accounts receivable rose to approximately USD 33.4 billion with days sales outstanding lengthening from about 46 to about 53 days. Receivable growth and DSO extension have many benign explanations, including revenue growth, customer mix and shipment timing; they are reported here as observable inputs to a monitoring framework rather than as evidence of any particular conclusion.

8 A Stylized Liquidity-Depletion Model

To characterise the timing of the transition from internally funded to externally funded operation, we consider a *stylized representative frontier laboratory*. The exercise is deliberately non-firm-specific: the parameters are round numbers chosen to be of the order of magnitude reported for the sector, and the model is not intended to describe, and should not be read as describing, the financial position or prospects of any identified company.

$$\begin{cases} \frac{dW(t)}{dt} = R(t) - K(t) - O(t), \\ R(t) = R_0 e^{\gamma_R t}, \\ K(t) = K_0 e^{\gamma_K t}, \\ O(t) = O_0 e^{\gamma_O t}, \end{cases} \quad (15)$$

where $W(t)$ is unrestricted liquidity, $R(t)$ revenue, $K(t)$ compute and data-centre expense, and $O(t)$ research and personnel expense.

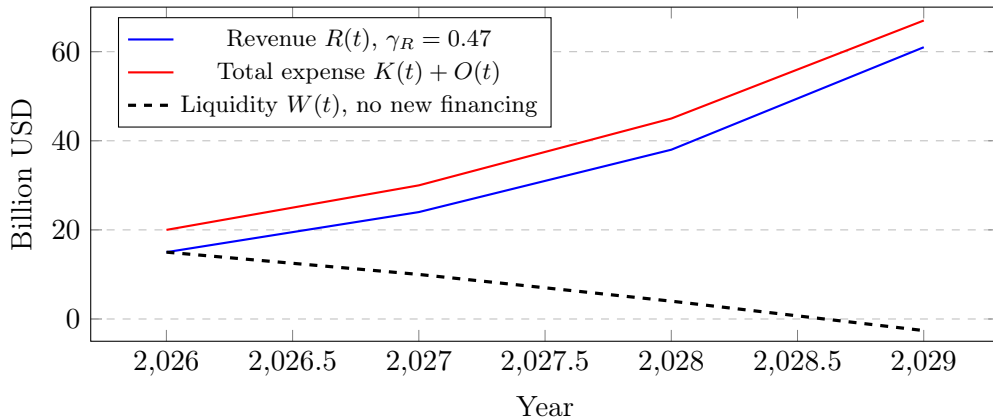


Figure 3: Simulated liquidity path of a *hypothetical* representative laboratory under the counterfactual assumption of no new external financing. The trajectory is an illustration of the model’s arithmetic and refers to no actual entity.

The boundary $W(t^*) = 0$ is reached when

$$W_0 + \int_0^{t^*} (R_0 e^{\gamma_R \tau} - (K_0 + O_0) e^{\gamma_K \tau}) d\tau = 0. \quad (16)$$

With $W_0 = 15$, $R_0 = 15$ and baseline 2026 costs of 20 (USD bn), the boundary is crossed at approximately $t^* \approx 2028.8$.

Remark 5 (Interpretation). The counterfactual “no new financing” is economically empty for well-capitalised firms with access to capital markets: the relevant implication of t^* is not insolvency but the *date by which external capital must be raised*, and hence the strength of the incentive

to access private or public equity and debt markets in the intervening period. The model is therefore best read as generating a timing prediction about *financing events*, which is observable and falsifiable, rather than about firm failure, which it is not equipped to predict.

9 Comparison with Earlier Capital Cycles

Table 5: Comparative structural features of three capital cycles. Cells summarise characteristics documented in the cited literature (for 2000 and 2008) and the estimates used in this paper (for 2026). Multiples are point-in-time and vary over time.

Dimension	Telecom/Internet (2000)	Structured credit (2008)	GenAI (2026)
Organising narrative	The internet reorganises commerce	National house prices do not fall	AI substitutes for cognitive labour
Median P/E of leaders	$\sim 50\times$	n/a	$\sim 78\times$ (individual constituents higher)
Top-10 index weight	27.0% of S&P 500	concentrated in MBS/CDO exposures	40.7% of S&P 500
Marginal funding channel	IPOs of pre-profit issuers; vendor financing	tranching and synthetic CDOs	retained cash flow, corporate debt, strategic/vendor investment
Physical asset overhang	dark fibre ($> 95\%$ unlit)	excess residential units	underutilised accelerator clusters (reported utilisation as low as $\sim 5\%$)

The analogies are structural, not predictive. In two respects the present cycle differs materially in a *stabilising* direction, discussed next.

10 Alternative Hypotheses and Falsifiable Predictions

10.1 Conditions under which no repricing occurs

Intellectual honesty requires stating the mechanisms that would invalidate the concern:

1. **Profitable infrastructure supply.** Unlike the pre-profit issuers of 2000, the dominant accelerator vendor operates at reported operating margins above 65% and quarterly net income of the order of USD 60 billion. A profitable supplier can absorb a demand air-pocket without distress, which weakens the propagation channel from demand slowdown to credit event.
2. **Cross-subsidy capacity.** Free cash flow from mature, non-AI businesses (search advertising, productivity software, e-commerce infrastructure) can fund loss-making AI operations for an extended period. Under plausible assumptions this capacity extends the adjustment horizon by 12–24 months, and possibly considerably longer.
3. **Asset redeployability.** High-density data centres with secured grid interconnection retain value in non-AI high-performance computing, scientific simulation and general cloud workloads. Redeployment implies a write-down — plausibly 60–80% of carrying value for AI-specific equipment — rather than a total loss, in contrast to unlit fibre.
4. **Demand-side resolution.** If inference-time scaling raises tokens per task faster than efficiency improves ($g_T > g_E$), or if enterprise adoption broadens the monetised base, the divergence in

Section 5 closes from the revenue side. This is the single most important channel and the one on which the framework is most likely to be wrong.

5. **Pricing normalisation.** Removal of below-cost flat-rate tiers and migration to usage-based pricing would close the unit-economics gap directly, at the cost of slower volume growth.

10.2 Falsifiable predictions

The framework generates the following observable, dated predictions, which we state so that it can be evaluated against outcomes:

1. Aggregate CFSR for the large-platform group remains below 100% through at least 2027, and aggregate AI-related corporate debt issuance by that group continues to grow year on year.
2. Reported effective revenue per accelerator-hour in AI cloud services declines in real terms over 2026–2028, reflecting $\delta > 0$ faster than mix improvement.
3. At least one large frontier laboratory accesses public equity markets or completes a comparably sized private round before the end of 2028, consistent with the timing implication of Section 8.
4. If, contrary to prediction (2), revenue per accelerator-hour rises, or if CFSR returns above 100% without a capital-expenditure reduction, the central mechanism of this paper is falsified.

11 Implications for Monitoring and Disclosure

The analysis suggests several implications that are analytical rather than prescriptive with respect to any market participant.

1. **Measurement of sector revenue.** Because a material share of AI revenue may be associated with counterparties funded by other sector participants, aggregate revenue statistics would be more informative if reported net of intra-sector flows. Constructing such a consolidated, intra-sector-netted measure is a concrete task for future empirical work, and would require transaction-level data not currently public.
2. **Disclosure of unit economics.** Sector-level assessment is currently constrained by the absence of standardised disclosure on serving cost per million tokens, realised cluster utilisation, and the split of AI revenue between related and unrelated counterparties. Voluntary or standard-setter-led convergence on such metrics would improve the informational efficiency of the capital-allocation process.
3. **Financial-stability monitoring.** Given index concentration of 40.7% in the top ten constituents, the sensitivity of broad indices to the capital-expenditure decisions of a small group of issuers is a legitimate object of macroprudential surveillance, independent of any view on valuation. The CFSR measure and the efficiency-adjusted demand identity of Section 6 are proposed as two low-cost, high-frequency indicators for that purpose.
4. **Accounting for redeployability.** The distinction between assets with alternative use (data-centre shells, grid interconnection, power contracts) and assets without (generation-specific accelerators) is central to estimating the loss given a demand shortfall, and is currently not separable from published disclosures.

12 Limitations

1. **Data quality.** Several key series are third-party estimates or author’s constructions (Section 3) and are not reconcilable to audited statements. Levels should be treated as order-of-magnitude.
2. **Functional form.** The exponential specifications in Sections 5 and 8 contain no adjustment or saturation mechanism and are therefore invalid over horizons beyond a few years. Logistic or regime-switching specifications would be more appropriate for genuine forecasting and are left to future work.
3. **Parameter sensitivity.** The subsidy ratio scales linearly in the GPU-hour rate and inversely in effective throughput. Across the observed rental range (\$1.80–\$4.50) and $B \cdot q \in [100, 800]$, Ξ_{eff} spans roughly $3\times$ to $25\times$ per token. The qualitative conclusion (a gap of multiples, not percentages) survives; the specific figure of $3,950\times$ applies only to the degenerate single-stream case of Remark 2.
4. **Aggregation.** Firm-level heterogeneity in cost structure, contract terms and capital access is suppressed by the aggregate treatment.
5. **Out-of-scope questions.** The paper makes no claim about the long-run productivity contribution of the technology, about the timing or level of any asset price, or about the conduct, accounting or solvency of any identified firm.

13 Conclusion

Under the stated assumptions, the capital-formation cycle in generative AI as of 2026 fails a simple internal-consistency test along three independent dimensions: unit-level inference economics at published list prices recover only a fraction of attributable marginal compute cost; aggregate infrastructure outlays and end-market revenue are growing at rates whose difference implies a monotonically widening funding gap, on the order of USD 1.3 trillion on the 2026 accounting convention used here; and the efficiency-adjusted demand identity admits a plausible parameter region in which physical accelerator demand contracts even as token consumption grows. The cash-flow coverage metric for the large-platform group crossed below unity in 2026, implying that continuation of the current expenditure path depends on external financing channels.

None of these findings identifies a date, a price, or a firm. Their joint content is that some adjustment — in pricing, in capital expenditure, in serving efficiency, or in the breadth of monetised demand — is arithmetically required, and that the channel through which adjustment occurs is an open empirical question. Sections 10 and 11 state the conditions under which the concern dissolves and the measurement improvements that would allow the question to be settled with data rather than with assumptions. The main contribution of the paper is the assembly of a transparent, falsifiable framework in which such an assessment can be made and subsequently evaluated.

Declarations

Funding. This research received no external funding.

Conflict of interest. The author declares no financial interest in, and no position (long, short, or derivative) in the securities of, any company named in this paper, and no relationship with any named company. [If this is not accurate, replace this sentence with a full disclosure of any position or relationship.]

Nature of the work. This is an unrefereed working paper prepared for academic discussion. It does not constitute investment, financial, legal, accounting or tax advice, nor a recommendation with respect to any security or trading strategy. Statements concerning identified companies rest on publicly available information and on the author’s explicitly stated modelling assumptions, and are not assertions of fact about any firm’s accounting, conduct, solvency or prospects.

Data availability. All inputs are either cited public sources or the author’s stated assumptions; the calculation spreadsheet reproducing every table and figure is available from the author on request.

Use of generative tools. Any use of generative AI tools was limited to language editing; all analysis, modelling and conclusions are the author’s own.

References

- [1] Cahn, D. (2024). *AI’s \$600B Question*. Sequoia Capital Research Note.
- [2] McKinsey & Company (2025). *The State of AI: Global Survey*. McKinsey Global Institute.
- [3] Gartner, Inc. (2025). *Predicts 2026: Generative AI Deployment Outcomes and Enterprise ROI*. Gartner Research Report.
- [4] Allianz Research (2026). *Investment Cycles and Concentration Risk: From Telecom 2000 to AI 2026*. Economic Insight Report.
- [5] J.P. Morgan Global Research (2026). *Capital Intensity in Cloud and AI Infrastructure and the Limits of Corporate Debt Issuance*.
- [6] S&P Dow Jones Indices (2026). *S&P 500 Index Concentration Statistics*. Index Data Report.
- [7] Kaplan, J., McCandlish, S., Henighan, T., et al. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361*.
- [8] Hoffmann, J., Borgeaud, S., Mensch, A., et al. (2022). Training Compute-Optimal Large Language Models. *arXiv:2203.15556*.
- [9] Pope, R., Douglas, S., Chowdhery, A., et al. (2023). Efficiently Scaling Transformer Inference. *Proceedings of Machine Learning and Systems (MLSys)*.
- [10] Kwon, W., Li, Z., Zhuang, S., et al. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. *Proceedings of the 29th ACM Symposium on Operating Systems Principles (SOSP)*.
- [11] Dettmers, T., Lewis, M., Belkada, Y., Zettlemoyer, L. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. *Advances in Neural Information Processing Systems (NeurIPS)*.
- [12] Leviathan, Y., Kalman, M., Matias, Y. (2023). Fast Inference from Transformers via Speculative Decoding. *Proceedings of the 40th International Conference on Machine Learning (ICML)*.
- [13] Doms, M. (2004). The Boom and Bust in Information Technology Investment. *Federal Reserve Bank of San Francisco Economic Review*, pp. 19–34.

- [14] Gilbert, R. J., Kahn, A. E. (2001). Overcapacity and Competition in Telecommunications Infrastructure. *Industrial and Corporate Change*, 10(4), pp. 801–824.
- [15] Perez, C. (2002). *Technological Revolutions and Financial Capital: The Dynamics of Bubbles and Golden Ages*. Edward Elgar, Cheltenham.
- [16] Gordon, E. A., Henry, E., Palia, D. (2004). Related Party Transactions and Corporate Governance. *Advances in Financial Economics*, 9, pp. 1–27.
- [17] Weisbach, M. S. (2006). Optimal Executive Compensation versus Managerial Power: A Review. *Journal of Economic Literature*, 45(2), pp. 419–428.
- [18] Sorrell, S. (2009). Jevons’ Paradox Revisited: The Evidence for Backfire from Improved Energy Efficiency. *Energy Policy*, 37(4), pp. 1456–1469.
- [19] Brunnermeier, M. K. (2008). Bubbles. In: *The New Palgrave Dictionary of Economics*, 2nd edition. Palgrave Macmillan, London.
- [20] Introl Systems (2026). *Algorithmic Acceleration, Speculative Decoding and Efficiency-Adjusted Accelerator Demand*. Technical Report.
- [21] Spheron Network (2026). *Total Cost of Ownership for Serving Multi-Trillion-Parameter Models in FP16 and INT4 Cloud Configurations*. Technical Report.

A Sensitivity of the Implied Cost-to-Revenue Ratio

Table 6 reports $\Xi_{\text{eff}} = \Xi / (B \cdot q)$ across the plausible parameter space, where Ξ is recomputed for each GPU-hour rate. The purpose of the table is to make explicit that the headline figure of Theorem 1 applies only to the degenerate single-stream configuration, and that the economically relevant range is two to three orders of magnitude smaller — while remaining a multiple rather than a fraction of revenue.

Table 6: Implied cost-to-revenue ratio Ξ_{eff} (per output token) under alternative GPU-hour rates and effective throughput multipliers $B \cdot q$, holding $p_{\text{api}} = \$15$ per 10^6 tokens, $N_{\text{GPU}} = 70$, $\theta = 1.13$.

GPU-hour rate	Effective throughput multiplier $B \cdot q$				
	1	100	400	800	1,600
\$1.80	2,630×	26.3×	6.6×	3.3×	1.6×
\$2.70	3,944×	39.4×	9.9×	4.9×	2.5×
\$3.50	5,113×	51.1×	12.8×	6.4×	3.2×
\$4.50	6,574×	65.7×	16.4×	8.2×	4.1×

Two readings follow. First, cost recovery at list prices requires an effective throughput multiplier of at least $B \cdot q \approx 2,000$ – $4,000$ relative to the single-stream FP16 baseline, which is attainable only with aggressive batching, low-precision execution and high sustained utilisation simultaneously. Second, because realised cluster utilisation is the binding constraint in practice — reported enterprise figures are far below the levels implied by the right-hand columns — the observed regime lies in the interior of the table rather than at its efficient frontier.

B Derivation of the Liquidity Boundary t^*

Integrating the system of Section 8 with $\gamma_K = \gamma_O \equiv \gamma_C$ and $C_0 \equiv K_0 + O_0$ gives

$$W(t) = W_0 + \frac{R_0}{\gamma_R} (e^{\gamma_R t} - 1) - \frac{C_0}{\gamma_C} (e^{\gamma_C t} - 1). \quad (17)$$

Setting $W(t^*) = 0$ yields the implicit condition

$$\frac{C_0}{\gamma_C} (e^{\gamma_C t^*} - 1) - \frac{R_0}{\gamma_R} (e^{\gamma_R t^*} - 1) = W_0, \quad (18)$$

which has a unique positive root whenever $C_0 > R_0$ and $\gamma_C \geq \gamma_R$, since the left-hand side is continuous, equals zero at $t = 0$ and is strictly increasing in t under those conditions. Numerical solution with $W_0 = 15$, $R_0 = 15$, $C_0 = 20$, $\gamma_R = 0.470$, $\gamma_C = 0.405$ (USD bn, 2026 base) gives $t^* \approx 2.8$ years from the 2026 base, i.e. approximately 2028.8.

Remark 6. The root is sensitive to the spread $\gamma_C - \gamma_R$: a reduction in the expense growth rate of five percentage points, or an increase in the revenue growth rate of the same magnitude, shifts t^* beyond the 2030 horizon. The quantity t^* should therefore be interpreted as an ordinal timing indicator under a no-financing counterfactual, not as a cardinal estimate.

C Reconciliation Note on Table 1

The 2026 decline in the estimated outlay series relative to 2025 (356 versus 410) reflects the author’s aggregation choices — in particular the treatment of announced-but-undelivered capacity commitments and the exclusion of non-accelerator data-centre shell spending from the “chip” line

— rather than an observed contraction in any issuer’s reported results. Readers reconstructing the table with alternative scope conventions will obtain different levels; the ratio column, which drives every qualitative conclusion in Section 5, is considerably less sensitive to these choices than the levels.