AI Value-Chain Analysis: Dynamics & Capex Investment

Deep-Dive Strategy Report (v.12)

Author: Dr. Angel Gavieiro Besteiro

LLMs: Research-Gemini 3.6 Flash; Notebook LM. Verification-ChatGPT 5.6 Luna; Claude Sonnet 5

Collaboration: our thanks to the AI Markets Group for their review and feedback

Classification: Public | Date: 17/09/2026

Executive Summary

AI infrastructure investment is currently running materially ahead of directly observable end-user AI commercial revenue. The magnitude of the eventual "Capital Payback Gap" will depend on the degree and speed of AI adoption across enterprises and where economic value is captured across the AI value chain as a result of its "Barbell Polarization".

"Barbell Polarization"

*Capital and economic profit are concentrating at the physical foundation (silicon foundries, memory suppliers, cloud providers) and at the downstream distribution layer (legacy SaaS enterprise and vertical AI apps). Conversely, the middle layer (proprietary LLM developers, thin-wrapped AI apps) is trapped in a structural **"Commoditization Squeeze"**, pressured from above by non-negotiable compute and infrastructure costs and from below by open-source models.*

*A **"DeepSeek Death Zone"** formed by open-source models executing tasks 25x cheaper and reaching performance breakthroughs, combined with the **"Physical Gridlock"** from energy constraints starting to substantially delay new data centres, are accelerating the squeeze. This could lead to potential **Geopolitical Sub-scenarios** over 2027-28, from a US nationalization of burned-out closed-LLM labs, to dual US/China bans on respective LLMs, and/or to the surge of on-premise LLM deployment by enterprises on the back of new, powerful local hardware.*

"Capital Payback Gap"

*Hyperscaler total CapEx is running at $725bn+ by end-2026, while directly observable end-user AI commercial revenue is only $160-175bn p.a. At 60% gross margin, to generate a 15% unlevered project return (IRR) on the projected **$5.0 Tr cumulative AI CapEx (2024-2030)** and associated AI Opex, our model implies that AI vendors would need **end-user commercial revenue of $450 Bn in 2026 and $2.4 Tr p.a. by 2030**.*

*This creates a revenue gap of $275-290 Bn in 2026, which would widen to $2.0 Tr by 2030 unless enterprise AI revenue jumps c.14.3x over 2026 levels to produce $2.4 Tr p.a. For that, enterprises must escape "Pilot Purgatory" (c.95% of pilots not scaling to produce measurable impact) by (1) delivering massive productivity offsets/efficiency gains ($1.4 Tr) from the labour knowledge-workforce, and (2) IT vendors cannibalizing IT Services & BPO budgets into AI-driven SaaS expenditure ($1.0 Tr.). A lower $2.15 Tr p.a. revenue would yield 10% IRR, utility-like returns but still within a **"Capacity Digestion"** scenario.*

*Illustratively, if realized 2030 revenue were only **$1.2 Tr p.a.** (i.e. 50% of the baseline $2.4 Tr) the sensitivity matrix indicates severe value destruction, including an **implied Project-IRR c. -25%.***

Table of Contents

IMPORTANT NOTICE & DISCLAIMER:
INFORMATIONAL AND STRATEGIC RESEARCH PURPOSES ONLY

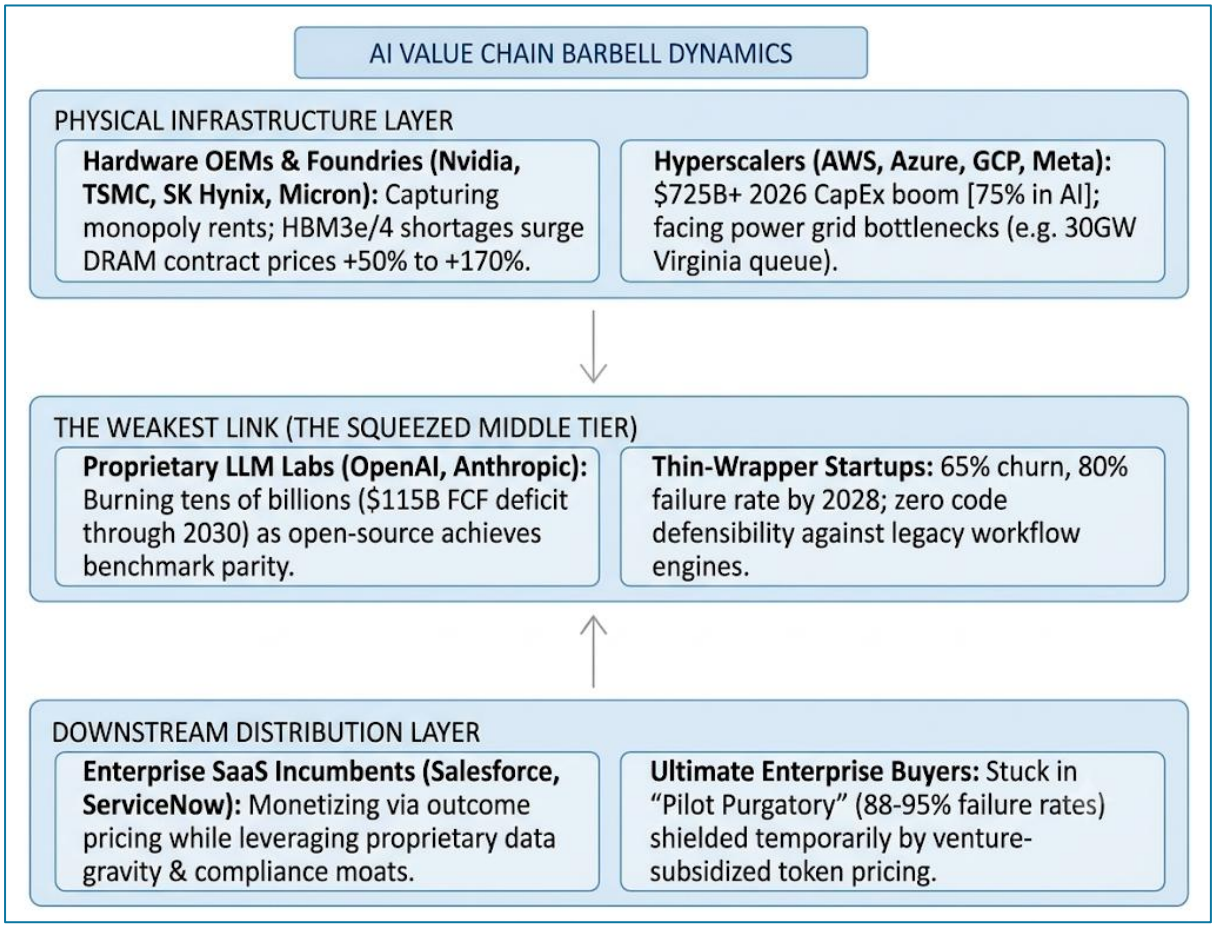
No Investment Advice or Solicitation
The analysis, scenario frameworks, valuation models, capital expenditure projections, and sector/asset ratings (e.g., "Neutral", "Positive", "Negative") contained in this report are prepared solely for strategic executive advisory, economic research, and general informational purposes. Nothing in this document constitutes, or is intended to constitute, financial, investment, legal, tax, or regulatory advice. This document does not constitute an offer, solicitation, invitation, or recommendation to purchase, subscribe for, hold, or sell any security, derivative, commodity, private credit instrument, or other financial asset, nor does it constitute an endorsement of any corporate entity, fund, or investment strategy.

Model Assumptions & Forward-Looking Uncertainty
The financial simulations, internal rate of return (IRR) estimates, discounted cash flow (DCF) models, terminal value calculations, and geopolitical sub-scenarios presented herein reflect theoretical stress-testing and analytical scenario modelling based on publicly available data, corporate earnings releases, consensus estimates, and specific mathematical assumptions (including gross margin hurdles, depreciation lifecycles, and capital expenditure amortization paths). Forward-looking statements and projections are inherently subject to significant macroeconomic, geopolitical, technological, supply-chain, regulatory, and market uncertainties, many of which are beyond control. Actual future financial results, infrastructure costs, enterprise adoption trajectories, and equity/debt market returns may differ materially from those expressed or implied in this research.

Independent Due Diligence Required
Any investment or capital allocation decision involves substantial risk, including the potential loss of principal. Readers and institutional decision-makers must perform their own independent due diligence, feasibility studies, and risk assessments, and should consult licensed and qualified financial, legal, compliance, and tax advisors before making any commercial, operational, or investment commitments.

Limitation of Liability & Absence of Warranties
While information has been obtained from sources believed to be reliable at the time of publication, neither the author, AG Strategy & Partners Ltd, nor any of their respective partners, affiliates, or contributors make any representation or warranty, express or implied, as to the accuracy, completeness, timeliness, or fitness for a particular purpose of the information or calculations contained herein. To the maximum extent permitted by applicable law, the author and AG Strategy & Partners Ltd expressly disclaim any and all liability for any direct, indirect, incidental, consequential, or special damages or losses arising from any reliance on or use of the contents of this report.

1. AI Value-Chain Competitive Dynamics



1.1 Chip Producers & Hardware Ecosystem:

Hardware OEMs capture Monopoly Rents as Bottlenecks Shift to High-Bandwidth Memory

Hardware OEMs (Nvidia, TSMC, SK Hynix, Micron) continue to command unmatched pricing power, as reflected in Nvidia's $130.5 Bn (+114%) in revenue and $72.9 Bn in net income (+145%) (FY25).[S13]

- **Accelerated Silicon Depreciation:** Hardware providers aggressively compress product lifecycles to force capital replacement cycles. The Blackwell architecture rollout (B100/B200), delivering 2.5x performance improvements at a 25% cost premium, rapidly depreciates existing Hopper (H100) clusters (and Vera Rubin shipping by Q3 2026).[S25]
- **High-Bandwidth Memory (HBM) Crisis:** The critical supply bottleneck has shifted from raw logic fabrication to HBM capacity. DRAM contract prices surged between 50% and 170%. NVIDIA B200 unit manufacturing costs stand at $6,400, with HBM memory alone representing nearly 50% ($2,900) of total bill-of-materials costs. Key memory producers Micron and SK Hynix reported HBM production fully booked through 2026, enabling SK Hynix to achieve 58.4% quarterly operating margin while raising HBM3E prices by 20%. [S2][S23]
- **Strategic Supply Chain Pivots:** Hardware constraints are forcing consumer ecosystem operators to seek alternative suppliers. For example, Apple qualified Chinese memory maker CXMT to secure cost-effective hardware for localized, on-device inference in iPhones and MacBooks.[S22]

1.2 Hyperscalers & Data Centres:

Asymmetric Risk Drives a $725B CapEx Boom Against Potentially Severe Power Grid Bottlenecks

Cloud hyperscalers (AWS, MS Azure, GCP, Meta and Oracle) are executing the largest technological infrastructure expansion in history, with aggregate 2026 CapEx reaching $725-$738 Bn (mid-point 65% YoY), on track to approach $1 Tr p.a. by 2027. Roughly 75% ($550+ Bn in 2026) is dedicated strictly to AI infrastructure (GPUs, custom silicon, and specialized data centres). [S4][S3]

- **Asymmetric Risk Management:** Tech giants treat over-building as a manageable, temporary hit to operating margins, whereas under-investing presents an existential threat to cloud market share. This shift has pushed CapEx-to-revenue ratios to historic utility-like levels of 21% to 35%, funded in 2025 with $75-85 Bn from net cash reserves, $140 Bn from operational cashflow, and $108 Bn in corporate debt issued in 2025, and YTD 2026 via c.$200 Bn on-balance sheet debt plus $150-220 Bn SPV commitments. [M1]
- **Elimination of Cash Buffers & Private Credit SPVs:** Excess cash balance buffers have been depleted in 2025 to fund initial CapEx. Hyperscalers are increasingly utilizing off-balance-sheet Special Purpose Vehicles (SPVs) financed by Private Credit funds (with insurance companies underwriting senior facilities) but at higher borrowing costs.
- **Circular Financing & Synthetic Demand Loops[1]:** Hyperscalers and hardware OEMs increasingly finance their main clients via equity investments, convertible facilities, and compute credits (e.g., Microsoft with OpenAI, Amazon and Alphabet with Anthropic, and Nvidia with CoreWeave). This vendor-financing is contractually tied to multi-year, multi-billion-dollar cloud capacity and hardware purchase agreements, recycling invested cash back into reported cloud and chip revenue. This dynamic accelerates cluster utilization but creates an "ouroboros" demand loop, masking the structural deficit in real AI commercial adoption and concentrating systemic credit risk if LLM developers face refinancing cliffs. [S29]
- **Physical Infrastructure Gridlock:** Expansion is colliding with power generation and transmission limits. AI data centres are projected to absorb 15% to 20% of total U.S. electricity demand by 2030 (vs 5.9% in 2025). Regional utilities face unprecedented queue backlogs (e.g. Dominion Energy's 30GW data centre capacity queue in Northern Virginia, which exceeds the state's total existing generating capacity) forcing Big Tech to negotiate direct nuclear Power Purchase Agreements (PPAs). [S6][S7]

1.3 LLM Developers (Proprietary vs. Open-Source):

The "DeepSeek Death Zone" Destroys Closed Unit Economics

The foundational model layer is undergoing rapid commoditization as open-source models, mainly Chinese, eliminate performance differentials relative to closed proprietary LLMs.

- **Unprecedented Capital Burn:** Frontier proprietary LLMs face severe structural cash burn; OpenAI is projected to burn $115 Bn before achieving free-cashflow profits by 2030. [M7]
- **Marginal-Cost Pricing Strategy:** Open-source labs actively price API access near the marginal cost of compute. Inference costs for frontier models dropped by 95% (2023-26), removing closed-model pricing power and leaving infrastructure efficiency as the sole moat. [S19][S21]

[1] Refer to Appendix 1 about Circular Financing & Synthetic Demand Loops

- **"DeepSeek Death Zone" Price Shock:** The rapid release of five open-weight/high-efficiency models in June-July (DeepSeek V4 Flash/Pro, Qwen3.8-Max, Kimi K3, GLM-5.2, LongCat 2.0), showing a very similar performance to latest proprietary LLMs, has substantially altered inference economics (e.g., DeepSeek vs Anthropic[2]):
  - **Frontier Tier Disparity:** running DeepSeek-V4 Pro at peak hours remains 12.5× cheaper than Claude Fable 5, widening to 25× cheaper during off-peak hours.
  - **Workhorse Tier Convergence:** The gap DeepSeek-V4 Flash vs Claude Sonnet 5 has narrowed. Flash is 6x cheaper than Sonnet at peak-hours, 20x cheaper at off-peak hrs.

1.4 AI Software Vendors (AI Apps vs Enterprise SaaS):

Zero-Barrier Code Crushes Thin-AI Wrappers, While Vertical AI Apps and Enterprise SaaS Retain Workflow Moats

AI automated software generation has removed traditional barriers to entry, triggering an operational split within the software application layer.

- **Thin-AI Wrappers Shakeout:** Startups relying on generic foundational APIs without proprietary data integrations face severe customer turnover (65% churn rate), with industry projections indicating that 80% will fail or be liquidated by 2028.

  This commoditization is accelerated by the rapid maturation of Computer-Use Agent (CUA) capabilities. As benchmarked on *OSWorld-Verified*, frontier model desktop navigation performance jumped from 42% in early 2025 to 85% in mid-2026 (led by Claude Fable 5), crossing the 72% human baseline. With UI interaction abstracted at the model layer, thin-wrapper startups aiming to monetize basic execution face full disintermediation. [S8]

- **Context-Native & Vertical AI Apps Emerging:** Distinct from generic AI wrappers, specialized vertical platforms and agentic harness platforms are establishing defensible moats.

  Because raw UI interaction is now commoditized at the model layer, these platforms build defensibility above the model by (1) ingesting proprietary company runbooks, (2) establishing deterministic run-caching architectures (dropping execution costs from $6-$8/hr to <$3.00/hr), and (3) designing multi-agent verification loops for asynchronous, offline edge cases. By remaining model-agnostic (swapping underlying frontier and open-weight models to minimize inference costs) they capture the economic surplus from BPO cannibalization without bearing foundation model research and infrastructure burn.

- **Enterprise SaaS Incumbent Dominance:** Legacy SaaS software platforms (e.g. Salesforce, ServiceNow, SAP) are securing downstream value capture. Their competitive moats rely on workflow integration, regulatory compliance, customer distribution, and proprietary data.

  Also, the AI competitive moat has shifted from raw UI navigation to the context layer. Enterprises and specialized platforms survive by embedding hard-earned company context (e.g. tribal knowledge, deterministic run-caching, and multi-agent verification harnesses). Production demonstrates that long-term margins depend on 'design for failure' architectures (i.e., caching initial loops into cheap deterministic code, re-invoking LLMs solely for self-healing error handling when UIs drift) rather than using continuous, high-cost token loops.

[2] Refer to Appendix 2 about Proprietary vs Open-Source Benchmark Analysis

SaaS incumbents are shifting away from per-seat SaaS licensing (e.g. $X/user/month) toward usage-based, API-metered, and outcome-based pricing structures that monetize autonomous agents as direct productivity units.

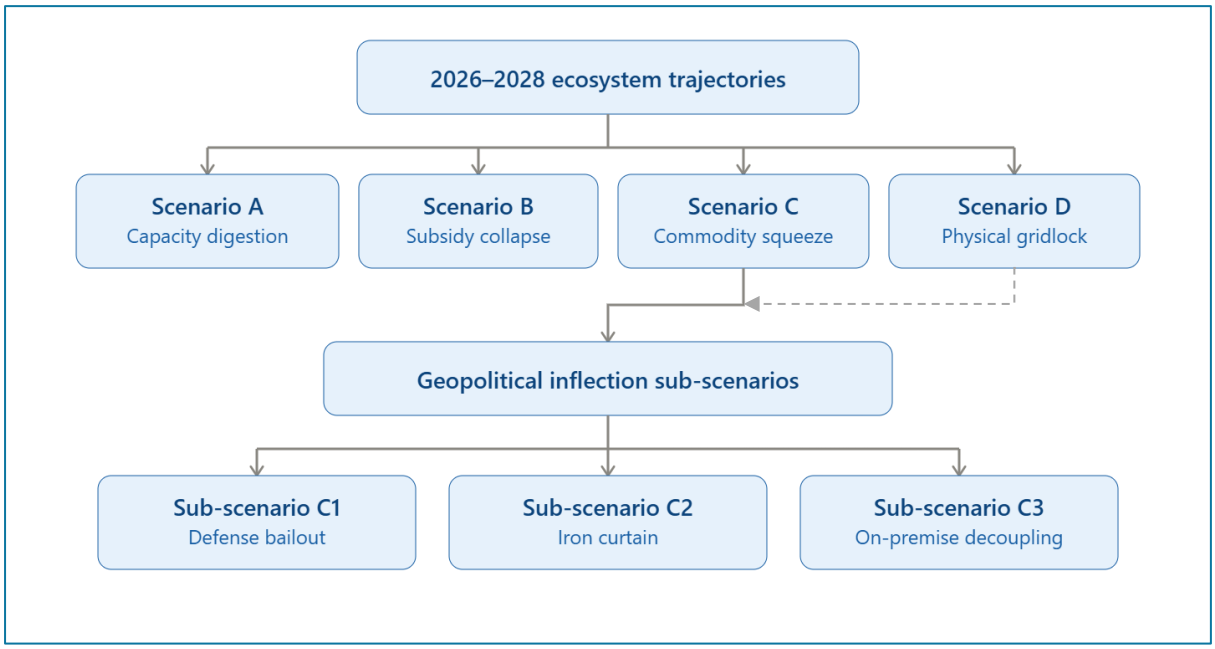
1.5 Enterprise AI End-Users:

Venture LLM Subsidies Disguise Execution Failures in "Pilot Purgatory"

Enterprise adoption of agentic AI remains largely stalled prior to broad organizational deployment.

- **Systemic Implementation Failure**: *MIT/NANDA* research finds only c. 5% of GenAI pilots delivered measurable enterprise impact (i.e., c. 95% fell short under the study definition). Beyond departmental silos (33%) and fragmented data governance (31%), agents stall on verification gaps and asynchronous ground truth. Workflows fail in production when: (1) outputs look visually plausible but contain extraction errors (e.g., misreading contract terms as 'Net 30' vs 'Net 60' into an ERP), or (2) success signals rely on offline human actions (e.g., insurance claim submitted via portal that stalls days later due to a missed phone inquiry). Without bespoke verification harnesses for these cases, frontier LLMs alone cannot push projects out of “Pilot Purgatory”.[S9]
- **Token Subsidy Distortions:** Enterprise buyers have been insulated from actual AI compute costs by a "Token Subsidy", where foundational LLM providers absorb operational losses to capture market share. Because autonomous agentic loops execute continuous multi-step operational queries (consuming tokens 100x to 1000x vs GenAI human tasks), enterprise deployment economics remain highly sensitive to underlying token pricing and volume scalability.

2. Ecosystem Scenarios (2026-28 Horizon)



2.1 Deep-Dive Main Scenarios

Scenario A: "Capacity Digestion" (⇒ Soft-Landing Utilization Trough)

Hyperscalers' infrastructure investment continues ahead of enterprise workflow transformation, creating an operational capacity overhang.

- **Driver:** Infrastructure supply outpaces enterprise process redesign towards agentic adoption during 2026-2028, causing a temporary utilization trough.
- **Impact:** Cloud providers face margin compression, and chip procurement growth moderates. Systemic risk remains contained because the buildout is mostly funded by Big Tech balance sheets and private credit rather than riskier high-yield speculative debt.

Scenario B: "Token Subsidy Collapse" (⇒ SaaS & AI Apps Consolidation)

Tightening funding markets force foundation LLM providers (two of them seeking an IPO in 2026) and AI apps to eliminate subsidized pricing and charge full cost recovery.

- **Driver:** The unwinding of LLM subsidies exposes true compute costs to end-users.
- **Impact:** Unit economics for multi-step autonomous agentic loops invert. Enterprises halt agentic processes, returning to deterministic SaaS tools and localized Small Language Models (SLM). Pure-play AI startups without proprietary data would face liquidation.

Scenario C: "Commoditization Squeeze" (⇒ Proprietary LLM Shakeout & DeepSeek Death Zone)

Rapid open-source LLM releases permanently depress the market price of raw intelligence to the marginal cost of compute, while increasing performance at par with state-of-the-art closed LLM.

- **Driver:** High-performing open-weight models (mainly Chinese) remove the capability advantage once commanded by closed proprietary LLM providers.

- **Impact:** Enterprises swing towards cheaper open-source AI configurations. Closed-model developers burdened by infrastructure commitments experience declining margins and severe cash burn. Value flows upward to silicon vendors and power-secured hyperscalers, and downward to legacy SaaS incumbents holding distribution and enterprise context.

Scenario D: "Physical Gridlock" (⇒ Infrastructure Bottleneck)

Severe electrical grid constraints and transformer shortages, along with community-driven lawsuits, delay cloud data centre completions across major Western markets.

- **Driver:** Interconnection and construction delays prevent centralized hyperscaler data centres from reaching planned compute capacity.
- **Impact:** Compute demand shifts toward energy-efficient architectures, on-device local inference, small edge models, and sovereign, state nuclear energy data clusters.

2.2 Deep-Dive Geopolitical Sub-Scenarios

Scenario C, even reinforced by a subsequent Scenario D, could trigger a geopolitical response by the US and China governments impacting the AI value-chain ecosystem with 3 potential sub-scenarios (i.e. second order tail risks dependent on Scenario C materializing):

Sub-Scenario C1: "Defense Bailout" (⇒ US Proprietary Lab Insolvency & Sovereign Nationalization)

- **Driver:** revenue decay acceleration for US closed LLMs, trapped between compute commitments and big cash flow deficits, could trigger insolvency. Some hyperscalers could try to acquire them, especially if contractually linked (e.g. Microsoft-Open AI), but they might not be in a solid footing to do so. Viewing frontier AI capabilities as vital for national security, the US federal government intervenes under Defense Production Act mandates.
- **Impact:** Commercial venture equity is restructured (or even fully written down) as frontier labs convert into cost-plus government defense utility contractors, capping long-term corporate equity returns while preserving physical compute operations.

Sub-Scenario C2: "Iron Curtain" (⇒ Open-Source Iron Curtain & Global Stack Bifurcation)

- **Driver:** Simultaneous regulatory bans from US sanctions prohibiting domestic corporate deployment of Chinese open-source models, combined with Chinese state export licensing on top-tier open-source to the Global South.
- **Impact:** Western enterprises would likely be locked into higher-cost, regulated model stacks (closed-LLM), while non-aligned markets (LatAm, Southeast Asia, Middle East, Africa) would likely adopt unrestricted, ultra-low-cost models. Global BPO and software execution hubs in non-sanctioned regions would capture a cost advantage over Western competitors.

Sub-Scenario C3: "On-Premise Decoupling" (⇒ End-Users' AI On-Premise & Edge Inference Surge)

- **Driver:** Breakthrough quantization allows near-state-of-the-art reasoning models (13B-27B active parameters) to execute locally on enterprise workstations and edge GPU clusters (i.e., on-premise inference modestly starting now with the Apple's Mac Studio M5 Ultra).
- **Impact:** Corporate buyers would exit public cloud AI APIs to avoid geopolitical compliance risks and recurring token fees. Value would shift away from hyperscaler API revenue toward on-premise hardware OEMs, edge chip designers, and local cybersecurity integrators.

3. AI Revenue, CapEx, Capital Structure & Free Cash Flow Simulation

3.1 AI Revenue Generation Across Value Chain Layers

To establish an observable AI commercial revenue baseline, we identify actual AI-generated revenue disclosures and run-rates across the three primary layers of the ecosystem, so to isolate the last layer "End-User Commercial Revenue".

A) Hardware OEMs [e27 Revenue: min. $400 Bn]

- **NVIDIA Data Centre:** FY2027 revenue consensus reached $343.4 Bn, driven by Blackwell and Vera Rubin architecture shipments. [S13]
- **Custom Enterprise Silicon:** Amazon's in-house Trainium and Graviton chip division reached a $20 Bn annual revenue run rate in Q1 2026. [S11]
- **High-Bandwidth Memory (HBM):** SK Hynix recorded a 58.4% operating margin on HBM3E contracts, with global HBM revenue exceeding $35 Bn in e2026. [S23]

B) Cloud Hyperscalers [e26 Revenue: $80-100 Bn]

- **Microsoft Azure AI:** Annualized AI revenue crossed $37 Bn run-rate in Q2 2026 (+123% YoY growth) (including OpenAI on Azure AI), supported by 60,000+ Azure AI enterprise customers. [S14]
- **Amazon AWS AI:** Disclosed an AI cloud run rate exceeding $15 Bn in Q1 2026, (including Anthropic on AWS Bedrock) with a contracted backlog of $244 Bn (+40% YoY). [S11]
- **Google Cloud Platform AI** (including Anthropic on GC Vertex AI) **+ Meta AI:** estimated combined $30 Bn.

C) LLM Developers & Enterprise SaaS [e26 Revenue: gross $160-175 Bn ⇒ net $128-143 Bn][3]

- **Anthropic:** Annualized revenue of $47 Bn gross run-rate in May 2026 (Series H disclosure), and secondary market reports indicate $60-$65 Bn gross run-rate by end-July 2026 (driven by enterprise adoption of Claude Code). Conservative enterprise tracking indicates 30-35% of Anthropic's gross revenue originates through hyperscaler marketplace agreements, subtracting c.$20 Bn (midpoint) yields a net run-rate of $40-$45 Bn. [S10]
- **OpenAI:** After tracking $20-$25 Bn gross run-rate in early 2026, secondary estimates and indicate $6.7 Bn in 2Q26 revenue, annualizing to $35-$40 Bn gross run-rate by July/August 2026 (driven by 50M+ subscribers, ChatGPT Plus/Pro subscriptions, direct ChatGPT Enterprise seats, first-party API billing, $2.5 Bn inaugural ads). 25-28% of OpenAI volume is intermediated by Azure, subtracting c.$10 Bn (midpoint) leaves $25-$30 Bn net run-rate. [S12]
- **Enterprise SaaS Integrations:** Direct AI revenue generated from corporate end-users (i.e. SaaS productivity tools, AI copilots, and agentic workflows) $65-70 Bn gross run-rate in aggregate revenue in 2026, subtracting[4] c.$2 Bn yields $63-68 Bn net run-rate.

[3] Gross annualized run-rate revenue includes Cloud Hyperscalers' compute distribution (e.g., via AWS Bedrock, Google Cloud Vertex, and Azure OpenAI Service). To maintain non-overlapping accounting integrity against Layer B, inter-company cloud pass-through revenue is excluded from net annualized run-rate revenue for Layer C.

[4] To avoid double-counting: (a) included: software-layer licensing fees, workflow automation seat add-ons, agentic outcome pricing, and application-level usage credits paid by business buyers to application vendors; (b) excluded:

3.2 AI Infrastructure CapEx Actuals & Forecast

Data centre expansion across the Big-5 hyperscalers (Amazon, Microsoft, Alphabet, Meta, and Oracle) has transitioned from asset-light cloud scaling to capital-intensive utility infrastructure.

According to *Morgan Stanley* and consensus market data, total annual hyperscaler CapEx grew c.65% YoY from $443 Bn in 2025 to $725-738 Bn in 2026. Approximately **75% of this capital ($550+ Bn in 2026)** is allocated directly to AI-specific compute, custom silicon, high-bandwidth memory, and power-backed data centres. [S4][S3]

- 2024-2027 Actuals & Estimates: 2026 and 2027 estimates as per analyst consensus.
- 2027-2030 Extension: Modelling Total CapEx with decelerating YOY growth rates (17.2%, 13.8%, and 11.7% respectively), resulting in a moderate Total CapEx CAGR of 14.2%.

Table 3.2: Annual Total vs AI-specific CapEx Schedule ($ Bn)

Year	Total Hyperscaler CapEx	AI Infrastructure Share (%)	Annual AI CapEx
2024	$256	74%	**$190**
2025	$443	74%	**$330**
2026	$732 ($725-$738)	75%	**$550**
2027	$960 ($905-$1,015)	78%	**$750**
2028	$1,125	80%	**$900**
2029	$1,280	82%	**$1,050**
2030	$1,430	84%	**$1,200**

Cumulative AI CapEx buildout (2024-2030) is estimated to be c.$5.0 Tr, roughly 0.59-0.64% of the global cumulative GDP estimates. As a reference, the 'Telecom & Internet Backbone boom' (1996-2001) invested $1.1 Tr (2026 USD: $2.1 Tr), representing 0.56% of the respective cumulative GDP (now, while dark fibre and conduits' useful life is 20-30 years, compute servers' is just 3-5 years). [S30]

3.3 Capital Structure & Financing Parameters[5]

Hyperscalers and project-finance SPVs fund data centre construction through on/off balance-sheet equity and project debt.

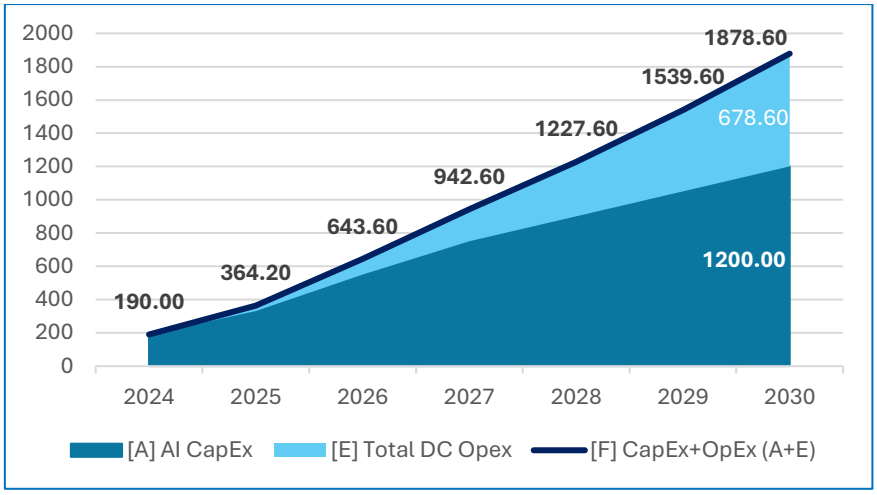
- **Data Centre Opex Factor[6]:** 18% of cumulative CapEx cost annually, 60% power, grid access & energy, 40% facilities and cooling.

Chart 3.3A: AI CapEx vs Total DC Opex ($ Bn)

cloud compute infrastructure and foundation API AI consumption hosted on AWS Bedrock, Google Cloud Vertex, or Microsoft Azure AI (accounted for in Layer (B))

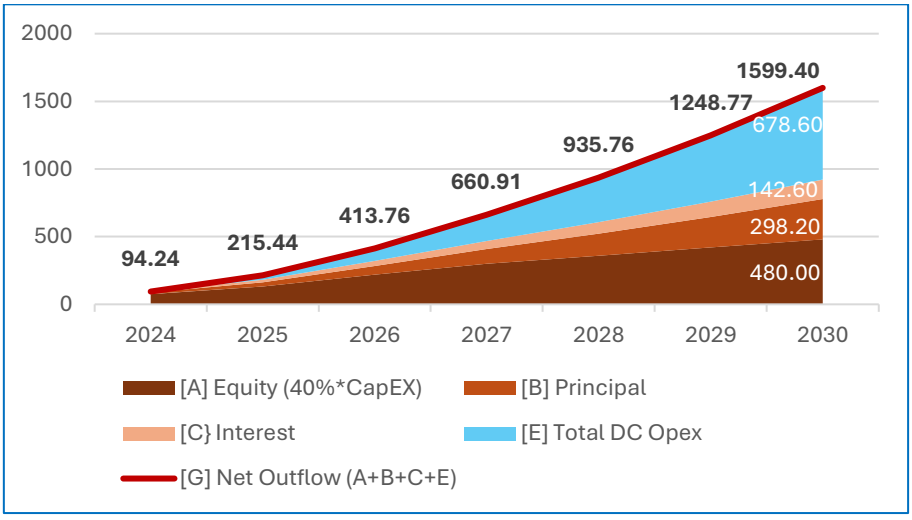
[5] Refer to Appendix 3 about Free Cash Flow & Financing Schedules

[6] Refer to Appendix 4 about Data centre Opex Factor Estimation



- **Debt / Equity Ratio (60 / 40):** Infrastructure shell, power, and rack builds are financed at 60% debt / 40% equity via 144A notes, private credit, and corporate bonds.
- **Cost of Debt ($K_d = 6.0\%$):** Blends Investment-Grade tech bonds (5.25%–5.50%) with data centre project finance notes (6.50%-7.25%).
- **Tenor & Amortization (10 Years):** Matches 10 to 15-year Power Purchase Agreements and reflects the weighted composite lifespan of short-lived servers/chips (4-5 years) and long-lived shell/power infrastructure (15-20 years).

Chart 3.3B: Net Outflow Breakdown ($ Bn)



Key Operational Takeaways:

1. **Power Consumption Escalation:** By 2030, direct annual electricity and power purchases exceed $407 Bn per year, representing 60% of all data centre operating costs. This underscores why hyperscalers are entering direct long-term Power Purchase Agreements with nuclear and utility providers.
2. **Opex Compounding vs CapEx Growth:** While annual CapEx increases by 6.3x from 2024 to 2030 ($190B to $1,200B), annual OpEx scales by 20x ($34.2B to $678.6B, 2025-2030) as cumulative installed physical assets require continuous power, cooling, and maintenance.
3. **Debt Service Burden:** Under a 60% debt-financing model, interest payments reach $142.60 Bn by 2030, creating an annual debt-service obligation (interest + principal) of $440.80 Bn in 2030 that must be covered before equity holders realize target returns.

3.4 Required Annual Revenue Simulation by Project IRR[7]

For a given target annual **Project IRR (r)**, the model calculates for 2024-2030 the annual **End-user Commercial Revenue (R_t)** fitted to a S-Curve for its Revenue Growth Rate (g_t). Then, it solves iteratively to ensure that the present value of realized **Gross Margin (GM)** (assumed at 60% for downstream AI software[8]) covers the present value of the required investment in **AI CapEx and Data Centre OpEx ($Outflow_t$)** offset by the Terminal Value (TV_{2030}, which embeds a perpetual **TV growth rate (g)** calibrated to balance both sides of the equation).

$$R_t = R_{2024} \times (1 + gt)^{t-2024}$$

$$\sum_{t=2024}^{2030} \frac{R_t \times GM}{(1+r)^{t-2024}} = \sum_{t=2024}^{2030} \frac{\text{Outflow}_t}{(1+r)^{t-2024}} - \frac{TV_{2030}}{(1+r)^6}$$

Table 3.5: End-User Realized vs. Required Commercial Revenue vs Project IRR Scenarios ($ Bn)

Scenario / Metric	2024	2025	2026	2027	2028	2029	2030
Actual Realized Revenue	**$35**	**$75**	**$160-175**	--	--	--	--
Required Rev @ 10% IRR	$65	$210	**$400**	$720	$1,120	$1,600	**$2,150**
Required Rev @ 15% IRR	$65	$210	**$450**	$820	$1,300	$1,800	**$2,400**
Required Rev @ 20% IRR	$65	$210	**$600**	$1,100	$1,700	$2,250	**$2,670**

At **15% Project IRR (Baseline)**, the ecosystem requires **$450 Bn in End-user Commercial Revenue by 2026**. This aligns directionally with the $600 Bn revenue requirement identified by *Sequoia Capital* and *Gartner*. With AI spending expected at **$160-175 Bn**, there is a **$275-290 Bn annual revenue gap**.

Crucially, **End-User Revenue required by 2030 is $2,400 Bn, x14.3 the estimated Realized Revenue by 2026 ($160-175 Bn)**. Even for a 10% Project IRR, still $2,150 Bn would be required by 2030.

3.5 Feasibility Assessment & Sensitivity Analysis

Table 3.5A: Sensitivity Matrix A-2030 End-user Revenue vs Project IRR vs Gross Margin ($ Bn)

Gross Margin on AI Services \ Project IRR	-25% IRR (Value Destruction)	8.0% (WACC=IRR)	10.0% IRR	15.0% IRR	20.0% IRR
50% Gross Margin (High Compute Cost)	$1,440	$2,400 *(Floor)*	$2,580	**$2,880**	$3,204
60% Gross Margin (Baseline)	**$1,200**	**$2,000 *(Floor)***	**$2,150**	**$2,400**	**$2,670**
70% Gross Margin (Low Compute Cost)	$1,029	$1,712 *(Floor)*	$1,843	**$2,057**	$2,289

Note: at 60:40 Debt:Equity ratio, at 6% cost of debt, assuming 11% cost of equity, WACC = 8% (pre-tax)

[7] Refer to Appendix 5 about Free Cash Flow Mathematical Formulation

[8] Refer to Appendix 6 about Gross Margin Calibration

For 60% GM, at $2 Tr Revenue by 2030 IRR=WACC (8.0%), reflecting the minimum structural revenue ('Floor') required to sustain zero post-2030 cash burn (EBITDA ≥ $520B Maintenance CapEx[9]).

For each GM scenario, **Revenue below the 'Floor' level generate negative post-2030 free cashflow, making the investment case economically unviable**. If for the 60% GM scenario at 15% IRR, we were to assign **50% of success** to achieving $2.4 Tr Revenue by 2030, the resulting **expected Revenue of $1.2 Tr would be well below the 'Floor', yielding a value destruction of -25% IRR**.

Holding the core Baseline parameters of a 15% Project IRR and a 60% GM constant, a final sensitivity analysis isolates the impact of the perpetual **Terminal Value growth rate (*g*)**. While the Baseline model solves for an implied post-2030 TV growth rate of 8.00% to achieve break-even (NPV = 0), this rate is optimistic compared to 'standard utility' or nominal GDP growth benchmark. The underlying strategic rationale for this elevated Baseline is that successful, enterprise-wide transition to institutional AI will continue to drive compounding demand for computational infrastructure beyond 2030, even as the initial adoption curve fits a standard technology S-curve.

However, to align with conservative valuation practices, the table below models the required 2030 downstream commercial revenue levels across a range of lower TV growth rates (down to a 3.50% 'utility standard') while strictly maintaining the break-even condition (NPV = 0).

Table 3.5B: Sensitivity Matrix B-Required 2030 Revenue vs Terminal Growth Rate (15% Project IRR)

Terminal Growth Rate (*g*)	Required Steady-State FCF ($ Bn)	Required 2030 EBITDA ($ Bn)	Required 2030 Revenue ($ Bn)	Implied Exit Multiple (EV/EBITDA)
8.00% (Baseline)	**$241.51**	**$761.51**	**$2,400.18**	**4.89x**
7.50%	$259.96	$779.96	$2,430.94	4.78x
6.50%	$297.39	$817.39	$2,493.32	4.56x
5.50%	$335.53	$855.53	$2,556.88	4.36x
4.50%	$374.40	$894.40	$2,621.66	4.17x
3.50% (Utility Standard)	**$414.02**	**$934.02**	**$2,687.70**	**3.99x**

Note: Calculations still apply 60% GM, 2030 Operating Costs of $678.60 Bn, and $520.00 Bn post-2030 Maintenance CapEx.

Constraining the **TV growth rate to "utility standard" 3.50% requires c. $2.7 Tr Revenue in 2030**, 12.0% more than the $2.4 Tr Baseline. This would further **widen the "Capital Payback Gap" and raises the AI revenue multiple required by enterprises from 14.3x to c.16x levels**.

In terms of **implied Exit Multiple**, even the Baseline case shows **compression of a large magnitude representing a dramatic de-rating from where hyperscalers and AI-exposed software companies trade today**. At a lower 3.50% TV growth rate, it drops to c.4x EV/EBITDA, logically a more conservative valuation floor.

[9] Refer to Appendix 7 about Maintenance CapEx at Terminal Value

4. End-user Commercial AI Revenue's Dual Engine

To justify the projected **$5.0 Tr cumulative CapEx buildout** by 2030, the AI ecosystem must generate **$2.4 Tr in annual End-user Commercial Revenue** (at a standard 15% IRR and 60% gross margin).

Because current end-user enterprise IT software spending cannot absorb this scale of capital on its own, corporate buyers must fund AI through two distinct economic mechanisms:

- **Labour Productivity Offset** (Enterprises switching white-collar payroll budgets into AI compute, or generating extra productivity to offset AI extra costs; AI vendors taking a part).
- **Software Market Expansion** (SaaS companies monetizing new agentic capabilities while cannibalizing legacy IT Services & Business Process Outsourcing service lines).

Table 4.1: Potential Sources of Annual AI Revenue Requirement (target: $2.4 T by 2030)

ENGINE 1: LABOUR PRODUCTIVITY OFFSET	ENGINE 2: SOFTWARE MARKET CANNIBALIZATION
(End-user: cost reduction from payroll and/or productivity increase / SaaS captures a fraction)	(SaaS: IT service cannibalization / End-user: IT budget shift from Service & BPO to AI SaaS)
• Global Knowledge-W Payroll Base: ~$30.0 Tr • Required Labour Offset[1]: 32% (= $9.6 Tr in white-collar hours captured) • SaaS AI vendor take-rate: 25% (20-33% range) • Economic Impact: SaaS AI generates **$2.4 Tr** revenue from the enterprise's savings	• Global Base Software Market: ~$1.0 Tr • Required Expansion Multiple: 2.4x • Economic Impact: Expands software market by **$2.4 Tr** to $3.4 Tr

1. Net Productivity Gain/Cost Saving over the Total Payroll Base of Knowledge-Workers (front, middle and back)

Why the Revenue Burden Falls on Enterprise Budgets ("Consumer TAM Ceiling")?

Consumer monetization (i.e., individual $20/month subscriptions and digital advertising) is accounted for within foundational lab run-rates (e.g., ChatGPT Plus seats and emerging ad inventory within OpenAI's baseline). However, consumer discretionary spending ($150-$200 p.a. Bn global digital subscription TAM) and global digital ad budgets ($750-$800 Bn p.a. zero-sum marketing spend) cannot mathematically support a $2.4 Tr annual revenue requirement. Enterprise balance sheets, drawing specifically from these above two engines, represent the only capital pools of sufficient macroeconomic scale to fund the infrastructure buildout. [S27] [S28]

4.1 Labour Productivity Offset

A) Global Payroll Base

The **$30.0 Tr global knowledge-worker compensation pool** is derived from macro-labour benchmarks set by *International Labour Organization, World Bank, and McKinsey Global Institute*:

- **Global Labour Compensation Pool:** Total global labour income across all sectors is approximately **$50.0 to $55.0 Tr** annually. [S16]
- **Knowledge Worker Allocation:** White-collar, cognitive, professional services, management, software, and admin. workers represent **55%-60% of total global wage value (~$30.0 Tr).** [S16]
- **Worker Demographics:** This pool comprises approximately 300 million high-wage workers in OECD economies (averaging $80-$100k in fully loaded annual compensation including

benefits, taxes, and overhead) alongside 700 million white-collar workers in emerging economies (averaging $10-$15k annually).

B) Required Labour Offset Benchmark

- For corporate CFOs to justify spending $2.4 Tr annually (by 2030) on AI models, infrastructure, and agentic subscriptions purely as an operational cost-saving measure, the technology must deliver a quantifiable **return against this payroll base**.
- This return must take into account which fraction of this saving will be captured by the SaaS AI vendors enabling it: **25% would be standard take-rate**, within a wider 20%-33% range.

$$\textbf{Labour Offset \%} = \frac{(\text{Required AI Revenue Target} \,/\, \text{SaaS Take-Rate})}{\text{Global Knowledge Worker Payroll}} = \frac{(\$2.40\ \text{Tr} \,/\, 0.25)}{\$30.00\ \text{Tr}} = \mathbf{32.0\%}$$

C) Real-World Economic Rationale

- **Reallocating Operating Budgets:** Enterprises currently spend 70% of operating expenses on labour and 5% on IT/software. CFOs will not approve multi-million-dollar AI deployments out of traditional IT budgets. Instead, **AI spend is approved as a payroll displacement line item**.
- **Capacity Expansion vs. Headcount Reduction:** A **32.0% labour offset** does not imply an immediate 32% global layoff rate. Instead, it represents either:
    1. **Direct Headcount Rationalization:** Eliminating low-complexity cognitive roles (e.g., customer support, document processing, paralegal screening, initial code reviews).
    2. **Capacity Gain Without Hiring:** Absorbing higher output per employee (productivity) without increasing headcount, suppressing future hiring growth.
- **SaaS AI vendor value capture:** A **25% take-rate** over the value of the labour offset enabled by their AI service to the enterprise represents a reasonable compromise for its CFO (implied ROI = 1/25% = 4x), and it could still be possible within 20% (ROI = 5x) to 33% (ROI = 3x). To do so, vendors are moving from seat licenses ($30/month) to outcome/resolution-based billing ($2/ticket or % of recovered revenue), directly taxing the output rather than the seat. [M3]
- **Feasibility Sanity Check:** *MGI* and *Goldman Sachs* estimate GenAI can automate 25% to 30% of work tasks across cognitive occupations over a 10-year horizon. Realizing a **32% net structural gain by 2030** requires executing on above this theoretical potential, making it **feasible ONLY IF agentic AI adoption is wide across all the organization.** [S16]

4.2 Software Market Cannibalization

A) Baseline Software Market

$1.0 Tr base software market is anchored in global IT market tracking data from *Gartner* and *IDC*:

- **Total Global IT Spend: $5.0 Tr annually**, divided across Data Centre IT ($300Bn), IT Services ($1.5Tr), Telecom Services ($1.5Tr), Devices ($700Bn), and Enterprise Software ($1.0Tr). [S15]
- **Enterprise Software Base:** The core software market includes Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), Supply Chain Management (SCM), Security, Analytics, and Collaboration platforms (e.g., Salesforce, ServiceNow, SAP, Microsoft 365, Adobe).

B) Required IT Software Market Expansion

If enterprise software buyers attempt to fund the required **$2.4 Tr AI revenue target** strictly through software and IT budget growth (without taking money from payroll or external services), the enterprise software market must expand dramatically:

$$\textbf{Required Market Expansion Multiple} = \frac{\text{Required AI Revenue}}{\text{Base Enterprise Software Market}} = \frac{\$2.40\text{ Tr}}{\$1.00\text{ Tr}} = \mathbf{2.4x}$$

$$\text{Total Post-AI Software Market (2030)} = \text{Base Software (\$1.0T)} + \text{AI Spend (\$2.4T)} = \$3.40\text{ Tr}$$

C) Real-World Economic Rationale

- **The IT Budget Limitation:** Baseline enterprise software spending grows at a historical compound rate of **10-12% p.a.**, expecting to reach **$1.6 Tr by 2030**. Adding another $2.4 Tr on top of baseline growth would require software spend as a percentage of corporate revenue to double from 3.5% to +8.0%, which is unsustainable for non-tech enterprises.
- **Cannibalization of IT Services & BPO:** To bridge this gap, AI software vendors are targeting the **$1.5 Tr IT Services & Business Process Outsourcing (BPO) market**. Autonomous agents directly displace outsourced human services (call centres, consulting, managed IT services).
- **Shift to Outcome-Based Monetization:** Software vendors are abandoning per-seat licensing (e.g., $30/user/month) in favour of **metered utility and outcome-based pricing** (e.g., $2 per completed customer support resolution, or 10% of revenue recovered by an automated AI sales agent). This shifts software pricing from a fixed overhead expense to a variable revenue-generating production input.
- **Feasibility Sanity Check**: Computer-Use Agents (CUA) vs. Human Labor Costs operational data from mid-2026 production deployments confirm the cost arbitrage:
    - **US Back-Office Labor: $30-$45/hr** (fully loaded, based on *BLS* median wage of $20.59/hr plus benefits/overhead). [S17]
    - **Offshore BPO: $10/hr** ($8-$15/hr range; India, fully loaded). [S8]
    - **Computer-Use Agent: $6-$8/hr** ($3-$15/hr range for inference, depending on screenshot frequency and context retention). [S8]

  Even under 'worst-case' unoptimized screenshot-heavy inference loops ($8/hr), CUAs operate at parity with offshore BPO labour ($10/hr) while yielding a 70%-80% gross margin advantage vs US back-office labour. When wrapped in run-caching harnesses (converting repeatable UI steps into deterministic code), blended execution costs drop below $3.00/hr.

  Enterprise evidence confirms this cannibalization velocity: SaaS platform vendors report live CUA deployments processing 1.5-2.1k IT tickets daily across managed-services contracts with explicit targets to redeploy 20%-25% of human headcount; consumer platforms report running 15M-20M automated monthly portal interactions while reducing dedicated maintenance teams by 50%. [S8]

4.3 The Dual-Engine Budget Reallocation Model

In practice, the **$2.4 Tr required annual revenue** will not come exclusively from cost reduction or IT budget expansion. Instead, we assume enterprise spending would follow a **60 / 40 blended model**.

Table 4.3: Blended Sources of Annual AI Revenue Requirement (target: $2.4 T by 2030)

ENGINE 1: LABOUR PRODUCTIVITY OFFSET (60%)	ENGINE 2: SOFTWARE MARKET CANNIBALIZATION (40%)
(End-user: cost reduction from payroll and/or extra productivity revenue)	(SaaS: IT service cannibalization / End-user: IT budget shift from Service & BPO to AI SaaS)
• Global Knowledge-Worker Payroll Base: ~$30.0 Tr • Required Net Gain[1]: 19.2% (= $5.76 Tr in white-collar hours captured) • SaaS AI vendor take-rate: 25% • Economic Impact: SaaS AI generates **$1.44 Tr** revenue from the enterprise's savings	• Global Base Software Market: ~$1.0 Tr • Required Expansion Multiple: 0.96x • Economic Impact: Expands software market by **$0.96 Tr** to c.$2.0 Tr
Probability Assessment (as of Aug. 2026)	
Medium-Low • 19.2% Net Gain by 2030, on track to 25%-30% over 10-year horizon (MGI/Goldman Sachs), but feasible ONLY IF agentic AI adoption is wide across all the organization.	**Medium-High** • SaaS enterprise evidence confirms this cannibalization velocity is happening

1. Net Productivity Gain/Cost Saving over the Total Payroll Base of Knowledge-Workers (front, middle and back-office)

5. Assessment of the AI Ecosystem's Dynamics & Scenarios' Impact

5.1 Strategic Conclusions for AI Ecosystem Dynamics & Scenarios

Key Takeaway 1: "Pilot Purgatory" extension to 2030 would mean "Agentic AI stagnation"

- **"Pilot Purgatory":** If enterprise **AI adoption remains stuck (with 95% of pilots not delivering measurable enterprise impact)** and fails to deliver demonstrable labour productivity gains, enterprises will keep AI expenditures confined to minor IT software add-on budgets.
- **Restricted Revenue Scenario:** Enterprise spending would **top out at $300-$500 Bn p.a. by 2030**, resulting in a severe **c.$2.0 Tr systemic revenue shortfall** that would trigger an aggressive margin contraction across LLM developers and hyperscalers, leading to the **"Token Subsidy Collapse" (Scenario B).**

Key Takeaway 2: $2.4 Tr 2030 Revenue would demand 14.3x enterprise AI revenue jump

- **Capital Outpacing Realized Revenue:** Hyperscaler data centre expansion is being funded by Big Tech cash flows under an asymmetric risk doctrine (i.e. the existential threat of missing the foundational compute shift outweighs the temporary margin hit of over-investing). However, **current annual commercial revenue of $160-175 Bn covers less than 35-39% of the $450 Bn required in 2026 to generate a 15% IRR**.
- **The Escalating 2030 Deficit: Without an exponential jump in adoption**, the gap between annualized revenue and required return expands from $275-290 Bn today to **$2.0 Tr per year by 2030**, risking severe valuation write-downs across the infrastructure stack, leading to a combination of **"Token Subsidy Collapse" (Scenario B) and "Commoditization Squeeze" (Scenario C)**.

Key Takeaway 3: Bridging the gap would require large corporate payroll & IT budgets reallocation

- **The IT Budget Ceiling:** Global enterprise software spending stands at $1.0 Tr. Enterprise IT budgets cannot absorb a **$2.4 Tr p.a. AI software expansion** on their own without expanding total software spend by an unrealistic **2.4x**.
- **Engine 1 - Labour Productivity Offset (Cost Reduction/Productivity Increase):** Enterprise CFOs must fund AI spend by reducing operating white-collar payroll budgets. Generating **$1.44 Tr p.a. in revenue** by 2030 requires AI agents unlocking a 19.2% net structural productivity gain/cost saving across the $30.0 Tr global knowledge-worker compensation base (assuming a standard 25% IT vendor take-rate).
- **Engine 2 - Software Market Cannibalization (IT Service Cannibalization/IT Budget Shift):** Software vendors are transitioning from seat-based SaaS subscriptions to metered utility and outcome-based pricing. This enables vendors to partly cannibalize $1.5 Tr enterprise IT services & BPO market, generating additional **$0.96 Tr p.a. in revenue** by 2030 from AI SaaS.
- Ultimately, only the hybrid contribution of both engines would ensure a credible path to $2.4 Tr p.a. AI revenue by 2030, leading to a **"Capacity Digestion" (Scenario A).**

Key Takeaway 4: Barbell polarization solidifies the middle layer as the "weakest link"

- **Barbell Margin Capture:** Economic **returns are concentrating at the physical infrastructure layer** (i.e. hardware foundries, memory OEMs -SK Hynix, Micron-, and power-secured

hyperscalers) **and at the downstream enterprise distribution layer** (i.e. legacy SaaS incumbents like Salesforce and ServiceNow holding workflow context and data gravity).

- **"Commoditization Squeeze" (Scenario C): Proprietary LLM labs** (facing very large cash burns through 2030) and **thin-wrapper AI startups** (with 65% churn) remain trapped between non-negotiable hardware costs from above and near-zero marginal cost open-weight models from below, making this middle layer **the weakest link in the AI ecosystem**.

5.2 Sub-Scenarios' Impact Across the AI Value-Chain Layers

The **"Commoditization Squeeze" (Scenario C)** on major LLM developers driven by the open-source competition is beginning to manifest in pricing policy (**"DeepSeek Death Zone"**).

- As of the closure of this report, overall token pricing at leading US proprietary LLM labs has declined by nearly 25% within weeks of the announcement of stellar performance results from the Chinese open-source models.
- While the Tier-1 frontier models maintain flat-to-rising premiums, competition in the mid-tier/workhorse API segment has evolved into an aggressive race: Open AI slashed pricing for its GPT-5.6 Luna lightweight model by 80% (cutting to $0.20/M input and $1.20/M output), and Anthropic launched Claude Opus 5 offering near-flagship intelligence at 50% the price of its top-tier Fable 5 model (at $10/M input and $50/M output), cancelling a scheduled price increase on Sonnet 5. [S18][S19][S20]

This spiral might as well get worse due to grid constraints and other physical restrictions that may substantially delay data centres deployment, leading to a **"Physical Gridlock" (Scenario D)**.

- The *International Energy Agency* now expects global data centre's electricity consumption to roughly double from 485 TWh in 2025 to 950 TWh by 2030. [S6]
- *Wood Mackenzie* recently reported US grid operators and utilities likely committing to just about 28% of the 1,066 gigawatts requested for data centre projects. The explosion of requests is stretching approval timelines, with some developers pitching the same project to multiple utilities. Texas alone has tracked c.474 gigawatts of connection requests (90% from data centres), around 5x Texas system's record peak demand. [S7]

The combination of both scenarios is threatening to undermine US efforts to compete in the global AI race, which might trigger a chain of **government geopolitical interventions** potentially leading to the **3 Sub-scenarios** with different impact across the AI Value-Chain layers.

Table 5.2: Investment Thesis Re-Rating Matrix & Geopolitical Sub-Scenarios[10]

Value-Chain Layer	Sub-Scenario C1 (Defense bailout)	Sub-Scenario C2 (Iron curtain)	Sub-Scenario C3 (On-premise decoupling)
Chip Producers & Hardware	**Neutral** (Cost-Plus)	**Positive** (Edge Hardware Surge)	**Very Positive** (Local Compute Surge)
Cloud Hyperscalers	**Neutral** (Defense Utility)	**Negative** (Lost Global BPO Revenue)	**Very Negative** (Cloud Disintermediation)

[10] Refer to page 8 for Geopolitical Sub-Scenarios' description

Value-Chain Layer	Sub-Scenario C1 (Defense bailout)	Sub-Scenario C2 (Iron curtain)	Sub-Scenario C3 (On-premise decoupling)
Proprietary LLM Labs	**Very Negative** (Equity Restructure)	**Negative** (Trapped in High-Cost Stack)	**Negative** (Model Arbitraged by On-premise)
Thin-Wrapper AI Apps Start-ups	**Very Negative** (Liquidation)	**Very Negative** (Margin Collapse)	**Very Negative** (Disintermediated)
Context-Native & Vertical AI Apps	**Positive** (Govt & Enterprise Agility)	**Very Positive** (Model Arbitrage Swapping)	**Positive** (Local / Hybrid Appliance Deployments)
Enterprise SaaS Incumbents	**Positive** (Govt Defense Work)	**Mixed** (Dual-Stack Overhead / Leverage)	**Positive** (Hybrid/On-Prem Execution)
Enterprise End-Users	**Neutral** (Regulatory Tax)	**Positive** (Global South AI Cost Edge)	**Positive** (OpEx Cost Control)

Note: the content in this table and narrative is for informational purposes only and does not constitute investment advice

The **aggregate impact across the three Sub-Scenarios** will define winners vs losers in the AI value-chain.

1. **Chip Producers & Hardware (Positive):** Physical hardware remains highly defensible. Sub-Scenarios C2, and especially C3, drive massive demand for on-premise enterprise GPUs, workstations and high-bandwidth memory.
2. **Cloud Hyperscalers (Negative):** In C2, hyperscalers lose AI revenue across the Global South. In C3, become vulnerable to cloud disintermediation if corporate inference moves locally on-premises running efficient open weights, impairing returns on projected annual CapEx.
3. **Proprietary LLM Labs (Negative):** Trapped in the "DeepSeek Death Zone". Commercial equity faces severe restructuring, even write-downs or nationalization under government's Defense Production Act mandates (Sub-Scenario C1).
4. **Thin-Wrapper AI Apps Start-ups (Very Negative):** Likely disintermediated across all scenarios. Lacking proprietary data, custom execution harnesses, or deterministic run-caching, they are crushed between commoditized model-layer computer-use capabilities and the distribution gravity of legacy incumbents.
5. **Context-Native & Vertical AI Apps (Positive):** Highly agile and defensively positioned. In C1, re-routing to sovereign-backed LLMs or local open weights. In C2, gross margin expansion by swapping Western LLMs for ultra-cheap open-source models. In C3, lightweight orchestration and deterministic run-caching harnesses can be deployed directly on-premise.
6. **Enterprise SaaS Incumbents (Positive):** Legacy software platforms leveraging proprietary data gravity, compliance frameworks, and flexible hybrid/on-premise deployment options capture long-term downstream AI surplus. In C2, a material risk is whether those acquired via PE leveraged buy-outs (i.e., 10-11x leverage in 2020-22) can avoid debt restructurings. [S26]
7. **Enterprise End-Users (Selective Positive):** In C2, global enterprises with presence in the Global South deploying low-cost open-weight models capture an immediate 25x operational cost advantage over Western peers sticking to closed LLMs. In C3, all would be able to own compute on-premise, drastically controlling AI OpEx costs.

Appendix 1 – Circular Financing & Synthetic Demand Loop

Recent evidence of circular financing within the AI Value Chain:

A) Scale of Capital Recycling:

- **Microsoft ↔ OpenAI:** Microsoft's cumulative investments (c.$13.8 Bn) are paired with commitments by OpenAI to purchase tens of billions in Azure compute capacity through 2030+.
- **Amazon / Alphabet ↔ Anthropic:** Amazon's committed capital (c.$8 Bn) and Alphabet's minority stakes are tied directly to multi-year infrastructure agreements on AWS (Trainium/Bedrock) and Google Cloud Vertex.
- **Nvidia ↔ Neoclouds & Labs:** Nvidia participates as a strategic equity investor in GPU cloud aggregators (i.e., CoreWeave, Nebius) and LLM labs (including OpenAI), which concurrently commit their raised equity and debt proceeds to purchasing Nvidia silicon.

B) Credit & Accounting Scrutiny:

- The **Bank for International Settlements (BIS)** and institutional credit desks have raised concerns regarding the economic substance of these transactions: an investment outflow recorded on the cash flow/investing ledger returns on the income statement as high-margin operational cloud revenue.
- The **US Federal Trade Commission (FTC)** and the **UK Competition and Markets Authority (CMA)** launched formal inquiries into foundation lab partnerships, scrutinizing whether these joint structures distort market demand signals and lock customers into proprietary infrastructure rails.

C) Systemic Credit Risk:

- If a LLM lab's burn rate triggers a liquidity crisis, hyperscalers face a double impairment: equity markdowns on their balance sheets alongside the sudden evaporation of contracted cloud backlog revenue.

Appendix 2 – Proprietary vs Open-Source Benchmark Analysis

The like-for-like comparisons below break down the models into two tiers: Tier 1 (Frontier & Autonomous Reasoning) and Tier 2 (Enterprise Workhorse & High-Efficiency). All DeepSeek figures reflect the new Peak / Off-Peak pricing schedule (since Aug. 16th).

Table 1: Frontier & Autonomous Reasoning (Tier 1)

Like-for-like comparison for deep multi-step reasoning, complex codebase orchestration, and autonomous agent loops.

Metric / Parameter	Claude Fable 5 (Anthropic)	DeepSeek-V4 Pro (DeepSeek)	Delta / Cost Multiple
Model Category	Mythos-class Frontier	1.6T Mixture of Experts	—
Context Window	1,000,000 tokens	1,000,000 tokens	Parity
Standard Input *(Cache Miss / 1M)*	$10.00	$0.66 *(Off-Peak)* $1.32 *(Peak)*	DeepSeek is 7.6× to 15.2× cheaper
Cached Input *(Cache Hit / 1M)*	$1.00 *(90% discount)*	$0.022 *(Off-Peak)* $0.044 *(Peak)*	DeepSeek is 22.7× to 45.5× cheaper
Standard Output *(per 1M)*	$50.00	$1.98 *(Off-Peak)* $3.96 *(Peak)*	DeepSeek is 12.6× to 25.3× cheaper
Primary Strength	Autonomous multi-file refactoring, zero-shot tool reliability	High-volume competition math, symbolic logic, low-cost evals	—

Table 2: Enterprise Workhorse & High-Efficiency (Tier 2)

Like-for-like comparison for balanced daily tool calling, structured data extraction, summarization, and high-throughput pipelines.

Metric / Parameter	Claude Sonnet 5 (Anthropic)	DeepSeek-V4 Flash (DeepSeek)	Delta / Cost Multiple
Model Category	Balanced Workhorse	284B Mixture of Experts	—
Context Window	1,000,000 tokens	1,000,000 tokens	Parity
Standard Input *(Cache Miss / 1M)*	$2.00	$0.22 *(Off-Peak)* $0.44 *(Peak)*	DeepSeek is 4.5× to 9.1× cheaper
Cached Input *(Cache Hit / 1M)*	$0.20 *(90% discount)*	$0.007 *(Off-Peak)* $0.014 *(Peak)*	DeepSeek is 14.3× to 28.6× cheaper
Standard Output *(per 1M)*	$10.00	$0.66 *(Off-Peak)* $1.32 *(Peak)*	DeepSeek is 7.6× to 15.2× cheaper
Primary Strength	Native adaptive thinking, instruction adherence, complex workflows	Ultra-low latency, commodity classification, scalable batch ingestion	—

Summary Takeaways

1. **Frontier Tier Disparity:** Even following the peak-hour price quadrupling, running DeepSeek-V4 Pro at peak hours remains **12.5× cheaper** than Claude Fable 5, widening to **25× cheaper** during off-peak hours.
2. **Workhorse Tier Convergence:** The gap between Claude Sonnet 5 and DeepSeek-V4 Flash has narrowed significantly. At peak rates Flash is now within **6x cheaper** price multiples of Sonnet, widening to **20x cheaper** during off-peak hours, making reliability and tool-use precision the primary decision criteria for enterprise routing.

Appendix 3 – Free Cash Flow & Financing Schedules

$$\text{Total Unlevered Outflow}_t = \text{Annual AI CapEx}_t + \text{Total Data Centre Opex}_t$$

$$\text{Total Data Centre Opex}_t = 18\% \times \text{BoY Cumulative CapEx Base}_{t-1}$$

$$\text{Power \& Energy Opex}_t = 60\% \times \text{Total Data Centre Opex}_t$$

$$\text{Facilities \& Cooling Opex}_t = 40\% \times \text{Total Data Centre Opex}_t$$

Table A: Comprehensive Unlevered Cash Outflow Schedule ($ Bn)

Year	[A] AI CapEx	[B] BoY CumCapEx Base	[C] Power (60%*E)	[D] Facility (40%*E)	[E] Total DC OpEx (18%*B)	[F] CapEx+OpEx (A+E)
2024	$190.00	$0.00	$0.00	$0.00	$0.00	**$190.00**
2025	$330.00	$190.00	$20.52	$13.68	$34.20	**$364.20**
2026	$550.00	$520.00	$56.16	$37.44	$93.60	**$643.60**
2027	$750.00	$1,070.00	$115.56	$77.04	$192.60	**$942.60**
2028	$900.00	$1,820.00	$196.56	$131.04	$327.60	**$1,227.60**
2029	$1,050.00	$2,720.00	$293.76	$195.84	$489.60	**$1,539.60**
2030	$1,200.00	$3,770.00	$407.16	$271.44	$678.60	**$1,878.60**
Total	**$4,970.00**	--	**$1,089.72**	**$726.48**	**$1,816.20**	**$6,786.20**

Note: BoY = Beginning of Year

Table B: Capital Structure & Levered Financing Cash Outflow Schedule ($ Bn)

Year	[A] AI CapEx	[B] Equity (40%*A)	[C] Debt Drawn (60%*A)	[D] Debt EoY (D_{t-1}+C-F)	[E] Interest (6%*(D_{t-1}+C))	[F] Principal	[G] Levered Outflow (B+E+F+ DC OpEx[1])
2024	$190.00	$76.00	$114.00	$102.60	$6.84	$11.40	**$94.24**
2025	$330.00	$132.00	$198.00	$269.40	$18.04	$31.20	**$215.44**
2026	$550.00	$220.00	$330.00	$535.20	$35.96	$64.20	**$413.76**
2027	$750.00	$300.00	$450.00	$876.00	$59.11	$109.20	**$660.91**
2028	$900.00	$360.00	$540.00	$1,252.80	$84.96	$163.20	**$935.76**

Year	[A] AI CapEx	[B] Equity (40%*A)	[C] Debt Drawn (60%*A)	[D] Debt EoY (D_{t-1}+C-F)	[E] Interest (6%*(D_{t-1}+C))	[F] Principal	[G] Levered Outflow (B+E+F+ DC OpEx[1])
2029	$1,050.00	$420.00	$630.00	$1,656.60	$112.97	$226.20	**$1,248.77**
2030	$1,200.00	$480.00	$720.00	$2,078.40	$142.60	$298.20	**$1,599.40**
Total	**$4,970.00**	**$1,988.00**	**$2,982.00**	--	**$460.48**	**$903.60**	**$5,168.28**

Note: 60:40 Debt:Equity, 6.0% interest rate, and 10-year linear amortization tenor. EoY = End of Year. 1. Refer to Table A.

Appendix 4 – Data Centre OpEx Estimation

The **Data Centre OpEx (18% of CapEx)** factor represents the annual cash cost required to power, cool, secure, and maintain high-density accelerated AI infrastructure.

This factor is derived from empirical Total-Cost-of-Ownership (TCO) engineering models and institutional infrastructure benchmarks (Epoch AI, SemiAnalysis, Morgan Stanley, Uptime Institute), reflecting the shift from low-density CPU enterprise cloud (5–10 kW/rack) to ultra-high-density GPU/accelerator clusters (40–120+ kW/rack).

1. Structural Breakdown of the DC OpEx factor

Annual Data Centre OpEx is split into two primary operational categories:

A. Power & Energy Consumption (10.8% of CapEx; 60% of DC Opex):

Electricity is the single largest variable cash operating expense in high-density AI clusters:

- **High Rack Power Density:** Modern GPU systems (e.g., NVIDIA NVL72 racks) draw 120 kW to 140 kW per rack, compared to traditional cloud server racks operating at 5 kW to 10 kW.
- **Continuous Load Utilization (85%–90%):** Unlike enterprise IT servers operating at 30%–50% average utilization, frontier LLM training and distributed agentic inference clusters run near continuous 24/7 duty cycles.
- **PUE (Power Usage Effectiveness):** Assumes a modern, optimized liquid-cooled data centre with a PUE of 1.15 to 1.25 (meaning 15%–25% of overhead energy is consumed by heat rejection, pumps, and power distribution units).
- **Blended Electricity Rates ($75–$110/MWh):** Reflects wholesale industrial electricity tariffs ($0.075 to $0.11/kWh) blended with dedicated nuclear, geothermal, and solar-plus-storage Power Purchase Agreements (PPAs), long-term capacity reservation charges, and transmission upgrade tariffs.
- **Worked Example (1 GW Facility):** A 1-gigawatt AI cluster supported by $38B–$40B in cumulative infrastructure and server CapEx consumes $600M–$900M annually in direct grid power, matching the 10.8% annual Energy-to-CapEx ratio.

B. Facilities, Liquid Cooling & Site Operations (6.5% to 8.3% ⇒ 7.2% of CapEx; 40% of DC Opex):

It covers physical facility overhead, environmental compliance, and specialized engineering:

- **Direct-to-Chip Liquid Cooling Infrastructure (2.5%-3.0% of CapEx):** Closed-loop direct-to-chip cooling, Coolant Distribution Units, dielectric fluid maintenance, pump replacements, and chemical water-treatment filtration require specialized ongoing maintenance.
- **High-Bandwidth Network Transit & Dark Fiber (1.5%-2.0% of CapEx):** Leased dark fiber corridors, Dense Wavelength Division Multiplexing (DWDM) optical transceiver maintenance, and redundant carrier-neutral internet exchange fees connecting distributed clusters.
- **Property Leases, Taxes & Insurance (1.5%-1.8% of CapEx):** High-voltage substation interconnection rights-of-way, municipal property taxes (or PILOT agreements), and commercial property/casualty insurance on multi-billion-dollar rack installations.

- **Specialized Engineering & Physical Security (1.0%-1.5% of CapEx):** 24/7 Network Operations Center (NOC) site reliability engineers, thermal technicians, on-site hardware swap technicians, and multi-tier physical and biometric security compliance (SOC 2, ISO 27001).

2. Industry Comparison: Traditional Cloud vs. High-Density AI

The 18% ratio reflects the physics and economics of accelerated compute:

Metric / Parameter	Traditional Cloud (CPU-Centric)	Accelerated AI Compute (GPU-Centric)
Average Rack Density	5 – 12 kW / rack	**40 – 140+ kW / rack**
Cooling Architecture	Standard Air / CRAC Chillers	**Direct-to-Chip Liquid + Rear-Door Heat Exchangers**
Average Duty Cycle	35% – 50% utilization	**80% – 90%+ continuous duty**
Facility OpEx-to-CapEx Ratio	**7.0% – 10.0% / year**	**15.0% – 22.0% / year**
Energy Share of Cash OpEx	25% – 35%	**55% – 70% (Modelled: 60.0%)**

Appendix 5 – Free Cash Flow Mathematical Formulation

How much End-user AI Commercial Revenue is required for the aggregate AI infrastructure investment to achieve a specified Return on the Total Capital deployed?

A) Standalone Single-Year Capital Recovery Yield (r)

Assuming a 60% Gross Margin (*GM*) on downstream AI enterprise software and agentic services (accounting for developer salaries, enterprise sales, and platform integration costs), the required annual End-user Commercial Revenue (R_t) needed to validate the investment trajectory as per the Net Outflow ($Outflow_t$): [S31]

$$\text{FCF}_t = (R_t \times GM) - \text{Outflow}_t$$

To generate a target annual Project IRR (r) over the cumulative installed asset base (K_{t-1} = BOY Installed CapEx Base), the annual Free Cash Flow (FCF_t) must equal the target capital yield:

$$\text{FCF}_t = r \times K_{t-1}$$

Substituting this capital return requirement into the previous formula gives the annual End-user Commercial Revenue in terms of IRR Capital Recovery for a given year:

$$R_t = \frac{\text{Outflow}_t + (r \times K_{t-1})}{GM}$$

Variable Definitions

- R_t: Required End-User Commercial Revenue in year t
- Outflow_t: Total Unlevered Cash Outflow in year t (Annual AI CapEx_t + Total Data Centre OpEx_t) as per Appendix 3- Table A. Financing flows in Appendix 3-Table B (i.e., debt drawn, principal repayment and interest) are excluded because this calculation evaluates project returns on total capital invested.
- K_{t-1}: Cumulative Installed CapEx Base at Beginning of Year t.
- r: Target Internal Rate of Return (unlevered project/asset-level return).
- GM: Downstream Enterprise Software Gross Margin (Base assumption = 60%).

B) Multi-Year DCF Cumulative IRR

The model must evaluate R_t over a continuous multi-year period (2024-2030) matching a target Project IRR (r), Net Present Value (NPV) of all FCF, plus the Terminal Value (TV_{2030}) must equal zero:

$$\text{NPV} = \sum_{t=2024}^{2030} \frac{(R_t \times GM) - \text{Outflow}_t}{(1+r)^{t-2024}} + \frac{\text{TV}_{2030}}{(1+r)^{2030-2024}} = 0$$

Isolating the discounted revenue stream:

$$\sum_{t=2024}^{2030} \frac{R_t \times GM}{(1+r)^{t-2024}} = \sum_{t=2024}^{2030} \frac{\text{Outflow}_t}{(1+r)^{t-2024}} - \frac{\text{TV}_{2030}}{(1+r)^{6}}$$

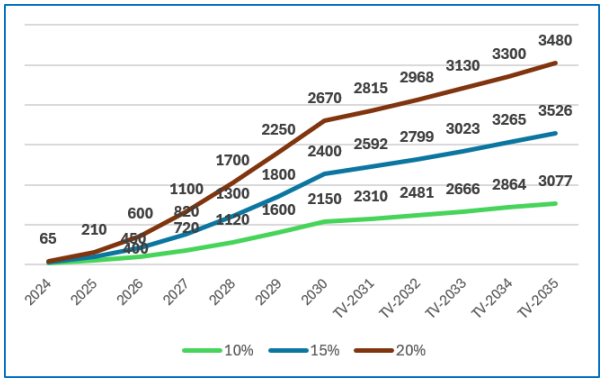
Embedding an exponential S-Curve adoption in the Revenue Growth Rate (g_t), $R_t = R_{2024} \times (1+gt)^{t-2024}$, R_{2030} is solved iteratively to ensure the present value of realized gross margins covers the present value of capital outlays and operating expenses. The S-Curve flattens after 2030 during the Terminal Value.

S-Curve Decay Multipliers & Annual Revenue Growth Rates (R_t, 2024-2030, $ Bn)

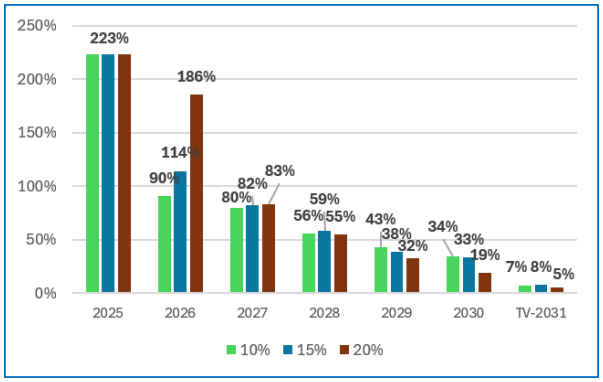
		IRR = 0.1			IRR = 0.15			IRR = 0.2		
Year	[A] Revenue (R_{t-1})	[B] S-Curve Decay Multiplier	[B] Revenue Growth	[C] Revenue (R_t)	[B] S-Curve Decay Multiplier	[B] Revenue Growth	[C] Revenue (R_t)	[B] S-Curve Decay Multiplier	[B] Revenue Growth	[C] Revenue (R_t)
2024	0.00			65.00			65.00			65.00
2025	65.00		223.08%	210.00		223.08%	210.00		223.08%	210.00
2026	210.00	0.40559	90.48%	**400.00**	0.51232	114.29%	**450.00**	0.83252	185.72%	**600.00**
2027	400.00	0.88419	80.00%	**720.00**	0.71943	82.22%	**820.00**	0.44871	83.33%	**1100.00**
2028	720.00	0.69444	55.55%	**1120.00**	0.71194	58.54%	**1300.00**	0.65455	54.55%	**1700.00**
2029	1120.00	0.77143	42.86%	**1600.00**	0.65705	38.46%	**1800.00**	0.59313	32.35%	**2250.00**
2030	1600.00	0.80210	34.38%	**2150.00**	0.86666	33.33%	**2400.00**	0.57699	18.67%	**2670.00**

Note 1: 2024 and 2025 are actual Revenue numbers.
Note 2: S-Curve Decay Multipliers are proxies for the logit function using a Limiting Factor (Target Addressable Market) of $3.4 Tr [refer to section 4: $1.0 Tr from Base Software Market + $2.4 Tr from IT & BPO Services]

S-Curve Annual Revenue (R_t, 2024-30 & TV, $ Bn)



Annual Revenue growth rates (2025-2030 & TV)



C) Terminal Value Methodology

The **Terminal Value at year 2030 (**TV_{2030}**)** represents the residual economic value of the cumulative physical infrastructure asset base (data centre shells, power grid interconnections, land, and compute clusters) beyond 2030.

The model evaluates TV_{2030} using two complementary corporate finance methods: the Perpetual Growth (Gordon Growth) Method (primary approach) and the Exit Multiple Method (cross-check).

1. Primary Method: Perpetual Growth (Gordon Growth)

The Perpetual Growth method assumes that after 2030, the AI infrastructure market transitions from hypergrowth mode into a mature utility asset growing at a long-term perpetual growth rate (g).

$$TV_{2030} = \frac{FCF_{2031}}{r - g} = \frac{FCF_{2030,normalized} \times (1 + g)}{r - g}$$

Parameter Assumptions:

- r = Target Internal Rate of Return (e.g. 15%).
- $EBITDA_{2030} = (R_{2030} \times GM) - Opex_{2030}$.
- Maintenance $CapEx_{2030} \approx \$520.00$ Bn (c.10.5% of cumulative $4.97 Tr installed base)[11].
- $FCF_{2030,normalized} = EBITDA_{2030} - Maintenance\ CapEx_{2030}$.
- g: Long-term perpetual Terminal Growth Rate (aligned with expected long-term global nominal GDP growth for (a) utility infrastructure or (b) digital infrastructure; its value will depend on the simultaneous calibration of gt from 2024 to 2030 so that ensures NPV = 0).

[11] Refer to Appendix 5

2. Cross-Check Method: Exit Multiple (EV / EBITDA)

To validate the Gordon Growth result, the Exit Multiple Method cross-checks TV_{2030} using an Enterprise Value to EBITDA exit multiple benchmarked against mature utility/software infrastructure.

$$TV_{2030} = EBITDA_{2030} \times \text{EV/EBITDA Exit Multiple}$$

Target Exit Multiple Range:

- 3.0x-6.0x EV/EBITDA => Reflects lower TV growth rate, more of a utility infrastructure company multiple.
- 6.0x-10.0x EV/EBITDA => Reflects higher TV growth rate, more of a mature digital infrastructure company multiple.

Summary Table: Terminal Value across Project IRR Scenarios ($ Bn)

Metric / Parameter	10.0% IRR	15.0% IRR (Baseline)	20.0% IRR
Historical Anchors (R_{2024}/ R_{2025})	$65.00 / $210.00	$65.00 / $210.00	$65.00 / $210.00
2030 Required Revenue Target (R_{2030})	$2,150.00	$2,400.00	$2,670.00
2030 Operational EBITDA	$611.40	$761.40	$923.40
Normalized Steady-State FCF	+$91.54	+$241.54	+$403.54
PV of 2024–2030 Operating Cash Flows	-$2,156.33	-$1,610.94	-$978.93
Break-Even Terminal Value (TV_{2030})	$3,826.48	$3,726.16	$2,922.34
PV of Break-Even TV (NPV = $0.00)	+$2,159.95	+$1,610.92	+$978.69
Terminal Growth Rate (g)	7.43%	8.00%	5.44%
Resulting Exit Multiple (EV/EBITDA)	6.26x	4.89x	3.16x

Note 1: All IRR scenarios represent unlevered project/asset-level target returns on total AI CapEx and associated operating cash outflows. They are not levered equity IRRs. The 8% WACC represents the project's estimated cost of capital and therefore its minimum economic break-even hurdle; the 10%, 15% and 20% scenarios test progressively higher project-level return requirements.

Note 2: The mathematically solved Terminal Growth Rate (g) represent the required break-even trajectory to achieve NPV=0 against peak 2024–2030 CapEx outlays. If Terminal Growth Rate were constrained to a utility standard rate (e.g., $g = 3.5\%$), the required 2030 operational revenue target must scale higher to compensate for lower Terminal Value capitalization.

Appendix 6 – Gross Margin Calibration

In traditional enterprise SaaS, gross margins reliably range between **75% and 85%** due to near-zero marginal distribution costs and semi-fixed database hosting. For downstream AI applications and agentic software (Layer C), empirical venture capital data and institutional equity research establish that gross margins compress to a **50%–65% corridor (anchoring our 60% baseline)** due to three structural cost drivers:

- **Variable Inference as Direct COGS:** Unlike traditional deterministic software where an extra user query costs fractions of a cent, generative and agentic workflows require continuous model inference (token consumption, context caching, and GPU cluster time). Accounting standards require customer-facing API token bills and dedicated cloud model hosting to be booked directly under Cost of Goods Sold (COGS), immediately lopping 15 to 25 percentage points off gross margin.
- **The Multi-Step Agentic Multiplier:** Autonomous Computer-Use Agents (CUAs) and multi-agent reasoning loops execute dozens to hundreds of inference steps per business task. Even with prompt caching and local SLMs, serving complex workflows creates a linear cost-to-serve that behaves more like a utility service than traditional zero-marginal-cost code.
- **Maintenance, Verification, and Context Harnesses:** Production-grade enterprise AI requires ongoing domain fine-tuning, deterministic run-caching maintenance, prompt engineering, and human-in-the-loop verification for edge cases. These operational expenses sit inside the cost of delivery, cementing a structural margin ceiling below legacy SaaS.

Overall, the Gross Margin represents downstream AI vendor economics after vendor-level delivery COGS (the separate AI CapEx and Data-centre OpEx in Appendices 3 & 4 represent the upstream physical capital required to support the ecosystem and are therefore assessed separately in the aggregate Project-IRR framework).

Key Empirical Benchmarks [S31]

Research Source	Benchmark / Finding	Coverage & Sample
Andreessen Horowitz (a16z)	**50%-60% Gross Margin**	Foundational benchmark for AI applications and foundation-model API layers, vs. 80%+ for legacy software.
ICONIQ Growth / State of AI	**52%-58% Average Gross Margin**	Tracked across c.300 enterprise software executives, showing AI-native products averaging c.52% in 2026.
Bessemer Venture Partners (BVP)	**c.65% Gross Margin Ceiling**	Cloud Index and State of AI tracking; highlights the "blend lag" where heavy AI usage pulls blended SaaS margins down toward 60%.
Boston Consulting Group (BCG)	**50%-60% Short-to-Medium Term Margin**	Enterprise software analysis modelling the margin dilution of agentic software architectures.

Appendix 7 – Maintenance CapEx at Terminal Value

In corporate finance and digital infrastructure modelling, the **Maintenance CapEx at Terminal Value** can be estimated as the **weighted composite annual replacement rate** across two distinct asset classes in an AI data centre:

$$\text{Composite Replacement Rate} = \sum \left(\text{Asset Class Weight\%} \times \frac{1}{\text{Useful Life in Years}} \right)$$

A) Short-Lived Silicon & IT Hardware (35% of Total Base):

- **Components**: GPUs (Blackwell/Rubin), High-Bandwidth Memory (HBM3e/4), enterprise switches, liquid cooling racks.
- **Replacement Life**: 4.5 years (Depreciation / Obsolescence Rate = 1/4.5 = 22.22% per year).
- **Weighted Contribution**: 35% * 22.22% = 7.78%

B) Long-Lived Physical Infrastructure (65% of Total Base):

- **Components**: Land, concrete building shells, high-voltage substations, power grid interconnects, water chilling plants.
- **Replacement Life**: 24.3 years (Depreciation / Maintenance Rate = 1/24.3 = 4.12% per year).
- **Weighted Contribution**: 65% * 4.12% = 2.68%

C) Total Blended Rate = 7.78% + 2.68% = 10.46%

Applying this **Blended Rate** to the **Cumulative AI CapEx**[12] ($4,970 Bn by2030) yields the annual steady-state **Maintenance CapEx needed post-2030**:

$$\textbf{Maintenance CapEx}_{\textbf{Post}-\textbf{2030}} = \mathbf{10.46\%} \times \mathbf{\$4,970.00\ Bn} = \mathbf{\$520.00\ Bn\ per\ year}$$

[12] Refer to Appendix 2

Sources

Source Mapping

Source markers [S1]-[S30] in the body identify external evidence supporting the principal factual and market-data claims. Model markers [M1]-[M7] identify material model assumptions or analytical choices, not external facts. Where a statement is scenario analysis or strategic inference rather than an externally observable fact, no source marker is used.

S1 Sequoia Capital / David Cahn, "AI's $600B Question", 2024. Supports: the historical AI infrastructure investment-vs-revenue framing and the original "$600B" revenue-requirement debate.

S2 Epoch AI, "NVIDIA B200 Manufacturing Cost and Bill-of-Materials Breakdown", December 2025. Supports: B200 estimated manufacturing cost and HBM share of bill-of-materials.

S3 Morgan Stanley Research, "AI's Silicon Backbone", 2025. Supports: semiconductor / accelerator supply-chain structure and the strategic importance of memory and compute infrastructure.

S4 Morgan Stanley Research, "AI CapEx 2026: $740 Bn Signals Bank Tailwinds", Alvaro Serrano & Manan Gosalia, March 11, 2026. Supports: 2026 hyperscaler CapEx estimates and the scale of the infrastructure investment cycle.

S5 S&P Global Ratings, "Global Infrastructure and Project Finance: Key Data Center Financing Takeaways From PPIF 2026", February 6, 2026. Supports: data-centre project-finance structures, debt capacity and infrastructure financing context.

S6 International Energy Agency, "Key Questions on Energy and AI", April 16, 2026. Supports: global data-centre electricity demand trajectory, power-system constraints and AI/energy nexus. https://www.iea.org/reports/key-questions-on-energy-and-ai

S7 Bloomberg, "Most Power Sought for US Data Centers Will Never Materialize", Gabriel Levin & Emily Forgash, August 12, 2026. Supports: US data-centre interconnection-request / grid-queue evidence and the high attrition of requested capacity.

S8 a16z, "Can Agents Use a Computer Yet? We've Got The Data", Andreessen Horowitz, Fabrizio Serafini, Seema Amble & Eric Zhou, August 10, 2026. Supports: OSWorld-Verified computer-use-agent performance and enterprise/BPO execution context. https://www.a16z.news/p/can-agents-use-a-computer-yet-weve

S9 MIT NANDA / Pythian, "Corporate AI Implementation Failure: Why 95% of Projects Never Reach Production", February 2026. Supports: the cited finding that only c.5% of studied enterprise GenAI initiatives delivered measurable business impact, under the study methodology.

S10 Anthropic company disclosures, Bloomberg, and SemiAnalysis:

- Anthropic, "Series H Disclosures & Capitalisation Update", 28 May 2026. Supports: the $47 Bn annualized run-rate baseline and Series H equity valuation.
- Bloomberg, "Anthropic's Revenue Run Rate Tops $65 Billion as Claude Adoption Surges", 17 August 2026. Supports: the updated mid-year gross annualized revenue run-rate (~$65 Bn).

- SemiAnalysis, "Amazon's AI Comeback: By Selling Claude Tokens, AWS's Profit Leverage", Jeremie Eliahou Ontiveros, 28 May 2026 (supplemented by Value Add VC, June 2026). Supports: the wholesale cloud distribution dynamic whereby AWS Bedrock and GCP Vertex account for ~30%–35% of Anthropic's gross enterprise token consumption, and the Token-as-a-Service (TaaS) revenue recognition rules separating Layer B from Layer C.

S11 Amazon, Andy Jassy shareholder communications / AWS disclosures, April 9, 2026. Supports: AWS AI revenue run-rate above $15 Bn and Amazon custom-chip / Trainium-Graviton disclosures.

S12 Reuters, Axios, Bloomberg, and Microsoft / Financial Times reporting:

- Reuters, "OpenAI Projects $2.5 Billion in Inaugural Ad Revenue as 2026 Sales Target $25 Billion", Krystal Hu et al., 9 April 2026. Supports: reported OpenAI 2026 revenue baseline ($25 Bn run-rate), inaugural advertising revenue targets ($2.5 Bn).
- Axios, "Inside OpenAI's 2030 Revenue Roadmap: Subscriptions, Ads, and the Enterprise Push", Ina Fried & Sara Fischer, 9 April 2026. Supports: business-model segmentation between consumer subscriptions, developer API metering, and ad-supported tiers.
- Bloomberg / Financial Times, "OpenAI Revenue Run-Rate Surpasses $35 Billion on Enterprise API Surge", August 2026. Supports: updated late-summer 2026 gross annualized revenue estimates ($35-$40 Bn).
- Microsoft Corporation / Financial Times, "Azure AI Workload Sourcing & Restructured Partnership Disclosures", May–August 2026. Supports: the contractual 20% revenue-share arrangement between Microsoft and OpenAI, the estimate that OpenAI-derived models generate ~70% of Azure AI's reported run-rate, and the resulting $8-$12 Bn wholesale cloud overlap deducted to prevent double-counting across Layers B and C.

S13 NVIDIA company financial disclosures / FY2025 results and S&P Global Market Intelligence (GMI) / Visible Alpha consensus:

- NVIDIA Corporation, "Fourth Quarter and Fiscal 2025 Financial Results" (Form 10-K / Form 8-K), 26 February 2025. Supports: FY2025 total revenue ($130.5 Bn), net income ($72.9 Bn), and segment breakdown.
- S&P GMI / Visible Alpha, "Nvidia Earnings Preview: Q1 2027 & Data Center Segment Forecast", Melissa Otto et al., May 2026. Supports: FY2027 Data Centre revenue consensus ($343.4 Bn) and Blackwell/Rubin architectural volume projections.

S14 Microsoft earnings disclosures / Azure AI disclosures, 2026. Supports: Azure AI run-rate and enterprise-customer metrics cited in the report.

S15 Gartner and International Data Corporation (IDC), global IT spending and enterprise software market tracking:

- Gartner, "Worldwide IT Spending Forecast", John-David Lovelock, 2024–2025. Supports: IT spending segmentation ($5.0 Tr total; $1.0 Tr enterprise software; $1.5 Tr IT services).
- IDC, "Worldwide Black Book: Live Edition & Software Tracker", 2024–2025. Supports: core enterprise software category sizing.

S16 McKinsey Global Institute (MGI) and Goldman Sachs Global Economics Research (GSGER), research on generative AI exposure, automation potential and productivity:

- MGI, "The Economic Potential of Generative AI: The Next Productivity Frontier", Michael Chui et al., June 2023. Supports: cognitive task automation potential (25%–30%) and enterprise use-case value.
- GSGER, "The Potentially Large Effects of Artificial Intelligence on Economic Growth", Joseph Briggs & Devesh Kodnani, March 2023. Supports: 25% cognitive task substitution benchmark and macroeconomic labor productivity impact.

S17 U.S. Bureau of Labor Statistics, wage and occupational employment data, 2025–2026. Supports: the US back-office wage benchmark used in the CUA-versus-labour comparison.

S18 OpenAI, "Advancing the price-performance frontier with GPT-5.6", July 30, 2026. Supports: GPT-5.6 Luna price reduction and the $0.20 / $1.20 per million input/output-token pricing cited in the report. https://openai.com/index/advancing-the-price-performance-frontier-with-gpt-5-6/

S19 Financial Times, "OpenAI and Anthropic in price war as Chinese AI rivals gain ground", August 14, 2026. Supports: broad evidence of model-price competition and pricing pressure following Chinese model advances.

S20 Bloomberg, "DeepSeek Quadruples Peak Prices of Flagship AI Models Before IPO", Saritha Rai, August 13, 2026. Supports: DeepSeek peak/off-peak pricing changes and associated competitive pricing dynamics.

S21 DeepSeek, official model / pricing disclosures, August 2026. Supports: DeepSeek V4 pricing schedules used in Appendix 1.

S22 TrendForce / Barrack AI, "The 2026 GPU Memory Crisis: What the Data Actually Shows", March 2026. Supports: HBM / memory-market tightness and pricing observations.

S23 SK Hynix and Micron company disclosures, 2026. Supports: HBM capacity, production-booking and margin / pricing observations cited in the hardware section.

S24 SemiAnalysis / Uptime Institute / Morgan Stanley (MS):

- SemiAnalysis, "AI Datacenter Power Crisis: NVL72 Rack Economics, Liquid Cooling, and Networking TCO Models", Dylan Patel et al., 2024–2025. Supports: high-density rack draw (120–140 kW), direct-to-chip liquid cooling overhead (2.5%–3.0%), and networking transit benchmarks.
- Uptime Institute, "Global Data Center Survey" & "Impact of AI Workloads on Infrastructure", Andy Lawrence et al., 2024–2026. Supports: liquid-cooled PUE efficiency (1.15–1.25), 85%–90% continuous duty cycles, and facility operations metrics.
- MS Research, "Data Centers & Power: Sizing the AI Infrastructure Cash OpEx Burden", Stephen Byrd et al., 2025–2026. Supports: the 18% annual DC OpEx-to-CapEx factor, the 60% power share of operating expenses, and wholesale PPA pricing structures.

S25 NVIDIA official Blackwell architecture disclosures, 2025–2026. Supports: Blackwell performance / platform claims and lifecycle context cited in Section 1.1.

S26 Bloomberg, "AI Looms Over Software Companies –and the Investors Who Piled Into Them", Paula Seligson and Michelle Cheng, August 12, 2026. Supports: x10-11 pro-forma leverage ratios (total debt/EBITDA) in the US leverage loan market of software LBOs.

S27 GroupM, "This Year Next Year: Global Media Forecast", 2025–2026. Supports: global digital advertising expenditure benchmarks (~$750–$800 Bn) and media budget reallocation dynamics.

S28 PwC, "Global Entertainment & Media Outlook", 2025–2029 / 2026 Edition. Supports: global consumer digital subscription market sizing (~$150–$200 Bn TAM across SVOD, digital audio, gaming, and consumer software subscriptions).

S29 Bank for International Settlements (BIS) / Financial Times / Capacity Media:

- BIS, "The AI Investment Race" (Working Paper 1367) & "Annual Economic Report 2026", June–July 2026. Supports: circular equity ties, over-investment contest modelling, and private credit financial stability risks.
- Financial Times, "Big Tech's Circular AI Economy: Vendor Financing and Regulatory Scrutiny", Richard Waters et al., 2025–2026. Supports: hyperscaler back-to-back equity-for-compute commitments (Microsoft/OpenAI, Amazon/Anthropic) and antitrust inquiries.
- Capacity Media, "What is Circular Financing in AI Infrastructure?", August 2026. Supports: neocloud GPU-collateralised borrowing, wholesale data-centre counterparty concentration, and infrastructure PPA credit exposure.

S30 International Telecommunication Union (ITU) / The World Bank / International Monetary Fund (IMF):

- ITU, "World Telecommunication Development Report 2002: Reinventing Telecoms". Supports: cumulative global telecom and internet backbone infrastructure CapEx exceeding $1.10 Tr (1996–2001).
- The World Bank, "World Development Indicators (WDI) Database: GDP, Current US$" (Indicator: NY.GDP.MKTP.CD) & "Global Economic Prospects", 2025–2026. Supports: historical cumulative nominal world GDP of $195.10 Tr across 1996–2001, the baseline telecom capital intensity ratio (0.56% of cumulative global output).
- IMF, "World Economic Outlook (WEO) Database: Gross Domestic Product, Current Prices" & "Navigating Global Divergence and Capital Realignment", 2025–2026. Supports: the projected 2024-2030 cumulative world GDP corridor of $780-$850 Tr (baseline c.$820 Tr), and the resulting 0.59%-0.64% cumulative macroeconomic absorption ratio for a $5.0 Tr AI infrastructure deployment.

S31 Andreessen Horowitz (a16z), Bessemer Venture Partners (BVP), and ICONIQ Growth:

- a16z, "The New Business of AI: And How It's Different From Traditional Software", Martin Casado & Matt Bornstein (supplemented by "How Enterprise CIOs Buy AI", 2025–2026). Supports: the structural compression of software gross margins from 80%+ down to 50%-60% driven by variable inference compute COGS and context maintenance.
- BVP, "State of the Cloud" & "State of AI: Sizing the AI Application Layer", 2024–2026. Supports: the 60%-65% gross margin benchmark for AI-native platforms and the enterprise software transition to outcome-based consumption models.
- ICONIQ Growth, "State of AI: Enterprise AI Product Benchmarks", 2025–2026. Supports: the empirical survey benchmark of 52%-58% realized gross margins for customer-facing generative and agentic enterprise deployments.

Model Assumptions / Analytical Choices

M1 2024–2030 CapEx path, AI-specific CapEx share, 18% annual DC Opex factor, 60/40 debt-equity structure, 6% cost of debt, 10-year amortization, 60% gross margin and target Project IRR hurdles are model assumptions / scenario parameters unless explicitly linked above.

M2 The 60/40 labour-productivity versus software-cannibalization mix is an analytical assumption, not an externally observed market split.

M3 The 25% AI-vendor take-rate on enterprise productivity value is an analytical assumption; the 20%–33% range is presented as a sensitivity rather than an empirical industry average.

M4 The terminal maintenance-CapEx rate of 10.46% is a modelled weighted replacement-rate proxy based on the asset mix and useful lives set out in Appendix 7.

M5 The $3/hr cached CUA cost is a modelled production-economics assumption / observed deployment benchmark rather than a standardized industry tariff.

M6 The geopolitical C1–C3 outcomes are scenario analysis / tail-risk hypotheses, not forecasts or sourced probabilities.

M7 The $115 Bn OpenAI cumulative cash-burn figure is treated as a reported/market estimate used for scenario analysis, not as an audited company guidance figure.

Secondary Sources (ordered by date)

- Technology: Hyperscaler CapEx 2026 Estimates, CreditSights Research, November 10, 2025.
- Surviving the AI CapEx Boom, Sparkline Capital, 2025.
- Capital Allocation in the AI Era, GQG Partners Report, 2025.
- Are we in a bubble? The AI boom in context, BlackRock Investment Institute, 2025.
- AI Impact on Technology M&A, Lazard Advisory, 2025/2026.
- Physical Constraints Threaten US and European AI Ambitions, BlackRock Investment Institute, 2025/2026.
- Hyperscaler CapEx Hits $600B in 2026: The AI Infrastructure Debt Wave, Introl Research, January 2026.
- The $600 Bn Question: Is AI Actually Worth It?, Reliable Data Engineering & Sequoia Capital Analysis, March 15, 2026.
- Nvidia Stock Prediction: The Path to a $20 Tr Market Cap, IO Fund Analysis, March 26, 2026.
- The Token Subsidy Stack: Why the AI price signal is fictional, Marco Masotto, April 28, 2026.
- The S&P 500 Is Forecast to Climb as Earnings Growth Powers Stocks Higher, Ben Snider, Goldman Sachs Research, May 28, 2026.
- AI in Finance: The Complete Guide, Truthifi, May 2026.
- Big Tech's $725B AI CapEx in 2026 — Up 77% From 2025, Value Add VC, May 2026.
- DeepSeek V4 vs GPT-5.5 vs Claude Opus 4.7: Pricing and Benchmark Performance, MindStudio, May 2026.

- 2026 Midyear Investment Outlook: AI Drives Resilient Growth, Seth B. Carpenter, Morgan Stanley, June 2026.
- AI Infrastructure Bubble or Build-out, Gemini & OpenAI, July 11, 2026.
- Fitch Expects to Rate Galaxy Helios Data Centers II LLC "BB-(EXP)", Fitch Ratings, July 23, 2026.
- DeepSeek Strategic Playbook, Gemini, July 24, 2026.
- The Opportunity in the AI Infrastructure Stock Pullback, Morgan Stanley, July 28, 2026.
- Amazon's AWS AI Outlook: $244B Backlog and Margin Expansion, Investing.com Analysis, July 28, 2026.
- The AI CapEx Cycle: $725B Hyperscaler Buildout & Five High-Conviction Positions, AL Capital Advisory, February 1, 2026 / July 2026 Update.
- Is AI Profitable Yet? Apple Is the King of AI and Nobody Knows It, Financial Analysis Report, 2026.
- Open-Source vs Proprietary AI: Cost Comparison 2026 (Llama, DeepSeek vs GPT-5, Claude), AI Cost Check, 2026.
- The Industrialization of Cognitive Capital: Generative and Agentic AI Adoption in Corporate & Investment Banking, Banking Strategy Report, 2026.

Source hierarchy note: Primary sources are used for the principal externally verifiable claims. Secondary sources are retained for context, triangulation or idea generation and are not relied upon as standalone evidence for the core conclusions where a primary source is available.