MONITORING PROTOCOL V1.2.1 | FINAL BASELINE

Too Indexed, Too Embedded, Too Strategic

A prospective research and adjudication protocol for frontier-firm concentration, hardware, infrastructure, and control

CENTRAL PROPOSITION

Frontier firms can become macro-financially consequential, operationally critical, or strategically important through distinct channels that must remain disaggregated. Public valuation can finance infrastructure and acquisitions, but value creation depends on capital discipline, product neutrality, customer retention, and independently supported demand. Hardware concentration can migrate among accelerators, foundries, memory, packaging, cloud control planes, power, and financing. Version 1.2 specifies hypothesis-level decision rules and conditional scenario bands; it does not create a composite systemic score or claim that the theory has already been tested.

Baseline evidence cutoff: July 24, 2026 | Protocol v1.2.1 Final Baseline fixed: July 24, 2026 | First quarterly observation update: October 2026

Abstract

This monitoring protocol examines what the public-market emergence of trillion-dollar frontier firms could mean for American finance, hardware supply chains, electricity, enterprise organization, labor, and geopolitical position. It begins with event discipline. SpaceX completed an exceptionally large initial public offering; OpenAI and Anthropic confidentially submitted draft registration statements; and SpaceX signed, but had not closed, an all-stock agreement to acquire Anysphere. The protocol then asks how valuation, index exposure, debt, accelerator supply, semiconductor fabrication, data-center construction, power, workflow ownership, and substitutability could transmit consequences beyond common shareholders.

The thesis rejects a single umbrella category for all forms of importance. It separates hard financial systemic risk, macro-financial amplification, operational criticality, strategic importance, and political salience. Index ownership can spread wealth losses and weaken capital expenditure without creating bank-style contagion. A firm is financially systemic only when distress can materially impair credit, funding, payments, clearing, or major intermediaries. Operational and strategic importance can justify continuity planning, but neither implies that common shareholders should be protected.

High valuation is treated as valuation-dependent acquisition capacity rather than proof of industrial dominance. The pending SpaceX-Cursor agreement is retained as a boundary case, but its full analytical treatment appears once, in Chapter 5; other chapters use cross-references rather than counting the same transaction repeatedly. Energy analysis incorporates sequential agentic compute, latency service levels, local physical ceilings, and rebound demand: efficiency can improve output per unit of energy while total electricity use still rises.

Most enterprises will not build and operate cross-model control planes themselves. The likely architectures are single-vendor consolidation, hyperscaler-managed plurality, independent managed control planes, and self-managed deployments for a minority of governments and unusually capable firms. Hardware is treated as a migrating set of control points rather than a permanent monopoly: NVIDIA combines accelerator, networking, software, and rack-scale advantages; Huawei is building a sovereign alternative; custom silicon can reduce accelerator dependence while deepening hyperscaler dependence; and TSMC, HBM, advanced packaging, data-center commissioning, and power remain shared constraints. Open weights can reduce model-layer dependence while leaving the substrate concentrated or shifting rents to adjacent layers.

The resulting document is Monitoring Protocol v1.2.1 - Final Baseline: a prospective research and adjudication protocol, not a tested theory or calibrated predictive model. It preserves separate qualification tests for financial-system transmission, macro-financial amplification, operational criticality, strategic importance, and resolvability. The final patch fixes verdict precedence, defines how event results roll into hypothesis-level conclusions, separates support-pattern simulation from evidence availability, completes the H1-H7 variable dictionary, relabels the historical exercise accurately, adds source-symmetry and revision ledgers, and clarifies the two-tier company taxonomy and migrating-constraint model. No composite systemic score is used.

PROTOCOL THESIS
The protocol separately tests issuer-specific financial transmission, common hardware-substrate shocks, infrastructure-finance stress, valuation-funded integration, power-latency constraints, hyperscaler intermediation, managed sovereignty, and open-weight rent migration. Each hypothesis can be supported, rejected, mixed, or not adjudicable under its own fixed rule. No cross-hypothesis probability or systemic score is calculated.

How to read this document

This document is a frozen baseline protocol, not a results paper. Words such as "may," "could," and "plausible" identify open questions. Observed events populate the baseline; hypotheses remain untested until their trigger and event window occur. Quarterly reviews update the event ledger and evidence state, but do not repeatedly re-test endpoints or select favorable interim results. Formal adjudication follows the precedence, event-level, hypothesis-level, mixed, and missing-data rules in Chapter 12. Any change must be versioned prospectively and cannot rescue a claim after an outcome is known.

A result may be "not adjudicable with available evidence." Private retention, model mix, utilization, contract, or margin data cannot be inferred from management narratives. Missing data support neither the hypothesis nor the null. The protocol contribution is a clean factual scaffold, disaggregated concepts, a reproducible support-pattern generator, and an executable real-world adjudication design. The empirical program begins after this baseline. Simulation is conditional on complete measurement; evidence availability is tracked separately.

PROTOCOL STATUS
Baseline evidence and decision rules are frozen at July 24, 2026. Version 1.2.1 closes the final implementation gaps: verdict precedence, event-to-hypothesis aggregation, full H1-H7 measurement units, simulation/observability separation, exact holdings traceability, contract-denominator reconciliation, source symmetry, two-tier company taxonomy, migrating constraint regimes, and a complete revision ledger. No primary hypothesis has been adjudicated. No composite systemic score is used.

Research propositions and rival explanations

Research proposition	Affirmative rival or null	Distinguishing measure	Executable rule
H1A. Issuer-specific transmission	A qualifying material AI-system issuer drawdown remains an equity event without material credit, funding, collateral, clearing, or forced-flow effects.	Matched intermediary spreads, funding, collateral, dealer flows, payments and clearing.	Support requires the credit-spread breach plus either funding/collateral or flow amplification. Null requires all three to remain below threshold.
H1B. Common hardware-substrate shock	A foundry, HBM, packaging, interconnect, or platform shock impairs operations but is absorbed without material financial spillover.	Multi-issuer operating shock, broad index drawdown, credit/funding effects.	Support requires all three conditions. A common operating shock without credit/funding transmission supports the null.

Research proposition	Affirmative rival or null	Distinguishing measure	Executable rule
H2. Infrastructure-finance bridge	Cancelled or underused AI infrastructure is redeployed or absorbed without material lender, utility, supplier, municipal, or ratepayer loss.	Principal impairment, spreads, guarantees, rate effects, supplier revenue and redeployment.	Support requires lender/project-credit breach, or both utility/ratepayer and supplier/project-finance breaches.
H3. Valuation-funded integration	Stock-financed application ownership destroys neutrality, retention, or economic value.	Retention, neutrality, voluntary owned-model use, margin, ROIC versus WACC, impairment.	Support requires retention, neutrality, no material impairment, and ROIC or operating improvement. Failed, blocked, or abandoned deals remain recorded outcomes.
H4. Power-latency constraint	Efficiency and supply outpace demand while latency and service levels remain intact.	Demand versus efficiency, P95/P99 latency, SLA misses, multi-site constraint.	Support requires demand to outrun efficiency plus latency/SLO breach and multi-site evidence.
H5. Hyperscaler intermediation	Frontier labs retain direct contracts, data authority, and durable margins despite cloud routing.	Spend share, master-contract ownership, substitutions, channel control and margin.	Support requires at least two of three pre-specified control-plane thresholds.
H6. Managed sovereignty response	Portability investments do not improve continuity or cost-adjusted recovery.	Architecture portability, recovery time, fallback success and TCO premium.	Support requires portability, RTO improvement, and TCO premium no greater than 15%.
H7. Open-weight rent redistribution	Open weights remain economically marginal or fail to alter concentration and rent location.	Production share, secure task cost, closed-model premium and adjacent-layer rent shift.	Support requires at least three of four thresholds; failure of model share, premium and rent-shift tests supports the null.

Contents

Method and scope

This is a pre-registration baseline for a theory-building research program, not an investment recommendation or a prediction that any specific rescue, market collapse, merger outcome, or vertical-integration strategy will occur. It distinguishes observed evidence, prospective cases, and conjecture. Public filings establish event facts, capital structure, governance, and disclosed obligations. Company statements establish strategic intent, not the soundness or future success of that strategy. Independent operational evidence is required before a transaction, product, or infrastructure program is used as confirmation of a general mechanism.

Industry coalition statements are treated as evidence of policy preference, commercial alignment, and strategic narrative. A broad signatory list can demonstrate that firms across chips, cloud, models, applications, cybersecurity, and capital share an interest in an open ecosystem. It cannot establish that the coalition's claims about cost, safety, competition, sovereignty, or national leadership have been realized.

The unit of analysis is the system formed by frontier models, hyperscalers, accelerator and equipment suppliers, foundries, memory and packaging providers, electricity and data-center infrastructure, capital markets, enterprise workflows, and the state. SpaceX remains an extreme boundary and mechanism-revealing case, not a representative blueprint. The protocol distinguishes primary frontier firms from material AI-system issuers. NVIDIA is analyzed symmetrically as both a hardware supplier and a material AI-system issuer. Generalization requires comparable behavior across firms, explicit common-shock controls, and causal mechanisms that survive failed transactions, neutral applications, custom-silicon substitution, and hyperscaler-controlled distribution.

Related work and benchmarking

Version 1.2 positioned this protocol against four external frameworks rather than presenting its method as sui generis. The IMF identifies vendor concentration, common models and data, herding, market fragility, and cyber risk as capital-market concerns, based partly on outreach to 27 industry stakeholders; its unit of analysis is mainly the provider category rather than named cross-layer firms. JPMorgan Chase evaluates U.S.-China AI competition across seven dimensions and explicitly rejects a single metric. The Cloud Security Alliance maps concentration and supply-chain-security risks across hardware, memory, packaging, frameworks, model distribution, and cloud. When We Crash provides a non-peer-reviewed structural analog for scenario generation, sensitivity analysis, and historical testing. This protocol uses only a historical applicability and rule-coverage audit; it does not claim backtest performance (IMF 2024; JPMorgan Chase Center for Geopolitics 2026; Cloud Security Alliance 2026; When We Crash 2026).

Work	Primary contribution	Method or evidence	How this protocol differs
IMF GFSR, Ch. 3 (2024)	AI in capital markets; vendor concentration, herding, market fragility and cyber risk.	Market analysis plus detailed responses from 27 stakeholders.	Names firms and follows capital, hardware, infrastructure and workflow links, while retaining narrower financial-system tests.

Work	Primary contribution	Method or evidence	How this protocol differs
JPMorgan Chase Center for Geopolitics (2026)	Seven-dimensional U.S.-China assessment: policy, hardware, models/software, energy, finance, socioeconomics, military/security.	Systemic country comparison using quantitative and qualitative indicators.	Adapts the dimensions to firms and regions without totals or a composite rank.
Cloud Security Alliance (2026)	Concentration and security risk across the AI development stack.	Research note drawing on public and secondary estimates plus incident evidence.	Uses its figures as reported estimates, separates market concentration from verified exposure, and adds portfolio and continuity tests.
When We Crash (2026 working paper)	Scenario generator, sensitivity ranges, selected-episode dashboard audit.	Factor-copula Monte Carlo and preliminary backtests; not peer reviewed.	Uses hypothesis-specific generators and fixed decision states; does not copy its probabilities or aggregate H1-H7.

Table M.2. External benchmarks clarify the protocol's scope, method, and limits.

Revision record and selection correction

External review identified a directional omission in the prior baseline: it discussed SpaceX scale and acquisition capacity without presenting the company's 2025 loss, third-party valuation countermarks, actual index weight, an independent estimate of pre-existing claims on proceeds, or the inaugural bond issuance. The v1.0 baseline restored those observations before the first quarterly adjudication. Recording the correction prevents later versions from silently changing the baseline after outcomes are known.

Protocol v1.1 added the hardware-substrate amendment. Version 1.2 corrected TSMC's announced U.S. investment baseline, replaced the 100,000-GPU figure as a current SpaceXAI anchor, added the disclosed Anthropic and Google compute agreements, split H1 into issuer-specific and common-substrate tests, admitted NVIDIA to the issuer test, and added the six-state logic, support-pattern simulation, and related-work benchmark.

Version 1.2.1 is a bounded final patch. It corrects adjudication precedence, defines event-level and hypothesis-level aggregation, separates simulation from evidence availability, completes missing H1-H7 units and denominators, relabels the historical exercise, adds exact NVIDIA holdings traceability and contract-ratio caveats, requires source symmetry, adopts a two-tier company taxonomy, reconciles hardware and power through migrating constraint regimes, and records every material factual and methodological change in Appendix D. The baseline is frozen after this release; later changes require a prospective versioned amendment.

Evidence class	Permitted use	Non-permitted inference
Public filings and transaction notices	Offering status, terms, share classes, governance, obligations, and disclosed risk factors	A signed deal or stated rationale is not evidence of successful integration or industry-wide inevitability.
Company disclosures	Management intent, product direction, reported scale, and claimed strategic logic	Self-description does not validate returns, customer value, neutrality, safety, or causal effect.

Evidence class	Permitted use	Non-permitted inference
Industry and coalition advocacy	Commercial alignment, policy preference, ecosystem strategy, and the claims participants want policymakers to accept	The number or prominence of signatories does not validate lower cost, stronger security, decentralization, or realized diffusion.
Regulators, agencies, and index providers	Rules, system exposure, infrastructure status, and policy response	A dated aggregate statistic is not a stable structural parameter; source, date, and methodology must travel with the number.
Independent empirical and customer evidence	Productivity, retention, reliability, labor, security, and process economics	Early or context-specific results require replication and matched comparisons.
Comparative cases and counterfactuals	Testing whether a mechanism appears beyond SpaceX and against neutral or failed alternatives	Case analogy cannot substitute for identification; selection and endogeneity remain.
Scenario analysis	Defining coherent outcomes and observable signposts	Scenarios do not validate a thesis unless classification rules and disconfirming observations are specified in advance.

Table M.1. Evidence hierarchy and limitations.

Analytical boundaries

The protocol evaluates channel-specific transmission and strategic incentives. It does not predict a specific offering price, assume announced infrastructure is completed, or aggregate heterogeneous forms of importance into one score. Financial transmission, macro amplification, operational criticality, strategic importance, and resolvability remain disaggregated; the remedy for vagueness is a threshold within each channel, not a composite index. Concentration statistics are dated observations rather than constants. The SpaceX-Cursor agreement remains pending at the evidence cutoff and cannot count as successful integration until it closes and produces independently observable post-close results. Chapter 5 contains the one canonical treatment; all other mentions are event facts or cross-references and create no additional evidentiary weight.

Limits of inference and case selection

SpaceX combines founder control, launch, communications, defense, compute, a frontier model, and a pending application acquisition in a structure that is unusually difficult to replicate. The case is valuable because it exposes mechanisms at their most concentrated, but it creates a selection risk: a theory built around the most extreme firm can mistake anomaly for equilibrium.

Evidence is classified as observed, prospective, or conjectural. Observed evidence records completed events and measured outcomes; prospective evidence records signed transactions,

filings, or contracts with unknown effects; conjecture identifies mechanisms for later comparison. Outcomes cannot be counted as support by relabeling the channel, scenario, or perspective.

CHAPTER 1

The event: one completed IPO, one signed merger, and two listing scenarios

A thesis about systemic importance must separate valuation from capital raised, a signed merger from a completed acquisition, investable float from headline market capitalization, and application-layer integration from aspirational strategy.

The motivating intuition is powerful: firms associated with launch infrastructure, global connectivity, frontier artificial intelligence, and enterprise automation can become unusually consequential when public valuation, dedicated infrastructure, and customer dependence reinforce one another. But size alone does not transform a growth company into a financial-system institution. The analysis must distinguish market exposure from financial linkage, a pending acquisition from successful integration, and a strategically important service from a shareholder claim that merits support.

What has actually happened

SpaceX completed its initial public offering in June 2026. It priced 555,555,555 Class A shares at $135 and granted the underwriters an option for an additional 83,333,333 shares. After full exercise, 638,888,888 shares were sold. Simple multiplication produces an offer value of $86.25 billion, while the company's closing release reported approximately $85.7 billion in "gross proceeds"; the prospectus had also estimated approximately $85.7 billion of net proceeds after underwriting discounts and expenses. Because the company documents use inconsistent labels, this protocol records both the mechanical offer value and the company-reported proceeds rather than treating them as identical (SpaceX 2026a; SpaceX 2026b; SpaceX 2026f). Trading began on Nasdaq under SPCX, and SpaceX joined the Nasdaq-100 on July 7, 2026 (Nasdaq 2026a).

On June 16, 2026, SpaceX, its wholly owned merger subsidiary X67 Inc., and Anysphere entered into a definitive merger agreement. At closing, Anysphere would become a wholly owned SpaceX subsidiary, and its outstanding shares would be exchanged for SpaceX Class A common stock based on an implied $60.0 billion equity value and SpaceX's seven-trading-day volume-weighted average price. SpaceX expected a third-quarter 2026 close, but the transaction remained subject to closing conditions, including regulatory approvals, at this thesis's evidence cutoff (SpaceX 2026d). The precise language matters: SpaceX has agreed to acquire Cursor; the acquisition is not yet completed.

OpenAI confidentially submitted a draft registration statement in June 2026. Anthropic did the same earlier that month. A confidential submission is a procedural step, not a completed offering. Neither company had publicly fixed its share count, price range, listing date, free float, or final

valuation by the evidence cutoff used here (OpenAI 2026a; Anthropic 2026a). OpenAI had disclosed a private financing at an $852 billion post-money valuation, and Anthropic had been reported at a $965 billion post-money valuation. These numbers make trillion-dollar public valuations plausible scenarios, but they are not completed IPO facts.

Firm	Status at July 24, 2026	What is known	What remains uncertain
SpaceX	Public; Nasdaq-listed; Nasdaq-100 constituent; definitive Cursor merger agreement signed	638.9M shares sold at $135. Mechanical offer value: $86.25B; company-reported proceeds: about $85.7B. On June 16, SpaceX agreed to acquire Anysphere for an implied $60B in SpaceX Class A stock; a Q3 close was expected subject to conditions.	Regulatory clearance and closing; final share issuance and dilution; integration economics; Cursor model neutrality and user retention; effect on Grok; cross-business capital allocation.
OpenAI	Confidential draft S-1 submitted	Private financing disclosed at $852B post-money; large consumer, API, and enterprise operations; filing process underway.	Offering timing, price, share count, float, governance rights, and final public valuation.
Anthropic	Confidential draft S-1 submitted	Filing process underway; large private valuation and multi-provider compute commitments.	Offering timing, price, share count, float, governance rights, and final public valuation.

Table 1.1. Event status. A private valuation, a confidential filing, a signed merger, and a completed IPO are economically different states.

SpaceX financial profile at the evidence cutoff

The profile below includes evidence that cuts both for and against the thesis. It disciplines the “too indexed” metaphor by separating company scale from float-adjusted index weight, and it treats third-party valuation and proceeds-use estimates as counter-evidence rather than company fact.

Metric	Baseline observation	Interpretive constraint	Evidence status
2025 revenue	$18.674B	Large operating scale does not establish profitability or financial-system transmission.	Company disclosure
2025 net loss	$(4.937)B; operating loss $(2.589)B	Valuation and capital access coexist with substantial reported loss.	Company disclosure
IPO consideration	638,888,888 shares x $135 = $86.25B mechanical offer value; company-reported proceeds about $85.7B	Do not equate offer value, net proceeds, market capitalization or index weight.	Company filings; terminology discrepancy noted
Opening valuation	Approximately $1.75T at the offer price	Headline value is not cash raised and is sensitive to float and marginal price.	Offer-price arithmetic
Independent valuation countermark	Morningstar fair value estimate: about $780B, or $63 per share	A credible outside estimate was far below the offer price; it is an estimate, not an observed market outcome.	Independent analyst estimate
Nasdaq-100 weight at inclusion	Approximately 1.34%	Material for the protocol’s >=1.0% trigger, but far smaller than headline market-cap share.	LSEG estimate reported by Reuters

Metric	Baseline observation	Interpretive constraint	Evidence status
Pre-existing claims on proceeds	New Constructs estimated about $62.6B, or 78.3% of an assumed $80B net raise, was economically spoken for by disclosed leases, debt and vendor obligations	This is an outside reconstruction, not a company-designated use-of-proceeds schedule.	Independent estimate
Inaugural bond issuance	$25B across five senior unsecured tranches due 2031-2056; coupons 5.35%-6.65%	Creates an observable corporate-credit channel, but not proof of systemic contagion.	Company disclosure

Table 1.2. SpaceX financial profile at the July 24, 2026 baseline. Dollar figures are nominal; third-party estimates are identified as such.

Four numbers that should never be confused

The public discussion tends to compress four different quantities into one headline. Market capitalization is the share price multiplied by shares outstanding. Offering proceeds are the cash raised from shares sold in the transaction. Investable float is the portion of equity available to public investors after insider, government, strategic, or otherwise restricted holdings are excluded. Index weight is the security's representation after the index provider applies eligibility, float adjustment, caps, and rebalancing rules. Each number answers a different question.

Market capitalization: How large is the equity claim at the current price?

Offering proceeds: How much new or secondary capital changed hands in the IPO?

Float-adjusted capitalization: How much of the company is practically available to public-market investors?

Index weight: How much exposure is mechanically or benchmark-driven through a particular index methodology?

A trillion-dollar market capitalization does not mean a trillion dollars has flowed into the stock. A relatively small marginal trade can set the price for all outstanding shares. Likewise, a large company need not immediately receive its full headline weight in a major index. Seasoning requirements, profitability tests, liquidity, governance, domicile, float, and concentration rules can delay or limit inclusion. S&P Dow Jones Indices, for example, declined in June 2026 to create an automatic exemption from its standard eligibility requirements solely for mega-cap companies (S&P DJI 2026a).

> **THE FACTUAL CORRECTION**
> **The economically relevant scenario is not "three firms raise $3.5 trillion." It is that three frontier firms could command several trillion dollars of public equity value, use that equity as acquisition currency, attract substantial benchmark demand and debt, and build or buy the infrastructure, applications, data feedback, and customer relationships that make valuation operationally consequential.**

Why these three firms belong in one thesis

SpaceX, OpenAI, and Anthropic are not interchangeable. SpaceX is an extreme infrastructure-to-application boundary case spanning launch, communications, defense, compute, xAI, Grok, and a pending application acquisition. OpenAI and Anthropic began with models and are moving into

agents, applications, enterprise deployment, and dedicated infrastructure. These trajectories expose incentives to cross layers, but they do not establish a common endpoint or show that other firms can reproduce SpaceX's capital access, government relationships, physical assets, and founder-controlled structure. Chapter 5 contains the canonical SpaceX-Cursor analysis.

The common research question is conditional rather than categorical: when does cross-layer expansion create operating value, and when does it destroy neutrality, partner support, focus, or financial discipline? Company descriptions of strategic logic establish intent only. Closing, post-close user behavior, application economics, and comparable cases are required before generalization.

A high valuation can lower the percentage dilution required to purchase a given target, but an all-stock acquisition remains dilutive and can transfer overvaluation from acquirer to target holders. The mechanism is two-sided and is specified generically here; Chapter 5 applies it once to the pending boundary case.

Analytical case	Starting control point	Prospective movement	What would make it informative
Boundary case: SpaceX	Launch, communications, defense, compute, xAI, Grok, founder-controlled equity	Pending move into a model-neutral developer application through Cursor	Post-close evidence on neutrality, retention, model mix, margins, integration, and separability.
Model-origin cases: OpenAI and Anthropic	Frontier models, APIs, consumer or enterprise distribution, evaluation and trust	Agents, applications, enterprise workflows, dedicated infrastructure, and selected services	Evidence that workflow ownership improves durable margins rather than creating liability, partner conflict, or customer resistance.
Counter-model: hyperscaler intermediation	Identity, cloud, security, systems of record, procurement, and enterprise contracts	Managed routing across many models while retaining the customer relationship	Whether labs become interchangeable suppliers behind a cloud control plane or retain direct billing, data, and workflow authority.

Table 1.3. Different origins, a boundary case, and a competing intermediation path.

The research question

> *Under what measurable conditions do trillion-dollar frontier firms become financially systemic, macro-financially consequential, operationally critical, or strategically important - and when do vertical integration, power scarcity, hyperscaler intermediation, enterprise capability, and sovereign alternatives weaken rather than reinforce their power?*

This question contains six subsidiary questions. First, does index concentration transmit into credit, funding, payments, or only wealth and capital expenditure? Second, does a high valuation create durable acquisition capacity or subsidize overpayment and neutrality loss? Third, do efficiency and new supply keep pace with agentic compute, latency, and rebound demand? Fourth, do frontier labs own the workflow, or do hyperscalers, neutral applications, systems of record, and business-process firms retain the control point? Fifth, do enterprises obtain portability through managed architectures or accept single-vendor consolidation? Sixth, does leadership remain distributed across the United States, China, Europe, India, Gulf states, and allied supply chains?

The thesis claims

Claim 1 - categories must remain separate: Index exposure can create wealth and capital-expenditure effects. Financial-systemic status requires evidence of transmission into credit, funding, payments, clearing, or major intermediaries. Operational and strategic importance are different claims with different evidence and remedies.

Claim 2 - valuation creates conditional acquisition capacity: A high share price can reduce percentage dilution and finance applications, talent, and distribution, but it can also facilitate overpayment, goodwill, cross-subsidy, and value-destroying self-preferencing.

Claim 3 - hardware, power, and latency are physical constraints: Accelerator availability, HBM, advanced packaging, interconnect, software compatibility, cooling, and energized capacity jointly bound useful throughput. Efficiency, routing, and higher-value work can improve economic output per megawatt, while sequential agentic workloads, refresh cycles, local grids, and facility design preserve hard physical limits.

Claim 4 - the workflow control point is contested: Frontier firms may move into applications and outcomes, but hyperscalers, model-neutral tools, systems of record, integrators, and customers can retain distribution and orchestration.

Claim 5 - sovereignty is a capability and operating-model continuum: Most enterprises will buy managed control, portability, and private deployment rather than build a bespoke multi-model platform or train a frontier model.

Claim 6 - geopolitical leadership is multipolar and hardware-contingent: U.S. capital, software, and chip-design strengths coexist with NVIDIA platform concentration, a Chinese sovereign stack led by Huawei Ascend, and shared dependence on TSMC, Taiwan-based packaging and system partners, Korean memory, Japanese materials, and European equipment. National leadership must therefore be evaluated across design, fabrication, packaging, memory, software, data centers, power, deployment, trust, and third-market adoption.

Why the timing matters

Four transitions are occurring at once: public markets are absorbing very large technology firms; high valuations can finance stock-funded acquisitions; electricity planning is adjusting to rapid load growth; and enterprise AI is moving toward systems that act under permissions. These facts support rival outcomes. Public equity can finance useful integration or empire building. Large-load commitments can build productive infrastructure or strand regulated assets. Application ownership can deepen distribution or destroy neutrality. Hyperscalers can enable model plurality while concentrating procurement and identity. This is not a rescue forecast. Continuity may be protected while shareholders absorb losses.

CHAPTER 2

A transmission taxonomy: financial risk, macro amplification, operational criticality, and strategy

The traditional banking analogy captures political salience but obscures the distinct mechanisms through which a frontier firm can affect portfolios, credit, critical operations, and state capacity.

"Too big to fail" emerged from institutions whose sudden collapse could interrupt payments, destroy short-term funding, freeze credit, and impose losses on counterparties unable to replace them. Frontier technology firms usually have different liabilities and failure dynamics. A large equity drawdown can weaken wealth and capital expenditure without becoming a financial crisis. Conversely, a smaller provider can be operationally critical without being financially systemic. The categories must be tested independently rather than combined into a score that can absorb any outcome.

Four categories and one cross-cutting condition

The framework uses four outcome categories. Hard financial systemic risk exists when distress materially impairs credit creation, funding markets, payments, clearing, or major intermediaries. Macro-financial amplification operates through wealth, collateral, investment, employment, suppliers, utilities, and regional tax bases. Operational criticality exists when a service cannot be replaced within the required recovery time. Strategic importance concerns defense, communications, science, industrial capacity, or state autonomy. Resolvability and governance cut across all four by determining whether a service can continue while equity and management change.

> **NO COMPOSITE SYSTEMIC SCORE Market exposure, hard financial transmission, macro-financial amplification, operational criticality, strategic importance, and resolvability are assessed separately. The framework does not assign weights, aggregate them into CSI, or infer shareholder protection from operational or political importance.**

Category	Qualification test	Primary transmission	Required evidence
Hard financial systemic risk	Would distress materially impair credit, funding, payments, clearing, or major intermediaries?	Defaults, liquidity stress, margin spirals, counterparty losses, payment or clearing disruption	Intermediary exposures, leverage, runnable funding, collateral, guarantees, stress losses, market-function evidence.

Category	Qualification test	Primary transmission	Required evidence
Macro-financial amplification	Would an equity or capex shock materially affect consumption, investment, employment, suppliers, utilities, or regions?	Wealth effects, collateral, capex cuts, project cancellation, supplier and municipal stress	Ownership, balance-sheet incidence, capex multipliers, project obligations, regional dependence, matched event studies.
Operational criticality	Can critical users substitute or recover within their required time and loss tolerance?	Service interruption, correlated error, cyber propagation, unavailable data or workflow authority	Critical-workflow share, RTO/RPO, tested fallback, portability, concentration, incident and recovery data.
Strategic importance	Would failure materially degrade defense, communications, scientific capacity, or state autonomy?	Emergency procurement, continuity intervention, industrial-policy response, allied disruption	Unique capability, government use, alternatives, lead times, security classification, allied dependency.
Resolvability and governance	Can services, data, contracts, and assets continue or transfer while ordinary losses are allocated?	Delay, key-person failure, related-party conflict, inseparable assets, blocked restructuring	Voting rights, board control, succession, entity map, transfer rights, continuity plan, separability tests.

Table 2.1. Different forms of importance require different qualification tests, evidence, and remedies.

Market importance is not financial-system importance

A market-capitalization loss is principally an asset-side event. Its immediate effects are lower portfolio values, weaker collateral, reduced risk appetite, and possibly lower consumption through wealth effects. A banking crisis is often a liability-side event in which runnable funding, maturity transformation, and payment obligations force rapid asset sales. The channels can intersect, especially when nonbank intermediaries use leverage or derivatives, but they should not be conflated.

This distinction changes the expected rescue mechanism. In a bank crisis, authorities may provide liquidity, guarantees, or capital because the liabilities themselves are part of the monetary and payments system. In a frontier-platform crisis, authorities are more likely to protect continuity: maintaining satellite service, preserving secure model access, facilitating a sale, transferring government contracts, ring-fencing data, or supporting critical suppliers. Shareholders can be diluted or wiped out while the service survives.

The categories are not a ladder

Hard financial systemic risk

A firm is financially systemic only when distress can impair core financial functions, not merely because many investors own the stock. The test focuses on bank and nonbank credit, short-term funding, collateral, derivatives, clearing, payments, and forced deleveraging. A drawdown that remains inside diversified portfolios is a market loss, even when politically painful.

Macro-financial amplification

A frontier firm can be macro-financially consequential without threatening the payments system. A valuation or investment shock can reduce household wealth, pension funding, infrastructure spending, supplier revenue, employment, and regional tax receipts. The claim is supported only when those effects are material relative to matched market shocks and can be traced to actual ownership, contracts, or project finance.

Operational criticality and strategic importance

Operational criticality depends on substitution time and tolerated loss. Strategic importance depends on unique relevance to defense, communications, science, industry, or sovereignty. Either can motivate the state to preserve a service or facilitate a transfer while allowing shareholders and creditors to absorb losses. Political pressure is not itself evidence of hard financial contagion.

Valuation-dependent acquisition: two competing branches

A stock-financed application acquisition exposes two rival mechanisms. In the value-creating branch, a high valuation reduces percentage dilution, the owner preserves product choice and quality, users remain, and workflow feedback improves task economics. In the value-destroying branch, ownership produces self-preferencing, partner conflict, user flight, overpayment, impairment, and weaker future acquisition capacity. Chapter 5 contains the canonical prospective application of this branch model.

Branch	Mechanism	Observable implication
Value creating	Preserve meaningful model choice while adding compute, distribution, and engineering resources.	Retention and model diversity remain stable or improve relative to pre-deal cohorts and neutral competitors.
Value creating	Workflow feedback improves the owner's model without degrading the application.	Voluntary owned-model adoption, successful tasks per compute, reliability, and gross margin improve.
Value creating	Operating cash flow validates the consideration paid.	Return on invested capital exceeds the standalone and matched acquisition alternatives.
Value destroying	Defaults, pricing, or access privilege the owner's model at the expense of neutrality.	Rival-model access worsens, users multi-home or leave, and partner support declines.
Value destroying	Overpayment, cross-subsidy, and integration complexity obscure economics.	Goodwill or intangibles are impaired; segment transfers rise; task margin and cash conversion weaken.

Branch	Mechanism	Observable implication
Value destroying	Valuation reversal makes future stock financing more expensive.	Deal capacity contracts, dilution rises, and the integration narrative ceases to support the share price.

Figure 2.1. Two-branch acquisition reflexivity. Ownership is the start of the test, not evidence of successful integration.

Governance, insider control, and resolvability

The ability to resolve a firm depends on voting control, board independence, succession, related-party transactions, asset partitioning, and transfer rights. SpaceX's prospectus states that Elon Musk would hold approximately 82.4 percent of the voting power after the offering and that the company qualifies as a controlled company under Nasdaq rules. It also describes dual-class shares, performance-based restricted shares, option and equity-plan supply, and extensive related-party relationships. OpenAI's nonprofit-controlled public-benefit structure and Anthropic's Public Benefit Corporation and Long-Term Benefit Trust create different governance tradeoffs. These structures affect key-person risk, public-shareholder influence, crisis decision rights, and the feasibility of separating critical services (SpaceX 2026f; OpenAI 2025a; Anthropic 2026b).

A complete analysis should therefore ask not only how large the firm is, but who can decide, who can replace the decision maker, and whether critical services can be transferred. Can models, weights, data, satellites, spectrum, data-center contracts, and customer obligations be separated? Are related-party agreements replaceable? Can regulators or courts maintain service during restructuring? Do dual-class voting rights, founder dependence, mission structures, or restricted awards obstruct ordinary resolution?

Channel-specific decision rules

No arbitrary count of conditions converts a firm from “soft” to “hard” systemic importance. Each claim must satisfy the outcome test for its own channel. Financial-systemic language is appropriate only when distress threatens core financial functions. Operational-criticality language is appropriate only when tested substitution or recovery cannot meet required time and loss tolerances. Strategic language requires a defined state capability at risk and a comparison with realistic alternatives.

Evidence supporting a hard financial claim: material leveraged exposure, runnable funding, correlated collateral, margin spirals, intermediary losses, or payment and clearing disruption that exceeds matched market controls.

Evidence supporting an operational or strategic claim: critical users, long tested substitution times, unique capability, weak separability, and continuity plans that cannot preserve the function through ordinary restructuring.

Evidence weakening all concentration claims: low leverage, modular contracts, multiple suppliers, robust portability, independent customer choice, tested recovery, separable services, and ordinary loss allocation. Policy should protect functions and competition, not infer capital support from size.

CHAPTER 3

Index inclusion, concentration, and financial transmission

Indexes broaden ownership and can magnify common exposures, but they do not create a guaranteed buyer at any price or convert every equity shock into a banking crisis.

The strongest part of the original intuition is that index membership changes who is exposed. A major constituent appears not only in direct portfolios, but also in retirement plans, insurance accounts, exchange-traded funds, institutional mandates, options, structured products, and active strategies measured against the same benchmark. S&P Global estimated that approximately $13 trillion was directly indexed to the S&P 500 at the end of 2024, with additional assets benchmarked to it. Nasdaq reports that more than $800 billion tracks the Nasdaq-100. At that scale, a new mega-cap constituent can matter even to investors who never choose the stock individually (S&P DJI 2025a; Nasdaq 2026a).

What passive ownership does - and does not do

Index funds purchase securities to minimize tracking error, not because a portfolio manager independently concludes that each security is fairly valued. Inclusion therefore creates a one-time and then ongoing source of rules-based demand. But the effect is bounded. Funds buy the float-adjusted weight specified by the index; creations and redemptions respond to investor flows; index providers rebalance; and arbitrageurs can supply shares. Passive ownership can transmit a price movement broadly without making the price immune to fundamentals.

The more important system-level effect is common ownership of the same factor. When a small number of firms account for a large share of index capitalization, market returns, options exposure, and expected earnings growth become dependent on a narrower set of assumptions. Nasdaq reported that the ten largest Nasdaq-100 constituents represented 52 percent of the index at year-end 2025. S&P Global reported that the ten largest S&P 500 constituents represented 36.4 percent as of May 29, 2026 (Nasdaq 2026b; S&P DJI 2026b). Adding another enormous technology constituent can increase concentration even if the index methodology caps or stages the weight.

Snapshot volatility and measurement discipline

Top-ten concentration is a dated, provider-dependent statistic, not a fixed parameter. Prices, reconstitutions, float assumptions, caps, and measurement dates can move the reported share by several percentage points over short periods. The dashboard should therefore use a monthly time series, a rolling range, contribution to index earnings and returns, and the exact source methodology rather than treating one snapshot as a stable structural input.

What the index and ETF literature implies

The bank-run analogy is weakened by fund structure. Federal Reserve research notes that in-kind creation and redemption can make exchange-traded funds less exposed to the liquidity-transformation risk of conventional open-end funds, although concentration among authorized participants and underlying-market liquidity remain relevant. BIS research finds that passive flows can affect prices and comovement, while Federal Reserve work on leveraged ETFs cautions that simple mechanical amplification estimates can overstate actual market impact. ECB analysis nevertheless treats concentration in U.S. technology holdings as a common-exposure channel worth monitoring (Federal Reserve Board 2014; Federal Reserve Board 2023; BIS 2021; ECB 2024).

The implication is deliberately narrow: “too indexed” names a transmission channel to test, not an established amplifier. The protocol therefore tracks authorized-participant concentration, securities lending, options, structured products, and leveraged or inverse-product rebalancing separately from ordinary unlevered index ownership.

A concentration claim should also distinguish market capitalization from investable float, voting control, free-float turnover, and the availability of shares around lockup expiry, registrations, employee vesting, option exercise, acquisitions, and follow-on offerings. A security can be enormous, tightly controlled, and thinly available at the same time.

Five transmission mechanisms

Mechanism	How the shock travels	Why it matters	What prevents overstatement
Portfolio wealth	A drawdown lowers retirement, household, institutional, and sovereign portfolio values.	Consumption, donations, pension funding, and risk appetite can weaken.	Losses are distributed; unlevered investors do not face runs.
Benchmark feedback	Active managers and derivatives desks adjust exposure around benchmark weights and volatility.	Crowded positioning can amplify flows and correlations.	Arbitrage, valuation discipline, and rebalancing can absorb flows.
Collateral and leverage	Shares and derivatives support margin, structured products, and nonbank financing.	Falling prices can force deleveraging and sales.	The channel depends on actual leverage and haircuts, not market cap alone.
Corporate finance	A lower share price raises the cost of equity, weakens acquisition currency, and reduces employee-compensation value.	Infrastructure investment and supplier demand can fall rapidly.	Profitable firms can finance internally or shift investment pace.

Mechanism	How the shock travels	Why it matters	What prevents overstatement
Confidence and narrative	A flagship failure revises expectations for AI, space, or productivity across the market.	Valuation multiples and capital availability compress together.	Other firms can benefit if the failure is company-specific rather than thematic.

Table 3.1. Equity concentration can transmit stress through multiple channels without resembling a deposit run.

NVIDIA as a material AI-system issuer

The hardware thesis must be applied symmetrically to NVIDIA rather than treating it only as a supplier beneath the focal firms. State Street's dated SPY daily-holdings workbook reports NVIDIA at 7.93528 percent of fund assets on July 23, 2026; the fund page rounds that observation to 7.94 percent. The exact source file is holdings-daily-us-en-spy.xlsx, and the weight is a dated observation that changes with prices and portfolio holdings. NVIDIA therefore exceeds the protocol's 1.0 percent material-index threshold by a wide margin and belongs in H1A's material AI-system issuer universe and H1B's common-substrate analysis (State Street Global Advisors 2026a; State Street Global Advisors 2026b).

The dependence is two-sided. Customers depend on CUDA, networking, rack designs and scarce accelerated compute, while NVIDIA depends on a small set of large direct and indirect buyers, external foundries, HBM suppliers, advanced packaging, system manufacturers, data-center commissioning and customer capital expenditure. NVIDIA reported two direct customers at 22 percent and 14 percent of fiscal 2026 revenue. It also stated that revenue is concentrated among a limited number of indirect customers, some individually representing at least 10 percent, and that one AI research and deployment company contributed a meaningful amount through cloud purchases; those indirect figures are company estimates rather than audited customer attribution. A shock can therefore originate with the issuer, a major customer, or the shared substrate (NVIDIA 2026a).

Table 3.1A. Issuer-specific and common-substrate shocks require separate causal tests.

Test	Trigger	Primary outcome	Control
H1A - issuer-specific	A qualifying drawdown in NVIDIA or another material AI-system issuer without a common hardware disruption.	Credit/funding, collateral, forced-flow, clearing or intermediary effects beyond equity.	Synthetic basket matched on sector, beta, leverage, size and index exposure.
H1B - common substrate	A foundry, HBM, packaging, interconnect, platform-security or power shock affecting multiple issuers.	Correlated operating loss, index drawdown and subsequent credit/funding transmission.	Unaffected sectors, alternate hardware ecosystems and pre-specified common-shock controls.

The first observable equity-to-credit bridge: SpaceX's $25 billion bond

On June 23, 2026, SpaceX priced $25 billion of senior unsecured notes in five tranches: $7.0 billion at 5.350% due 2031, $6.0 billion at 5.650% due 2033, $6.0 billion at 5.875% due 2036, $2.5 billion at 6.600% due 2046, and $3.5 billion at 6.650% due 2056. The company said proceeds would repay a

bridge loan and support general corporate purposes (SpaceX 2026c). This is the first direct observation in the protocol of equity-scale ambition creating a large public corporate-credit exposure.

The issuance strengthens the case for monitoring an equity-to-credit bridge, but it does not establish systemic risk. The notes are long-dated and unsecured rather than runnable money-like liabilities. H1 therefore tracks their secondary spreads, holder concentration, derivatives, collateral treatment, covenant events and spillovers during a qualifying equity shock; H2 separately tests whether debt or guarantees connect the issuer to data-center, generation, transmission, supplier or municipal losses.

Transmission channel	Baseline observation	What would make it material
Corporate credit	$25B public senior unsecured notes now create a measurable creditor class.	Large spread widening, losses or forced sales relative to matched issuers.
Infrastructure finance	Use of proceeds is broad; project-specific linkage is not yet established.	Guarantees, SPVs, take-or-pay contracts or creditor dependence on specific projects.
Market concentration	Bond issuance is distinct from passive equity ownership.	Correlated equity-credit selling, collateral calls or intermediary impairment.
Operational continuity	Long maturity may improve funding resilience.	Contracts or public functions become inseparable from refinancing support.

Table 3.2. The bond is an observable credit channel, not a finding of financial-system importance.

Reflexivity, acquisition currency, and the cost of capital

A high public valuation can finance infrastructure, talent, and acquisitions, but a stock-funded agreement is evidence of acquisition capacity rather than observed evidence of successful integration. Equity consideration dilutes existing holders; the percentage dilution depends on relative valuation, while the economic outcome depends on retention, product quality, cash flow, and the opportunity cost of the shares issued. Chapter 5 contains the one canonical application of this mechanism to SpaceX-Cursor.

Valuation-dependent acquisition creates two possible loops. Preserved choice and improved product economics can convert shares into distribution, feedback, and voluntary demand. Self-preferencing, overpayment, user exit, and channel conflict can destroy the acquired asset. An impairment generally recognizes a decline in expected value; it does not mechanically cause the underlying economic loss. The decisive evidence is post-close retention, product mix, application margin, tasks per compute, integration cost, and return on the shares issued.

The crucial test is whether the assets financed by valuation can generate durable cash flows. Data centers, generation, transmission, launch facilities, and semiconductor capacity have long construction periods and limited alternative uses in some locations. If equity enthusiasm supports debt or take-or-pay contracts, a later revaluation can migrate from shareholders to lenders, utilities, suppliers, and municipalities. This is the bridge from asset-price concentration to balance-sheet stress.

Lockups, dilution, and insider supply

IPO and post-IPO risk depends on the future supply of shares as well as index demand. Analysts should map lockups and waivers, registration rights, options, restricted awards, employee vesting, acquisition consideration, convertible instruments, and follow-on issuance. SpaceX's prospectus disclosed dual-class control, substantial restricted Class B awards, equity-plan reserves, outstanding options, and potential additional dilution associated with transactions. These instruments do not predict a sell-off, but they invalidate any assumption that the initial float is permanent or that passive demand faces fixed supply (SpaceX 2026f).

Voting power and economic ownership must also be separated. Public investors can bear price risk while exercising limited influence over the board, related-party arrangements, succession, or capital allocation. Founder control can support long-horizon strategy, but it can also increase key-person risk, reduce corrective governance, and complicate resolution when the same individual links several affiliated firms.

Why a 2000 analogy is more useful than a 2008 analogy - at first

The initial comparison is closer to the late-1990s technology boom than to the global financial crisis. Both involve extraordinary expectations, heavy infrastructure investment, high equity concentration, and the possibility that a genuine general-purpose technology is priced too far ahead of realized cash flow. A collapse can destroy wealth and financing capacity while leaving the underlying infrastructure available for later productive use. The analogy becomes more like 2008 only if leverage, opaque structured exposure, short-term funding, guarantees, or correlated nonbank balance sheets grow around the equity story.

This distinction also prevents a false dichotomy between “bubble” and “transformation.” The internet transformed the economy even though many securities were overvalued. AI and space infrastructure can produce large social returns while investors in particular projects or firms earn poor returns. A thesis about systemic importance must therefore separate technological value, firm-level profitability, and financial stability.

Who actually bears the risk

The incidence of an equity shock depends on ownership and financing structure. Public pension plans may hold the index while private infrastructure funds finance data centers; banks may provide revolvers while insurers buy project bonds; utilities may recover grid upgrades through rate bases; municipalities may provide tax incentives; and strategic suppliers may accept equity or long-term purchase commitments. “Every investor is exposed” is rhetorically effective but empirically imprecise. The thesis should instead map each claim to a balance sheet and identify who is first-loss, who is guaranteed, and who can pass costs to customers or taxpayers.

Households: Direct and retirement-account wealth exposure; labor-market exposure through employers and regional growth.

Pensions and insurers: Benchmark exposure plus long-duration debt and infrastructure holdings.

- **Banks and nonbanks:** Revolvers, underwriting, derivatives, project finance, warehouse facilities, and collateral links.
- **Utilities and ratepayers:** Interconnection, generation, and transmission costs that may be socialized if projects are speculative or contracts are weak.
- **Governments:** Tax incentives, procurement dependence, industrial-policy commitments, and continuity obligations.
- **Suppliers:** Concentrated revenue, inventory expansion, and take-or-pay or capacity-reservation agreements.

The concentration dashboard

The monitoring system should combine a dated concentration series, share-supply schedule, voting control, acquisition terms, balance-sheet exposures, and common hardware dependencies. Track float-adjusted weight with monthly ranges; contribution to index returns and earnings; options, lending, and margin; debt and project guarantees; lockups, registrations, vesting, and follow-ons; stock-financed deal value and implied dilution; application retention and model share; NVIDIA and other hardware-issuer index weights; large-customer concentration; and the percentage of capital expenditure financed or guaranteed by customers and counterparties.

Decision rule: Treat index concentration as a macro-financial vulnerability when a plausible equity drawdown would trigger forced deleveraging, materially impair infrastructure counterparties, or interrupt critical services. Otherwise, describe it as a concentration and wealth risk rather than a systemic-solvency risk.

CHAPTER 4

The physical substrate: hardware, data centers, electricity, and capital

AI is often described as weightless software, yet its economic frontier is determined by accelerators, host processors, high-bandwidth memory, advanced packaging, interconnect, server assembly, liquid cooling, semiconductor fabs, substations, turbines, transmission lines, and financing contracts.

Electricity is the most important correction to purely digital accounts of artificial intelligence. Models can be copied, software can be distributed globally, and demand can appear instantaneously. Power systems cannot expand at the same speed. A hyperscale data center may be designed and constructed in a few years, while major generation and transmission projects can require longer planning, interconnection, permitting, equipment, and construction cycles. The resulting mismatch makes power a real constraint - but not a simple ceiling on growth.

The scale of the load

The U.S. Department of Energy estimated that data centers consumed approximately 4.4 percent of U.S. electricity in 2023 and could consume 6.7 to 12 percent by 2028, with annual demand rising from about 176 terawatt-hours to 325-580 terawatt-hours (DOE 2024). The U.S. Energy Information Administration's AEO2026 likewise identifies data centers as a major driver of electricity-demand growth (EIA 2026). The International Energy Agency estimated global data-center electricity use at roughly 415 terawatt-hours in 2024 and projected about 945 terawatt-hours by 2030 in its base case (IEA 2025a). The ranges are wide because model efficiency, utilization, hardware, demand, and project completion remain uncertain.

The Federal Energy Regulatory Commission made the constraint institutionally visible in June 2026 by directing all six regional transmission organizations and independent system operators to justify or reform their treatment of large-load interconnection. The orders focused not only on speed, but also on cost allocation, reliability, speculative projects, and the need for financial commitments that protect existing customers (FERC 2026a; FERC 2026b). The policy problem is therefore not merely "build more power." It is "connect credible loads quickly without shifting unearned costs or reliability risk to everyone else."

Constraint layer	What becomes scarce	Typical lead-time problem	Economic response
Site	Land, water, fiber, permits, tax agreements, proximity to power	Local opposition and infrastructure coordination	Move workloads, pre-negotiate sites, co-locate with generation

Constraint layer	What becomes scarce	Typical lead-time problem	Economic response
Grid	Interconnection capacity, substations, transformers, transmission, firm service	Queues and equipment delivery exceed data-center build cycles	Flexible load, behind-the-meter generation, long-term power contracts
Generation	Reliable energy and capacity with acceptable cost and emissions	New plants and fuel infrastructure take years	Portfolio of gas, nuclear, renewables, storage, imports, and demand response
Compute	Advanced accelerators, networking, memory, packaging, cooling	Semiconductor and equipment capacity is concentrated	Multi-vendor hardware, co-design, model efficiency, inventory commitments
Capital	Long-duration financing for uncertain utilization and revenue	Assets arrive before demand is proven	Project finance, customer prepayments, vendor financing, equity, guarantees

Table 4.1. "Electricity" is a stack of interdependent physical and financial constraints.

The hardware substrate: chips are systems, not interchangeable units

An AI data center is not a warehouse of interchangeable processors. It is a tightly coupled production system. Accelerators must be paired with host CPUs, high-bandwidth memory, advanced packaging, scale-up links inside the rack or pod, scale-out networking across racks, storage, compiler and runtime software, power conversion, cooling, and a facility designed for the resulting density. A shortage or incompatibility in any layer can strand capital in the others. The economically relevant unit is commissioned, software-usable, networked compute delivered within a power, thermal, latency, and reliability envelope - not chips ordered or nameplate FLOPS.

This distinction changes the thesis. Model capability can diffuse while hardware rents remain concentrated. Open weights can improve model choice but still require a proprietary compiler, scarce HBM, a particular interconnect, or a liquid-cooled rack. Conversely, a weaker individual accelerator can remain strategically important if a domestic ecosystem can manufacture, network, operate, and improve it at sufficient scale. The protocol therefore treats hardware as a stack of measurable control points rather than a single chip-performance league table.

Hardware layer	Economic function	Primary concentration or failure mode	Protocol measure
Accelerator and host CPU	Execute tensor, reasoning, orchestration, retrieval, and tool workloads.	Allocation, architecture cadence, precision support, and dependence on one platform.	Installed and productive share by generation; matched goodput per watt and per dollar.
HBM and advanced packaging	Keep model state and data close to compute; join large dies and memory at usable bandwidth.	HBM supply, packaging yield, CoWoS-class capacity, substrate and thermal limits.	Qualified capacity, lead time, yield, memory bandwidth used, and packaging geography.
Scale-up interconnect	Makes many accelerators behave as one logical machine inside a rack or pod.	Proprietary fabrics, switch availability, collective-communication efficiency, fault domains.	Effective bandwidth, communication overhead, failure recovery, and portability.

Hardware layer	Economic function	Primary concentration or failure mode	Protocol measure
Scale-out network	Connects racks and sites for distributed training and inference.	Optics, switches, congestion control, topology, and network power.	Application goodput, tail latency, packet loss, optical power, and time to deploy.
Software and toolchain	Compilers, kernels, libraries, schedulers, security, observability, and model optimization.	CUDA, Neuron, CANN, or other ecosystem lock-in; skills and code portability.	Porting time and cost, performance retained after migration, developer availability, incident rate.
Rack, power, and cooling	Turns components into an operable high-density system.	Liquid-cooling readiness, busbars, transformers, power quality, serviceability, and spare parts.	Commissioned rack count, rack-ready MW, PUE, cooling availability, maintenance downtime.
Fabrication and equipment	Convert designs into leading-edge logic, memory, packaging, and tested systems.	Foundry and tool concentration, process yield, earthquakes, water, electricity, and geopolitics.	Production-qualified wafers and packages by geography; recovery time; supplier concentration.
Facility and grid	Provide land, power, water, fiber, permits, and continuity.	Interconnection queues, local opposition, firm power, construction, and financing mismatch.	Status ladder from announced to productive MW; curtailment, utilization, and contract exposure.

Table 4.2. AI hardware is a coupled production stack; the bottleneck can migrate between silicon, software, packaging, networking, cooling, and the site.

Concentration and security across the development stack

A May 2026 Cloud Security Alliance research note reports approximately 92 percent NVIDIA share of the discrete-GPU market; HBM supply of roughly 50 percent SK hynix, 40 percent Samsung and 10 percent Micron, with 2025-2026 capacity committed; a roughly 30 percent HBM price increase in late 2025; and more than 70 percent of TSMC CoWoS-L capacity reportedly secured by NVIDIA. These figures are useful directional estimates, not audited market statistics: the note itself relies partly on secondary sources. The protocol therefore records the source, date, denominator and uncertainty rather than treating each percentage as an immutable structural fact (Cloud Security Alliance 2026).

The same note connects economic concentration to supply-chain security. It cites the PyTorch nightly dependency attack, the May 2024 Hugging Face Spaces breach, more than 352,000 suspicious issues identified across approximately 51,700 models, and a reported 6.5-fold year-over-year rise in malicious model uploads through 2024. Concentration can therefore create a correlated security exposure even when no physical shortage exists. H1B and the dashboard track common software, model-distribution and hardware-platform incidents separately from firm-specific drawdowns (Cloud Security Alliance 2026).

Layer	Reported concentration or incident	Protocol interpretation	Required measure
Discrete GPUs	CSA reports NVIDIA at about 92%.	Potential platform and security common mode; denominator is broad discrete GPUs, not only frontier accelerators.	Productive workload share, substitutability, porting loss, exposure-adjusted incidents.
HBM	CSA reports SK hynix ~50%, Samsung ~40%, Micron ~10%; 2025-26 supply committed.	A three-supplier complement can bind otherwise available accelerator capacity.	Qualified supply, allocation, lead time, price, geography and recovery.

Layer	Reported concentration or incident	Protocol interpretation	Required measure
Advanced packaging	CSA reports NVIDIA >70% of TSMC CoWoS-L capacity.	Reservation power may constrain competitors while concentrating Taiwan exposure.	Customer allocation, package yield, alternative processes and time to qualify.
Framework/distribution security	PyTorch dependency attack; Hugging Face breach; malicious model-upload growth.	Ubiquity can amplify one compromise across many organizations.	Common dependency inventory, provenance, patch time, affected workload share.

Table 4.2A. Reported concentration is a starting point for exposure measurement, not a composite risk score.

NVIDIA: a full-stack AI-factory architecture

NVIDIA is best understood as both a system platform and a material AI-system issuer. Its fiscal 2026 filing describes data-center systems co-designed across GPUs, CPUs, NVLink switches, DPUs, NICs, scale-out networking, CUDA, libraries, models, and management software. The GB300 NVL72 illustrates the architecture: a liquid-cooled rack integrates 72 Blackwell Ultra GPUs and 36 Grace CPUs with NVLink and high-bandwidth memory. Vendor performance claims require matched workloads and facility-level power measurement. Separately, NVIDIA's 7.93528 percent SPY weight on July 23, 2026 and its large-customer concentration place it inside the same issuer and transmission tests applied to SpaceX and future public listings (NVIDIA 2026a; NVIDIA 2026b; State Street Global Advisors 2026a).

The platform can shorten the path from delivered components to productive clusters and reduce integration uncertainty. The same integration creates switching cost across CUDA, collective communication, networking, security, operators, facility design and model optimization. That power is not one-way: NVIDIA reported direct-customer concentration of 22 percent and 14 percent, plus estimated concentration among indirect customers. Large hyperscalers can use purchasing scale, custom silicon and enterprise distribution to bargain with or substitute for the platform even while remaining major buyers (NVIDIA 2026a).

NVIDIA is fabless and depends on external foundries, HBM, advanced packaging, contract manufacturers, power and customer capex. Its control point is therefore powerful but contingent. A common TSMC, HBM, packaging or platform-security shock can affect NVIDIA and multiple customers simultaneously; a customer-specific demand shock can flow in the opposite direction. H1A tests the issuer channel, while H1B tests the shared substrate.

A common seven-dimension cross-firm lens

JPMorgan Chase’s country framework evaluates policy, hardware, models and software, energy and resources, finance, socioeconomics, and military and security, while warning that the dimensions should not be reduced to one metric. Version 1.2 adapts those same dimensions to the four focal firms. The entries are structured descriptions, not scores; no row is weighted and no total is calculated (JPMorgan Chase Center for Geopolitics 2026).

Table 4.2B. The same seven dimensions are applied to each focal firm. The table has no totals and is not a composite systemic score.

Dimension	SpaceX	OpenAI	Anthropic	NVIDIA
Policy	Launch, spectrum, defense, export and merger approvals.	AI rules, procurement, safety, listing and infrastructure approvals.	AI rules, safety, procurement, investment and listing approvals.	Export controls, industrial policy, competition and supply-chain rules.
Hardware	Owned/hosted clusters plus launch, satellite and network assets; depends heavily on NVIDIA and external fabs.	Large NVIDIA and cloud commitments; capacity still partner-mediated and partly announced.	Diversified Trainium, TPU and SpaceX-hosted NVIDIA access; orchestration complexity rises.	Accelerator, networking, software and rack platform; fabless dependence on TSMC, HBM and assembly.
Models/software	Grok plus prospective Cursor application distribution.	Frontier models, consumer distribution, agents and enterprise platform.	Frontier models, coding/enterprise distribution and safety research.	CUDA, libraries, networking, systems software and model tooling.
Energy/resources	Large cluster loads, sites, cooling and connectivity; disclosed capacity contracts create utilization obligations.	Multi-GW development ambitions and supplier-specific power exposure.	Compute portfolio spans several providers and power systems.	Demand depends on customers commissioning power-dense sites; supply depends on materials and manufacturing ecosystems.
Finance	Public equity, $25B bond, stock acquisition currency and contracted compute revenue.	Large private capital, strategic-investor links, prospective listing and multi-GW commitments.	Large private financing, cloud partnerships, prospective listing and contracted compute.	Large public index weight, high cash generation, strategic investments and concentrated customers.
Socioeconomics	Connectivity, launch, defense, regional construction and developer-workflow exposure.	Consumer and enterprise adoption, labor substitution, software and service reorganization.	Enterprise and developer productivity, safety and labor effects.	Supplier employment, customer capex, market wealth, technology diffusion and regional build-outs.
Military/security	Launch, satellite, defense, communications, cyber and supply-chain continuity.	Model misuse, government deployments, cyber and information risks.	Safety, cyber capability, government use and concentration risk.	Export-controlled compute, platform vulnerabilities, defense relevance and common-mode supply-chain risk.

Huawei: a sovereign full-stack alternative

Huawei is not simply a lower-cost substitute for one NVIDIA accelerator. It is constructing a parallel stack around Ascend NPUs, Kunpeng CPUs, the CANN and MindIE software environment, UnifiedBus interconnect, CloudMatrix infrastructure, cloud services, and liquid-cooled SuperPoDs. Huawei states that its Atlas 900 A3 can connect up to 384 Ascend NPUs as one logical system and that more than 300 systems had been deployed to more than 20 customers. Those deployment and performance figures are company disclosures, not neutral benchmarks (Huawei 2025a; Huawei 2026).

Huawei has explicitly described cluster scale and interconnect as a response to restricted access to advanced process nodes. This is strategically important. Export controls can raise unit cost and delay leading-edge production, while also stimulating domestic substitution in chips, software, networking, memory, and operating practice. A platform that is weaker per device can still be viable if it is available, supported, and deployable at scale. The offset is physical: using more devices to reach a target can increase floor area, power, cooling, network complexity, failure domains, and maintenance burden. The relevant comparison is verified task goodput per megawatt, per dollar, and per unit of operator effort - not raw chip count (Huawei 2025a; BIS 2024).

For SpaceXAI, OpenAI, and Anthropic, Huawei is not a currently disclosed direct supply path. Its importance is competitive and geopolitical. Ascend can reduce Chinese dependence on U.S.-

controlled accelerators, create a second software ecosystem with its own switching costs, pressure global pricing, and alter the effectiveness of export controls. The thesis should therefore track Huawei as a sovereign industrial system, not as a product benchmark inserted into a U.S. procurement table.

Custom silicon and control-point migration

Google TPUs, AWS Trainium, AMD accelerators, Meta MTIA and Huawei Ascend can reduce dependence on one accelerator vendor. They do not necessarily decentralize the stack. A customer that moves from NVIDIA hardware into Trainium or TPU may exchange accelerator dependence for cloud, compiler, managed-service and contract dependence. The relevant question is not whether one vendor's share falls, but where portability, pricing power, security exposure and operating authority move.

The durable proposition is therefore control-point migration rather than a permanent NVIDIA monopoly. Bargaining power can move among accelerators, interconnects, compilers, clouds, foundries, HBM, packaging, power, application distribution and finance. The dashboard tracks layer-specific shares and switching costs so substitution at one layer cannot be misreported as full-stack decentralization.

TSMC and Taiwan: concentrated production under active geographic diversification

TSMC is a different kind of control point. It does not compete with NVIDIA or Huawei as a model platform; it turns designs into high-yield wafers and advanced packages. In 2025, TSMC shipped 15.0 million 12-inch-equivalent wafers, and technologies at 7 nanometers and below accounted for 74 percent of wafer revenue. The company continued preparing multiple phases of 2-nanometer fabs in Hsinchu and Kaohsiung and expanding leading-edge and advanced-packaging capacity in Taiwan (TSMC 2026a).

The Arizona build-out is meaningful diversification rather than immediate duplication of the Taiwan cluster. TSMC reports that its first Arizona fab entered high-volume N4 production in the fourth quarter of 2024, a second fab is moving toward production in the second half of 2027, and a third began construction in 2025. On July 16, 2026, the City of Phoenix announced an additional $100 billion plan, bringing total announced Arizona investment to $265 billion and describing ten fabs, two advanced-packaging facilities and an R&D center. The $265 billion figure is announced capital and facilities, not commissioned capacity; timing, yield, supplier density and production qualification remain separate measures (TSMC 2026a; City of Phoenix 2026).

Taiwan concentration extends beyond wafers, but it is no longer analytically sound to describe the location as a static, singular chokepoint. NVIDIA reported a dense Taiwan systems ecosystem participating in the Vera Rubin ramp; TSMC identifies earthquakes, water and electricity as operational risks; and leading-edge packaging and engineering cadence remain concentrated. At the same time, Arizona and other geographic investments are active diversification. The protocol therefore tracks announced, built, qualified, volume-producing and recoverable capacity separately by geography rather than assuming either complete dependence or completed duplication (NVIDIA 2026c; TSMC 2025b; City of Phoenix 2026).

The classification remains disaggregated. TSMC/Taiwan concentration is operationally and strategically critical because substitution and qualification take time. It becomes macro-financially consequential if delayed hardware cuts capital expenditure, revenue, utility load, supplier cash flow, or regional investment. It is financially systemic only if those effects materially impair credit, funding, payments, clearing, or major financial intermediaries. Geographic importance does not waive the protocol's channel-specific qualification tests.

Dimension	NVIDIA	Huawei	TSMC / Taiwan
Primary role	Global full-stack AI-compute platform and ecosystem.	Sovereign full-stack alternative centered on Ascend.	Leading foundry, advanced-packaging base, and supplier cluster.
Core control points	GPU/CPU, HBM integration, NVLink, networking, CUDA, rack designs, management.	NPU/CPU, UnifiedBus, CANN/MindIE, CloudMatrix, SuperPoD and cloud deployment.	Process technology, yield, CoWoS-class packaging, qualification, volume ramp, supplier density.
Strategic advantage	Installed software base, system codesign, time to productive deployment.	Domestic availability, stack control, substitution under technology restrictions.	Scale, yield, process cadence, customer diversity, and advanced packaging.
Main external dependency	Foundries, HBM, packaging, Asian assembly, power and customer capex.	Advanced-node equipment and memory access, software maturity, power and cooling at cluster scale.	Equipment, materials, water, electricity, natural-hazard resilience, and geopolitical continuity.
Lock-in vector	CUDA, interconnect, network operations, model optimization, facility design.	CANN, UnifiedBus, Ascend tooling, domestic cloud and procurement ecosystem.	Long design and qualification cycles, process rules, packaging and capacity reservations.
Do not compare by	Vendor peak FLOPS or projected token claims alone.	NPU count or vendor comparisons to NVIDIA alone.	Headline fab count or announced investment alone.
Decisive measures	Matched goodput/MW, utilization, porting cost, uptime, total system cost.	Matched goodput/MW, reliability, developer support, software burden, total system cost.	Qualified yield and capacity, package lead time, geographic share, recovery and time to volume.

Table 4.3. NVIDIA, Huawei, and TSMC occupy different hardware control points. Vendor claims require matched-workload and production evidence.

Hardware implications for SpaceXAI, OpenAI, and Anthropic

The three focal firms do not have the same hardware posture. Their dependence should be measured through contracts, productive fleet, software portability, power, and ownership of the facility rather than inferred from model rankings. The table records disclosed evidence available at the baseline; it does not assume that announced systems have shipped or that vendor-reported cluster size equals economically productive capacity.

Firm	Disclosed hardware and build-out evidence	Strategic implication	Critical measure
SpaceXAI	A 2024 NVIDIA disclosure described 100,000 Hopper GPUs as a historical milestone. 2026 SEC filings disclose capacity agreements tied to about 325,000 NVIDIA GPUs for Anthropic and about 110,000 for Google. Those contracts do not by themselves establish one non-overlapping productive fleet or internal allocation.	Direct infrastructure and long-term customer contracts can support scale, but create delivery, utilization, termination, platform, site, power and capital-allocation risk.	Installed, delivered, simultaneously available and utilized fleet; customer allocation; termination exposure; task goodput/MW; site diversity; refresh burden.
OpenAI	OpenAI disclosed a 2026 financing that included $30B from NVIDIA and next-generation NVIDIA inference capacity. A 2025 letter of intent contemplated at least 10 GW of NVIDIA systems and up to $100B of progressive NVIDIA investment as capacity is deployed.	Capital, compute and supplier relationships can accelerate deployment while creating negotiation, execution, concentration and demand-quality questions.	Signed versus deployed capacity; investment conditions; provider mix; commissioned MW; utilization; third-party revenue; margin after depreciation and power.

Firm	Disclosed hardware and build-out evidence	Strategic implication	Critical measure
Anthropic	A diversified but interdependent portfolio spans AWS Trainium, Google TPU access, a SpaceX agreement tied to about 325,000 NVIDIA GPUs, and an NVIDIA agreement to invest up to $10B subject to conditions.	Multiple accelerator paths can improve bargaining power and continuity, while increasing contractual, financing, orchestration and portability complexity.	Workload share by platform; effective cost; porting time; retained quality; availability; contract concentration; independent fallback; related financing.

Table 4.4. Hardware posture of the focal firms at the July 24, 2026 baseline. Sources are company and supplier disclosures; counts and performance are not independently normalized.

Migrating bottlenecks: hardware scarcity and grid stress can coexist

Hardware scarcity and power stress are not opposites. Scarcity is measured against desired demand, not against zero deployment. Hundreds of thousands of accelerators can remain scarce while their absolute load is large enough to strain a local grid. The binding constraint can move by quarter, site, workload and supply-chain stage.

> *PRODUCTIVE COMPUTE = min(accelerators, HBM and packaging, networking, energized power and cooling, software-ready capacity, operator capacity, economic demand). The minimum is the binding constraint for that site and period. Substitution at one layer may relieve one bottleneck while deepening dependence at another.*

Constraint regime	What binds	Observable symptom	What would move the bottleneck
Chip-bound	Accelerators, HBM, packaging or interconnect	Energized halls or signed demand wait for systems; hardware prices and delivery times rise.	New supply, better packaging yield, smaller models, or custom silicon.
Power-bound	Grid connection, generation, transformers or cooling	Delivered racks wait for power; sites curtail or miss commissioning dates.	New generation, transmission, load flexibility, denser cooling, or geographic relocation.
Software-bound	Portability, compilers, networking, schedulers or operators	Installed systems show low productive utilization or poor reliability.	Porting, orchestration, training, tooling, and matched-workload optimization.
Demand-bound	Weak customer ROI or overbuilt capacity	Available compute has falling utilization, prices or margins.	Real workflow value, lower task cost, new demand, or project restructuring.

Table 4.4A. Migrating constraint regimes reconcile hardware scarcity with grid stress without assuming that one bottleneck is permanent.

Protocol implication: H1B tests shared hardware shocks; H2 tests financing and redeployment; H4 tests demand, efficiency, commissioned supply and latency. A site can move from chip-bound to power-bound to software-bound and finally demand-bound. The protocol records the binding constraint at each observation instead of forcing all periods into one static chokepoint story.

Data-center build-outs: from announced gigawatts to productive compute

Data-center build-outs were already central to this chapter through demand forecasts, interconnection, financing, environmental burden, and the distinction between announced and energized capacity. The hardware amendment makes the status ladder more exact. A project can have land and a power agreement while lacking transformers, chillers, HBM, packaged accelerators, switches, optical modules, software qualification, or an operating team. It can be energized while still failing to deliver production workloads.

Announced: A company or government states a target, site, investment, chip count, or gigawatt figure. This is evidence of intent, not capacity.

Contracted and financed: Land, power, construction, equipment, tenant, and financing agreements are executed with enforceable deposits, guarantees, and cancellation terms.

Under construction: Civil works, substations, generation, cooling, networking, and data halls have observable progress; duplicated announcements are removed.

Energized and rack-ready: The site has usable power and cooling at the density required by the selected hardware, not merely a grid connection at the property line.

Commissioned and software-usable: Accelerators, memory, network, storage, security, schedulers, and compilers pass acceptance tests and can execute the intended workloads.

Utilized and economically productive: Capacity delivers successful, latency-compliant work at a measured margin after power, depreciation, networking, maintenance, and human operations.

The named cases occupy different rungs. OpenAI's multi-gigawatt figures remain capacity under development rather than an energized baseline. SpaceX's 100,000-GPU Colossus milestone is historical; the more current public evidence is two customer agreements tied to approximately 325,000 and 110,000 NVIDIA GPUs, subject to ramp, delivery, termination and allocation conditions. Amazon's Project Rainier disclosure describes operational Trainium capacity, while utilization and comparable task economics remain private. TSMC's first Arizona fab had reached volume production, while later fabs and packaging facilities remained construction or announced capacity. Quarterly reviews must move each project through the status ladder instead of summing announcements, contracts and productive systems (OpenAI 2025c; SpaceX 2026e; SpaceX 2026g; Amazon 2026a; TSMC 2026a).

Power is not a fixed revenue ceiling - but it is a physical and latency constraint

A firm earns revenue from useful work produced by an energized system, not from electricity in the abstract. Chip efficiency, architecture, routing, cache reuse, utilization, task success, latency, reliability, price, and the share of customer value captured determine economics within a physical envelope. Better economics can raise revenue per megawatt; they cannot make a fixed local power and cooling system perform unlimited sequential computation.

> CAPACITY AND ENERGY PRODUCTIVITY - MEASUREMENT WINDOW T
> Economic output rate ($/hour) = verified outcomes delivered within the required service level x average economic value per outcome / hours in T.
> Capacity productivity ($/MW-hour of energized capacity) = economic output in T / (average energized MW x hours in T).
> Energy productivity ($/MWh consumed) = economic output in T / electricity consumed in T.
> “Verified outcome” already includes correctness, reliability and latency compliance; do not multiply those factors again.

Illustrative worked example (hypothetical, not an observation): a system averages 20 MW of energized capacity for 24 hours, consumes 360 MWh, and delivers 120,000 verified outcomes worth $2.50 each. Economic output is $300,000; capacity productivity is $625 per MW-hour; energy productivity is about $833 per MWh. Outcomes missing the service-level deadline are excluded rather than discounted a second time.

Input	Hypothetical value	Calculation
Energized capacity window	20 MW x 24 hours	480 MW-hours of available capacity
Electricity consumed	360 MWh	75% average electricity-to-capacity ratio in the illustrative window
Verified outcomes	120,000 x $2.50	$300,000 economic output
Capacity productivity	$300,000 / 480 MW-hours	$625 per MW-hour
Energy productivity	$300,000 / 360 MWh	$833 per MWh

Table 4.5. Worked example showing consistent units for capacity and energy productivity.

The original claim should therefore be modified in both directions. Power scarcity increases the premium on successful work per unit of compute, but high-value tasks are not infinitely compressible. A provider can improve revenue through routing, utilization, smaller models, higher success, and higher-value outcomes. It can still hit local limits when workloads require many serial reasoning and tool-use steps, strict service levels, geographic residency, or continuous firm capacity.

The rebound effect

Efficiency does not necessarily reduce total electricity use. Lower cost per inference can make AI viable in more applications, increase frequency of use, and enable longer or more autonomous tasks. The result resembles a rebound effect: each task uses less compute, but the number and scope of tasks expand faster. The IEA's wide scenarios reflect this uncertainty. Energy efficiency is essential for affordability and reliability, yet it may accelerate aggregate demand rather than end the power build-out (IEA 2025a; IEA 2025b).

This is one reason the economic metric should be cost per successful task rather than cost per token. A system that generates many cheap tokens but requires repeated attempts, human correction, or downstream remediation can use more energy and labor than a more expensive

model that completes the task correctly. Energy productivity and business productivity must be measured together.

Latency, sequential reasoning, and the local physical ceiling

Agentic systems can consume many inference calls, retrievals, tool executions, verifications, and retries for one user intent. Some work can be batched or shifted; real-time settlement, operations, diagnostics, coding feedback, and customer interactions often cannot. P50 averages are therefore insufficient. P95 and P99 completion latency, deadline compliance, queueing, retry depth, and the number of sequential model and tool steps belong in the economic denominator.

The relevant distinction is local physical throughput versus global revenue. A firm may grow by moving work, adding sites, raising price, or selecting higher-value tasks, but a specific grid zone and facility have finite power, cooling, networking, and latency capacity. The thesis is weakened if efficiency and new supply consistently improve cost and latency faster than demand; it is strengthened when demand, sequential depth, and service requirements produce persistent scarcity rents or missed service levels.

Firm load, flexible load, and interruptible intelligence

Not every AI workload requires identical reliability. Online consumer inference, safety systems, and live enterprise agents may need low latency and high availability. Training runs, evaluation batches, synthetic data generation, and some research tasks can be delayed, moved, or interrupted. This creates a potential new class of flexible industrial load. Providers that can shift work across time and regions may receive lower power costs and help integrate variable generation, while providers that demand continuous firm capacity will require more generation and grid investment.

The distinction should be measured rather than assumed. A data center can advertise flexibility while preserving contractual service levels that make actual curtailment rare. Analysts should track the percentage of load that is technically and contractually interruptible, the duration of acceptable curtailment, geographic mobility, backup generation, and the compensation paid for grid services.

Environmental and community constraints

Electricity is only one local input. Data centers can affect water withdrawal and consumption, land use, construction, backup generation, noise, heat, air emissions, transmission corridors, and municipal infrastructure. These effects vary sharply by cooling technology, climate, fuel mix, siting, and whether waste heat or reclaimed water is used. A credible infrastructure thesis must measure the burden rather than treating community acceptance as a permitting delay.

The dashboard should therefore track water intensity, carbon intensity by hour and region, backup-generator use, heat reuse, land and transmission footprint, local rates, tax and employment benefits, complaints, and community-benefit commitments. A project that improves

national compute capacity while transferring unpriced environmental or infrastructure costs to one locality has not demonstrated socially efficient scale.

The financing bridge from AI optimism to the utility system

The physical build-out becomes a financial-system issue through contracts. Utilities may construct generation and transmission for a customer whose future demand is uncertain. Infrastructure funds and special-purpose vehicles may finance data centers under long-term leases. Cloud companies may sign take-or-pay agreements; chip suppliers may extend credit; strategic investors may invest in a lab that later spends the proceeds on the investor's infrastructure. Each arrangement reallocates risk.

The central questions are who guarantees the load, what happens if the customer delays or cancels, whether the asset has alternative users, and whether costs can be recovered from ordinary ratepayers. FERC Commissioner David Rosner emphasized cost-recovery agreements and protections against speculative projects in the 2026 large-load proceedings (FERC 2026b). This is the appropriate prudential lens: the danger is not high electricity use alone, but long-lived regulated or leveraged assets built against weakly secured demand.

Financing structure	Who receives the upside	Who may bear downside	Key diligence question
Balance-sheet capex	The frontier firm or hyperscaler	Equity and unsecured creditors	Can operating cash flow support depreciation, maintenance, and refresh cycles?
Project-financed data center	Developer, lenders, equity sponsors, tenant	Project lenders and sponsors; sometimes guarantor	Is the lease investment-grade, transferable, and matched to debt maturity?
Utility rate-base investment	Utility shareholders and large-load customer	Ratepayers if cost allocation or exit protection is weak	Does the customer fund dedicated facilities and termination risk?
Take-or-pay capacity	Supplier and customer through secured availability	Customer if demand disappoints; supplier if guarantee fails	Are commitments enforceable, and is capacity fungible?
Strategic/vendor financing	Both parties through ecosystem growth	Investor-supplier and firm if demand is circular	Is third-party demand independent of financing relationships?

Table 4.6. Infrastructure risk follows the contract, not the press release.

Circularity and demand quality

The frontier ecosystem contains observable capital-compute-revenue interdependence. OpenAI announced a $110 billion financing that included $30 billion from NVIDIA and said it secured next-generation NVIDIA inference compute. A prior OpenAI-NVIDIA letter of intent contemplated at least 10 GW of systems and up to $100 billion of progressive NVIDIA investment. NVIDIA

disclosed an agreement, subject to conditions, to invest up to $10 billion in Anthropic. SpaceX disclosed customer agreements tied to approximately 325,000 NVIDIA GPUs for Anthropic at $1.25 billion per month and approximately 110,000 GPUs for Google at $920 million per month after ramp-up. These facts do not prove improper circularity; they show that capital, supplier demand, compute contracts and reported revenue can be mutually dependent (OpenAI 2026g; OpenAI 2025d; NVIDIA 2025a; SpaceX 2026e; SpaceX 2026g).

Amazon provides another named example of the same linkage. It announced a $50 billion OpenAI investment, beginning with $15 billion and followed by $35 billion when stated conditions are met. The partnership also names AWS as the exclusive third-party cloud distribution provider for OpenAI Frontier, commits OpenAI to consume 2 GW of Trainium capacity, and expands the companies' infrastructure agreement by $100 billion over eight years. The investment, distribution right and compute commitment should be recorded as separate flows rather than treated as one proof of demand (Amazon 2026b; OpenAI 2026h).

Headline contract arithmetic also requires denominator discipline. Dividing the disclosed monthly payment by the disclosed GPU count produces about $3,846 per GPU-month for Anthropic and $8,364 for Google. Those ratios are not comparable unit prices. Public disclosures do not normalize GPU generation, delivery timing, guaranteed versus maximum capacity, networking, storage, power, software, support, priority rights, duration, utilization or financing services.

Customer	Disclosed monthly payment	Disclosed GPU count	Headline ratio	Interpretation
Anthropic	$1.25B	About 325,000	About $3,846 per GPU-month	Arithmetic ratio only; contract scope and ramp are not normalized.
Google	$920M after ramp-up	About 110,000	About $8,364 per GPU-month	Arithmetic ratio only; services, generation and rights may differ.

Table 4.6A. Headline payment-per-disclosed-GPU ratios are not normalized contract prices.

The relevant demand-quality tests are cash from end users unrelated to strategic financing; contract cancellation and delivery rights; capacity utilization; renewal without new investment; gross margin after depreciation, networking and power; and whether customer commitments are enforceable, transferable and matched to asset life. Analysts should map each flow separately: equity investment, hardware purchase, cloud or compute commitment, capacity-backed financing, guarantee, customer payment and independent end-user revenue. Appearance on both sides of a transaction is a reason for measurement, not a finding of misconduct.

What to monitor

Energized capacity: Megawatts actually connected and available, not merely announced or permitted.

Utilization: Productive accelerator hours divided by available hours, adjusted for maintenance and reserved capacity.

Revenue and gross profit per MW: A measure of whether economic value is scaling faster than the physical footprint.

Successful tasks per kWh and latency: Include retries, routing, tool calls, human remediation, and P50/P95/P99 completion against the required service level.

Interconnection quality: Firm versus interruptible service, queue position, completion milestones, and customer-funded facilities.

Financial exposure: Debt, guarantees, termination payments, take-or-pay obligations, and ratepayer cost allocation.

Refresh burden: The cadence at which accelerators and cooling systems become economically obsolete relative to financing maturity.

Rebound and environmental intensity: Compare growth in total task demand with efficiency gains, and track water, emissions, land, backup generation, and community cost per successful task.

Hardware platform mix and portability: Productive workload share by accelerator and generation; CUDA, Neuron, CANN, and other porting time; performance retained after migration; software and operator availability.

Fabrication, memory, and packaging: Production-qualified leading-edge and advanced-packaging capacity by geography; HBM allocation; yield, lead time, recovery time, and customer concentration.

Commissioning conversion: Ordered chips, delivered racks, rack-ready MW, commissioned systems, productive accelerator hours, and economically productive MW reported as separate stages.

> **PRIMARY INFERENCE**
>
> Power scarcity does not prove that frontier labs must become services firms. It strengthens the incentive to increase the value captured from each successful task. Whether that incentive produces vertical integration depends on model pricing, customer bargaining power, partner economics, liability, and organizational capability.

CHAPTER 5

From tokens to outcomes: the frontier-lab business model

The frontier lab can sell access to intelligence, a software seat, an autonomous worker, a transaction, or a completed business result. Each step captures more value and assumes more responsibility.

The frontier-lab revenue question is central because infrastructure scale alone does not determine returns. A lab that sells undifferentiated tokens into a competitive market may experience falling unit prices even as usage grows. A lab that owns consumer distribution, enterprise deployment, workflow context, and transaction authority can capture a larger share of the value created. The strategic pressure is therefore toward deeper embedding - but the economically optimal boundary remains unsettled.

The revenue ladder

Layer	What the customer buys	Pricing basis	Provider responsibility	Strategic trade-off
Model/API	Tokens, calls, embeddings, or model access	Usage	Availability, model quality, safety controls	Scalable and partner-friendly, but vulnerable to price competition
Seat/application	A packaged assistant or productivity tool	Subscription per user or tier	Interface, workflow features, support, data controls	Distribution and recurring revenue, but competes with software incumbents
Agent platform	A system that plans and takes actions through tools	Seat, usage, capacity, or hybrid	Orchestration, permissions, monitoring, evaluation	Higher switching costs, but larger security and reliability burden
Managed agent	Ongoing digital labor for a defined process	Capacity, work unit, service level	Operation, exception handling, continuous improvement	Captures service margin, but requires delivery operations
Transaction/outcome	A completed claim, sale, reconciliation, resolution, or other result	Per transaction, savings share, or outcome fee	Execution quality and sometimes financial or legal liability	Highest value capture and proof burden; most direct conflict with customers and partners

Table 5.1. The revenue ladder increases value capture and liability together.

Evidence of movement beyond model access

OpenAI's March 2026 financing announcement connected infrastructure to higher revenue per unit of compute and described demand moving from model access toward systems that reshape business operations. Its Frontier and Presence offerings emphasize context, permissions, evaluations, actions, and deployment inside workflows. These disclosures establish management intent and product direction. They do not independently validate economy-wide productivity, outcome margins, or the claim that direct workflow ownership is superior to partner distribution (OpenAI 2026b; OpenAI 2026c; OpenAI 2026d).

Anthropic has similarly developed managed-agent infrastructure and enterprise partnerships for longer-horizon tasks. Its agent-safety research identifies unintended action, malicious instruction following, and consequential failures as a new boundary. This is evidence that labs are building toward execution; it is not proof that a horizontal lab can operate thousands of domain processes more efficiently than hyperscalers, software vendors, integrators, or specialized service firms (Anthropic 2026c; Anthropic 2026d).

SpaceX's pending acquisition of Cursor is the protocol's one canonical prospective treatment of infrastructure-to-application integration. The company says the application can provide distribution, structured feedback, and recurring demand for compute and Grok. The signed agreement is evidence that management is attempting the strategy, not that it works. Closing, regulatory conditions, model-neutral access, user retention, data rights, integration cost, dilution and return on invested capital remain unresolved (SpaceX 2026d; SpaceX 2026e).

Why public markets may accelerate - or distort - the move

Public investors will evaluate growth, margins, capital intensity, free cash flow, and return on invested capital. Falling model prices and large infrastructure commitments can encourage applications, acquisitions, agents, and outcomes. The same pressure can reward revenue narratives before fully loaded margins are known, encourage stock-funded overpayment, and favor visible integration over neutral partner ecosystems. Public-market scale changes the feasible strategy; it does not select the correct boundary.

The pressure can be expressed as a margin problem. The provider must compare the price of an outcome with all costs required to deliver it reliably. Those costs include inference, storage, integration, monitoring, human exception handling, customer support, indemnity, insurance, compliance, and the expected cost of errors. The outcome model is attractive only when the provider can standardize the process and control variance.

> **TASK MARGIN**
>
> Task margin = outcome price - inference cost - integration cost - monitoring cost - human exception cost - compliance cost - expected liability. A lab should own the workflow only when this margin and the strategic data advantage exceed the value of remaining a neutral platform.

Three competing routes to the workflow control point

The control point can migrate through at least three routes. A model lab can move outward into agents and outcomes. An infrastructure owner can acquire models or applications. A hyperscaler can retain identity, security, data, procurement, billing, and the enterprise contract while routing requests among many model providers. Microsoft Foundry and Amazon Bedrock already demonstrate managed routing and governance services, making hyperscaler intermediation an observable alternative to lab-owned workflow integration (Microsoft 2026; Amazon Web Services 2026).

Route	Control point	Commercial consequence
Model lab moves outward	Agent runtime, application, workflow context, transaction authority	Higher value capture and liability; possible conflict with partners and customers.
Infrastructure owner moves upward	Compute plus acquired model or application distribution	Utilization and feedback may rise, or neutrality and focus may be destroyed.
Hyperscaler intermediation	Identity, cloud, data, security, router, procurement, billing	Labs can become substitutable suppliers behind one managed enterprise interface.
Neutral application multi-homes	User workflow and provider choice	Application preserves bargaining power and routes among models on quality, cost, or policy.
Enterprise-managed architecture	Customer policy, data, evaluation, and exit rights	Customer captures orchestration value but bears engineering and governance cost.
Vertical software or process provider	Domain workflow, data, accountability, and distribution	Specialist captures outcome value while buying models and compute as inputs.

Figure 5.1. Rival routes to the workflow control point: lab integration, infrastructure integration, hyperscaler intermediation, neutral applications, and customer control.

Cursor as a prospective natural experiment

The transaction can test whether model neutrality is an asset that survives ownership. Meaningful neutrality requires continued access to rival models; simultaneous or nondiscriminatory releases; transparent and overridable routing; no punitive markup; bring-your-own-provider options; portable workflow history; and data-use rules that do not disadvantage competitors. The thesis should not infer neutrality from a list of available models if defaults, latency, price, or context portability steer users elsewhere.

Post-close adjudication should compare pre-deal cohorts with matched neutral competitors under the H3 thresholds in Chapter 12. Relevant measures are active-developer and net-revenue

retention, model mix, voluntary Grok adoption, routing overrides, rival-provider terms, application gross margin, customer acquisition, integration cost, impairment, successful tasks per unit of compute, and return on the shares issued. If the merger is blocked or abandoned, it becomes evidence about regulatory or transaction feasibility rather than integration performance; H3 then rolls prospectively to the next qualifying stock-financed application deal through 2030.

Why the move is not inevitable

Neutral platforms can be more valuable than operators

A lab or infrastructure owner that competes with every customer, software vendor, consultancy, and business-process provider may weaken its distribution ecosystem. Cloud platforms became powerful partly by enabling partners rather than performing every end task. The Cursor case sharpens this tension: part of the application's value may come from model choice, while an integrated owner has incentives to privilege Grok and internal compute. If neutrality declines faster than integration improves performance, users can migrate and the acquisition can destroy the distribution asset it was intended to secure.

Domain work is operationally messy

Business processes contain exceptions, tacit knowledge, local regulation, legacy systems, interpersonal judgment, and accountability structures that do not disappear when model performance improves. The last ten percent of automation can consume most of the delivery effort. A frontier lab optimized for research and horizontal infrastructure may be structurally ill-suited to thousands of industry-specific service obligations.

Liability can overwhelm inference margin

A provider that executes a payment, denies a claim, changes a production system, or communicates with a customer is no longer selling only an informational tool. It may face contract damages, regulatory penalties, professional standards, discrimination claims, security obligations, and operational losses. Customers may demand indemnification precisely when the provider wants a high-margin, standardized product.

Customers will resist strategic dependency

Large enterprises may welcome managed automation while refusing to cede control over pricing, customer relationships, product decisions, or regulated judgments. They can use multi-model routing, internal control planes, system integrators, and open-weight models to retain workflow authority. The bargaining equilibrium may therefore be a modular architecture rather than full vertical integration.

Hyperscaler counter-leverage

Hyperscalers control identity, network policy, enterprise data, cloud commitments, systems of record, security tooling, marketplaces, procurement, and billing. They can bundle competing models, alter hosting economics, subsidize routing, or place a lab behind a managed endpoint. A

frontier model can remain technically differentiated while losing the customer relationship and a large share of the margin.

The empirical contest is therefore not only model versus model. Track the share of enterprise inference purchased through cloud routers, the party holding the master contract, direct versus intermediary billing, data and evaluation ownership, committed-cloud-spend incentives, model substitution frequency, and margin retained by the lab. Hyperscaler-managed plurality can reduce model lock-in while increasing cloud-level concentration.

Where frontier labs are most likely to own outcomes

The most attractive processes have high volume, digital inputs and outputs, measurable quality, limited physical-world variance, clear permissions, and a repeatable exception taxonomy. Coding, document processing, customer support triage, internal IT resolution, revenue operations, claims preparation, compliance review, and research synthesis fit parts of this profile. Processes with irreversible physical consequences, ambiguous fiduciary duties, high emotional stakes, or fragmented local rules are less likely to be fully owned by a horizontal lab.

Process characteristic	Favors platform/API	Favors managed outcome
Task variability	High and organization-specific	Low to moderate with repeatable exception classes
Measurability	Quality is subjective or delayed	Result, time, accuracy, or savings can be audited
Liability	Large, uncapped, or regulated	Contractible, insurable, and bounded
Data advantage	Customer-specific and nonportable	Feedback generalizes across many customers
Integration	Deep legacy customization	Common systems and standardized interfaces
Customer strategy	Workflow is core differentiation	Workflow is necessary but non-differentiating

Table 5.2. A boundary test for whether the lab should supply intelligence or own the work.

A likely hybrid equilibrium

The most plausible near-term structure is a layered market. Frontier labs will own consumer applications, developer platforms, general agent runtimes, and selected high-scale workflows. Enterprise software vendors will embed multiple models into systems of record. Consultancies and integrators will redesign processes and operate exceptions. Business-process providers will add AI to labor-intensive services. Large enterprises will retain orchestration, data, identity, and evaluation control. The boundaries will move process by process rather than through one universal replacement event.

CHAPTER 6

Enterprise reorganization and the contest for the workflow

The largest near-term economic event may be neither a model breakthrough nor an IPO, but a redistribution of value across software, services, labor, and the operating processes of the firm.

Enterprise AI is often evaluated as a feature: a copilot added to existing software, a chatbot connected to internal documents, or a coding assistant sold by the seat. The deeper transition occurs when the unit of adoption changes from the individual user to the workflow. At that point, the organization must decide who owns process design, permissions, data, evaluation, exception handling, and accountability. This is the contest for the workflow.

Why the system of record is no longer the only control point

Traditional enterprise software gained power by becoming the authoritative database for customers, employees, finance, service, or operations. Agents create a new layer above those systems. An agent can read from several systems, decide what to do, and write back across them. The control point can migrate from the application that stores the record to the orchestration layer that interprets the situation and initiates action.

This does not make systems of record obsolete. Their data integrity, permissions, audit trails, and domain logic remain essential. It does, however, pressure seat-based economics when fewer humans directly interact with the interface. Vendors may respond by pricing agents, transactions, data access, or outcomes. The result is not simply "software is replaced by AI." It is that software, labor, and services become recombined around a new execution layer.

Incumbent layer	Traditional source of value	AI-era pressure	Likely response
Enterprise SaaS	System of record, workflow UI, per-seat subscription	Agents reduce direct UI use and can route across applications	Agent pricing, transaction fees, proprietary data services, embedded orchestration
Consulting	Diagnosis, process design, implementation, change management	Models automate analysis and artifact production	Outcome contracts, proprietary accelerators, governance, transformation operations
Business-process outsourcing	Labor arbitrage, process scale, service levels	Digital labor compresses headcount-based revenue	Human-agent operations, exception specialization, transaction pricing

Incumbent layer	Traditional source of value	AI-era pressure	Likely response
Internal shared services	Control, institutional knowledge, standardized support	Automation challenges staffing and location models	AI operations center, retained process ownership, smaller expert teams
Frontier lab	Model capability, compute, API, consumer distribution	Model access commoditizes and customers demand proof of value	Agents, deployment platforms, managed workflows, commerce and transactions

Table 6.1. AI redistributes the control point rather than eliminating every incumbent layer.

Four enterprise operating models

A production agent is a socio-technical system with identity, authority, tools, memory, policy, evaluation, escalation, and a human owner. But most enterprises will not build every control component. The key choice is which operating model supplies those functions and which control rights remain with the customer.

Single-vendor consolidation: One application, model family, cloud, and governance suite minimize integration burden but maximize provider concentration.

Hyperscaler-managed plurality: One cloud, identity system, contract, and policy plane routes among several models. Procurement is concentrated even when model use is diverse.

Independent managed control plane: A neutral vendor or integrator operates routing, evaluation, observability, and fallback across providers and clouds.

Self-managed sovereignty: The enterprise operates the architecture internally. This is plausible for governments, regulated institutions, and unusually capable firms - not the default for the Global 2000.

Stage	Mechanism	Economic consequence
1	Observe	The agent receives events, data, or requests from approved sources.
2	Interpret	Models classify the situation, retrieve context, and propose a plan.
3	Authorize	Policy and identity systems determine whether an action is allowed.
4	Act	The agent writes to systems, communicates, purchases, changes code, or triggers a process.
5	Evaluate	Automated and human checks assess correctness, safety, cost, and outcome.

Stage	Mechanism	Economic consequence
6	Escalate and learn	Exceptions go to accountable humans; feedback updates policies, prompts, tools, or models.

Figure 6.1. The enterprise agent control loop. Durable advantage resides in the whole loop, not the model alone.

The durable enterprise moat

The phrase “proprietary data is the moat” is incomplete. Raw data can be low quality, legally constrained, stale, or disconnected from decisions. The more durable advantage is a governed feedback loop: rights-cleared data, authority to act in a workflow, human review, outcome labels, evaluation infrastructure, and distribution to the people or systems that use the result. A provider that owns this loop can improve reliability and economics faster than a provider with a static data repository.

A stock-financed application acquisition can be an attempt to acquire a governed feedback loop rather than only another model. The pending boundary case and its data requirements are treated once in Chapter 5; no enterprise moat is inferred before post-close evidence.

> **DURABLE DATA ADVANTAGE**
>
> Rights-cleared data + workflow authority + human feedback + evaluations + distribution. Remove any one element and the “data moat” is weaker than it appears.

Why many pilots fail to become economic systems

Pilot success is often measured by user satisfaction, benchmark quality, or time saved on a sample task. Production value depends on a broader denominator. The enterprise must pay for integration, permissions, data cleanup, evaluation, monitoring, security, change management, human review, and model usage. It must also account for work created by the system: exceptions, duplicated outputs, rework, new compliance reviews, and higher demand generated by lower unit cost.

The correct return-on-investment calculation compares the fully loaded cost and outcome quality of the redesigned process with the previous process. It should distinguish gross time saved from cash savings, capacity released from headcount eliminated, faster cycle time from additional demand, and local productivity from enterprise-wide financial performance. The ILO's review of empirical studies finds real but uneven productivity effects, limited evidence of large aggregate displacement so far, and substantial dependence on work design and institutional context (ILO 2026a).

ROI field	Common weak measure	Better measure
Productivity	Self-reported hours saved	Completed output per paid hour at equal or better quality
Quality	Model benchmark or user rating	Error rate, rework, customer impact, compliance, and downstream loss
Cost	Inference spend	Fully loaded integration, supervision, security, change, and liability cost
Adoption	Licenses issued or pilot users	Share of eligible workflow volume processed in production
Financial impact	Estimated annual benefit	Realized cash flow, avoided cost, revenue, working capital, or risk loss
Resilience	Normal-operation uptime	Recovery time, fallback success, vendor-exit test, and incident performance

Table 6.2. Enterprise AI ROI should be measured at the process level, not inferred from model usage.

The middle-tier compression thesis

The most exposed firms may be neither frontier labs nor deeply embedded systems of record, but thin layers of interface, prompting, or manual service between a broadly available model and a customer workflow. Labs can move into applications, incumbents can combine models with proprietary data and distribution, and customers can assemble open-weight alternatives. The middle tier survives through defensible workflows, regulated trust, unique distribution, proprietary outcome data, or operating excellence.

This compression extends beyond software. Advisory and outsourcing providers priced by hours or headcount must shift toward outcome accountability, specialized judgment, risk transfer, or proprietary operating platforms as AI reduces routine labor. The growth category is ongoing, technology-mediated execution of defined work - not simply more consultants using AI.

The operational capability constraint

Multi-model routing sounds modular at the architecture level but creates real operating burden: prompt and tool drift, provider-specific context and safety behavior, evaluation maintenance, identity, logging, data-zone rules, cost allocation, incident response, and version change. Many enterprise IT departments will rationally buy these capabilities from a hyperscaler or managed provider instead of operating them as a custom platform.

This yields an important measurement distinction. Procurement concentration can coexist with model diversity. An enterprise may use many underlying models while remaining dependent on one cloud, identity layer, router, contract, and billing relationship. Concentration must therefore be measured separately at the cloud/control-plane, model, application, data, and compute layers.

A board-level architecture for control rights and bargaining power

- **Own the control rights, not necessarily the software: Retain authority over identity, policy, approved models, evaluation, audit, cost limits, recovery, and exit even when a managed provider operates the plane.**
- **Separate model, control plane, and workflow: Avoid making one provider's agent representation, memory, or tool schema the only executable form of a critical process.**
- **Retain outcome data:** Contract for logs, corrections, evaluations, and process-performance data.
- **Design human accountability:** Name decision owners and the conditions that trigger human review.
- **Test exit before scale:** Prove fallback, model switching, or manual operation within a defined recovery time.
- **Price the whole process:** Negotiate against verified outcomes, not seats or token discounts.

Most production systems will combine purchased models, a managed control plane, internal policy and evaluation, third-party software, and human operations. The strategic decision is which rights and evidence the enterprise must retain - not whether it writes every routing component itself.

CHAPTER 7

Sovereign and open-weight AI as a counterstrategy

Sovereignty is not a binary choice between a public API and national frontier training. It is a verified capability to govern, continue, audit, adapt, and exit at the level required by the workload - whether the control plane is self-operated or purchased as a managed service.

The original theory correctly identifies sovereign data, private processing, open-weight models, and internal talent as responses to concentration. Open weights are also more than a hedge: they are a competing industrial architecture built around diffusion, specialization, and application-layer entry. The theory nevertheless overestimates the number of organizations that can operate a dynamic multi-model stack. For most enterprises, sovereignty will be procured through contract, dedicated deployment, managed routing, portable evidence, and tested recovery rather than built from first principles.

Open source, open weight, and sovereign are different claims

The terms are often used interchangeably but should not be. An open-weight model makes trained parameters available under a license, while training data, architecture details, code, or process may remain partly closed. Open source is a stronger claim about inspectable, modifiable, and reusable components under recognized licensing conditions. Sovereign AI concerns who can operate, govern, modify, audit, and sustain the system under the organization's legal and strategic authority. A proprietary model can support some sovereignty goals through dedicated regional deployment; an open-weight model can still create dependency on foreign chips, cloud tooling, or a small group of maintainers. Because releases differ materially, openness should be decomposed rather than treated as one binary label.

State of the open-weight frontier at the evidence cutoff

The open-weight ecosystem is not one market or one license. At the baseline, relevant families included Meta’s Llama, Mistral, Alibaba’s Qwen, DeepSeek, and Moonshot’s Kimi line, while Hugging Face functioned as a major distribution and tooling layer. Llama uses a community license rather than an OSI-style open-source license. Most Mistral open models use Apache 2.0, while some use a modified MIT license with a commercial threshold. Qwen3 published multiple checkpoints under Apache 2.0. DeepSeek releases weights but does not thereby disclose a complete training-data and training-process record (Meta 2025; Mistral 2026; Qwen 2025; DeepSeek 2026).

Moonshot announced Kimi K3 on July 16, 2026 and made hosted access available, while promising full weights for July 27. At the July 24 cutoff, the weights and final license were not yet available

for verification, so K3 is recorded as an announced open-weight release, not as a completed downloadable release. Its timing is nevertheless proximate context for the July 24 U.S. coalition statement and the associated policy dispute (Financial Times 2026a).

Family or layer	Origin	Baseline openness status	Protocol implication
Llama	United States	Downloadable weights under Meta community terms; not equivalent to fully open source.	Weight access can coexist with use restrictions and incomplete training transparency.
Mistral	France	Most open models Apache 2.0; some modified MIT terms with a revenue threshold.	License must be evaluated model by model.
Qwen3	China	Multiple open-weight checkpoints under Apache 2.0.	Strong evidence of broad model-layer entry and tooling support.
DeepSeek	China	Downloadable weights and technical disclosures; not a complete open training stack.	Open weights do not eliminate provenance, governance or supply-chain questions.
Kimi K3	China	Hosted at cutoff; full weights promised after the cutoff.	Do not count announced openness as delivered portability.
Hugging Face / infrastructure layer	Cross-border ecosystem	Distribution, hosting, libraries and evaluation rather than one model family.	Open diffusion can strengthen a concentrated or commercially important substrate.

Table 7.1. Open-weight status is model- and license-specific at the baseline.

The openness stack

A release can be open at one layer and closed at another. The following dimensions should be recorded separately before the thesis infers competition, portability, or sovereignty from an "open" label.

Weights and architecture: Whether trained parameters and sufficient architectural information are available for meaningful inspection and adaptation.

Code and reproducibility: Whether inference, evaluation, fine-tuning, and training code are available and whether an independent party can reproduce material behavior.

License rights: Whether modification, redistribution, commercial use, hosted service, and model-derived development are permitted without discriminatory restrictions.

Data and provenance: Whether the origin, rights, curation, and material limitations of training and adaptation data are documented.

Evaluation and safety transparency: Whether capabilities, limitations, incidents, red-team results, and mitigation evidence are available for independent scrutiny.

Deployment portability: Whether the model can run across viable hardware, cloud, and on-premises environments without losing essential workflow context or control.

No single layer establishes full openness, and no openness profile by itself establishes enterprise sovereignty. The relevant question is which rights, artifacts, operating capabilities, and alternatives are usable for the workload in question.

The sovereignty and operating-capability ladder

Level	Operating model	Control retained by customer	Capability burden	Typical adopter
1	Single-provider managed suite	Application policy, contract terms, basic data controls	Low; provider owns most technical control and change	Teams, startups, noncritical workflows
2	Hyperscaler-managed multi-model service	Approved model set, identity, data zone, policy, logs, routing mode	Moderate procurement and evaluation; cloud-level dependence	Large enterprises and public agencies
3	Dedicated or private managed deployment	Isolation, residency, capacity, logging, contractual continuity	Contract and integration capability; provider still operates stack	Regulated and latency-sensitive organizations
4	Independent managed control plane with open-weight options	Cross-provider routing, evaluation, portability, fallback, data and workflow evidence	Strong architecture and vendor-governance capability	Large enterprises and sector consortia
5	Self-managed domain or distilled stack	Deployment, behavior, update cadence, offline operation, intellectual property	Model serving, security, data engineering, continuous evaluation	Well-capitalized firms with repeated domain workloads
6	Full national or frontier-model stack	Research, training, infrastructure, supply chain, strategic autonomy	Gigawatt-scale capital, frontier talent, hardware and institutional depth	Major states, hyperscalers, and a small number of labs

Table 7.2. Sovereignty includes an operating-capability ladder; most firms will buy managed control rather than build the entire stack.

Why enterprises pursue sovereignty

Data control: Restrictions on residency, retention, secondary use, trade secrets, personal data, or classified information.

Operational resilience: The ability to continue critical work during provider outages, policy changes, export restrictions, or contract disputes.

Bargaining power: A credible alternative to a single vendor improves price and contract negotiations.

Behavioral control: The ability to evaluate, adapt, constrain, and document system behavior for a specific domain.

Strategic autonomy: Reduced exposure to a foreign jurisdiction, political decision, or concentrated supply chain.

Economic optimization: At sufficient volume, dedicated inference or smaller domain models may lower cost and latency.

These motives can conflict. A national model hosted on imported accelerators and built from foreign software may satisfy data residency while remaining dependent on the external hardware stack. An open-weight model may improve portability while increasing the organization's security and maintenance burden. A managed proprietary system may provide better audited controls than a locally operated system with weak engineering. Sovereignty should therefore be evaluated by threat model and operating capability rather than by deployment location alone.

Vertical ownership can turn an application-level procurement into a stack-level dependency. The risk is general: an enterprise adopting an application may also inherit the owner’s preferred model, data policies, compute, roadmap and capital-allocation choices. Contractual model choice, portable logs and evaluations, data-use limits, and tested exit are therefore economically important.

The economics of local operation

Local inference is attractive when workload volume is predictable, hardware can be well utilized, latency or privacy is valuable, and the model is small enough to operate efficiently. It is unattractive when demand is bursty, models change rapidly, specialized hardware sits idle, or the organization lacks model-serving and security talent. Cloud pricing includes a margin, but it also converts utilization, refresh, reliability, and staffing risks into a service.

> **TOTAL-COST TEST**
>
> Local AI cost = hardware depreciation + financing + power + cooling + networking + facilities + operations talent + model maintenance + security + evaluation + downtime + refresh risk. Compare this with managed-service cost at equal quality, latency, reliability, and control - not with token price alone.

The economics often favor a portfolio operated through a managed or customer-governed control layer. Stable, sensitive workloads may run privately; bursty frontier tasks may use managed providers; low-risk work may route to smaller models. The customer does not need to write the router, but it must retain policy authority, evaluation evidence, logs, approved-provider constraints, and a tested path to continue critical work.

Security and trust are not automatic properties of either model type

A closed service can concentrate vulnerability: one compromised provider, malicious update, outage, or policy failure can affect many customers. An open-weight model can be inspected and operated privately, but the customer assumes patching, hardening, access control, model provenance, monitoring, and incident response. The relevant comparison is the security of the complete deployed system, including data pipelines, tools, agent permissions, identity, code dependencies, and operators.

The National Institute of Standards and Technology's AI Risk Management Framework and Generative AI Profile provide a useful structure: govern, map, measure, and manage risks across the lifecycle rather than treating model provenance as a sufficient control (NIST 2023; NIST 2024).

For critical infrastructure, the central questions are failure containment, human override, monitoring, recovery, supply-chain assurance, and evidence that controls work under adversarial conditions.

Risk	Concentrated proprietary service	Self-hosted open-weight system	Hybrid mitigation
Provider outage	High correlated exposure across customers	Lower external dependence but local operations may fail	Multi-provider routing, fallback models, tested manual mode
Data exposure	Contractual and technical dependence on provider controls	Data remains local, but internal access may be weak	Minimize data, encrypt, segment, audit, and enforce purpose limits
Model compromise	Central update can propagate widely	Customer controls versions but must detect and patch issues	Signed artifacts, staged release, independent evaluations, rollback
Prompt injection and agent abuse	Provider offers controls but cannot know all workflows	Enterprise can customize controls but owns the burden	Least privilege, tool isolation, transaction limits, human approval
Vendor lock-in	Proprietary interfaces, memory, and evaluation data	Lower model lock-in; higher infrastructure specialization	Portable logs, common interfaces, model-agnostic control plane
Compliance evidence	Provider certifications may help	Enterprise must produce evidence itself	Shared control matrix and auditable responsibility allocation

Table 7.3. Security is an allocation of control and responsibility, not a synonym for closed or open.

Open-weight safety: distributed defense versus irreversible capability

The July 24, 2026 coalition statement concedes that released weights cannot easily be recalled and that modified versions can be difficult to trace or reverse. It nevertheless argues that broader access equips defenders, researchers, and operators to inspect behavior, identify vulnerabilities, red-team systems, and avoid a small number of opaque single points of failure (American Innovators Network et al. 2026). That is a serious safety theory, but the document is an advocacy statement rather than comparative outcome evidence.

The thesis therefore treats two rival mechanisms separately. Under the distributed-defense hypothesis, broad access increases vulnerability discovery, mitigation diversity, and resilience faster than it increases misuse. Under the distributed-capability-risk hypothesis, irreversible release, replication, and modification expand misuse and make provenance, containment, and coordinated remediation harder. Adjudication requires comparable incident rates, severity, time to detection and remediation, independent evaluation coverage, concentration-related outages, and operating burden normalized by deployment scale.

The strategic value of open-weight competition

Open-weight models can prevent a small number of labs from converting temporary capability leadership into permanent control of distribution and pricing. They can give enterprises a fallback, support regional and language-specific adaptation, expand application entry, and make smaller task-specific deployments viable outside the largest proprietary endpoints. The July 24 coalition statement argues that American leadership should be measured by economy-wide diffusion rather than ownership of one frontier model, and that organizations should match the right model to the right task instead of paying frontier-model prices for every workload (American Innovators Network et al. 2026).

The breadth of the coalition is itself informative. Its signatories span accelerator and hardware suppliers, cloud and enterprise software firms, model developers, cybersecurity companies, application builders, investors, and open-technology institutions. That demonstrates organized commercial support for an open-ecosystem strategy, but not neutral validation of its claimed outcomes: many participants benefit when more models and applications consume compute, cloud, security, and application services. The countervailing risk is fragmentation, inconsistent patching, and hidden technical debt. The strategic objective should not be maximum fragmentation; it should be contestability - enough interoperability and alternative capacity that no single provider can dictate access to essential intelligence.

Open ecosystem, concentrated substrate

The July 24 coalition includes chip, cloud, hardware, model, distribution, enterprise-software, cybersecurity and investment firms, including NVIDIA, Microsoft, Meta, IBM, Dell, Hugging Face, Mistral, ServiceNow, CrowdStrike, Palantir, Andreessen Horowitz and Y Combinator. Their support is evidence of organized strategy and policy preference. It is also consistent with a “commoditize the complement” incentive: cheaper and more plentiful model capability can increase demand for accelerators, cloud capacity, security, orchestration, enterprise integration and applications. That interpretation is an inference from signatory positions, not a proven coalition outcome (American Innovators Network et al. 2026).

Model-layer openness does not imply full-stack decentralization. Thousands of models and applications can coexist with a small number of accelerator designers, foundries, memory and packaging suppliers, hyperscalers, identity systems, managed control planes, and data-center operators. One plausible equilibrium is therefore an open application ecosystem operating on a concentrated physical and commercial substrate.

This shifts the competition question from “open or closed?” to “where do control and rents migrate?” The dashboard should compare production workload share, model-layer concentration and price premia, switching cost, fully loaded task cost, and the share of spending and gross margin captured by models, compute, cloud, control planes, data, and applications. Model-layer openness can discipline frontier labs while strengthening the bargaining position of infrastructure and distribution owners.

Distillation and model-derived training

The live dispute illustrates why technique and conduct must be separated. Anthropic reported that DeepSeek, Moonshot and MiniMax generated more than 16 million exchanges through approximately 24,000 fraudulent accounts, describing the conduct as illicit capability extraction while acknowledging that distillation itself is a legitimate and widely used method. These are Anthropic's allegations and evidence, not an adjudicated legal finding. July claims linking Moonshot's Kimi K3 specifically to Anthropic's Fable model remained contested at the cutoff and are not treated as established fact (Anthropic 2026f; Financial Times 2026a).

The policy question is therefore whether access was authorized, terms were violated, identity or geographic controls were evaded, trade secrets or protected expression were taken, and what remedy is proportionate. A rule that protects closed-model investment can also entrench incumbents or impair legitimate evaluation and follow-on innovation. H7 and exploratory proposition E3 track the economic and safety effects without presuming that either permissive or restrictive treatment is optimal.

The coalition statement asks policymakers to distinguish legitimate distillation and model improvement from unlawful extraction, and to prefer targeted legal and commercial remedies over sweeping restrictions on model-derived development (American Innovators Network et al. 2026). This is a contested policy position, not settled law or proof that all forms of output-based training are benign.

The conflict joins several legitimate interests: protecting frontier investment and trade secrets, preserving competition and follow-on innovation, respecting copyright and contract, documenting provenance, and preventing foreign or domestic appropriation that bypasses negotiated access. The thesis should therefore track license terms, output-use restrictions, provenance practices, extraction disputes, judicial or regulatory remedies, and whether rules preserve application innovation without eliminating incentives to fund frontier development.

What a sovereign-capable enterprise architecture looks like

Stage	Mechanism	Economic consequence
1	Classify	Segment workloads by sensitivity, latency, criticality, reversibility, and required capability.
2	Control	Centralize identity, policy, permissions, secrets, logging, evaluation, and cost limits.
3	Route	Select proprietary, open-weight, local, or specialized models based on policy and observed performance.

Stage	Mechanism	Economic consequence
4	Contain	Limit agent tools, transaction size, data scope, and blast radius; isolate critical systems.
5	Verify	Use continuous evaluations, red teaming, outcome measurement, and independent review.
6	Exit	Retain portable data, logs, prompts, evaluations, interfaces, and tested fallback procedures.

Figure 7.1. Sovereignty is implemented through an operating architecture, not declared through a procurement label.

A prospective sovereignty test

A deployment should be called sovereign-capable only if the organization has legal authority over the workload; can audit inputs, outputs, actions, and changes; can export the data, configuration, evaluations, and workflow context needed to migrate; can invoke an alternative within a defined recovery time; and can sustain critical operations for the required period if a supplier, jurisdiction, or network becomes unavailable.

This is a threshold for a specific workload, not a national marketing label. A locally hosted system that depends on one foreign software stack and cannot be patched or recovered may be less sovereign than a managed service with strong contractual continuity, audit, and exit. The dashboard should report which test failed rather than assigning a vague sovereignty score.

The talent and institution constraint

The original theory is right that hardware, data-center space, and talent are gates. The deepest constraint is institutional capability: the ability to turn technical control into safe, reliable operations. Many organizations can purchase accelerators but cannot maintain a model platform, build evaluations, secure agents, manage data rights, or recruit people who understand both the domain and the systems. Sector consortia, managed private-cloud services, shared evaluation institutions, and public research infrastructure may therefore be more important than isolated in-house model programs.

> **REFINED OPEN-WEIGHT AND SOVEREIGNTY THESIS**
> **Open weights are a coequal industrial architecture for diffusion, bargaining, and specialized deployment, not merely a hedge. They do not by themselves create sovereignty, lower cost, stronger security, or decentralization. Most organizations will buy managed control, and model-layer openness may shift rents toward chips, cloud, control planes, data, and applications. Success requires workload-level evidence on total cost, safety, portability, recovery, and where control and margin migrate.**

CHAPTER 8

The United States, China, and the multipolar sovereignty contest

The proposition that one country "leads AI" is too coarse, and a bilateral U.S.-China frame is incomplete. Models, chips, fabrication, packaging, power, capital, public compute, deployment, standards, security, and trusted access can produce different leaders and coalitions.

The original theory presents U.S. leadership in open source, commercial offerings, hardware, cost effectiveness, security, and trust as imperative to future growth. The imperative is reasonable; the descriptive claim is too broad. U.S. strengths coexist with Chinese capability convergence, European public-compute and regulatory strategy, Indian national infrastructure and data initiatives, Gulf sovereign-capital and full-stack programs, and deep dependence on allied fabrication, memory, packaging, and equipment.

A seven-dimension geopolitical baseline

A country cannot be assessed by one benchmark, one chip count or one spending total. JPMorgan Chase organizes U.S.-China competition across policy, hardware, models and software, energy and resources, finance, socioeconomics, and military and security. Version 1.2.1 uses the same disaggregated dimensions, then extends the comparison to Europe, India, Gulf states and allied manufacturing economies. The dimensions are not summed into a league-table score (JPMorgan Chase Center for Geopolitics 2026).

Dimension	U.S. baseline	China baseline	Multipolar implication
Policy	Export, investment, safety, procurement and industrial-policy tools.	State support, domestic substitution, standards and market access.	EU regulation/public compute, India mission procurement, Gulf capital and bilateral access shape third markets.
Hardware	Design, EDA, equipment, NVIDIA and allied fabrication strengths.	Huawei/Ascend substitution, manufacturing depth and restricted leading-edge access.	Taiwan, Korea, Japan and the Netherlands remain indispensable; Arizona and regional projects diversify selected stages.
Models/software	Frontier closed models, cloud and enterprise distribution.	Rapid capability convergence and strong open-weight diffusion.	European and Indian language/domain systems and Gulf-hosted capacity create bargaining options.
Energy/resources	Large installed capacity but higher power costs and interconnection friction.	Faster generation expansion and lower reported data-center-region power costs, with utilization concerns.	Gulf energy/capital and regional grids can attract workloads; physical delivery still depends on imported systems.
Finance	JPMorgan reports $490B U.S. AI spend and 75% of global AI VC in 2025.	Report estimates $98B spend and 5% of AI VC, supplemented by state support.	Sovereign capital and public compute can alter access without matching U.S. venture depth.
Socioeconomics	Private-sector diffusion, labor transition, community and ratepayer constraints.	Industrial diffusion, engineering scale, regional overbuild and utilization risk.	India emphasizes language and public access; EU emphasizes institutional capacity; Gulf states use infrastructure as diversification.

Dimension	U.S. baseline	China baseline	Multipolar implication
Military/security	Allied controls, cyber capability, defense procurement and trusted access.	Military-civil integration, cyber, domestic stack and security governance.	Partners choose interoperability, autonomy and access conditions rather than one binary bloc.

Table 8.1A. A common seven-dimension baseline without a composite national score.

Capability convergence

Stanford's 2026 AI Index reports two different convergence patterns. As of March 2026, the leading U.S. model was 2.7 percent ahead of the leading Chinese model on the report's Arena-based comparison, a gap that had fluctuated in the single digits. By contrast, the leading closed model was 3.3 percent ahead of the leading open model, up from 0.5 percent in August 2024, and six of the top ten Arena models were closed. The correct conclusion is therefore not that every gap is closing: U.S.-China capability had narrowed, while the top closed-open gap had reopened over that interval (Stanford HAI 2026a).

Layer	U.S. position	Chinese position	Strategic variable to watch
Frontier models	Large concentration of highly capable proprietary labs and global consumer distribution	Near-parity on many benchmarks; rapid iteration and strong open-weight releases	Reliability and cost on real workflows, not benchmark peaks
Open-weight ecosystem	Strong research, tooling, startups, and cloud distribution	Aggressive model publication, price competition, multilingual and industrial focus	Adoption outside home markets; license freedom; total operating cost
Chip design and platforms	Leadership in accelerator design, CUDA-class software, EDA and much semiconductor equipment; high NVIDIA platform concentration.	Huawei Ascend full-stack substitution, large domestic demand, and sustained ecosystem investment under controls.	Matched goodput per watt, memory and interconnect, software portability, developer adoption, and total system cost.
Fabrication, memory and packaging	Allied supply centered on TSMC/Taiwan, Korean HBM, Japanese materials, European tools, and expanding Arizona capacity.	Large mature-node base and domestic substitution; constrained access to parts of the leading-edge global tool and HBM stack.	Qualified yield and capacity, HBM and packaging lead times, share outside Taiwan, recovery time, and equipment access.
Cloud and capital	Deep equity and debt markets; global hyperscalers	Large platforms, state-guided finance, and domestic scale	Cost of capital, utilization, financing durability, access in third markets
Power and industrial build-out	Large resource base but slow interconnection and permitting in many regions	Rapid construction capability and integrated industrial planning, with regional constraints	Time to energized MW, reliability, fuel access, transmission expansion
Enterprise deployment	Global software and services distribution; strong multinational customer access	Manufacturing and domestic-platform integration; state and industrial use cases	Verified ROI, exportability, and ecosystem lock-in

Layer	U.S. position	Chinese position	Strategic variable to watch
Trust, security, and alliances	Allied relationships, standards influence, and enterprise-security brands	Cost and availability appeal; different governance model	Data governance, supply-chain assurance, political acceptance
Sovereign and regional capacity	Allied cloud, capital, research, and standards network	Large domestic platforms, state coordination, and infrastructure scale	EU public compute, IndiaAI access, Gulf full-stack build-out, and third-market adoption

Table 8.1. Leadership is a portfolio of positions across the stack.

The durable U.S. advantages

The United States combines frontier-lab density, advanced chip design, electronic-design automation, semiconductor equipment, global cloud platforms, venture and public capital, university research, and enterprise software distribution. The Semiconductor Industry Association describes continued U.S. leadership in chip design, EDA, and equipment, while public incentives are projected to restore a meaningful share of advanced logic fabrication by the early 2030s (SIA 2025). These advantages reinforce one another: capital funds compute, compute attracts talent, talent creates products, and distribution generates data and revenue.

The dollar financial system is another advantage. Frontier firms can raise private equity, public equity, syndicated credit, project finance, and infrastructure capital at a scale unavailable in most markets. Yet abundant capital is useful only when paired with disciplined allocation. A country can win the financing race and still earn poor returns if it builds low-utilization capacity or socializes project risk.

The durable Chinese advantages

China's advantages include manufacturing depth, a vast engineering workforce, large domestic demand, a strong publication and patent base, and the ability to integrate AI into industrial systems at scale. Stanford reports Chinese leadership in AI publication volume and citations as well as substantial patent activity, while the narrowing model gap demonstrates that export constraints have not ended capability development (Stanford HAI 2026b). Open-weight releases can also turn cost competition into geopolitical influence by making Chinese models attractive to countries and firms unable to afford premium proprietary services.

China may also benefit where infrastructure coordination and equipment manufacturing reduce the time from plan to energized capacity. This advantage should not be romanticized: local power constraints, capital misallocation, semiconductor bottlenecks, data controls, and trust concerns can slow deployment. But a U.S. strategy that assumes permanent model or open-weight superiority would be fragile.

Export controls: moat, delay, and stimulus

Advanced-semiconductor export controls are intended to restrict access to the compute required for military and frontier applications. The U.S. Department of Commerce continued to revise licensing treatment for high-end accelerators in 2026, illustrating that the control regime is dynamic rather than a one-time boundary (BIS 2024; BIS 2026). Controls can delay capability and raise cost. They also encourage stockpiling, domestic substitution, architectural efficiency, and alternative supply relationships.

The strategic measure is therefore not whether controls stop all progress. It is whether they preserve a meaningful advantage at acceptable cost to U.S. firms and allies while sustaining trust in the broader technology ecosystem. Overly broad restrictions can fragment markets and accelerate competing standards; overly narrow controls can transfer strategic compute with little friction. The policy problem is one of continuous calibration.

Cost effectiveness and security must be evaluated together

A low token price is not necessarily low economic cost. A system that requires more retries, human review, or integration may be more expensive per successful task. Likewise, a high-performing model that cannot meet security, residency, continuity, or audit requirements has limited value in critical sectors. The competitive frontier is secure total cost per successful task: the fully loaded cost of producing a verified result under the required controls.

> **NATIONAL COMPETITIVENESS METRIC**
>
> Secure total cost per successful useful task = compute + energy + integration + supervision + security + compliance + expected error loss, divided by verified outcomes. Add deployment speed and resilience as separate constraints.

Trust is an economic asset - and can be lost

The United States benefits when foreign enterprises believe that its technology is reliable, secure, contractually predictable, and governed by institutions that permit redress. Trust reduces the cost of adoption and expands the addressable market. It can be weakened by surveillance concerns, abrupt export restrictions, opaque model behavior, unreliable services, weak privacy protection, or the perception that strategic access can be withdrawn for political reasons. Chinese providers face their own trust constraints related to state access, governance, and geopolitical alignment.

Trust should not be treated as branding. It is produced by verifiable controls, transparent incident handling, independent testing, interoperability, legal commitments, supply-chain assurance, and continuity planning. A trusted ecosystem can command a premium even when a competing model is cheaper. A complacent ecosystem can lose that premium quickly.

Quantitative capacity and capital context

JPMorgan Chase reports installed data-center capacity of 53.7 GW in the United States, 31.9 GW in China, 11.9 GW in the European Union, 6.6 GW across Japan and Korea, 3.6 GW in India and 1.1 GW in the Middle East. Installed capacity is not equivalent to AI-ready, commissioned or utilized capacity. The report also cites 2025 AI spending of $490 billion by U.S. firms versus $98 billion by Chinese firms, U.S. receipt of 75 percent of global AI venture capital versus 5 percent for China, and electricity costs of roughly $0.07-$0.09/kWh in major Chinese data-center regions. Its statement that as many as 80 percent of Chinese data centers may be idle is an external estimate, not an official utilization census (JPMorgan Chase Center for Geopolitics 2026).

The same source supplies the counterweights. It reports a fourth-quarter 2025 national average of about $0.18/kWh in the United States versus $0.08/kWh in China, while noting wide U.S. regional variation: about $0.15-$0.22 in Silicon Valley and $0.06-$0.08 in Dallas-Fort Worth. It also reports DeepSeek's claim that V4 trails state-of-the-art frontier models by roughly three to six months. These figures complicate both simple cost-leadership and full-capability-parity claims, and they remain source-qualified estimates or company claims rather than matched independent measurements (JPMorgan Chase Center for Geopolitics 2026).

Region	Installed data-center capacity reported by JPMorgan	Strategic reading	Measurement caveat
United States	53.7 GW	Largest installed base, deep capital and cloud ecosystem.	Not all capacity is AI-ready, frontier-capable, energized for new loads or well utilized.
China	31.9 GW	Large base, lower reported power cost and rapid generation expansion.	Advanced accelerators and reported utilization constrain effective frontier capacity.
European Union	11.9 GW	Public-compute, regulatory and industrial-policy strategy.	Depends heavily on imported accelerators and allied manufacturing.
Japan and Korea	6.6 GW	Critical memory, materials, manufacturing and regional capacity.	Aggregate obscures different national roles and grid conditions.
India	3.6 GW	National compute access, language models and public procurement.	Mission capacity and installed commercial capacity must be separated.
Middle East	1.1 GW	Energy and sovereign capital support rapid planned build-out.	Announced gigawatts and bilateral projects are not yet equivalent to commissioned capacity.

Table 8.1B. Installed capacity is a baseline input, not a measure of productive frontier compute.

A multipolar sovereignty contest

Many countries will not replicate either the U.S. or Chinese stack, but they are not passive customers. Their strategies combine imported accelerators and models with domestic compute access, data authority, language assets, regulation, procurement and capital. These are active decoupling and bargaining strategies, not merely demand for U.S. exports.

European Union: The EU combines regulation with public and industrial compute. By April 2026 the Commission reported 19 AI Factories and 13 regional antennas, alongside the InvestAI facility and

an AI IPCEI candidate. This is not merely demand for U.S. exports: it is an attempt to preserve legal control, research access, industrial applications and bargaining power while remaining dependent on allied accelerators, fabrication and power (European Commission 2023; European Commission 2025; European Commission 2026a; European Commission 2026b).

Gulf states: Stargate UAE specifies a 1 GW Abu Dhabi cluster with 200 MW expected online in 2026 and reciprocal investment in U.S. infrastructure; Saudi Arabia's HUMAIN is building across data centers, cloud, models and applications. These states use energy, sovereign capital, procurement and cross-bloc partnerships to build independent leverage. Their success depends on commissioning, operator capability, hardware access and third-market services rather than announcements alone (OpenAI 2025b; PIF 2025).

India: The IndiaAI Mission operates a national compute portal with subsidized access to multiple accelerator types, AIKosh data and model infrastructure, and a domestic foundation-model program involving Sarvam AI. The strategy combines language capability, public procurement, shared compute and domestic deployment. It is an independent sovereignty model whose performance should be judged by sustained production workloads, secure cost, access and portability rather than GPU announcements alone (IndiaAI 2026a; IndiaAI 2026b; Sarvam AI 2026).

Allied manufacturing states: Taiwan, Korea, Japan, the Netherlands and other partners specialize in fabrication, HBM, advanced packaging, equipment, materials, servers and optics. TSMC remains the central leading-edge foundry and packaging node, while NVIDIA's own platform ramp depends on a dense Taiwan systems ecosystem. Arizona, Japan and European projects diversify selected stages but do not yet replicate Taiwan's complete cadence and supplier density. Chapter 4 provides the canonical hardware treatment. Supply-chain resilience and trusted access are sovereign variables in their own right, not background assumptions.

Smaller and nonaligned states: Regional infrastructure, open-weight models, procurement standards, data-residency rules and provider diversification can preserve bargaining power without full autarky. The relevant test is whether these arrangements support critical workloads through a supplier or geopolitical shock, not whether they reproduce a frontier stack end to end.

Two geopolitical stress cases

Chinese fabrication acceleration: If domestic advanced-node, memory, packaging, equipment, and software capabilities mature faster than expected, the U.S. chip-design and export-control advantage can narrow even without benchmark leadership changing immediately.

Allied supply disruption: If fabrication, memory, packaging, materials, or equipment flows from Taiwan, Korea, Japan, or Europe are interrupted, U.S. cloud capital and model distribution cannot substitute quickly for missing physical production. "U.S. leadership" is therefore partly an alliance and supply-chain property.

Two competing U.S. leadership strategies

A frontier-champion strategy concentrates capital, closed frontier capability, compute, and national-security controls in a small number of firms. Its advantages are coordinated scale, accountability at identifiable providers, and the ability to fund expensive research. Its risks are lock-in, concentrated failure, rent extraction, and industrial dependence on a few balance sheets and governance structures.

An open-ecosystem strategy expands compute access, shared datasets and evaluation assets, capable open weights, application-layer competition, and sector-specific deployment. The July 24 coalition statement presents this as leadership through diffusion into ordinary economic workflows rather than ownership of one leading model. Its advantages are specialization, contestability, and a broader developer base; its risks are irreversible capability release, operational fragmentation, and continued concentration at the chip, cloud, and control-plane layers (American Innovators Network et al. 2026).

The strategies are not mutually exclusive. A resilient U.S. portfolio can support frontier closed models and a capable open-weight ecosystem while applying different safeguards to different capabilities and deployment contexts. The adjudicating measures are secure task cost, independent adoption, switching power, safety performance, third-market diffusion, and the distribution of margins and control across layers.

The U.S. policy objective

The defensible objective is not to guarantee that every leading company is American or that every model is closed. It is to maintain an ecosystem in which frontier research, capable open weights, shared evaluation and compute assets, competitive applications, infrastructure, talent, finance, and trusted governance reinforce one another. Policy should distinguish legitimate model improvement from unlawful extraction through targeted, evidence-based rules rather than assuming that either openness or closure is inherently safe. Leadership should be measured by third-market adoption, secure task cost, infrastructure deployment, supply-chain resilience, switching power, and the ability of ordinary firms and public institutions to use AI productively - not by one benchmark or one champion.

CHAPTER 9

Productivity, labor, and the distribution problem

An economy can experience rapid firm-level productivity gains, disappointing aggregate growth, and severe distributional conflict at the same time.

The investment case for frontier AI ultimately depends on economic value outside the technology sector. If businesses cannot convert model capability into higher output, better quality, faster innovation, or lower risk, then infrastructure spending becomes difficult to sustain. Yet even successful productivity growth does not settle who benefits. The same system can raise output, compress some occupations, increase demand for others, and concentrate income among owners of capital and scarce skills.

What the empirical evidence already shows

A prominent field study of 5,179 customer-support agents found that access to a generative-AI assistant increased productivity by about 14 percent on average and by roughly 34 percent for novice and lower-skilled workers. The tool appeared to diffuse practices associated with higher performers (Brynjolfsson, Li, and Raymond 2023). This supports a complementarity perspective: AI can make less-experienced workers more capable rather than simply replace them.

A separate study spanning 66 firms and 7,137 knowledge workers found that users of generative AI spent about two fewer hours per week on email, but the researchers did not detect a significant change in the quantity or composition of tasks over the study window (Dillon et al. 2025). The result illustrates the adoption gap. Local time savings can be real while organizational structure, output, staffing, and financial performance change slowly.

The International Labour Organization's 2026 review concludes that measured productivity gains are real but uneven and that evidence of large aggregate employment displacement remains limited at this stage. It also emphasizes risks to job quality, early-career learning, inequality, and work organization (ILO 2026a). These findings are consistent with a transition in which tasks change before occupations disappear and in which institutions determine whether productivity becomes broad wage growth or concentrated profit.

Channel	Productivity-upside mechanism	Distributional risk	Institutional response
Skill diffusion	Novices receive guidance and approach expert practices	Expert tacit knowledge is extracted without sharing gains	Gain sharing, training, recognition, career pathways
Task automation	Routine cognitive work is completed faster or continuously	Entry-level work and apprenticeship tasks disappear	Redesign progression, supervised practice, credential alternatives

Channel	Productivity-upside mechanism	Distributional risk	Institutional response
Demand expansion	Lower cost creates new products, service levels, and customers	Work intensity rises and savings do not become compensation	Workload standards, bargaining, transparent performance measures
Quality improvement	Fewer errors, better consistency, faster access to expertise	Automated mistakes scale and accountability becomes diffuse	Human responsibility, audit, incident and appeal rights
Capital deepening	Workers gain powerful tools and firms expand output	Returns accrue mainly to model, data, and infrastructure owners	Competition policy, broad ownership, tax and social policy
Geographic reallocation	Digital capability reaches underserved regions	High-value work concentrates around compute and talent hubs	Infrastructure access, education, regional investment

Table 9.1. Productivity and distribution are separate questions.

The apprenticeship problem

Many occupations develop expertise through work that appears automatable: drafting, basic analysis, document review, first-line coding, routine customer interaction, and operational troubleshooting. If AI performs these tasks, firms may reduce entry-level hiring while still needing senior judgment later. The result can be a pipeline failure. Current experts remain productive, but fewer people accumulate the experience required to replace them.

A sustainable operating model must therefore distinguish low-value repetition from developmental practice. Organizations may need explicit apprenticeships, simulated cases, model-critique assignments, rotations, and supervised decision rights. Otherwise, the short-run efficiency gain can produce a long-run shortage of accountable experts.

The aggregation paradox

Firm-level gains do not automatically aggregate into measured macroeconomic gains. Time saved may be absorbed by additional meetings, higher output expectations, or more complex work. Firms may duplicate infrastructure and integration spending. AI may increase demand for verification, cybersecurity, and compliance. Productivity may appear first in quality or consumer surplus rather than revenue. And workers displaced from one activity may take time to move into another. The ILO describes an aggregation paradox in which widespread use can coexist with modest measured productivity if complementary organizational changes lag (ILO 2026c).

This matters for the frontier-lab valuation thesis. Revenue can grow rapidly because firms purchase AI as a strategic option even before aggregate productivity is visible. Public-market expectations can finance infrastructure in advance of measured returns. The system becomes vulnerable when financing commitments mature faster than organizations learn to redesign work.

Distribution feeds back into demand

If AI raises productivity while shifting income from labor toward capital, aggregate demand can weaken unless gains are reinvested, prices fall, new work expands, or policy redistributes purchasing power. Frontier platforms depend on consumer subscriptions, enterprise spending, advertising, commerce, and tax-supported government procurement. They cannot be analyzed separately from the demand base that purchases their services. A highly automated economy with concentrated income may produce strong corporate margins but weaker mass-market growth.

The macroeconomic outcome depends on elasticity. When AI sharply lowers the cost of a service, demand may expand enough to raise employment in complementary tasks. When demand is saturated or the service is internal overhead, labor savings are more likely to reduce staffing. When AI enables entirely new products, the employment effect can be positive but distributed across different skills and regions. Sector-level analysis is therefore more informative than one economy-wide displacement number.

Gender and occupational concentration

Exposure is not evenly distributed. The ILO reports higher generative-AI exposure in female-dominated occupations than in male-dominated occupations, reflecting the concentration of women in clerical and administrative work (ILO 2026b). Exposure does not equal job loss, but it identifies where task redesign, monitoring, and transition support are most urgent. A thesis focused only on national competitiveness can miss how gains and risks are allocated within the country.

What a serious productivity dashboard requires

Process output: Completed units, cycle time, quality, and downstream loss at the workflow level.

Labor incidence: Hiring, hours, wages, promotion, occupational entry, and demographic distribution.

Capital intensity: Infrastructure, integration, and software spending required to generate the gain.

Organizational change: Decision rights, span of control, team structure, training, and exception operations.

Consumer effects: Price, access, quality, product variety, and time saved outside paid work.

Aggregate effects: Productivity, labor share, investment, consumption, regional growth, and tax revenue.

The public bargain

Trillion-dollar frontier firms will face a legitimacy test. Their infrastructure can produce broad scientific, educational, medical, and productivity benefits. It can also draw power, water, land, tax incentives, and scarce capital while concentrating returns. The durability of the sector will depend on whether communities see credible benefits: reliable and fairly allocated grid investment, quality employment, lower prices or better services, tax contributions, security, and pathways for workers to share in productivity gains.

> **REFINED LABOR THESIS**
>
> The central risk is not a single unemployment number. It is a mismatch between the speed of task automation, the speed of organizational redesign, and the speed at which income, skills, and demand adjust. That mismatch can turn a productive technology into a source of political and financial instability.

CHAPTER 10

Competing explanations, external critiques, and adjudication tests

A useful thesis must expose rival mechanisms to tests it does not control. Listing a critique does not rebut it, and assigning every critic a "blind spot" can immunize the document against challenge. Each explanation below therefore states a mechanism, a distinctive prediction, and evidence that would disconfirm it.

The frontier-firm debate contains substantive disagreements and methodological challenges. Substantive explanations differ about cash flow, infrastructure, competition, labor, and security. Methodological challenges ask whether the thesis itself has valid constructs, representative cases, independent evidence, stable thresholds, and prospective tests. The latter cannot be absorbed as one more viewpoint within the market debate.

Competing-explanation map

Competing explanation	Mechanism	Unique prediction	Disconfirming evidence
Productive diffusion	Learning, lower task cost, complementary investment, and demand expansion spread benefits.	Independent enterprise ROI, utilization, productivity, and consumer value improve across many sectors.	Fully loaded returns and output per MW remain weak after broad deployment.
Bubble and capex reset	Valuations, circular demand, and infrastructure commitments outrun cash flow.	Price declines, impairments, cancellations, and restructurings precede broad realized ROI.	Sustained free cash flow, independent demand, and high utilization at disciplined capital intensity.
Hard financial transmission	Leverage, collateral, guarantees, and common intermediaries turn equity stress into core financial dysfunction.	Frontier-firm shocks widen funding and credit measures beyond matched equity controls.	Large drawdowns remain contained within portfolios and ordinary corporate finance.
Grid and environmental constraint	Physical lead times and local externalities bind faster than digital demand can be served.	Missed energization, latency, reliability, water, emissions, or rate impacts persist despite efficiency.	Supply, efficiency, flexibility, and community arrangements keep cost and service improving.
Frontier-platform integration	Compute, models, applications, and feedback create a compounding operating flywheel.	Voluntary owned-model use, retention, task economics, and direct customer control rise after integration.	Users leave, neutrality falls, rivals remain essential, and integration fails to improve returns.

Competing explanation	Mechanism	Unique prediction	Disconfirming evidence
Hyperscaler intermediation	Cloud, identity, data, procurement, and routers abstract models behind one managed control plane.	Enterprise AI spend and billing concentrate at hyperscalers while model suppliers become substitutable.	Labs retain direct contracts, data, margins, and workflow authority at scale.
Competition, founder, and conglomerate governance	Self-preferencing, dual-class control, related parties, cross-subsidy, and inseparability weaken discipline.	Model choice and separability decline while transfer pricing, governance risk, and integration opacity rise.	Independent governance, interoperability, transparent economics, and service separability remain robust.
Managed sovereignty	Customers buy portability, private deployment, and controlled routing rather than full autarky.	Exit tests and managed plurality spread after outages, policy shocks, or bargaining disputes.	Single-vendor concentration persists without meaningful customer concern or self-managed stacks dominate.
Open-ecosystem industrial policy	Capable open weights reduce model scarcity, enable specialization, and shift innovation and rents toward applications and infrastructure.	Open-weight production share, application formation, and sector diffusion rise while model price premia fall and adjacent-layer value capture increases.	Open weights remain marginal or costlier fully loaded, fail to improve switching or entry, and do not change concentration or rent distribution.
Labor and public interest	Work design and bargaining determine who captures productivity and who bears transition costs.	Firm gains coexist with weakened entry paths, labor share, job quality, or regional legitimacy.	Productivity diffuses into wages, prices, access, mobility, and broad-based employment.
Frontier AI safety	Misuse, malfunction, autonomy, and loss-of-control risks grow as systems gain access and permissions.	Capability and deployment expand faster than reliability, containment, evaluation, and incident controls.	Risk-relevant capabilities plateau or verifiable controls keep severe incidents and exposure bounded.
National security	Chips, compute, power, models, communications, and allies form a strategic industrial base.	States support continuity and redundancy even when private financial returns are weak.	Capabilities remain easily substitutable and state intervention does not track strategic dependence.

Table 10.1. Twelve competing explanations with distinct predictions and disconfirming evidence.

The productivity-bull perspective

Mechanism

The bull argues that artificial intelligence is not another application category but a general-purpose technology comparable to electrification, computing, and the internet. Model capability, hardware efficiency, and distribution improve together. Each decline in the cost of intelligence expands the set of economically viable tasks. Frontier firms require large capital bases because the market is correspondingly large. Index inclusion is a sign of successful diffusion, not a systemic defect. Energy demand is evidence that the technology is useful enough to justify a new industrial build-out.

From this perspective, vertical integration is productive. Agents reduce the transaction cost of converting capability into results. Frontier labs, software vendors, and service firms will discover

business models through competition. Infrastructure that appears excessive during construction may become the platform for later innovation, just as fiber and cloud capacity enabled unanticipated services. The greatest policy error would be to constrain financing, permitting, or experimentation before the benefits become measurable.

Distinctive prediction

The distinctive prediction is broad, independent diffusion: successful task cost falls; utilization and customer cash rise; process-level productivity appears outside subsidized pilots; and social value grows even if some early investors lose money.

Disconfirming evidence

This explanation is disconfirmed if fully loaded enterprise ROI, successful tasks per kWh, utilization, and revenue per energized megawatt fail to improve after broad deployment, or if gains remain confined to strategic financing and a narrow set of firms.

The bubble-skeptic perspective

Mechanism

The skeptic argues that a real technology can support a financial bubble. Capital providers extrapolate model improvements and user growth into revenues that may not cover compute, depreciation, integration, and customer acquisition. Strategic investors finance labs that purchase their infrastructure, making demand appear more independent than it is. Announced gigawatts, contracted chips, and private valuations reinforce one another before end customers demonstrate willingness to pay for completed work.

The skeptic expects price competition to erode model rents, open-weight systems to weaken scarcity, and enterprises to discover that integration and human supervision consume much of the apparent productivity gain. Public listings may transfer late-stage risk into indexes and retirement portfolios just as the capital cycle peaks. The resulting correction need not invalidate AI; it may simply reveal that infrastructure owners and model providers cannot all earn the margins embedded in their valuations.

Distinctive prediction

The distinctive prediction is that prices, depreciation, weak utilization, circular demand, impairments, cancellations, and refinancing stress appear before broad independent customer cash and process-level return.

Disconfirming evidence

This explanation is disconfirmed by sustained free-cash-flow growth, independently financed customer demand, high utilization, improving return on invested capital, and verified productivity across sectors despite falling model prices.

The financial-stability-regulator perspective

Mechanism

The regulator does not need to predict an equity collapse. The task is to identify plausible channels through which a shock would impair core functions. The Financial Stability Board has highlighted AI-related third-party concentration, market correlations, cyber risk, model risk, and dependence on a small number of providers (FSB 2024; FSB 2025). A regulator therefore asks which banks, insurers, funds, clearing members, utilities, and critical institutions rely on the same firms, clouds, models, or data centers.

From this view, the problem is opacity. Equity investors see public filings; regulators may not see the full network of capacity reservations, guarantees, derivatives, cloud commitments, project vehicles, and operational dependencies. The state should collect exposure data, stress-test common shocks, require continuity planning for critical AI services, and coordinate financial, energy, cyber, competition, and national-security oversight.

Distinctive prediction

The distinctive prediction is measurable spillover from a frontier-firm shock into funding, credit, clearing, margin, or material intermediary losses beyond matched market controls.

Disconfirming evidence

The explanation is disconfirmed if large equity and operational shocks remain contained, substitution works within required recovery times, and exposure maps show little leveraged or common-provider concentration.

The grid-planner and consumer-advocate perspective

Mechanism

The grid planner sees both opportunity and asymmetry. A large, creditworthy customer can justify generation and transmission that benefits a region. But the customer can also announce several possible sites, reserve capacity, delay construction, or leave after dedicated infrastructure is built. Data centers arrive on technology timelines while regulated assets are financed over decades. If contracts are weak, households and ordinary businesses can inherit the cost.

The preferred policy is milestone-based interconnection, transparent queue rules, customer funding of dedicated facilities, termination security, flexible-load incentives, and protection against speculative duplication. Power should be allocated through credible commitments and reliability value rather than through political enthusiasm for any particular industry.

Distinctive prediction

The distinctive prediction is persistent local scarcity: delayed energization, service-level misses, rising rate or reliability costs, water or emissions conflict, and weak transferability of dedicated assets.

Disconfirming evidence

The explanation is disconfirmed where supply, flexibility, efficiency, customer security, and community agreements keep rates, reliability, latency, and environmental burdens within pre-specified limits.

The frontier-platform-strategist perspective

Mechanism

The platform strategist believes that capital, compute, models, consumer distribution, developer adoption, enterprise context, applications, and feedback form a flywheel. Lower cost expands use; more use produces revenue and evaluation data; better products deepen distribution; deeper distribution justifies more infrastructure. Public capital can strengthen the flywheel by funding construction and acquisitions. Chapter 5 applies this mechanism once to the pending boundary case.

The strategist does not necessarily seek to replace all software or services. The objective is to own the orchestration layer: identity, memory, model selection, permissions, tools, evaluation, and billing. Partners can deliver domain work as long as the frontier platform remains the default runtime. In this view, the highest strategic risk is becoming an interchangeable model behind someone else's customer relationship.

Distinctive prediction

The distinctive prediction is that integrated firms gain voluntary user retention, direct billing, owned-model use, better task economics, and proprietary feedback while maintaining application quality and partner participation.

Disconfirming evidence

The explanation is disconfirmed by user departure, dependence on rival models, deteriorating task margins, weak cash conversion, high exception cost, impaired acquisitions, or customer migration to neutral and hyperscaler-managed alternatives.

The hyperscaler-intermediation perspective

Mechanism

The hyperscaler explanation argues that frontier labs do not necessarily own the most valuable layer. Cloud providers already control enterprise identity, data, networking, security, committed spend, marketplaces, systems of record, and procurement. Managed model routers can make several frontier and open models appear behind one endpoint, allowing the cloud to optimize quality, cost, latency, policy, and failover while retaining the master customer relationship.

The result can be model plurality with cloud concentration. A lab may remain scientifically important but commercially interchangeable, while the hyperscaler captures the control plane, billing, observability, and bargaining power. Labs that attempt to bypass the cloud can face hosting-margin changes, bundled alternatives, or restricted enterprise distribution.

Distinctive prediction

The distinctive prediction is rising enterprise spend through managed cloud routers and agent platforms, increasing substitution among underlying models, and a declining share of direct lab contracts, data ownership, and margin even as total inference grows.

Disconfirming evidence

This explanation is disconfirmed if frontier labs retain direct enterprise contracts, workflow data, billing, durable price premia, and customer authority at scale despite hyperscaler routing and bundling.

The competition and conglomerate-governance perspective

Mechanism

The competition and conglomerate-governance perspective argues that cross-layer ownership can convert technical efficiency into foreclosure. A firm controlling compute, a frontier model, and a widely used application can privilege its own model, bundle capacity, restrict interoperability, cross-subsidize entry, and collect workflow data unavailable to rivals. Chapter 5 contains the canonical pending-case test; this perspective supplies the general rival mechanism.

This perspective focuses less on headline purchase price than on post-merger conduct and separability: continued rival access, nondiscriminatory defaults and pricing, data rights, transparent intercompany economics, and the ability to separate critical services in restructuring. Those variables are operationalized in Chapter 5 and H3 rather than repeated here.

Distinctive prediction

The distinctive prediction is observable foreclosure or governance weakness: discriminatory defaults or pricing, reduced rival access, opaque transfer pricing, cross-subsidy, declining separability, or founder decisions that ordinary governance cannot correct.

Disconfirming evidence

The explanation is disconfirmed if model choice, interoperability, independent board oversight, user retention, segment economics, and service separability remain robust and performance improves without exclusionary conduct.

The enterprise-sovereignty perspective

Mechanism

The sovereignty advocate views intelligence as an enterprise dependency comparable to identity, data, payments, or cloud infrastructure. A firm should not let one provider determine which workflows operate, what data is retained, how performance is measured, or whether prices can be changed. Proprietary models can be used, but only behind a customer-controlled architecture with portable context, evaluation, and fallback.

This perspective favors multi-model routing, open-weight alternatives, private inference for sensitive workloads, contractual data rights, and tested exit. It interprets sovereign investment not

as an attempt to duplicate every frontier lab, but as insurance against concentration and geopolitical disruption.

Distinctive prediction

The distinctive prediction is that outages, policy changes, price disputes, or ownership changes lead enterprises to buy managed portability, private deployment, approved model subsets, and tested recovery rather than accept unqualified lock-in.

Disconfirming evidence

The explanation is disconfirmed if single-vendor systems remain reliably substitutable and economically dominant after real dependency shocks, or if managed and self-managed alternatives fail to reduce recovery time, legal exposure, or bargaining risk.

The open-ecosystem industrial-policy perspective

Mechanism

This perspective argues that advanced model capability should become a broadly usable input rather than remain a scarce rent controlled by a few frontier providers. Open weights lower the fixed cost of experimentation and adaptation, allow smaller or specialized models to serve routine work, and move competition toward applications, data, integration, and sector knowledge. A cross-layer commercial coalition can support this strategy because chip, cloud, security, software, and application firms may gain from more total deployment even when model-layer rents fall.

The policy program includes access to compute, shared datasets and evaluation tools, plural frontier capability, portable deployment, and targeted rather than sweeping treatment of distillation and model-derived training. Its evidentiary burden is to distinguish coalition advocacy from realized diffusion and to identify whether lower model concentration merely shifts dependency to the substrate.

Distinctive prediction

The distinctive prediction is that capable open-weight models gain production share across smaller firms, public institutions, and specialized workflows; model-layer price premia and concentration fall; application formation and task specialization increase; and a larger share of spending and gross margin moves toward compute, cloud, control planes, data, and applications.

Disconfirming evidence

The explanation is disconfirmed if open-weight systems remain economically marginal, unreliable, or more expensive after security and operations are included; fail to improve switching, application entry, or public-sector access; or leave both model-layer concentration and adjacent-layer value capture materially unchanged.

The labor and public-interest perspective

Mechanism

This perspective begins with incidence. Productivity gains are not socially neutral if workers lose entry paths, communities subsidize infrastructure, ratepayers absorb grid costs, and a small group of equity owners captures the return. Automation can intensify work, expand surveillance, and make performance management less contestable. Errors in benefits, hiring, credit, health, or public services can scale faster than appeals and accountability.

The public-interest response includes worker participation in system design, transparency about automated evaluation, appeal rights, investment in apprenticeship, gain sharing, community-benefit agreements, and competition policy. The objective is not to prevent automation, but to ensure that productivity becomes lower prices, better services, higher wages, or broadly shared ownership rather than only higher rents.

Distinctive prediction

The distinctive prediction is a divergence between firm-level gains and social incidence: reduced entry-level hiring or apprenticeship, weaker labor share, surveillance, regional cost shifting, or unequal access despite rising output.

Disconfirming evidence

The explanation is disconfirmed if productivity gains consistently reduce prices, improve access and job quality, raise wages or mobility, preserve expertise formation, and distribute benefits without extensive corrective policy.

The frontier AI safety and catastrophic-risk perspective

Mechanism

This explanation begins with misuse, malfunction, and the possibility of loss of control rather than valuation or market structure. As general-purpose systems gain autonomy, tool access, persistence, and permissions in critical environments, cyber, biological, fraud, reliability, and control failures can scale through the same concentrated infrastructure and workflows that create economic value. The 2026 International AI Safety Report finds growing real-world evidence for misuse and reliability failures while treating loss-of-control scenarios as uncertain but potentially severe (International AI Safety Report 2026).

This perspective changes the meaning of vertical integration. Owning compute, models, applications, and action channels can improve monitoring and accountability, but it can also concentrate correlated failure and give one system broader access to critical resources. Economic success can increase exposure faster than safeguards mature.

Distinctive prediction

The distinctive prediction is that autonomy, access, and deployment criticality grow faster than reliable evaluation, containment, monitoring, and recovery; severe near misses, restricted releases, safety cases, or mandatory controls become economically material.

Disconfirming evidence

This explanation is weakened if risk-relevant capabilities plateau, tool and permission controls remain effective under adversarial testing, severe incidents remain rare relative to deployment, and independently verifiable safety cases keep exposure bounded.

The national-security perspective

Mechanism

The national-security strategist sees an integrated technology-industrial base. Chips, models, data centers, power, launch, secure communications, talent, and allied adoption shape military readiness, intelligence, science, and economic influence. Frontier firms can become strategic assets before they become financially systemic; failure, cyber compromise, hostile acquisition, or supply interruption can have state-level consequences.

The response is resilient domestic and allied capacity, targeted export controls, continuity planning, secure procurement, supply-chain diversification, and protection of critical talent and infrastructure. Some redundancy and higher cost may be justified because resilience has option value in crisis.

Distinctive prediction

The distinctive prediction is that states finance redundancy, impose access controls, or protect continuity according to defined strategic capabilities even when market returns alone would not justify the action.

Disconfirming evidence

The explanation is disconfirmed if capabilities remain readily substitutable across allies and suppliers, continuity plans work without incumbent protection, and state intervention does not follow the claimed strategic dependency.

Methodological challenges not resolved by the perspective map

Construct validity: A six-input composite without weights, calibration, thresholds, or outcome mapping cannot function as a predictive index. Channel labels therefore require distinct qualification tests and cannot be aggregated to preserve the argument after an adverse result.

Case dependence: SpaceX is an extreme boundary case, and Cursor is an unconsummated transaction at the evidence cutoff. Neither can demonstrate a general pattern without post-close results and comparable cases.

Endogeneity and selection: High valuations, acquisitions, infrastructure, and product success can cause one another. Event studies, matched cases, pre-deal trends, and failed or neutral alternatives are required before assigning causality.

Self-disclosure circularity: A company's rationale shows intent, not validity. Strategic claims require customer, operating, regulatory, or audited corroboration.

Scenario elasticity: A scenario map is useful only if expected indicator signs are recorded in advance and the best-fitting scenario is reported with residuals. Multiple scenarios cannot be declared simultaneously correct whenever the evidence changes.

Prior and threshold discipline: Each research proposition is paired with a rival or null explanation and a decisive observation. Primary hypotheses specify units, horizons, metrics, nulls, and disconfirming outcomes; exploratory propositions are labeled separately.

Synthesis and adjudication rule

The competing explanations demand many of the same measurements but predict different signs, timing, and control points. Enumerating an explanation is not evidence for or against it. Each quarterly review must identify which primary hypothesis changed, compare the observation with its null, report the best-fitting rival explanation, and state which part of the thesis was weakened or rejected. A critique of the method cannot be dismissed as the "blind spot" of a substantive perspective.

CHAPTER 11

Four scenarios for 2030

Scenarios organize uncertainty without pretending to know a single outcome. The four below differ along two axes: the strength of realized AI productivity and the degree of platform concentration.

The future will likely contain elements of more than one scenario. Different sectors and regions can occupy different states. The purpose is to identify coherent combinations of capital markets, energy, enterprise adoption, labor, and geopolitics - and to specify the signs that one combination is becoming dominant.

Scenario typology, not a 2x2

The four scenarios are a coherent typology selected to expose distinct mechanisms, not the exhaustive cells of a two-axis matrix. Productivity realization, control-point concentration, geopolitical fragmentation, finance and regulation interact in more than two dimensions. The protocol therefore classifies each annual observation by the scenario whose distinctive signposts fit best and reports residual evidence rather than forcing every world into a nominal 2x2.

Scenario classification rule

For each annual review, compare realized indicators with the expected sign in Table 11.2, identify the scenario with the best overall fit, and publish the largest residuals. Do not declare all scenarios compatible with the evidence. A transition between scenarios is permissible only when the dated indicator record shows the change.

Scenario	Core configuration	Distinguishing mechanism	Failure mode or rival
A. Managed multi-model diffusion	High or rising productivity; customer control mediated by hyperscalers, neutral applications and managed plurality.	Task costs fall, production diffusion broadens and exit options remain credible.	Control silently concentrates in one cloud or integration burden prevents diffusion.
B. Capacity overbuild and capex reset	Productivity or monetization lags announced infrastructure; finance and construction retrench.	Utilization, cash conversion and independent demand fail to support the build-out.	Capacity is absorbed after a lag and becomes productive infrastructure.
C. Managed-digital-labor oligopoly	High productivity with concentrated workflow, application and outcome control.	A few platforms combine models, distribution, permissions, data and liability-bearing execution.	Neutral tools, hyperscalers, regulation or customer resistance keep the control point contestable.
D. Sovereign fragmentation	Productivity and concentration vary by bloc; states sponsor regional stacks and restrictions.	Trust, security, data authority and industrial policy outweigh global efficiency.	Interoperability and economics preserve a more integrated global market.

Table 11.1. Four internally coherent 2030 scenarios.

Scenario A: managed multi-model diffusion

Model capability, efficiency, and infrastructure improve, but the dominant enterprise control point sits with hyperscalers, systems of record, managed control-plane providers, and capable customers rather than one frontier lab. Models compete behind common endpoints, and capable open weights become a substantial production layer for specialized, cost-sensitive, or private workloads. Enterprises use private inference for sensitive tasks and managed frontier systems for bursty or difficult tasks.

Productivity diffuses because organizations can buy managed plurality rather than build it themselves. The market can be open at the model and application layers while remaining concentrated in chips, cloud, identity, and control planes. The scenario succeeds when secure task cost falls, open and proprietary systems remain substitutable, enterprises retain control rights, and gains spread through competition and organizational redesign rather than being absorbed entirely by the substrate.

Leading signs: Falling cost per successful task; high production conversion; rising model substitution behind managed endpoints; stable direct customer control over data and evaluation; successful exit tests; broad productivity; and cloud concentration that does not eliminate model or application competition.

Financial outcome: Large but differentiated winners; infrastructure resembles a competitive utility and technology stack rather than a single bubble.

Policy priority: Interoperability, skills, grid acceleration, competition, and targeted safeguards for high-risk uses.

Scenario B: capacity overbuild and capex reset

Enterprises continue purchasing pilots and capacity, but process redesign, data quality, reliability, and liability slow conversion to economic outcomes. Model prices fall faster than serving costs. Strategic financing and long-term commitments obscure weak independent demand. Data centers, generation, and semiconductor capacity arrive as revenue growth decelerates. Public listings broaden exposure near the top of the cycle.

The correction begins as a valuation event and migrates into project finance, suppliers, utilities, and regional development plans. Some firms fail or consolidate; equity holders absorb large losses; government intervention focuses on critical services and infrastructure rather than restoring old valuations. The technology survives, and distressed capacity later lowers prices, but the capital cycle imposes a recessionary drag in exposed regions and portfolios.

Leading signs: Falling utilization, repeated project delays, rising cancellations, weak cash conversion, increasing vendor financing, enterprise renewals below pilot growth, and negative free cash flow despite scale.

Financial outcome: Technology-bust dynamics with pockets of credit and utility stress; systemic severity depends on leverage and guarantees.

Policy priority: Exposure transparency, ratepayer protection, orderly resolution, continuity plans, and avoiding indiscriminate bailouts.

Scenario C: managed-digital-labor oligopoly

A small number of frontier platforms combine capital, compute, superior models, consumer distribution, enterprise applications, agents, identity, memory, payment, and transaction systems. They use high valuations to acquire application-layer control points and then route demand toward owned models and infrastructure. Customers increasingly purchase outcomes rather than software seats or labor hours. Smaller providers become complements, resellers, or acquisition targets. Enterprise systems of record remain, but the agent runtime and application interface become the primary economic control points.

Productivity is high, but rents and dependency are concentrated. A platform outage or policy change can interrupt many sectors. Workers experience rapid task displacement, and the platforms capture a large share of the surplus. Governments treat continuity as strategically essential, creating a form of soft systemic protection. Competition policy struggles because efficiency and concentration arise from the same integrated system.

- **Leading signs: Rising outcome-based revenue; repeated stock-financed acquisitions of application distribution; increasing owned-model share inside acquired products; declining direct use of enterprise software interfaces; reduced model neutrality; a rising share of critical workflows on one platform; and high margins despite base-model price competition.**
- **Financial outcome:** Sustained mega-cap concentration and strong cash flows, accompanied by operational and political systemic importance.
- **Policy priority:** Portability, non-discrimination, critical-service continuity, data rights, liability, and limits on self-preferencing.

Scenario D: sovereign fragmentation

Export controls, cyber incidents, data-localization rules, industrial policy, and political distrust fragment the global stack. The United States, China, Europe, India, Gulf states, and other regions build or sponsor distinct compute and model ecosystems. Enterprises maintain separate deployments to serve different jurisdictions. Open-weight models spread, but licenses, standards, hardware, and safety requirements diverge.

Redundancy improves some forms of resilience while raising cost and reducing scale economies. Countries with cheap power, capital, and trusted political relationships attract infrastructure. Smaller countries depend on regional alliances or shared platforms. Global technology firms face duplicated compliance and capital expenditure. Productivity gains continue but diffuse unevenly, and national-security logic increasingly shapes commercial architecture.

- **Leading signs:** Regional model mandates, incompatible standards, localized training and inference, restricted chip and model flows, duplicated corporate stacks, and politically conditioned cloud access.
- **Financial outcome:** Lower global margins but protected regional champions; infrastructure is financed partly for strategic redundancy rather than utilization.
- **Policy priority:** Allied interoperability, minimum cross-border standards, supply-chain resilience, and disciplined strategic subsidies.

Scenario signposts

Indicator	Managed diffusion	Capex reset	Managed-labor oligopoly	Sovereign fragmentation
Model/API prices	Fall with equal or faster cost improvement	Fall faster than cost and revenue	Fall at base layer; platform captures outcome rent	Diverge by region and access
Enterprise ROI	Broad, verified, and repeatable	Pilot-heavy and inconsistent	High on platform-owned workflows	Mixed; duplicated compliance costs
Revenue per MW	Rises steadily with latency compliance	Falls or remains weak	Rises sharply for leaders	Varies with subsidy and local protection
Market concentration	Chip/cloud remain high while model/application layers are contested	Falls after correction	Rises and persists across layers	High within regional blocs
Open-weight role	Major routed and specialized production layer; lower model rents with concentrated substrate	Accelerates price pressure and asset repricing	Commoditized input, niche alternative, or absorbed by integrated platforms	Strategic national and regional infrastructure
Labor outcome	Task change with demand expansion and managed transition	Investment slowdown and transition stress	High productivity with concentrated displacement	Uneven by bloc and industrial policy
Policy posture	Interoperability, portability, competition, and managed infrastructure	Resolve projects and protect counterparties	Regulate critical platforms and self-preferencing	Subsidize, restrict, and align
Vertical integration	Selective deals coexist with neutrality and hyperscaler routing	Deals impair or unwind	Repeated acquisitions consolidate applications and data	Regional champions acquire or sponsor domestic stacks
Customer control point	Hyperscaler or enterprise-managed plane	Fragmented after retrenchment	Frontier platform runtime	Regional sovereign or state-aligned plane

Table 11.2. Pre-specified indicator signs for adjudicating the four scenarios.

What would make the worst outcomes systemic

The capex-reset scenario becomes a financial crisis only if leverage, guarantees, collateral, funding structures, and intermediary exposure transmit losses into core financial functions. The workflow-oligopoly scenario becomes an operational crisis only if tested substitutes and resolution cannot meet required recovery times. Sovereign fragmentation becomes a macroeconomic crisis if duplication and restrictions materially reduce productivity and investment. Scenario severity is adjudicated by channel-specific outcomes, not by an aggregate systemic score.

CHAPTER 12

Pre-specified hypotheses, nulls, and the monitoring protocol

The thesis becomes research only when the unit, horizon, metric, null, and disconfirming observation are specified before the outcome. Broad propositions remain exploratory until they meet that standard.

Seven primary hypotheses are designed for review through 2030. A primary hypothesis can be supported, unchanged, weakened, or rejected; it cannot be rescued by switching to another channel or scenario. The event definitions below identify observations that trigger measurement, not outcomes that automatically confirm the claim.

Executable hypotheses and adjudication states

Protocol definitions remain fixed through the first quarterly cycle. Primary frontier firms are SpaceX/xAI, OpenAI, Anthropic, and future firms whose principal strategic identity is a frontier model, frontier application, or integrated frontier platform. Material AI-system issuers are public companies whose index weight and direct control of a critical AI-stack layer can transmit AI-related shocks; NVIDIA qualifies at the baseline. H1A uses the material AI-system issuer universe, not the narrower primary-firm label. A material index weight is at least 1.0 percent in the S&P 500 or Nasdaq-100. A major infrastructure project is at least 1 GW or $10 billion of committed capital. Comparator membership, thresholds, event IDs and endpoints are frozen before each trigger.

Quarterly reviews are observation updates, not repeated endpoint tests. Temporal gating occurs first: no qualifying event means not triggered; a qualifying event whose observation window remains open is reported as window open. At the endpoint, verdict precedence is mandatory-evidence failure -> not adjudicable; simultaneous support and null satisfaction -> mixed; support only -> support; null only -> null; neither -> mixed/indeterminate. Missing treated or comparator data never create support. Formal statistical inference, if later added, must pre-specify the test family and repeated-look control.

State	Meaning	Permitted conclusion
1. Not triggered	No qualifying event has occurred.	Baseline monitoring only.
2. Triggered; window open	A qualifying event occurred, but the formal observation horizon has not ended.	Interim facts may be reported; no support or rejection verdict.
3. Support threshold met	All mandatory support conditions satisfy the fixed Boolean rule at the endpoint.	Support for that hypothesis only; no inference to other channels.
4. Null threshold met	The affirmative rival/null conditions satisfy their fixed rule.	The tested mechanism is weakened or rejected for that event.

State	Meaning	Permitted conclusion
5. Mixed or indeterminate	At the endpoint, support and null both hold, conflict, or neither complete rule is met.	Retain uncertainty. Do not score partial movement as support or null.
6. Not adjudicable	Mandatory treated or comparator evidence is unavailable, invalid, or outside the frozen unit.	No verdict. Record the missing evidence and next feasible date.

Table 12.1. Six-state adjudication machine. Temporal gating occurs before endpoint precedence; missing evidence outranks all substantive verdicts, and overlapping support/null evidence is mixed.

Decision precedence and event-to-hypothesis aggregation

Every qualifying event receives its own immutable event ID, trigger date, comparator, evidence state, endpoint and six-state verdict. Event results are never averaged into a systemic score. The rules below determine when several events are enough to say something about a general hypothesis rather than one case.

Level	Frozen rule	Permitted conclusion
Single event	Apply the hypothesis Boolean rule once at the fixed endpoint. Report support, null, mixed/indeterminate, or not adjudicable.	A result about that event only.
Canonical H3A case	SpaceX-Cursor remains a permanent record: completed-success, completed-failure, blocked, abandoned, delayed, mixed, or not adjudicable.	The case cannot be erased by substituting a later deal.
Replicated event hypotheses	For H1A, H1B, H2, H3B and H4: require >=3 adjudicable events from >=2 firms, projects or regions. Support requires >=60% support and <=20% null; null requires the mirror image. Otherwise mixed.	General mechanism support or null only after minimum replication.
Panel hypotheses	H5-H7 use one pre-registered panel endpoint. Annual snapshots are window-open observations. A later independent panel is a replication, not pooled evidence.	A panel verdict at the specified endpoint.
Missing and mixed evidence	Report all triggered events. If >50% are not adjudicable, no hypothesis-level verdict is issued. Mixed events remain in the ledger but outside support/null shares.	Evidence insufficiency cannot be hidden by the available subset.

Table 12.1A. Event verdicts and hypothesis-level conclusions are separate layers of the protocol.

Hypothesis	Trigger, unit, horizon and data	Support rule	Null, mixed and missing-data rule
H1A issuer shock	Material AI-system issuer falls >=20% over <=20 trading days without a qualifying common-substrate event; +60 trading days. Market, TRACE/CDS, fund, clearing and regulatory data.	Credit spread widens >=75 bp versus synthetic control AND either: short-term funding spread widens >=50 bp or collateral haircut rises >=5 pp; OR forced sales reach >=0.5% of free float in 5 days or clearing margin rises >=25%.	Null: credit <25 bp, funding <20 bp, haircut <2 pp, forced sales <0.1% and clearing margin <10%. Both support and null -> mixed. Not adjudicable if exposed-intermediary mapping is unavailable.
H1B common substrate	Foundry, HBM, packaging, interconnect, platform-security or power shock affecting >=3 material AI-system issuers; +120 trading days.	At least 3 issuers show >=10% production/capacity/revenue-guidance impact or >24-hour critical service loss AND relevant index falls >=15% AND credit/funding widens >=50 bp versus unaffected controls.	Null: operating shock occurs, but credit/funding stays <25 bp and no core-market disruption. Missing issuer exposure map -> not adjudicable.

Hypothesis	Trigger, unit, horizon and data	Support rule	Null, mixed and missing-data rule
H2 infrastructure finance	Cancellation, delay or utilization failure for project >=1 GW or $10B; 24 months. Project, utility, lender, municipal and supplier data.	Lender/project loss >=5% of committed principal OR both utility/ratepayer cost >=2% and exposed supplier revenue loss >=10%.	Null: >=80% of failed nameplate MW or committed capital is re-contracted within 24 months at >=80% of original contracted economics, while creditor/rate effects stay below half the support thresholds. Not adjudicable without financing and replacement-customer maps.
H3 integration	Completed stock-financed application acquisition >=1% of acquirer market cap; 12 and 24 months. Cohort, product, accounting and market data.	Retention no more than 5 pp below neutral comparator AND neutrality deterioration <10 pp AND impairment <10% of purchase consideration AND either 24-month ROIC >= WACC or verified operating improvement >=10%.	Null: ROIC <= WACC-5 pp OR two of retention loss >15 pp, neutrality deterioration >=20 pp, impairment >=10%, or operating underperformance. Blocked/abandoned H3A remains recorded; H3B may use later qualifying deals.
H4 power-latency	Matched sites/workloads over four quarters after >=25% increase in service-level-compliant task demand or agentic depth. Demand and supply are accelerator-equivalent compute-hours at a frozen workload mix.	Required compute-hour demand growth exceeds efficiency improvement by >=15 pp AND P95 latency or SLA miss rate worsens >=20% AND constraint appears at >=2 sites.	Null: compute-hours per verified task improve >=20%, P95/SLA worsens <=5%, and commissioned SLA-capable compute-hours meet >=90% of demand growth. Not adjudicable without workload and site telemetry.
H5 hyperscaler intermediation	Stratified panel of >=30 enterprise contracts across >=3 sectors; no provider >40% of sample; annual endpoint through 2030.	At least 2 of 3: hyperscaler receives >=60% of total AI platform/model/cloud spend; owns master contract/control plane in >=60% of deployments; model substitution occurs without customer contract change in >=40%.	Null: at least 2 of 3 remain below 35%. Mixed between thresholds. Not adjudicable if recruitment, spend denominator or contract access fails.
H6 managed sovereignty	Dependency shock or exit test in >=20 pre-selected critical workflows with affected/unaffected matches; 12 months.	Fallback meets frozen output-quality and security tolerance in >=80% of workflows, median RTO improves >=30%, and fully loaded TCO premium <=15% versus single-provider control.	Null: fallback success <40% and RTO improves <10%; TCO remains reported separately. Mixed otherwise. Not adjudicable without tested failover, RTO and cost data.
H7 open-weight rents	Matched production-workload and layer-spend panel through 2030. Production share is compute-weighted verified tasks; secure task cost uses the same workload and service level.	At least 3 of 4: open-weight production share >=30%; secure task cost <=90% of closed comparator; closed-model price premium falls >=25%; adjacent-layer share of AI spend/gross profit rises >=10 pp.	Null: production share <10%, closed premium decline <5%, adjacent-layer shift <3 pp, and secure task cost remains >110% of closed comparator. Mixed otherwise; not adjudicable without matched workload and layer-spend data.

Table 12.2. Executable H1A, H1B and H2-H7 rules. Thresholds are frozen protocol choices, not natural constants.

Frozen measurement units and denominators

The terms below close the remaining dimensional gaps. A future review may revise a threshold prospectively, but it may not change the unit or denominator after an event is known.

H	Frozen unit or denominator	Completed variable definition	Mandatory evidence
H1A	Issuer and exposed-intermediary spread change versus synthetic control	Funding breach, collateral haircut, forced-sale share of free float, and clearing-margin change each have numeric thresholds.	Issuer debt/CDS, dealer/fund flows, collateral and clearing data.
H1B	Common event across >=3 material AI-system issuers	Operating impact means >=10% capacity/production/revenue-guidance change or >24-hour critical service loss.	Supplier exposure, issuer operations, index and credit controls.
H2	Failed MW or committed capital and replacement economics	Redeployment is >=80% re-contracted at >=80% of original economics within 24 months.	Original and replacement contracts, creditor, utility and supplier maps.
H3	Deal-date value of issued shares and post-close operating cohorts	Material impairment is >=10% of consideration; ROIC includes issued equity at deal-date value; neutrality	Cohort retention, model mix, accounting and integration cost.

H	Frozen unit or denominator	Completed variable definition	Mandatory evidence
		is measured by availability, defaults, price and overrides.	
H4	Accelerator-equivalent compute-hours at frozen workload mix and SLA	Demand, efficiency and commissioned supply share one unit; latency is P95 and SLA miss rate.	Site telemetry, workload depth, compute-hours and commissioned capacity.
H5	Total contracted AI platform/model/cloud spend in a stratified contract panel	Provider, sector and spend denominators are frozen before recruitment; model substitution is measured without customer contract change.	Contracts, invoices, control-plane ownership and routing changes.
H6	Twenty pre-selected critical workflows	Fallback success requires output, security, RTO and recovery-point tolerance; TCO includes integration and standby cost.	Executed exit tests, logs, costs and matched controls.
H7	Compute-weighted verified tasks and matched fully loaded task cost	Open-weight share, closed premium and adjacent-layer rent shift use one frozen workload/spend panel.	Production logs, prices, security cost, human review and layer economics.

Table 12.2A. H1-H7 measurement units, denominators and required evidence.

Hypothesis-specific scenario generator

To expose how support-side assumptions interact before sufficient data exist, v1.2.1 retains a simplified Monte Carlo support-pattern generator modeled on the transparency discipline of the When We Crash working paper, not on its parameters or reported probabilities. For each hypothesis, it samples the disclosed component probabilities, generates correlated Bernoulli outcomes, and applies the fixed support Boolean pattern. The generator is conditional on a qualifying event and complete measurement. It does not estimate event incidence, null probability, public observability, or the chance of a not-adjudicable verdict. Those belong to the real-world event ledger (When We Crash 2026).

Frozen implementation: random seed 20260724; 200,000 base trials per hypothesis; base within-hypothesis equicorrelation 0.35; 400 sensitivity batches of 4,000 trials; correlation varied from 0.10 to 0.60; component probabilities sampled from the low/mode/high ranges in Table 12.3. The simulation evaluates support patterns only, so it contains no support-versus-null precedence rule and cannot output "not adjudicable." Real-world endpoint precedence is governed by Table 12.1.

H	Component priors: low / mode / high	Support Boolean pattern
H1A	Credit 0.10/0.25/0.45; funding/collateral 0.05/0.15/0.35; flow 0.15/0.35/0.60.	Credit AND (funding OR flow).
H1B	Multi-issuer operating 0.35/0.60/0.82; index 0.25/0.50/0.75; credit 0.08/0.22/0.45.	Operating AND index AND credit.
H2	Lender 0.20/0.40/0.65; utility 0.15/0.30/0.55; supplier 0.20/0.45/0.70.	Lender OR (utility AND supplier).
H3	Retention 0.45/0.65/0.82; neutrality 0.40/0.65/0.85; ROIC 0.25/0.45/0.65; operating 0.30/0.55/0.75; no impairment 0.70/0.85/0.95.	Retention AND neutrality AND no impairment AND (ROIC OR operating).
H4	Demand>efficiency 0.45/0.65/0.85; latency/SLO 0.20/0.40/0.65; multi-site 0.25/0.50/0.75.	All three.

H	Component priors: low / mode / high	Support Boolean pattern
H5	Spend 0.45/0.65/0.80; contract 0.40/0.60/0.78; channel/margin 0.35/0.55/0.75.	At least two of three.
H6	Portability 0.25/0.45/0.65; RTO 0.20/0.40/0.60; TCO 0.30/0.55/0.75.	All three.
H7	Share 0.30/0.50/0.70; cost 0.30/0.55/0.75; premium 0.40/0.60/0.80; rent shift 0.35/0.55/0.75.	At least three of four.

Table 12.3. Frozen component-prior ranges and Boolean support patterns for the illustrative scenario generator.

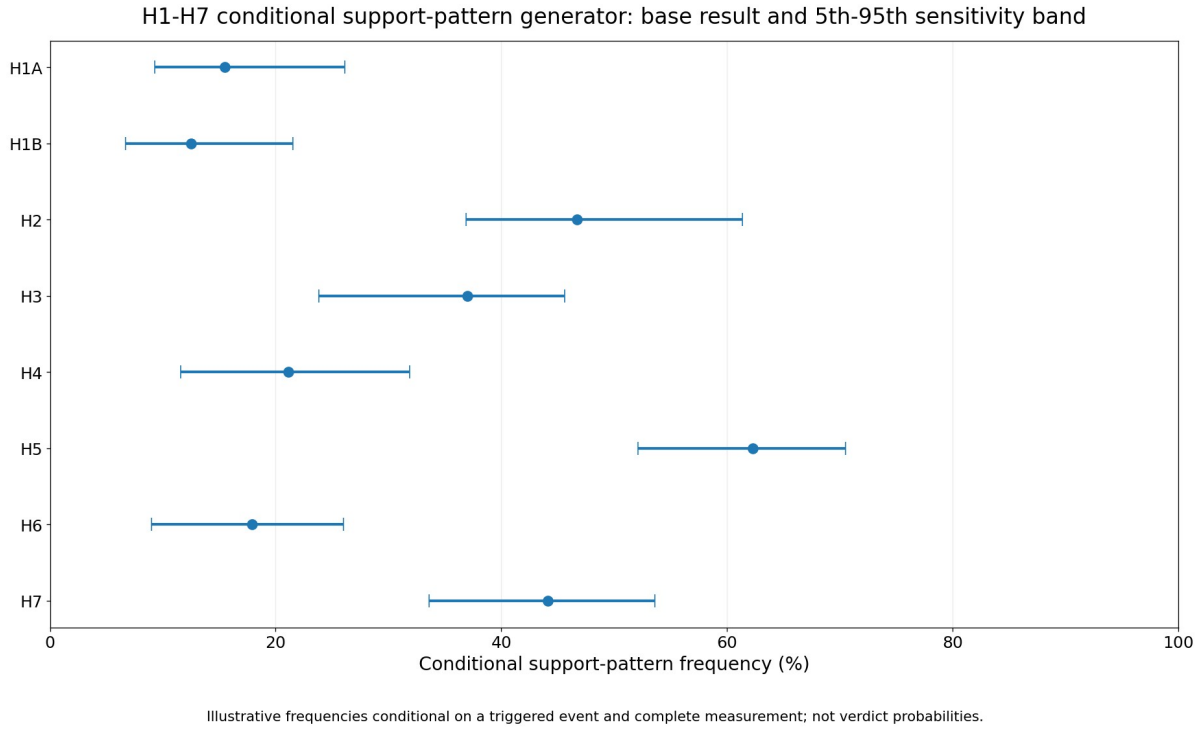

Figure 12.1. Conditional support-pattern frequency and 5th-95th sensitivity band. These are assumption-driven scenario outputs, not calibrated probabilities or a cross-hypothesis score.

H	Base conditional support	5th-95th sensitivity band	Null / missing simulated?	Interpretation
H1A	15.5%	9.3%-26.1%	No	Conditional on trigger and complete measurement.
H1B	12.5%	6.7%-21.5%	No	Conditional on trigger and complete measurement.
H2	46.7%	36.9%-61.4%	No	Conditional on trigger and complete measurement.
H3	37.0%	23.8%-45.6%	No	Conditional on trigger and complete measurement.
H4	21.1%	11.6%-31.9%	No	Conditional on trigger and complete measurement.
H5	62.3%	52.1%-70.5%	No	Conditional on trigger and complete measurement.
H6	17.9%	9.0%-26.0%	No	Conditional on trigger and complete measurement.
H7	44.1%	33.6%-53.6%	No	Conditional on trigger and complete measurement.

Table 12.4. Conditional support-pattern outputs. These frequencies assume a triggered event and complete measurement; they are not verdict probabilities.

Historical applicability and rule-coverage audit

The historical exercise is not a backtest. It does not compute detection rates, lead time, false positives, precision or sensitivity. It asks only whether selected past episodes expose useful channel analogs and false comparisons for the current rules. The episode set is selected rather than exhaustive, so no predictive-performance claim is permitted. A genuine backtest would require a frozen episode universe, negative windows, dated proxies, complete decision rules and scored outcomes before results are inspected (When We Crash 2026).

Episode	Channel analog	Protocol lesson
LTCM 1998	Funding, leverage, counterparty and liquidity transmission.	A credit/funding rule should detect intermediary stress even when the originating asset is not a material AI-system issuer.
Dot-com 2000-02	Equity concentration, telecom/fiber overbuild and capex reset.	Technology value can survive while securities and projects fail; H1A and H2 must remain separate.
Global financial crisis 2008	Runnable funding, collateral and core intermediary failure.	Benchmark for hard financial-system transmission; an equity loss alone should not satisfy the rule.
Europe 2011	Sovereign-bank credit loop.	Common macro shock requires controls; do not mislabel it as one issuer's effect.
Q4 2018	Fast equity correction with limited systemic spillover.	Useful null analog for H1A.
COVID-19 2020	Exogenous common operational and market shock.	Tests common-shock exclusion and the need to separate source from transmission.
2022 cycle	Rates, valuation compression and financing repricing.	Tests whether spreads and project stress exceed broad rate controls.
August 2024	Short, fast market shock.	Tests whether windows and duration filters avoid treating every drawdown as systemic.

Table 12.5. Selected historical analogs provide rule coverage and false-analogy checks; they are not a backtest.

Backtest readiness by hypothesis

v1.2.1 freezes what can and cannot be tested historically. The first true backtest should be published as a separate workpaper rather than inferred from this narrative audit.

Hypothesis	Historical backtest status	Reason
H1A / H1B	Partially feasible; not performed here	Market and credit proxies exist, but issuer and substrate exposure maps require a frozen sample and event dates.
H2	Partially feasible; not performed here	Telecom, energy and data-center overbuild analogs exist; project guarantees and redeployment economics are often private.
H3	Not comparable enough	Modern stock-financed frontier-model/application integrations lack a

Hypothesis	Historical backtest status	Reason
		stable historical peer set.
H4	Not currently feasible	No consistent historical panel links agentic workload depth, compute-hours, power, latency and SLA outcomes.
H5-H7	Not currently feasible	Comparable enterprise contracts, exit tests, production model share and layer-rent panels do not exist historically.

Table 12.5A. Backtest readiness is reported separately from historical narrative coverage.

The leading indicators

Leading indicators change before financial statements reveal the outcome. They include monthly concentration ranges, lockup and vesting supply, interconnection milestones, project finance, latency and sequential task depth, independent customer cash, production conversion, cloud-router share, stock-financed acquisition status, model neutrality, user retention, safety incidents, and environmental commitments. Announcements, confidential filings, signed deals, and pilot counts are states to monitor - not realized demand or confirmed mechanisms.

Domain	Leading indicators	Interpretation
Capital markets	Float, lockup expiry, index eligibility, option skew, securities lending, follow-on issuance	Tests whether ownership and leverage are broadening faster than earnings support
Infrastructure	Interconnection milestones, transformer/turbine orders, customer deposits, construction progress, power contracts	Separates credible capacity from duplicated announcements
Hardware substrate	Productive fleet by accelerator and generation; HBM and advanced-packaging lead times; porting cost across CUDA, Neuron and CANN; commissioned racks; qualified capacity by geography	Tests whether model competition reduces dependence or merely relocates concentration to chips, software, fabrication, packaging and facility design
Commercial	Production deployments, renewal, workflow volume, outcome contracts, independent customer cash	Distinguishes adoption and willingness to pay from pilots and strategic financing
Technical	Cost per successful task, reliability, human escalation, context length used, model routing	Shows whether capability becomes economical work
Enterprise architecture	Multi-model control planes, private inference, exit tests, retained logs and evaluations	Measures bargaining power and operational substitutability
Open ecosystem	Open-weight production share, matched secure task cost, deployment time, model-layer price and concentration, license rights, and margin by adjacent layer	Tests whether openness broadens diffusion and switching or merely shifts value and dependency to the substrate

Domain	Leading indicators	Interpretation
Labor	Entry-level hiring, job redesign, training hours, wage distribution, AI-related performance monitoring	Reveals transition before aggregate employment changes
Geopolitics	Export licenses, domestic chip substitution, regional model mandates, allied compute projects	Shows whether the market is integrating or fragmenting
Vertical integration	Stock-financed deal value, implied dilution, application ownership, default model share, interoperability, and user retention	Tests whether market value is becoming industrial power and whether acquisitions create operating value rather than only narrative scale

Table 12.6. Leading indicators for quarterly monitoring.

The coincident indicators

Coincident indicators describe the operating state: energized megawatts, utilization, successful tasks per kWh, P95/P99 latency compliance, gross margin after depreciation and power, workflow volume, agent success and rollback, model mix, retention, direct versus intermediary billing, exit-test performance, safety events, water and emissions intensity, and realized labor productivity. They must be segmented by workload and control layer.

The lagging indicators

Lagging indicators reveal whether the system-level thesis materialized: free cash flow, return on invested capital, defaults or restructurings, utility rate effects, pension and household wealth losses, aggregate productivity, labor share, regional employment, and the concentration of critical workflows. By the time these appear, strategic options may be narrower. Their value is calibration: they show which leading indicators were genuinely predictive.

Exploratory propositions and non-scored channel classification

The thesis no longer scores firms on a low-to-high composite systemic scale. Analysts classify evidence separately by hard financial transmission, macro-financial amplification, operational criticality, strategic importance, and resolvability. Broader propositions that lack thresholds or comparative data remain exploratory and carry less epistemic weight than the seven primary hypotheses.

Exploratory proposition	Why it is not yet primary	Evidence needed	Downgrade or upgrade rule
E1. U.S.-China leadership remains layer-specific and multipolar.	"Persistent dominance" lacks a stable horizon and comparable weights across layers.	Five-year panel of secure task cost, capability, chips, packaging, power, deployment, trust, and third-market adoption.	Upgrade after thresholds are fixed; reject if one ecosystem dominates the pre-specified majority of weighted layers for the full horizon.

Exploratory proposition	Why it is not yet primary	Evidence needed	Downgrade or upgrade rule
E2. Productivity depends on work design and institutions.	Firm heterogeneity is nearly guaranteed and does not identify the causal complement.	Matched business units using the same model with pre-specified redesign, training, quality, and output measures.	Upgrade if redesign treatment predicts durable incremental output; weaken if model capability alone explains results.
E3. Open weights improve distributed defense enough to offset irreversible capability and operating burden.	Comparable denominators for misuse, vulnerability discovery, patching, and deployment burden remain weak.	Matched open and closed deployments with incidents, severity, red-team coverage, time to detection and remediation, concentration outages, staffing, and scale.	Upgrade after comparable exposure-adjusted evidence; weaken if irreversible misuse or operational burden grows faster than defensive benefit.
E4. Model neutrality is a central application asset.	Neutrality lacks one universal threshold and varies by workflow.	Model availability, release timing, price, routing override, BYO credentials, context portability, retention, and counterfactual competitors.	Upgrade by application after pre-deal baselines; weaken where ownership improves value without restricting meaningful choice.
E5. Full-stack convergence generalizes beyond SpaceX.	Current evidence is concentrated in one extreme boundary case and company intent.	Comparable cross-layer acquisitions or organic moves by multiple independent firms, including failures and neutral alternatives.	Upgrade after repeated post-integration success; reject as a general pattern if the boundary case remains isolated.
E6. Frontier AI safety risk becomes economically load-bearing.	Severe misuse and control outcomes have uncertain probability and evolving measurement.	Capability, access, permissions, incidents, near misses, safety-case performance, insurance, release restrictions, and regulatory cost.	Upgrade when pre-specified risk thresholds affect deployment or finance; weaken if verified controls bound exposure as capability grows.
E7. Hardware-stack control predicts bargaining power and resilience independently of model quality.	Fleet allocation, effective pricing, utilization, porting effort, yield and package capacity are largely private and platform claims are not normalized.	Eight-quarter panel across at least three accelerator ecosystems: productive share, matched task goodput/MW, effective cost, porting time, HBM and package lead time, uptime, and facility conversion.	Upgrade after thresholds and matched workloads are fixed; weaken if hardware platform and geography cease to predict cost, deployment speed, continuity, or margin after model quality is controlled.

Table 12.7. Exploratory propositions are separated from primary hypotheses rather than given a composite score.

Evidence standards

Separate project status: Record announced, contracted, financed, under-construction, energized, utilized, and profitable capacity as different states.

Tag the evidence: Distinguish observed, prospective, and conjectural claims; a filing, signed merger, or management rationale is not a completed operating result.

Separate source from result: Use company claims as evidence of strategy or disclosed scale, then seek customer, regulator, audited, or independently observed confirmation.

Pre-register the test: Record the unit, horizon, baseline, counterfactual, metric, and disconfirming result before the event matures.

Measure workflows, not models alone: Include tools, data, permissions, retries, human review, service levels, and downstream outcomes.

Keep value, rescue, and disconfirmation distinct: Separate social value from investor return and continuity support from shareholder rescue; every review must record evidence that weakened the thesis.

Source symmetry: For every load-bearing source, record supportive, neutral, and disconfirming observations contained in that source. A relevant item excluded from the body must have a stated scope or relevance reason. Appendix D is the baseline ledger.

CHAPTER 13

Implications for investors, boards, policymakers, and workers

The same system produces different duties for capital providers, enterprise operators, frontier firms, utilities, regulators, governments, educators, and workers.

The thesis does not imply one universal action such as buying infrastructure, avoiding mega-cap equities, or building a local model. It implies a diligence framework. Each stakeholder should identify the dependency it is creating, the balance sheet that bears the downside, the evidence required to justify scale, and the fallback available if capability, price, power, or policy changes.

For public-equity and private-market investors

Investors should resist the comforting idea that every "picks-and-shovels" asset is safer than the application or model layer. Utilities, data centers, cooling providers, chip suppliers, and power producers can benefit from structural demand, but they can also be exposed to customer concentration, technological obsolescence, take-or-pay risk, rate disputes, or overbuilding. The safest-looking layer can become the most leveraged layer.

- **Underwrite independent demand:** Separate customer cash from strategic-investor funding, prepayments, barter, and related ecosystem transactions.
- **Underwrite acquisition currency: Compare stock-financed deal value with market capitalization, implied dilution, target cash flow, customer retention, model neutrality, integration cost, and expected return on the acquired workflow.**
- **Normalize capital intensity:** Treat accelerator refresh, depreciation, power, networking, and facilities as recurring economic costs rather than temporary growth spending.
- **Map contract duration to asset duration:** A 20-year infrastructure asset backed by a short or cancellable AI commitment has a maturity mismatch.
- **Track value per compute:** Revenue and gross profit per energized MW, successful task, and dollar of invested capital are more informative than tokens or users alone.
- **Track rent migration:** Open-weight adoption can compress model pricing while increasing demand and margin at the chip, cloud, control-plane, data, or application layer. Measure spending and gross margin by layer rather than calling the whole stack "open."
- **Stress the index channel:** Model drawdowns, lockup expiries, follow-on issuance, and rebalancing rather than assuming passive demand is permanent.
- **Map future share supply and control: Track lockups and waivers, registration rights, options, vesting, restricted awards, acquisition shares, follow-ons, dual-class votes, insider sales, and key-person dependence.**
- **Price governance and succession: Understand voting rights, controlled-company exemptions, board independence, mission structures, related-party arrangements, founder dependence, succession, and the ability to separate or sell critical assets.**

Distinguish the technology from the security: A transformative technology can coexist with overvalued equity and poor project returns.

Interpret impairment correctly: Goodwill and intangible write-downs usually recognize a deterioration in expected value; they do not by themselves create the underlying economic loss.

For boards, chief executives, and chief financial officers

Boards should treat production AI as an operating-model and concentration decision, not merely an IT procurement. The most important questions concern workflow authority, liability, resilience, and economics. A board that approves enterprise-wide agents without knowing their permissions, failure modes, fallback, and fully loaded process return is making an unpriced operational bet.

Create a critical-workflow inventory: Identify which processes use which models, clouds, tools, data, and human escalation paths.

Set dependency limits by layer: Define acceptable concentration separately for cloud/control plane, model, application, data, identity, and compute.

Own evaluation: Retain customer-specific test sets, outcome labels, red-team results, and acceptance thresholds.

Contract for portability: Secure access to logs, configuration, prompts, tools, memory, and process data required to migrate.

Protect model choice: When an application is owned by a model or infrastructure provider, contract for transparent defaults, access to alternatives, data-use limits, and the ability to preserve workflow history during migration.

Choose an operating model explicitly: Decide whether single-vendor consolidation, hyperscaler-managed plurality, an independent managed control plane, or self-management best fits the organization's capability and risk.

Treat open weights as an option, not a label: Compare license rights, provenance, portability, deployment burden, security evidence, and secure task cost with proprietary alternatives at equivalent quality and service levels.

Measure realized economics: Require process-level cash, quality, cycle-time, risk, and labor metrics after integration and supervision costs.

Govern action rights: Use least privilege, transaction limits, separation of duties, approvals, and revocation for agents.

Fund the human system: Preserve expertise, apprenticeship, exception operations, and accountable owners rather than assuming automation removes them.

For frontier labs and platform companies

Frontier firms that seek public-market scale will need a disclosure and governance model suited to infrastructure-like and conglomerate dependence. Investors and regulators will ask not only about model capability, but also about capacity commitments, stock-funded acquisitions, related-party and intersegment economics, customer concentration, power availability, depreciation, application neutrality, safety incidents, and outcome liability. The companies can reduce the risk of punitive regulation by making economics, substitutability, and continuity visible before a crisis.

Disclose the capacity ladder: Announced, contracted, financed, under construction, energized, utilized, and economically productive capacity.

Clarify revenue quality: Separate consumer, API, enterprise seat, agent, transaction, advertising, and outcome revenue, including related-party activity.

Disclose acquisition integration: Report consideration and dilution, retention, model mix, product neutrality, intercompany pricing, cross-subsidy, integration costs, and the operating return attributed to acquired applications.

Define the partner boundary: State which workflows the platform will own and where software vendors, integrators, and service providers remain primary.

Build resolution architecture: Separate critical services and data sufficiently to preserve continuity during restructuring or ownership change.

Price liability explicitly: Use bounded action scopes, service levels, insurance, and contract design rather than hiding delivery risk inside gross margin.

Support portability: Common interfaces and exportable logs can increase trust and market size even when they reduce short-run lock-in.

State model-derived training terms precisely: Distinguish evaluation, adaptation, and legitimate follow-on development from unauthorized extraction through clear licenses, provenance, technical controls, and proportionate remedies.

Report failure: Publish meaningful incident, reliability, and evaluation information rather than capability metrics alone.

Publish safety cases proportionate to access: Report misuse, malfunction, autonomy, cyber, biological, and control evaluations together with permissions, containment, incident response, and release decisions.

For utilities, grid operators, and state regulators

Large AI loads can finance valuable infrastructure and improve regional economics. They can also create stranded costs and reliability pressure. The correct response is neither an automatic subsidy nor a blanket moratorium. It is contract and queue design that rewards credible, flexible projects while assigning cost to the party that creates it.

Use milestone-based queue rights: Require deposits, site control, equipment progress, and escalating financial security.

Protect ordinary customers: Large loads should fund dedicated facilities, exit risk, and incremental capacity unless broader benefits are demonstrated.

Price environmental and community effects: Measure water, emissions, backup generation, land, heat, transmission, local infrastructure, tax benefits, and community commitments instead of treating them only as permitting risk.

Value flexibility honestly: Compensate technically and contractually verifiable curtailment, geographic shifting, and demand response.

Plan portfolios, not single solutions: Use transmission, generation, storage, efficiency, firm fuel, and flexible demand together.

Coordinate across jurisdictions: Prevent duplicated reservations and ensure that simultaneous large-load plans are physically consistent.

Publish standardized status: Distinguish inquiries, studies, agreements, construction, energization, and actual load.

Test failure cases: Model customer bankruptcy, downsizing, or relocation before approving socialized investment.

For financial and operational regulators

The Financial Stability Board's work on AI concentration and responsible adoption points toward a cross-sector approach rather than a new capital regime based on firm valuation (FSB 2026). Authorities should map both financial exposures and critical-service dependencies. The objective is to detect when market concentration is becoming leveraged or operationally irreplaceable.

Collect exposure maps: Bank and nonbank credit, derivatives, collateral, project vehicles, guarantees, insurer holdings, and pension concentration.

Identify critical third parties: Models, clouds, data centers, identity services, and agent platforms whose failure would affect many regulated institutions.

Review vertical conduct: Monitor tying, bundling, default placement, discriminatory model access, data combination, intercompany pricing, and whether acquisitions make critical services harder to separate or resolve.

Run channel-specific combined stress tests: Pair equity drawdowns with credit and funding exposures, operational outages, cyber incidents, power constraints, and project cancellations, then report which core function is impaired.

Require recovery and exit: Regulated institutions should demonstrate migration or fallback for critical AI services.

Prepare resolution playbooks: Protect service continuity and customer data while preserving ordinary loss allocation.

Coordinate mandates: Financial, energy, cyber, competition, labor, and national-security agencies should share dependency information.

Avoid incumbent entrenchment: Scale obligations to criticality and systemic linkage so compliance does not become an entry barrier unrelated to risk.

For competition authorities and corporate-governance regulators

Cross-layer mergers should be evaluated as platform and vertical conduct problems, not only as overlaps between products. Under Section 7 of the Clayton Act, the central U.S. merger question is whether the transaction may substantially lessen competition or tend to create a monopoly. The 2023 U.S. Merger Guidelines direct agencies to examine competition between, on, and to displace multi-sided platforms and whether a merger can limit access to products, services, or routes to market. The relevant evidence for a compute-model-application combination includes defaults, ranking, price, latency, data access, interoperability, bundling, transfer pricing, and the ability of users and rival models to multi-home (U.S. Department of Justice and Federal Trade Commission 2023).

Do not presume either prohibition or clearance: Possible outcomes include unconditional approval, delay, nondiscrimination obligations, continued rival-model access, routing auditability, data-use separation, interoperability, governance separation, structural separability, divestiture, or

prohibition. Remedy intensity should follow evidence of foreclosure and the feasibility of monitoring.

Review founder and related-party control: Require clear disclosure of voting power, board exemptions, succession, affiliated transactions, segment transfers, and conflicts that can move resources or risk across the group.

Preserve service separability: A merger should not make a critical application, model, communication system, or data asset impossible to transfer or continue during restructuring.

Measure post-merger conduct: Track model availability, defaults, pricing, release timing, routing overrides, data combination, customer retention, partner access, and the independent economics of each layer.

Map concentration beyond the model layer: Open-weight availability does not establish a competitive stack. Track chips, cloud, identity, control planes, data access, application distribution, and the location of margin and bargaining power.

For national economic and security policy

The United States should pursue a portfolio objective: abundant and reliable power, advanced semiconductor capacity, frontier research, capable open weights, shared datasets and evaluation infrastructure, secure cloud and networking, talent, competitive capital markets, trusted governance, and diffusion into ordinary firms. No single champion can substitute for the ecosystem. Protecting an incumbent from competition can weaken the very innovation and trust that create strategic advantage, while releasing capability without measured safeguards can create irreversible externalities.

Accelerate credible infrastructure: Modernize interconnection, permitting, transmission, equipment supply, and generation while protecting reliability and communities.

Build a plural open ecosystem: Support interoperability, open research, capable open-weight options, compute access, shared datasets and evaluations, cloud competition, and application-layer entry while measuring operational and safety outcomes.

Target strategic subsidies: Tie public support to measurable capacity, resilience, knowledge spillovers, and domestic or allied supply-chain value.

Invest in measurement: Build public statistics for compute, energy, AI prices, adoption, productivity, labor transitions, and incidents.

Coordinate with allies: Treat trusted compute, chips, standards, and research as a shared network rather than a purely national stockpile.

Engage multipolar sovereignty strategies: Build interoperable relationships with European public compute, IndiaAI, Gulf sovereign infrastructure, and allied manufacturing rather than treating every non-U.S. initiative as alignment with China.

Calibrate controls: Protect genuinely strategic capabilities while monitoring substitution, allied costs, and market fragmentation.

Calibrate model-derived training rules: Use targeted legal and commercial remedies for misappropriation while preserving legitimate evaluation, distillation, interoperability, and follow-on innovation where rights and safeguards permit.

Maintain legitimacy: Ensure that infrastructure benefits, environmental costs, data rights, and productivity gains are distributed visibly enough to sustain public consent.

For workers, educators, and professional institutions

The practical response to AI is not to compete with a model at every routine task. It is to develop judgment, domain accountability, system understanding, and the ability to supervise and improve automated work. Yet individuals cannot solve a structural transition alone. Employers and institutions must preserve pathways through which novices become experts and workers share in the value of the tools they help train and operate.

- **Teach verification and systems thinking:** Workers need to understand evidence, model limits, tool permissions, downstream effects, and when to escalate.
- **Preserve deliberate practice:** Replace disappearing junior tasks with simulations, rotations, critique, and supervised decisions.
- **Build portable skills:** Data literacy, process design, security, evaluation, and domain expertise transfer better than mastery of one vendor interface.
- **Negotiate data and monitoring rights:** Clarify how worker interactions, performance, and corrections are used to train or manage systems.
- **Share gains:** Use wages, bonuses, reduced hours, employee ownership, and advancement to connect productivity with worker benefit.
- **Track differential exposure:** Direct transition support toward occupations, regions, and demographic groups facing concentrated change.

A sequence for action

Table 13.1. A common action sequence across stakeholders.

Stage	Mechanism	Economic consequence
1	Map dependencies	Identify the firms, models, clouds, power assets, contracts, and people that a critical outcome requires.
2	Measure economics	Calculate fully loaded cost, success, quality, utilization, and risk at the workflow and infrastructure levels.
3	Price failure	Assign the cost of cancellation, outage, error, migration, and restructuring to explicit parties.
4	Preserve options	Build portability, alternative supply, flexible load, human fallback, and separable critical services.
5	Scale on evidence	Expand only when independent demand, utilization, ROI, and controls are demonstrated.
6	Recalibrate	Update the thesis when price, performance, concentration, policy, or geopolitical conditions change.

CHAPTER 14

Conclusion

The frontier-firm question is not whether the companies are large. It is what their size allows them to connect - and whether those connections remain substitutable, financeable, and governable.

The public-market emergence of SpaceX and possible listings of OpenAI and Anthropic connect capital markets with infrastructure, models, applications, and enterprise execution. But the baseline does not demonstrate one stable “frontier firm” equilibrium. SpaceX is an extreme, founder-controlled boundary case; Cursor is a signed but unclosed transaction at the evidence cutoff; and the labs face powerful alternatives in hyperscaler-managed distribution and open-weight diffusion. This document is therefore a disciplined, adjudicable research protocol with a clean factual scaffold - not a tested general law of trillion-dollar firms.

What the evidence supports

Index inclusion broadens exposure and can amplify wealth, confidence, and capital-expenditure effects. It does not make a firm financially systemic. That claim requires observable impairment of credit, funding, payments, clearing, or major intermediaries. Operational criticality and strategic importance remain analytically useful, but they require their own evidence and do not imply shareholder rescue.

Electricity, interconnection, equipment, cooling, water, and local acceptance are genuine constraints. Efficiency can raise successful work per megawatt, while agentic sequential depth, strict latency, and rebound demand can make total power and local throughput more binding. The question is whether supply and efficiency outpace demand while maintaining service levels and social cost.

Frontier firms have incentives to move into applications and outcomes, but the control point is contested. Stock-financed integration can preserve neutrality and improve economics, or it can produce overpayment, self-preferencing, user flight, and impairment. Hyperscalers, model-neutral applications, systems of record, integrators, and specialized process firms are coequal alternative explanations.

Open-weight and sovereign architectures remain credible bargaining and resilience strategies, and open weights now warrant treatment as a coequal industrial architecture for diffusion and specialization. The July 24 coalition statement demonstrates organized cross-layer support for that strategy, not proof of its outcomes. Most enterprises will still procure managed control rather than operate a bespoke multi-model stack. Open weights must be evaluated by an explicit openness profile, fully loaded secure task cost, safety performance, portability, and whether model-layer competition shifts concentration and rents toward the substrate. Sovereignty must be

demonstrated through legal authority, audit, portability, recovery, and sustained operation for a defined workload.

U.S. leadership is substantial but neither uniform nor self-sufficient. Chinese capability and industrial scale, European public compute, Indian national infrastructure, Gulf full-stack investment, and allied fabrication and packaging make the competitive system multipolar and interdependent. Claims of layer-specific leadership remain exploratory until measured against a fixed horizon and comparable weights.

The central causal mechanism

Table 14.1. Conditional research program: market, integration, power-latency, intermediation, and substitution mechanisms lead to rival outcomes rather than one predetermined loop.

Stage	Mechanism	Economic consequence
1	Public-market scale	Large valuations and index membership broaden exposure and lower the apparent cost of capital.
2	Acquisition currency	Highly valued equity can purchase application distribution, workflow data, talent, and customer relationships with comparatively limited dilution.
3	Infrastructure and application expansion	Capital finances compute, data centers, power, semiconductors, suppliers, and the software interfaces that create recurring demand.
4	Price and return pressure	Model competition compresses basic access prices while investors demand returns on construction and acquisitions.
5	Workflow capture	Providers seek more value per compute through applications, agents, transactions, and completed outcomes.
6	Dependency response	Enterprises and states increase portability, model choice, open-weight capacity, private inference, competition scrutiny, and continuity planning.

Stage	Mechanism	Economic consequence
7	System outcome	Productivity, concentration, fragility, and geopolitical power depend on acquisition discipline, infrastructure economics, workflow performance, and substitutability.

Contributions of the thesis

The protocol contributes eight disciplined shifts. First, it replaces a composite systemic score with channel-specific qualification thresholds. Second, it treats valuation as conditional acquisition capacity with value-creating and value-destroying branches. Third, it adds latency, sequential compute, rebound demand, and environmental incidence to the power model. Fourth, it makes hyperscaler intermediation a rival control-point theory. Fifth, it defines sovereignty as a verified capability that can be managed rather than universally self-built. Sixth, it treats open weights as a coequal industrial architecture whose effect must be measured through diffusion, safety, total cost, and rent migration. Seventh, it uses one canonical SpaceX-Cursor treatment plus a blocked-or-abandoned contingency rather than repeated evidentiary counting. Eighth, it specifies symmetric nulls, thresholds, windows, comparisons, sources and observability rules before the first review.

What would prove the thesis wrong

The protocol should be judged first against H1-H7 and then against the following canonical disconfirmation conditions. They are not rhetorical caveats; an observed rejection threshold requires an explicit weakening or rejection of the corresponding claim in the next version.

Frontier-firm drawdowns remain contained within ordinary portfolio and corporate-finance channels, with no material credit, funding, payment, clearing, or intermediary effect.

Infrastructure delays or cancellations produce little loss for lenders, suppliers, utilities, ratepayers, insurers, municipalities, or regions.

Stock-financed application acquisitions do not preserve retention and neutrality or improve task economics, cash flow, model performance, and return on invested capital.

Power supply and efficiency consistently improve successful-task cost and latency faster than demand, making scarcity rents and missed service levels immaterial.

Frontier labs sustain durable margins and direct customer control without owning workflows, while hyperscaler routing does not commoditize them.

Enterprises continue to accept single-vendor concentration after real dependency shocks and show little demand for managed portability, private deployment, or tested exit.

Open-weight diffusion fails: Capable open weights do not achieve meaningful production share, reduce model-layer price or concentration, improve switching or access, or redistribute spending and margins after full operating and security costs are included.

The SpaceX boundary case remains isolated and comparable frontier firms do not converge across infrastructure, models, applications, and workflows.

One geopolitical ecosystem achieves durable dominance under the pre-specified cross-layer measures, or multipolar strategies fail to create meaningful capability and bargaining power.

The policy principle

The state should protect functions, competition, and resilience - not valuations. It should accelerate credible infrastructure while assigning cancellation and reliability costs to the parties that create them. It should require continuity and recovery where AI becomes critical, without turning compliance into an unnecessary moat. It should support frontier capability and a plural open ecosystem through compute access, shared evaluation assets, interoperability, and targeted rules for model-derived training, while measuring security and rent migration rather than assuming openness is inherently beneficial. And it should ensure that productivity produces a visible public return.

> *The defining question of the frontier economy is not whether intelligence becomes abundant. It is who finances the abundance, who controls the workflows it enters, who bears the failures, and whether the system remains open enough to adapt when today's leaders, technologies, or geopolitical arrangements change.*

Final formulation

PROTOCOL THESIS
Trillion-dollar frontier firms may become macro-financially consequential, operationally critical, or strategically important when valuation, infrastructure, and workflow dependence reinforce one another. They become financially systemic only when distress impairs core financial functions. High valuation creates conditional acquisition capacity; integration succeeds only if neutrality, retention, economics, governance, and separability survive. Power is a local and latency-sensitive constraint whose unit efficiencies may intensify aggregate demand. Most enterprises will obtain plurality and sovereignty through managed control planes, while hyperscalers compete with frontier firms for the customer control point. Capable open weights may reduce model-layer scarcity and broaden diffusion, but can shift dependency and rents to a concentrated substrate. These propositions have not yet been tested; adjudication begins with the first quarterly review under H1-H7.

APPENDIX A

Definitions, qualification tests, and analytical identities

A compact vocabulary for maintaining conceptual discipline as the market and technology evolve.

Term	Definition
Frontier firm	A company operating near the leading edge of a strategically important technology and capable of financing or coordinating infrastructure at unusually large scale.
Frontier platform	A frontier firm whose models, networks, distribution, interfaces, or services become a shared operating layer for other organizations.
Hard financial systemic risk	Distress that materially impairs credit, funding, payments, clearing, or major financial intermediaries.
Macro-financial amplification	Transmission from valuation or investment shocks into consumption, collateral, capital expenditure, suppliers, utilities, employment, or regional public finance without necessarily impairing core financial functions.
Operational criticality	Reliance on a service whose interruption cannot be absorbed or replaced within the required recovery time and loss tolerance.
Strategic importance	Relevance to defense, communications, science, industrial capacity, public administration, or national autonomy.
Resolvability	The ability to continue or transfer critical services while ordinary losses, ownership change, and restructuring occur.
Boundary case	An extreme case used to reveal a mechanism; it is not representative evidence unless the mechanism survives comparison with ordinary, failed, and countervailing cases.
Valuation-dependent acquisition capacity	The ability to use highly valued equity as consideration; a higher relative valuation can reduce percentage dilution but does not eliminate dilution, overpayment, or integration risk.
Model neutrality	Meaningful, auditable choice among model providers across availability, release timing, price, latency, routing override, credentials, data use, and portable workflow context.
Hyperscaler intermediation	A market structure in which cloud, identity, data, security, routing, procurement, and billing abstract multiple model providers behind one managed enterprise control plane.
Managed control plane	Routing, policy, evaluation, observability, failover, cost, and governance functions operated by a provider while specified decision rights and evidence remain with the customer.
Sovereign-capable deployment	A workload-specific architecture in which the organization has legal authority, audit evidence, portable context, a tested alternative within its recovery objective, and the ability to sustain critical operation through a defined disruption.
Open-weight model	A model whose trained parameters are available under a license, even if training code, data, or process are not fully open.
Managed digital labor	An ongoing service in which a provider operates AI agents and human exception processes to execute defined work.
Outcome-as-a-service	A commercial model priced against completed transactions, savings, service levels, or business results rather than tokens or seats.
Secure latency-compliant cost per successful task	Fully loaded compute, energy, integration, supervision, security, compliance, expected-error, and delay cost divided by verified outcomes delivered within the required service level.

Term	Definition
Founder and key-person control risk	Dependence created when voting power, technical authority, strategic relationships, succession, or affiliated transactions are concentrated in one person or a small insider group.
Openness stack	The separate dimensions of weights, architecture, code, license rights, data provenance, evaluation transparency, reproducibility, and deployment portability; no single dimension establishes full openness.
Open ecosystem, concentrated substrate	A market structure in which models and applications are plural or portable while chips, fabrication, cloud, identity, control planes, or data-center infrastructure remain concentrated.
Model-derived training and distillation	Use of model outputs to evaluate, adapt, or train another system; its legal and competitive treatment depends on license, contract, provenance, extraction method, and applicable law.
Primary frontier firm	A firm whose principal strategic identity is a frontier model, frontier application, or integrated frontier platform. Baseline cases: SpaceX/xAI, OpenAI, and Anthropic.
Material index weight	At least 1.0% float-adjusted weight in the S&P 500 or Nasdaq-100 on the trigger date.
Major infrastructure project	A data-center, generation, transmission or equipment program with at least 1 GW of planned load/capacity or at least $10B of committed capital.
Matched comparison	A pre-specified control selected on observable size, sector, beta, leverage, contract structure, region, workload or other variables named in the relevant hypothesis before the event window.
Verified outcome	A completed task that meets pre-specified correctness, safety and latency requirements. Reliability and latency are included in the outcome definition rather than multiplied again.
Not observable	A formal protocol state used when required treated and comparison data are unavailable. It supports neither the hypothesis nor the null.
AI factory	A commissioned production system combining accelerators, host processors, memory, advanced packaging, interconnect, network, storage, software, power, cooling, facility, and operators. Ordered chips or announced MW are not an AI factory.
Hardware portability	The demonstrated ability to move a defined workload between accelerator ecosystems within a stated time, cost, performance-loss, security, and recovery threshold.
Economically productive MW	Energized capacity that is rack-ready, commissioned, software-usable, utilized, and delivering successful latency-compliant work after power, depreciation, networking, maintenance, and operating cost.
Adjudication state	One of six protocol outcomes: not triggered; triggered/window open; support met; null met; mixed/indeterminate; not adjudicable.
Common hardware-substrate shock	A foundry, memory, packaging, interconnect, platform-security, power or other shared-input event that affects multiple material issuers and therefore cannot be treated as an idiosyncratic firm shock.
Conditional support-pattern frequency	The share of simulated draws satisfying a support Boolean under disclosed priors and dependence assumptions, conditional on a qualifying event and complete measurement. It is not a verdict probability, calibrated forecast, null estimate, or observability estimate.
Material AI-system issuer	A public issuer with material index weight and direct control of a critical AI-stack layer whose distress could transmit an AI-related shock. NVIDIA qualifies at the baseline; other issuers enter only through frozen inclusion criteria.
Event-level verdict	The six-state result for one qualifying event at its fixed endpoint. It is not automatically a conclusion about the general hypothesis.
Hypothesis-level verdict	A replicated conclusion formed only under the minimum-event, diversity, support-share, null-share and missing-data rules in Table 12.1A.

Table A.1. Core definitions.

Rule 1: channel qualification - no composite score

CHANNEL QUALIFICATION Assess hard financial transmission, macro-financial amplification, operational criticality, strategic importance, and resolvability separately. Do not assign weights or aggregate them. A claim fails when its channel-specific outcome is absent, even if another form of importance is present.

Formula 1: capacity and energy productivity with explicit units

CAPACITY AND ENERGY PRODUCTIVITY - WINDOW T
Economic output rate ($/hour) = verified outcomes within the service level x average value per outcome / hours in T. Capacity productivity ($/MW-hour) = economic output / (average energized MW x hours). Energy productivity ($/MWh) = economic output / electricity consumed. “Verified outcome” already includes correctness, reliability and latency, preventing double counting. These are accounting identities; causal interpretation requires H4’s matched site design.

Formula 2: task margin

TASK MARGIN
Task margin = outcome price - inference - integration - monitoring - human exception handling - compliance - support - expected liability. Outcome pricing is attractive only when the provider can standardize the process, control variance, and bound liability.

Formula 3: managed or sovereign deployment total cost

DEPLOYMENT TCO Total deployment cost = provider charges or hardware depreciation + financing + power + cooling + networking + facilities + operations talent + model and router maintenance + security + evaluation + downtime + migration and refresh risk. Compare alternatives at equivalent quality, latency, reliability, legal authority, and recovery performance.

Formula 4: secure latency-compliant cost per successful useful task

COMPETITIVE METRIC Secure latency-compliant task cost = (compute + energy + integration + supervision + security + compliance + expected error loss + delay loss) / verified outcomes delivered within the required service level. Customer control, resilience, environmental incidence, and geopolitical availability remain additional constraints.

Formula 5: hypothesis-specific scenario generator

CONDITIONAL SUPPORT-PATTERN GENERATOR
For each hypothesis, sample support-component probabilities from frozen low/mode/high priors, draw correlated Bernoulli components under the disclosed dependence assumption, apply the fixed support Boolean, and report the conditional support frequency and sensitivity band for that hypothesis only. The generator assumes a qualifying event and complete measurement. Real-world null, mixed and not-adjudicable states are governed by the event ledger and endpoint rules. Never aggregate across H1A, H1B, or H2-H7.

APPENDIX B

Monitoring dashboard

A practical set of indicators for quarterly review, now populated with the July 24, 2026 baseline where observable. "Not yet observable" is an explicit instrument state, not missing decoration. Definitions must remain consistent across periods.

Domain	Indicator	Why it matters	Baseline: July 24, 2026	Cadence
Event and equity	Offering proceeds, primary vs. secondary shares, float, lockups, follow-ons	Separates capital raised from valuation and future supply	SpaceX: 638.9M shares at $135; $86.25B mechanical offer value; about $85.7B company-reported proceeds. Full future-supply schedule remains to be maintained.	Event/ quarterly
Event and equity	Float-adjusted index weight and top-ten concentration	Measures portfolio transmission rather than headline market cap	SpaceX about 1.34% of Nasdaq-100 at inclusion; NVIDIA 7.93528% of SPY on 2026-07-23 (rounded fund-page display 7.94%); NDX top 10 = 52% at YE2025; S&P 500 top 10 = 36.4% on 2026-05-29.	Monthly
Event and equity	Options, securities lending, margin, structured-product exposure	Identifies leverage and forced-flow channels	No complete public aggregate baseline; first review will compile options OI, securities lending and fund filings.	Monthly/ quarterly
Event and equity	Leveraged and inverse ETFs, options and structured-product rebalancing exposure	Separates flow amplification from ordinary unlevered index ownership	No standardized constituent-level baseline; compile fund assets, leverage targets, option OI and dealer gamma in Q3 2026.	Monthly/ event
Capital structure	Corporate debt, revolvers, maturities, covenants, interest burden	Shows whether equity risk is migrating to creditors	$25B SpaceX senior unsecured notes priced June 23, 2026; five tranches, 5.35%-6.65%, due 2031-2056.	Quarterly
Capital structure	Project debt, SPVs, guarantees, take-or-pay and purchase obligations	Maps off-balance-sheet and infrastructure linkage	Consolidated SPV, guarantee and project-debt exposure not publicly observable at cutoff.	Quarterly
Demand quality	Cash from unrelated end customers	Tests whether demand is independent of strategic financing	Not separately disclosed across the three firms.	Quarterly
Demand quality	Production conversion, renewal, expansion, churn, workflow volume	Distinguishes pilots from durable adoption	No comparable public production-conversion, renewal and churn baseline.	Quarterly
Demand quality	Revenue mix: consumer, API, seat, agent, transaction, outcome, ads	Shows movement up the value-capture ladder	Not publicly disaggregated on a comparable basis.	Quarterly
Compute	Accelerator fleet by generation, provider, and location	Measures concentration, refresh risk, and flexibility	No reconciled fleet baseline. Historical: xAI Colossus 100k Hopper GPUs in 2024. 2026 SpaceX filings disclose agreements tied to ~325k NVIDIA GPUs for Anthropic and ~110k for Google; do not sum as a verified non-overlapping productive fleet.	Quarterly
Compute	Utilization and productive accelerator hours	Tests whether capacity is economically used	Not publicly observable.	Monthly/ quarterly
Compute	Cost per successful task and retries	Links technical efficiency to business value	Public API prices exist; matched fully loaded cost baseline not yet compiled.	Monthly
Hardware platform	Productive workload share by accelerator, generation, provider and site	Measures NVIDIA, custom-silicon and alternative-stack concentration rather than ordered fleet alone	No comparable productive-fleet baseline. Disclosed anchors: xAI 100,000-Hopper Colossus in 2024; AWS Rainier 500,000+ Trainium2 in 2026; cross-firm fleet shares private.	Quarterly
Hardware portability	Porting time, engineering cost, retained performance, security and reliability across CUDA, Neuron, CANN and other stacks	Tests whether nominal multi-platform access creates real bargaining power and continuity	Not publicly observable. Establish matched workload and codebase panel before first portability adjudication.	Semiannual/event

Domain	Indicator	Why it matters	Baseline: July 24, 2026	Cadence
Fabrication and packaging	Leading-edge logic, HBM and advanced-packaging qualified capacity, yield, allocation, lead time and geography	Identifies common upstream chokepoints and whether geographic diversification is production-equivalent	TSMC: 15.0M 12-inch-equivalent wafers shipped in 2025; 74% of wafer revenue at 7nm and below; Taiwan remains central while Arizona expansion proceeds.	Quarterly/ annual
Data-center commissioning	Ordered chips, delivered racks, rack-ready MW, commissioned systems, productive accelerator hours and economically productive MW	Separates chip and gigawatt announcements from operable and revenue-producing infrastructure	OpenAI reported >5 GW Stargate capacity under development; standardized hardware-aware commissioning panel begins Q3 2026.	Monthly/ quarterly
Power	Announced, contracted, financed, construction, energized MW	Prevents announcement inflation	Large announced pipeline exists; standardized project-status panel begins in Q3 2026.	Monthly/ quarterly
Power	Firm vs. interruptible load; curtailment performance	Measures reliability burden and flexibility value	Not publicly observable across major sites.	Seasonal/ annual
Power	Revenue and gross profit per energized MW	Tests value creation against physical scale	Not publicly observable.	Quarterly
Power	Interconnection security, deposits, dedicated-facility funding	Shows who bears cancellation and grid costs	FERC show-cause orders issued June 18, 2026; project-level security not standardized.	Event/ quarterly
Model economics	API and effective inference price by quality tier	Measures commoditization and price pressure	Public schedules available; quality-matched baseline begins Q3 2026.	Monthly
Model economics	Serving cost, depreciation, power, gross margin	Tests whether cost falls faster than price	Not publicly observable.	Quarterly
Agent economics	Success, escalation, exception, rollback, and incident rates	Measures the true cost of autonomy	Not publicly observable with comparable denominators.	Monthly
Agent economics	Task margin and outcome-liability provisions	Tests whether workflow ownership is profitable	Not publicly observable.	Quarterly
Enterprise control	Critical workflows by model/cloud/provider	Measures operational concentration	No public cross-enterprise baseline.	Quarterly
Enterprise control	Exit tests, recovery time, fallback success, data portability	Measures substitutability and resilience	No public matched baseline; first enterprise panel planned for 2027.	Semiannual
Enterprise control	Private inference and open-weight share	Tracks sovereignty and bargaining response	Production share not publicly observable.	Quarterly
Security	Material incidents, prompt injection, unauthorized actions, data exposure	Tests trust and operational risk	Case reports exist; no exposure-adjusted denominator.	Continuous /quarterly
Security	Independent evaluation and red-team coverage	Shows whether controls are evidenced	Vendor and independent tests exist; coverage denominator incomplete.	Quarterly
Labor	AI-exposed hiring, entry-level share, hours, wages, promotions	Detects transition and apprenticeship effects	Public labor series exist; frontier-firm-linked baseline not yet compiled.	Quarterly/ annual
Labor	Output and quality per paid hour	Measures realized productivity rather than use	No standardized cross-firm workflow baseline.	Quarterly
Distribution	Labor share, prices, consumer access, regional effects	Shows who receives the economic gain	Annual public data available with lag; July 2026 baseline not yet compiled.	Annual
Geopolitics	Export licenses, restricted hardware flows, domestic substitution	Tracks effectiveness and fragmentation	Advanced-chip controls and revised 2026 China license-review policy active.	Event/ quarterly
Geopolitics	Allied compute, model, and standards adoption	Measures trusted ecosystem reach	JPMorgan reports installed DC capacity: U.S. 53.7 GW, China 31.9, EU 11.9, Japan/Korea 6.6, India 3.6, Middle East 1.1; installed is not equivalent to AI-ready or utilized. EU/India/UAE/Saudi programs tracked separately.	Quarterly/ annual
US-China	Secure task cost, deployment speed, reliability by layer	Replaces benchmark-only comparisons	No matched cross-country cost panel; Stanford U.S.-China capability gap reported at 2.7%.	Quarterly

Domain	Indicator	Why it matters	Baseline: July 24, 2026	Cadence
Systemic risk	Common-provider exposure among regulated institutions	Identifies correlated operational failure	Not publicly aggregated across regulated institutions.	Annual/ stress test
Systemic risk	Channel-specific combined stress loss: financial, macro, operational, and strategic outcomes reported separately	Prevents one form of importance from rescuing a failed claim in another channel	No qualifying baseline shock; outcomes remain disaggregated.	Annual
Resolution	Critical-service separability and continuity plan maturity	Shows whether failure can occur without service collapse	No standardized public disclosure.	Annual
Public legitimacy	Rate impacts, tax contributions, jobs, community benefits, complaints	Tracks political durability of the build-out	Project-specific data exist; no standardized national baseline.	Annual
Event and equity	Stock-financed acquisition value, implied dilution, deal spread, approval and closing status	Measures use of valuation as acquisition currency and separates signed transactions from completed integration	SpaceX-Anysphere: signed $60B all-stock agreement; expected Q3 close; pending approvals.	Event/ weekly
Vertical integration	Owned-model share inside acquired applications; model-neutral access; user and developer retention	Tests whether distribution converts into durable model demand without destroying application value	Pre-close; no post-close model-mix or retention baseline.	Monthly/ quarterly
Capital allocation	Segment transfers, cross-subsidy, acquisition costs, goodwill and intangible returns	Reveals conglomerate economics and whether risk is migrating across otherwise distinct businesses	Not publicly disclosed.	Quarterly/ annual
Event and equity	Monthly concentration range, index-return contribution, free float, lockups, registrations, vesting, insider sales, options and acquisition shares	Treats concentration and supply as time-varying rather than fixed	SpaceX initial NDX weight about 1.34%; future supply and lockups disclosed but require monthly schedule.	Monthly/ event
Governance	Voting control, controlled-company status, board independence, related-party transactions, succession, restricted awards	Measures founder and key-person risk and public-shareholder influence	SpaceX founder voting control reported at 82.4%; mission/control structures differ at OpenAI and Anthropic.	Quarterly/ annual
Latency and rebound	Sequential model/tool steps, retries, P50/P95/P99 completion, missed service levels, demand growth versus efficiency	Tests physical and commercial limits of agentic workloads	No matched production telemetry.	Monthly/ quarterly
Environment and community	Water, carbon intensity, backup generation, land, heat, rate effects, local infrastructure, benefits and complaints	Measures social cost and political durability of data-center growth	Project-level disclosures incomplete and non-comparable.	Quarterly/ annual
Hyperscaler intermediation	Managed-router share, direct versus intermediary billing, model substitutions, master-contract owner, margin by layer	Identifies the enterprise control point even when many models are used	Azure Foundry and AWS Bedrock routing available; contract and spend shares not public.	Quarterly
Model neutrality	Availability, release timing, default share, price, latency, overrides, BYO credentials, context portability, rival-provider terms	Operationalizes whether an acquired application retains meaningful provider choice	Cursor pre-close supports multiple providers; quantitative defaults and shares not public.	Monthly/ quarterly
Frontier AI safety	Misuse and malfunction incidents, autonomy and access evaluations, permissions, near misses, safety cases, release restrictions	Tracks whether risk management keeps pace with capability and deployment criticality	International and company safety reporting exists; no common exposure-adjusted denominator.	Continuous /quarterly
Multipolar sovereignty	EU public compute, IndiaAI access and domestic models, Gulf full-stack capacity, allied fabrication and packaging resilience	Tests whether non-U.S./China strategies create meaningful capability and bargaining power	EU 19 factories/13 antennas; IndiaAI portal and Sarvam program; Stargate UAE 1GW; HUMAIN full-stack program.	Quarterly/ annual

Domain	Indicator	Why it matters	Baseline: July 24, 2026	Cadence
Competition and remedies	Merger status, interoperability, nondiscrimination, data-use limits, governance separation, divestiture or compliance	Tracks regulatory circuit breakers and post-merger conduct	SpaceX-Anysphere pending; no public remedy or final decision at cutoff.	Event/ quarterly
Open-weight diffusion	Production inference share, matched secure task cost, deployment time, sector and public-institution adoption, model-layer concentration and price premium	Tests whether open weights create economically meaningful diffusion and model-layer competition	Unknown. Stanford closed-open gap 3.3%; Kimi K3 weights promised July 27 but not available at cutoff.	Quarterly
Rent migration	AI spending and gross-margin share by model, chip, cloud, control plane, data, and application layer	Tests whether model-layer openness decentralizes the stack or transfers value to a concentrated substrate	Not publicly observable by layer.	Quarterly/ annual
Openness profile	Weights, architecture, code, license rights, training-data provenance, evaluation transparency, reproducibility, and portability	Prevents an “open” label from collapsing distinct rights, artifacts, and operating capabilities	Model-family differences documented in Chapter 7; no single binary “open” baseline.	Release/ event
Open-weight safety	Misuse and malfunction incidents, vulnerability discovery, independent red-team coverage, patch time, concentration outages, and operating burden versus closed comparators	Tests distributed defense against irreversible capability and maintenance risk	International and company safety reporting exists; no common exposure-adjusted denominator.	Continuous /quarterly
Model-derived training	License and output-use terms, provenance, distillation or extraction disputes, legal rulings, technical restrictions, and remedies	Tracks the balance between investment protection, competition, follow-on innovation, and misappropriation	Anthropic alleges large-scale illicit distillation; coalition urges targeted remedies; no adjudicated ruling.	Event/ quarterly
Adjudication	H1A-H7 state, endpoint and missing evidence	Makes the decision engine executable and prevents repeated-look cherry-picking	All hypotheses not triggered or awaiting future data at baseline; no formal verdict.	Quarterly/ endpoint
Simulation	Hypothesis-specific conditional support-pattern bands	Shows assumption sensitivity under a triggered event and complete measurement; does not estimate observability or verdict probabilities	Base and 5th-95th support bands fixed in Table 12.4; no null/missing simulation and no cross-hypothesis aggregation.	Versioned amendment
Hardware concentration	Common dependency share and exposure-adjusted security incidents	Tests whether one platform or distribution layer creates correlated operational loss	CSA-reported figures entered as source-qualified estimates; organization-specific exposure not yet observable.	Quarterly/ event
NVIDIA issuer risk	Index weight, customer concentration, spreads, options and common-substrate exposure	Applies the same issuer and transmission tests to the largest public hardware firm	SPY daily workbook: 7.93528% at 2026-07-23; fund page rounds to 7.94%. FY2026 direct customers 22% and 14%; indirect concentration disclosed but estimated.	Monthly/ quarterly
TSMC diversification	Announced, built, qualified and volume-producing U.S. fabs and packaging	Separates geographic diversification from completed substitution	$265B total Arizona investment announced; commissioned contribution and timeline tracked separately.	Quarterly/ annual
Evidence integrity	Source-symmetry ledger completeness	Shows whether load-bearing sources are represented with supportive and disconfirming evidence	Appendix D baseline completed; every future quarterly review must append changes and exclusions.	Quarterly/ versioned

Table B.1. Frontier-firm monitoring dashboard.

Quarterly review questions

Before scoring: confirm that the trigger, threshold, comparison group and data source were fixed before the event. Record any unavailable measure as “not observable,” and publish versioned amendments prospectively.

What changed in the thesis? State the three strongest confirming observations and three strongest disconfirming observations.

Which announced projects became real? Move each capacity item through the status ladder and record cancellations.

Where did risk migrate? Identify whether exposure moved from equity to debt, utilities, suppliers, governments, or customers.

Did economics improve? Review task margin, revenue per MW, utilization, cash conversion, and verified enterprise ROI.

Where did openness move concentration: Compare open-weight production share and model prices with chip, cloud, control-plane, data, and application spending, margins, and switching power.

Did substitutability improve? Review model diversity, exit tests, control-plane portability, and recovery performance.

Who benefited and who paid? Review wages, hiring, prices, rate impacts, tax incentives, and regional effects.

Which scenario gained probability? Use signposts from Chapter 11 without assigning false precision.

APPENDIX C

Scenario-generator parameters and reproducibility specification

This appendix freezes the v1.2.1 conditional support-pattern generator. The values are author-specified priors, not estimated frequencies, and the outputs are not calibrated forecasts. The generator assumes a qualifying event and complete measurement. It does not simulate event incidence, public observability, null outcomes, mixed outcomes, or the probability of a not-adjudicable verdict. Channel outcomes remain disaggregated.

Component probability ranges and support Boolean patterns are frozen in Table 12.3. Appendix C separately records real-world observability so a clean simulation assumption cannot be mistaken for available evidence. A hypothesis can have a reproducible support-pattern range and still be not adjudicable in practice.

Table C.1. Real-world observability register. Simulation conditions on complete measurement; this table governs the missing-data state.

H	Mandatory evidence	Public status at baseline	Likely evidence owner	Earliest formal endpoint	Missing-data treatment
H1A	Issuer/intermediary spreads, funding, collateral, flows, clearing	Partial	Markets, dealers, funds, clearinghouses, regulators	60 trading days after trigger	Not adjudicable if exposure map or mandatory control is missing
H1B	Shared supplier exposure, issuer operations, index and credit controls	Partial	Issuers, suppliers, exchanges, regulators	120 trading days after trigger	Not adjudicable if common exposure cannot be mapped
H2	Project financing, guarantees, utility/rate, supplier and redeployment data	Limited	Sponsors, lenders, utilities, regulators, municipalities	24 months after trigger	Not adjudicable without financing and replacement maps
H3	Cohort retention, model mix, neutrality, integration cost, ROIC	Mostly private	Acquirer, target, customers, auditors	12 and 24 months post-close	Not adjudicable if mandatory cohort/accounting data remain private
H4	Workload depth, compute-hours, site power, latency, SLA and supply	Private	Operators, customers, utilities	Four quarters after trigger	Not adjudicable without matched site/workload telemetry
H5	Contract, spend, control-plane and substitution panel	Private/collectable	Enterprises, vendors, procurement owners	Annual panel endpoint	Not adjudicable if frozen recruitment or spend denominator fails
H6	Executed failover, RTO/RPO, output quality, TCO	Private/collectable	Enterprises and providers	12 months after shock/test	Not adjudicable without executed tests and costs
H7	Production share, matched task cost, price premium, layer economics	Mostly private	Enterprises, providers, researchers	2030 panel endpoint	Not adjudicable without matched production and spend panel

Four-factor loading rows are ordered market/capital, physical substrate, platform/enterprise, and geopolitics/policy. H1A [0.70, 0.15, 0.10, 0.10]; H1B [0.45, 0.55, 0.10, 0.20]; H2 [0.45, 0.55, 0.10, 0.20];

H3 [0.35, 0.15, 0.60, 0.10]; H4 [0.10, 0.75, 0.25, 0.15]; H5 [0.15, 0.10, 0.75, 0.20]; H6 [0.05, 0.20, 0.65, 0.35]; H7 [0.10, 0.35, 0.55, 0.40]. Hypothesis-specific normal noise supplies the remaining variance. These loadings affect support-component dependence only.

Table C.2. Conditional support-pattern sensitivity cases and fixed execution settings.

Case	Parameter change	Interpretation
Low support prior	Component probability modes x0.80; factor loadings x0.80	Lower conditional support pattern; not a thesis-null or missing-data scenario.
Base	Frozen Table 12.3 priors, seed 20260724, 200,000 trials, correlation 0.35	Reference conditional support-pattern generator.
High support prior	Component probability modes x1.20; factor loadings x1.20 with residual variance positive	Higher conditional support pattern; not an observed probability.
Sensitivity envelope	400 batches x 4,000 trials; correlation 0.10-0.60; triangular low/mode/high priors	Produces 5th-95th support bands only.

Reproduction sequence: (1) draw four standard-normal factors and hypothesis-specific idiosyncratic errors; (2) construct correlated latent variables from the frozen loading rows; (3) transform to component uniforms with the normal CDF; (4) compare each component with its triangularly sampled probability; (5) apply the hypothesis support Boolean; (6) repeat for base and sensitivity cases; and (7) report conditional support frequency only. Real-world event occurrence, observability, null, mixed and not-adjudicable states are handled by Tables 12.1, 12.1A, 12.2 and C.1. Later parameter changes require a prospective versioned amendment and do not overwrite v1.2.1.

APPENDIX D

Source symmetry and complete revision ledger

The final baseline records both sides of load-bearing sources and every material correction made before the first observation cycle.

Source-symmetry rule

A load-bearing source must be read for supportive, neutral, and disconfirming observations. Selection is allowed only for scope or relevance, and the reason must be recorded. The ledger below is the July 24, 2026 baseline; future reviews append rows rather than silently replacing them.

Source	Supportive evidence used	Counterevidence or qualification used	Treatment
JPMorgan U.S.-China report	U.S. capacity, AI spending and VC; lower Chinese power costs; seven-dimension framework.	U.S. national average about $0.18/kWh with regional variation; DeepSeek V4 claim of 3-6 month frontier lag; Chinese idle-capacity estimate is not an official census.	Both sides appear in Chapter 8; figures remain source-qualified.
NVIDIA FY2026 10-K	Direct customers at 22% and 14%; platform and supplier exposure.	Indirect-customer concentration is estimated; some customers may be >=10%; one AI research company contributed meaningfully through cloud purchases.	Direct and indirect concentration now appear together.

Source	Supportive evidence used	Counterevidence or qualification used	Treatment
Amazon-OpenAI partnership	$50B investment and broader distribution/compute relationship.	$35B is conditional; AWS gains exclusive third-party Frontier distribution; OpenAI commits to 2 GW Trainium and a larger infrastructure agreement.	Investment, distribution and compute are recorded as separate flows.
State Street SPY holdings	NVIDIA is a large index constituent.	Weight changes daily; fund page rounds the dated workbook.	Workbook file, as-of date, retrieval date and rounded page are all named.
SpaceX amended S-1	Anthropic and Google agreements disclose large payments and GPU counts.	Payment-per-GPU arithmetic is not a comparable unit price because scope, generation, ramp and services differ.	Chapter 4 shows the ratios and the denominator caveat.
Cloud Security Alliance note	Reported GPU, HBM, packaging and security concentration.	Market shares draw partly on public and secondary estimates; organization-specific exposure is not audited.	Used as reported estimates, never as a calibrated systemic measure.
Open-weights coalition statement	Broad coalition preference for diffusion, competition and control.	Weights are hard to recall; signatories benefit commercially from more compute, cloud, security and application demand.	Used as evidence of policy and commercial alignment, not realized outcomes.
Stanford AI Index 2026	Narrow U.S.-China capability gap in the cited comparison.	Closed-open gap widened from 0.5% to 3.3% over the cited interval; metrics and dates differ.	No blanket convergence claim is permitted.

Table D.1. Source-symmetry ledger for load-bearing evidence at the final baseline.

Complete revision ledger

The ledger distinguishes factual corrections, newly available events, methodological changes, scope expansions, and production fixes. A correction does not imply that the earlier claim was fabricated; it records why the baseline changed.

Version	Type	Material change	Reason and effect
v1.0	Factual	Restored SpaceX revenue/loss, proceeds-use estimate, independent valuation, actual index weight and $25B bond.	Removed directional selection and created a real equity-to-credit observation.
v1.0	Factual	Corrected Stanford closed/open gap direction and reconciled IPO share arithmetic.	Prevented a directional misstatement and kept offer value separate from reported proceeds.
v1.0	Method	Retired composite systemic scoring; separated financial, macro, operational, strategic and resolvability tests.	Made disaggregation the core contribution and moved rigor into channel thresholds.
v1.0	Method	Added numeric hypothesis thresholds, affirmative nulls, evidence states and baseline dashboard.	Converted broad claims into a prospective monitoring protocol.
v1.0	Structure	Made Chapter 5 the canonical SpaceX-Cursor treatment; retained blocked/abandoned contingency.	Preserved the boundary case without counting repeated references as evidence.
v1.0	Structure	Repaired scenario typology, formula units, open-weight ecosystem, index literature and multipolar cases.	Resolved internal dimensional and coverage gaps.
v1.0	Production	Added full source URLs and citation consistency sweep.	Made the evidence trail reproducible.
v1.1	Scope	Added NVIDIA, Huawei, TSMC/Taiwan, HBM, packaging, data-center commissioning and hardware portability.	Extended the thesis to the physical substrate and build-out ladder.

Version	Type	Material change	Reason and effect
v1.2	Factual	Updated TSMC U.S. announced investment to $265B; replaced 100k GPU baseline; added Anthropic/Google contracts.	Brought hardware facts to the July 24 cutoff and separated historical from current capacity.
v1.2	Taxonomy	Added NVIDIA to issuer testing and split H1 into issuer-specific and common-substrate shocks.	Closed the NVIDIA exemption and separated idiosyncratic from shared shocks.
v1.2	Method	Added six-state adjudication, frozen Boolean rules, support-pattern simulation, related work and observability flags.	Made the protocol independently reproducible while retaining explicit limits.
v1.2.1	Method	Corrected endpoint precedence; overlapping support/null is mixed and missing mandatory evidence is not adjudicable.	Removed the internal support-priority contradiction.
v1.2.1	Method	Added immutable event ledger and hypothesis-level replication rules.	Separated one-case results from general mechanism conclusions.
v1.2.1	Method	Separated support-pattern simulation from event occurrence and evidence availability.	Prevented a clean simulation from implying that private evidence will be observable.
v1.2.1	Method	Completed H1-H7 units, denominators, sample rules, redeployment, demand, RTO and task-cost variables.	Closed remaining executable and dimensional gaps.
v1.2.1	Method	Renamed historical exercise as applicability/rule coverage and added backtest-readiness table.	Removed an unsupported predictive-performance claim.
v1.2.1	Factual	Named dated SPY workbook and exact NVIDIA weight; added indirect-customer context and Amazon-OpenAI terms.	Improved traceability and source symmetry.
v1.2.1	Factual	Displayed Anthropic/Google payment-per-GPU arithmetic with non-comparability caveat.	Kept denominators with headline contract figures.
v1.2.1	Taxonomy	Separated primary frontier firms from material AI-system issuers.	Kept the founding cases coherent while testing infrastructure issuers consistently.
v1.2.1	Structure	Added migrating constraint regimes and source-symmetry ledger.	Reconciled hardware scarcity with grid stress and made evidence selection auditable.
v1.2.1	Governance	Declared final baseline freeze before the October 2026 observation cycle.	Future changes must be prospective and versioned.

Table D.2. Complete factual, methodological, structural and production revision ledger through v1.2.1.

Primary sources and research literature

Sources are grouped alphabetically and display full URLs. Company and coalition statements are used for event facts, disclosed strategy, policy preference, and evidence of commercial alignment; they are not independent validation of economy-wide cost, security, competition, or diffusion outcomes. Appendix D records supportive and countervailing evidence from load-bearing sources. The July 24 coalition statement is preserved in the research archive and linked through a signatory reproduction.

Amazon Web Services. 2026. Understanding intelligent prompt routing in Amazon Bedrock. AWS Documentation. URL: https://docs.aws.amazon.com/bedrock/latest/userguide/prompt-routing.html

Amazon. 2026a. Amazon.com Announces Fourth Quarter Results. Amazon Investor Relations. URL: https://ir.aboutamazon.com/news-release/news-release-details/2026/Amazon-com-Announces-Fourth-Quarter-Results/default.aspx

Amazon. 2026b. OpenAI and Amazon Announce Strategic Partnership. February 27, 2026. URL: https://press.aboutamazon.com/2026/2/openai-and-amazon-announce-strategic-partnership

American Innovators Network et al. 2026. Open Weights and American AI Leadership. Coalition statement, July 24, 2026. Author-supplied archival copy; reproduced by Emergence Capital (accessed July 24, 2026). URL: https://www.emcap.com/thoughts/why-we-signed-an-industry-wide-letter-on-open-weights

Anthropic. 2024. Powering the next generation of AI development with AWS. Anthropic. URL: https://www.anthropic.com/news/anthropic-amazon-trainium

Anthropic. 2026a. Anthropic confidentially submits draft S-1 to the SEC. Anthropic. URL: https://www.anthropic.com/news/confidential-draft-s1-sec

Anthropic. 2026b. Company and governance. Anthropic. URL: https://www.anthropic.com/company

Anthropic. 2026c. Building a managed agent system. Anthropic Engineering. URL: https://www.anthropic.com/engineering/managed-agents

Anthropic. 2026d. Building and evaluating trustworthy agents. Anthropic Research. URL: https://www.anthropic.com/research/trustworthy-agents

Anthropic. 2026e. Higher limits with SpaceX infrastructure. Anthropic. URL: https://www.anthropic.com/news/higher-limits-spacex

Anthropic. 2026f. Detecting and preventing distillation attacks. Anthropic (accessed July 24, 2026). URL: https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks?via=free

Bank for International Settlements. 2021. Passive funds affect prices: evidence from the most ETF-dominated asset classes. BIS Working Papers No. 952 (accessed July 24, 2026). URL: https://www.bis.org/publ/work952.htm

Brynjolfsson, Erik; Danielle Li; and Lindsey R. Raymond. 2023. Generative AI at Work, NBER Working Paper 31161. National Bureau of Economic Research. URL: https://www.nber.org/papers/w31161

Bureau of Industry and Security. 2024. Commerce strengthens export controls to restrict China's capability to produce advanced semiconductors for military applications. U.S. Department of Commerce. URL: https://www.bis.gov/press-release/commerce-strengthens-export-controls-restrict-chinas-capability-produce-advanced-semiconductors-military

Bureau of Industry and Security. 2026. Department of Commerce revises license review policy for semiconductors exported to China. U.S. Department of Commerce. URL: https://media.bis.gov/press-release/department-commerce-revises-license-review-policy-semiconductors-exported-china

City of Phoenix. 2026. TSMC Announces Additional $100 Billion Investment in Arizona. July 16, 2026. URL: https://www.phoenix.gov/newsroom/ced-news/tsmc-announces-additional--100-billion-investment-in-arizona.html

Cloud Security Alliance AI Safety Initiative. 2026. AI Development Stack Concentration Risk. May 3, 2026. URL: https://labs.cloudsecurityalliance.org/research/csa-research-note-ai-stack-concentration-risk-20260503-csa-s/

DeepSeek. 2026. Transparency and model-release materials. DeepSeek (accessed July 24, 2026). URL: https://www.deepseek.com/en/transparency/

Dillon, Eleanor W.; Sonia Jaffe; Nicole Immorlica; and Christopher T. Stanton. 2025. Shifting Work Patterns with Generative AI, NBER Working Paper 33795. National Bureau of Economic Research. URL: https://www.nber.org/papers/w33795

Emergence Capital. 2026. Why We Signed an Industry-Wide Letter on Open Weights. Emergence Capital (accessed July 24, 2026). URL: https://www.emcap.com/thoughts/why-we-signed-an-industry-wide-letter-on-open-weights

EuroHPC Joint Undertaking. 2026. EuroHPC JU's mandate expanded under new regulation amendment. European High Performance Computing Joint Undertaking. URL: https://www.eurohpc-ju.europa.eu/eurohpc-jus-mandate-expanded-under-new-regulation-amendment-2026-01-20_en

European Central Bank. 2024. Financial Stability Review, May 2024. ECB (accessed July 24, 2026). URL: https://www.ecb.europa.eu/press/financial-stability-publications/fsr/html/ecb.fsr202405~7f212449c8.en.html

European Commission. 2023. IPCEI Next Generation Cloud Infrastructure and Services. European Commission (accessed July 24, 2026). URL: https://competition-policy.ec.europa.eu/state-aid/ipcei/approved-ipceis/cloud_en

European Commission. 2025. Memorandum of Understanding on AI Gigafactories. European Commission (accessed July 24, 2026). URL: https://digital-strategy.ec.europa.eu/en/library/memorandum-understanding-ai-gigafactories

European Commission. 2026a. AI Continent Action Plan delivers major milestones. European Commission (accessed July 24, 2026). URL: https://digital-strategy.ec.europa.eu/en/news/ai-continent-action-plan-delivers-major-milestones

European Commission. 2026b. IPCEI candidates in the Design Support Hub: Artificial Intelligence and Compute Infrastructure Continuum. European Commission (accessed July 24, 2026). URL: https://competition-policy.ec.europa.eu/state-aid/ipcei/design-support-hub_en

Federal Energy Regulatory Commission. 2026a. FERC launches aggressive, targeted action to speed large-load integration. FERC. URL: https://www.ferc.gov/news-events/news/ferc-launches-aggressive-targeted-action-speed-large-load-integration

Federal Energy Regulatory Commission. 2026b. Commissioner Rosner's remarks on large-load show-cause orders. FERC. URL: https://www.ferc.gov/news-events/news/commissioner-rosners-remarks-large-load-show-cause-orders-e-7-e-12-june-18-2026

Federal Reserve Board. 2014. Are Concerns About Leveraged ETFs Overblown? Finance and Economics Discussion Series. Board of Governors of the Federal Reserve System (accessed July 24, 2026). URL: https://www.federalreserve.gov/econres/feds/are-concerns-about-leveraged-etfs-overblown.htm

Federal Reserve Board. 2023. New Insights from Form N-CEN: Liquidity Management at Open-End Funds and Primary Market Concentration of ETFs. FEDS Notes. Board of Governors of the Federal Reserve System (accessed July 24, 2026). URL: https://www.federalreserve.gov/econres/notes/feds-notes/new-insights-from-n-cen-liquidity-management-at-open-end-funds-and-pmc-of-etfs-20230111.html

Financial Stability Board. 2024. The Financial Stability Implications of Artificial Intelligence. FSB. URL: https://www.fsb.org/2024/11/fsb-assesses-the-financial-stability-implications-of-artificial-intelligence/

Financial Stability Board. 2025. Monitoring Adoption of Artificial Intelligence and Related Vulnerabilities in the Financial Sector. FSB. URL: https://www.fsb.org/2025/10/monitoring-adoption-of-artificial-intelligence-and-related-vulnerabilities-in-the-financial-sector/

Financial Stability Board. 2026. Sound Practices for the Responsible Adoption of Artificial Intelligence: Consultation Report. FSB. URL: https://www.fsb.org/2026/06/sound-practices-for-responsible-adoption-of-artificial-intelligence-ai-consultation-report/

Financial Times. 2026a. Nvidia and Palantir urge U.S. not to ban open AI models after China scare. Financial Times, July 24, 2026 (accessed July 24, 2026). URL: https://www.ft.com/content/3203fc9a-2321-44f8-8093-b7e16c8fc6d7

Huawei. 2025a. Groundbreaking SuperPoD Interconnect: Leading a New Paradigm for AI Infrastructure. Huawei Connect keynote. URL: https://www.huawei.com/en/news/2025/9/hc-xu-keynote-speech

Huawei. 2026. 2025 Annual Report. Huawei. URL: https://www.huawei.com/en/annual-report/2025

IndiaAI. 2026a. IndiaAI Compute Portal. Government of India (accessed July 24, 2026). URL: https://compute.indiaai.gov.in/login

IndiaAI. 2026b. AIKosh: India's national platform for AI datasets, models and tools. Government of India (accessed July 24, 2026). URL: https://aikosh.indiaai.gov.in/home/about-us/

International AI Safety Report. 2026. International AI Safety Report 2026. International AI Safety Report. URL: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

International Energy Agency. 2025a. Energy and AI: Executive Summary. IEA. URL: https://www.iea.org/reports/energy-and-ai/executive-summary

International Energy Agency. 2025b. Energy demand from AI. IEA. URL: https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai

International Energy Agency. 2026. Key Questions on Energy and AI: Executive Summary. IEA. URL: https://www.iea.org/reports/key-questions-on-energy-and-ai/executive-summary

International Labour Organization. 2026a. The impact of GenAI on jobs, productivity and work organization: a review of the empirical evidence. ILO. URL: https://www.ilo.org/publications/impact-genai-jobs-productivity-and-work-organization-review-empirical

International Labour Organization. 2026b. Gen AI, occupational segregation and gender equality in the world of work. ILO. URL: https://www.ilo.org/publications/gen-ai-occupational-segregation-and-gender-equality-world-work

International Labour Organization. 2026c. The aggregation paradox of AI: why do micro-economic productivity gains from AI disappear at scale. ILO. URL: https://researchrepository.ilo.org/esploro/outputs/encyclopediaEntry/The-aggregation-paradox-of-AI-why/995700572102676

International Monetary Fund. 2024. Global Financial Stability Report, Chapter 3: Advances in Artificial Intelligence: Implications for Capital Markets. October 2024. URL: https://www.imf.org/-/media/files/publications/gfsr/2024/october/english/ch3.pdf

JPMorgan Chase Center for Geopolitics. 2026. Beyond the Benchmarks: A Systemic View of U.S.-China AI Competition. May 2026. URL: https://www.jpmorganchase.com/content/dam/jpmorganchase/documents/center-for-geopolitics/cfg-us-china-report.pdf

Meta. 2025. Llama model and license materials. Meta AI (accessed July 24, 2026). URL: https://www.llama.com/

Microsoft. 2026. Model router for Microsoft Foundry concepts. Microsoft Learn. URL: https://learn.microsoft.com/en-us/azure/foundry/openai/concepts/model-router

Mistral AI. 2026. Under which license are Mistral's open models available? Mistral Help Center (accessed July 24, 2026). URL: https://help.mistral.ai/en/articles/347393-under-which-license-are-mistral-s-open-models-available

Morningstar. 2026. SpaceX IPO analysis and fair value estimate. Morningstar (accessed July 24, 2026). URL: https://www.morningstar.com/en-us/business/insights/research/spacex-ipo-analysis

Nasdaq. 2026a. Space Exploration Technologies Corporation to join the Nasdaq-100 Index beginning July 7, 2026. Nasdaq. URL: https://www.nasdaq.com/press-release/space-exploration-technologies-corporation-join-nasdaq-100-indexr-beginning-july-7

Nasdaq. 2026b. 2025 Nasdaq-100 reconstitution and performance highlights. Nasdaq. URL: https://www.nasdaq.com/articles/global-indexes/2025-nasdaq-100-reconstitution-and-performance-highlights

Nasdaq. 2026c. SpaceX makes history, raising $85 billion through Nasdaq listing. Nasdaq. URL: https://www.nasdaq.com/newsroom/spacex-makes-history-raising-85-billion-through-nasdaq-listing

National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST. URL: https://www.nist.gov/itl/ai-risk-management-framework

National Institute of Standards and Technology. 2024. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST. URL: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence

New Constructs. 2026. SpaceX (SPCX): Going Boldly Where No One Has Gone Before. New Constructs (accessed July 24, 2026). URL: https://www.newconstructs.com/spacex-spcx-going-boldly-where-no-one-has-gone-before/

NVIDIA. 2024. NVIDIA Ethernet Networking Accelerates World's Largest AI Supercomputer, Built by xAI. NVIDIA Newsroom. URL: https://nvidianews.nvidia.com/news/spectrum-x-ethernet-networking-xai-colossus

NVIDIA. 2025a. Quarterly Report on Form 10-Q for quarter ended October 26, 2025; strategic-investment commitments. U.S. Securities and Exchange Commission. URL: https://www.sec.gov/Archives/edgar/data/1045810/000104581025000230/nvda-20251026.htm

NVIDIA. 2026a. Annual Report on Form 10-K for fiscal year ended January 25, 2026. U.S. Securities and Exchange Commission. URL: https://www.sec.gov/Archives/edgar/data/1045810/000104581026000021/nvda-20260125.htm

NVIDIA. 2026b. NVIDIA GB300 NVL72. NVIDIA. URL: https://www.nvidia.com/en-us/data-center/gb300-nvl72/

NVIDIA. 2026c. NVIDIA Vera Rubin Ramps Into Full Production to Power Agentic AI Factories Worldwide. NVIDIA Newsroom. URL: https://nvidianews.nvidia.com/news/vera-rubin-full-production-agentic-ai-factory

NVIDIA. 2026d. NVIDIA Unveils Vera, the CPU for Agents. NVIDIA Newsroom. URL: https://nvidianews.nvidia.com/news/nvidia-unveils-vera-the-cpu-for-agents

OpenAI. 2025a. Evolving OpenAI's structure. OpenAI. URL: https://openai.com/index/evolving-our-structure/

OpenAI. 2025b. Introducing Stargate UAE. OpenAI (accessed July 24, 2026). URL: https://openai.com/index/introducing-stargate-uae/

OpenAI. 2025c. Stargate advances with 4.5 GW partnership with Oracle. OpenAI. URL: https://openai.com/index/stargate-advances-with-partnership-with-oracle/

OpenAI. 2025d. OpenAI and NVIDIA announce strategic partnership to deploy 10 gigawatts of NVIDIA systems. September 22, 2025. URL: https://openai.com/index/openai-nvidia-systems-partnership/

OpenAI. 2026a. OpenAI submits confidential S-1. OpenAI. URL: https://openai.com/index/openai-submits-confidential-s-1/

OpenAI. 2026b. OpenAI raises $122 billion to accelerate the next phase of AI. OpenAI. URL: https://openai.com/index/accelerating-the-next-phase-ai/

OpenAI. 2026c. Introducing OpenAI Frontier. OpenAI. URL: https://openai.com/index/introducing-openai-frontier/

OpenAI. 2026d. Introducing OpenAI Presence. OpenAI. URL: https://openai.com/index/introducing-openai-presence/

OpenAI. 2026e. Frontier Alliance partners. OpenAI. URL: https://openai.com/index/frontier-alliance-partners/

OpenAI. 2026f. Introducing the OpenAI Partner Network. OpenAI. URL: https://openai.com/index/introducing-openai-partner-network/

OpenAI. 2026g. Scaling AI for everyone. February 27, 2026. URL: https://openai.com/index/scaling-ai-for-everyone/

OpenAI. 2026h. OpenAI and Amazon announce strategic partnership. February 27, 2026. URL: https://openai.com/index/amazon-partnership/

Orrick. 2026. Anthropic raises $65 billion Series H at $965 billion post-money valuation. Orrick. URL: https://www.orrick.com/en/News/2026/05/Anthropic-Raises-65B-Series-H-at-965B-Post-Money-Valuation

Public Investment Fund. 2025. HRH Crown Prince launches HUMAIN as global AI powerhouse. PIF. URL: https://www.pif.gov.sa/en/news-and-insights/press-releases/2025/hrh-crown-prince-launches-humain-as-global-ai-powerhouse/

Qwen Team. 2025. Qwen3: Think Deeper, Act Faster. Qwen (accessed July 24, 2026). URL: https://qwenlm.github.io/blog/qwen3/

Reuters. 2026. Wall Street warms to SpaceX ahead of Nasdaq-100 inclusion. Reuters, July 7, 2026 (accessed July 24, 2026). URL: https://www.reuters.com/business/wall-street-warms-spacex-ahead-nasdaq-100-inclusion-2026-07-07/

S&P Dow Jones Indices. 2025a. Annual Survey of Assets. S&P Global. URL: https://investorfactbook.spglobal.com/sp-dow-jones-indices/sp-dow-jones-indices-annual-survey-of-assets/

S&P Dow Jones Indices. 2026a. Consultation on treatment of MegaCap companies: results. S&P Global. URL: https://press.spglobal.com/2026-06-04-S-P-Dow-Jones-Indices-Consultation-on-Treatment-of-MegaCap-Companies-Results

S&P Dow Jones Indices. 2026b. S&P 500 index facts and characteristics. S&P Global. URL: https://www.spglobal.com/spdji/tc/indices/equity/sp-500/

Sarvam AI. 2026. Sovereign-model and open-weight program materials. Sarvam AI (accessed July 24, 2026). URL: https://www.sarvam.ai/

Semiconductor Industry Association. 2025. Emerging Resilience in the Semiconductor Supply Chain. SIA. URL: https://www.semiconductors.org/emerging-resilience-in-the-semiconductor-supply-chain/

SpaceX. 2026a. Space Exploration Technologies Corp. Announces Pricing of Initial Public Offering. SpaceX Investor Relations. URL: https://ir.spacex.com/updates/releases-details/2026/Space-Exploration-Technologies-Corp--Announces-Pricing-of-Initial-Public-Offering/default.aspx

SpaceX. 2026b. Space Exploration Technologies Corp. Announces Closing of Initial Public Offering, Including Full Exercise of Underwriters' Option. SpaceX Investor Relations. URL: https://ir.spacex.com/updates/releases-details/2026/Space-Exploration-Technologies-Corp--Announces-Closing-of-Initial-Public-Offering-Including-Full-Exercise-of-Underwriters-Option-to-Purchase-Additional-Shares-2026-RgoR-Y1Vwh/default.aspx

SpaceX. 2026c. SpaceX Announces Pricing of $25 Billion Inaugural Bond Issuance. SpaceX Investor Relations. URL: https://ir.spacex.com/updates/releases-details/2026/SpaceX-Announces-Pricing-of-25-Billion-Inaugural-Bond-Issuance-2026-33VwNgsx3O/default.aspx

SpaceX. 2026d. Current Report on Form 8-K: Agreement and Plan of Merger with Anysphere, Inc. U.S. Securities and Exchange Commission. URL: https://www.sec.gov/Archives/edgar/data/1181412/000162828026043411/spaceexplorationtechnologi.htm

SpaceX. 2026e. Disclosure Summary: UK Retail Offer of Shares in SpaceX. U.S. Securities and Exchange Commission. URL: https://www.sec.gov/Archives/edgar/data/1181412/000162828026040874/spacexukfwp.htm

SpaceX. 2026f. Prospectus: Space Exploration Technologies Corp. U.S. Securities and Exchange Commission. URL: https://www.sec.gov/Archives/edgar/data/1181412/000162828026042639/spaceexplorationtechnologi.htm

SpaceX. 2026g. Free Writing Prospectus: Cloud Service Agreement with Google LLC. U.S. Securities and Exchange Commission. URL: https://www.sec.gov/Archives/edgar/data/1181412/000162828026041150/spacexagreementfwp.htm

Stanford Institute for Human-Centered AI. 2026a. AI Index 2026: Technical Performance. Stanford HAI. URL: https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance

Stanford Institute for Human-Centered AI. 2026b. AI Index 2026: Research and Development. Stanford HAI. URL: https://hai.stanford.edu/ai-index/2026-ai-index-report/research-and-development

Stanford Institute for Human-Centered AI. 2026c. AI Index 2026 Report. Stanford HAI. URL: https://hai.stanford.edu/ai-index/2026-ai-index-report%C2%A0

State Street Global Advisors. 2026a. SPY daily holdings workbook, as of July 23, 2026. File: holdings-daily-us-en-spy.xlsx; NVIDIA weight 7.93528%. Retrieved July 24, 2026. URL: https://www.ssga.com/library-content/products/fund-data/etfs/us/holdings-daily-us-en-spy.xlsx

State Street Global Advisors. 2026b. State Street SPDR S&P 500 ETF Trust fund page, fund top holdings as of July 23, 2026. Retrieved July 24, 2026. URL: https://www.ssga.com/us/en/individual/etfs/state-street-spdr-sp-500-etf-trust-spy

TSMC. 2025a. TSMC Intends to Expand Its Investment in the United States to US$165 Billion to Power the Future of AI. TSMC. URL: https://pr.tsmc.com/english/news/3210

TSMC. 2025b. 2024 Annual Report, operational-risk discussion. TSMC. URL: https://investor.tsmc.com/static/annualReports/2024/english/index.html

TSMC. 2026a. 2025 Annual Report. TSMC. URL: https://investor.tsmc.com/static/annualReports/2025/english/index.html

U.S. Department of Energy. 2024. DOE releases new report evaluating increase in electricity demand from data centers. DOE. URL: https://www.energy.gov/articles/doe-releases-new-report-evaluating-increase-electricity-demand-data-centers

U.S. Department of Justice and Federal Trade Commission. 2023. Merger Guidelines. U.S. Department of Justice. URL: https://www.justice.gov/atr/merger-guidelines

U.S. Energy Information Administration. 2026. Data centers are a major driver of U.S. electricity demand growth in AEO2026. EIA. URL: https://www.eia.gov/pressroom/releases/press587.php

UAE Government. 2025. AI-enabled digital transformation via public-private-people partnership. The Official Platform of the UAE Government. URL: https://u.ae/en/information-and-services/business/public-private-people-partnership/pppp/AI-enabled-digital-transformation-via-PPPP

When We Crash. 2026. A Systemic-Risk Monitoring Framework for High-Valuation Markets. Working paper, May 2026. Non-peer-reviewed. URL: https://www.whenwecrash.com/paper