

Quantifying Systemic Vulnerability in the Foundation Model Industry

Claudio Pirrone*

Stefano Fricano

Gioacchino Fazio

Conference Paper - SIEPI (29-30 January 2026) - Bari

Abstract

The foundation model industry exhibits unprecedented concentration in critical inputs: semiconductors, energy infrastructure, elite talent, capital, and training data. Despite extensive sectoral analyses, no comprehensive framework exists for assessing overall industrial vulnerability. We develop the Artificial Intelligence Industrial Vulnerability Index (AIIVI) grounded in O-Ring production theory (Kremer, 1993), recognizing that foundation model production requires simultaneous availability of non-substitutable inputs. Given extreme data opacity and rapid technological evolution, we implement a validated human-in-the-loop methodology using large language models to systematically extract indicators from dispersed grey literature, with complete human verification of all outputs. Applied to six state-of-the-art foundation model developers, AIIVI equals 0.82, indicating extreme vulnerability driven by compute infrastructure (0.85) and energy systems (0.90). While industrial policy currently emphasizes semiconductor capacity, energy infrastructure represents the emerging binding constraint. This methodology proves applicable to other fast-evolving, opaque industries where traditional data sources are inadequate.

Keywords: Foundation Models, Industrial Vulnerability, O-Ring Production, Composite Indices, Human-in-the-Loop Methodology

1 Introduction

The Transformer architecture (Vaswani et al., 2017) enabled a fundamental shift in artificial intelligence through foundation models: large-scale models trained on extensive datasets using self-supervised learning, adaptable to diverse downstream tasks (Schneider et al., 2024). Following the November 2022 release of ChatGPT, adoption accelerated dramatically. By mid-2025, 26.5% of internet users aged 16 and above used ChatGPT monthly, representing 557 million users (We Are Social and Meltwater, 2025).

This rapid growth raises fundamental questions about industrial sustainability. The foundation model industry attracts a substantial share of global venture capital investment,

*Corresponding author. Email: claudio.pirrone@unipa.it

raising concerns about bubble dynamics (Fang et al., 2025). More critically, the industry’s opacity and velocity create severe challenges for economic analysis. Traditional data sources prove inadequate. Companies disclose minimal operational details, market structures shift monthly, and current state-of-the-art models were all released after September 2025. This creates a methodological challenge for researchers seeking to assess an industry where traditional econometric data are missing, relevant information exists only in dispersed grey literature, and the time scale of change makes manual collection obsolete before completion.

We address this challenge through three contributions. First, we develop AIIVI, a composite vulnerability index grounded in O-Ring production theory (Kremer, 1993). Second, we demonstrate a transferable methodology for constructing composite indices in data-scarce environments through human-in-the-loop AI assistance. Third, we provide the first systematic assessment of supply chain vulnerabilities across the foundation model upstream value chain, identifying energy infrastructure, not just semiconductors, as the emerging binding constraint.

The paper proceeds as follows. Section 2 develops our theoretical framework applying O-Ring production theory to foundation model vulnerability. Section 3 defines AIIVI formally. Section 4 details our human-in-the-loop methodology and validation protocol. Section 5 presents pilot estimation results. Section 6 discusses findings and policy implications. Section 7 concludes with limitations and future research directions. Supplementary materials provide complete documentation of the estimation process.

2 Theoretical Framework

Kremer (1993) formalized how production processes requiring multiple complementary inputs create multiplicative output functions where the weakest component determines performance. A single defective O-ring destroyed the Space Shuttle Challenger; similarly, a single missing input can halt foundation model production entirely. This framework, originally explaining income inequality across countries with different skill levels, applies directly to industrial vulnerability when production requires simultaneously available, non-substitutable inputs.

Foundation model production exhibits extreme complementarity. Advanced semiconductors, massive energy infrastructure, elite talent, substantial capital, and quality training data cannot compensate for each other but must all be present. If Taiwan’s semiconductor fabrication capacity becomes unavailable, abundant energy elsewhere provides zero production capacity. If datacenters lack power, stockpiled GPUs remain idle. This differs fundamentally from manufacturing processes where labor can partially substitute for capital, or quality can trade against quantity.

We analyze vulnerability at the upstream value chain of foundation models for three reasons. Foundation models constitute the core technology enabling all downstream applications, so their vulnerability directly affects the entire AI ecosystem. Semiconductor vulnerability exists independently of AI; analyzing foundation model-specific vulnerabilities isolates AI industry risks. Demand-side application vulnerability depends on foundation model availability; supply-side constraints represent the binding factor. Based on industry analyses (Competition and Markets Authority, 2023; Maslej et al., 2025), we specify a

general production function:

$$FM_{output} = f(\text{Compute, Data, Talent, Capital, Energy}) \quad (1)$$

These inputs are identified from the Competition and Markets Authority value chain analysis (Competition and Markets Authority, 2023), augmented with capital (Fang et al., 2025) and energy (de Vries, 2023; Maslej et al., 2025) based on emerging bottlenecks.

Foundation model production requires massive computational resources, creating dependency on concentrated upstream markets. The supply chain exhibits extreme concentration at multiple levels. NVIDIA holds 86% of the AI accelerator market, TSMC controls 70% of leading-edge semiconductor fabrication, and 92% of advanced chip manufacturing capacity is located in Taiwan (Maslej et al., 2025; Papadopoulos and Magafas, 2025). This creates both market power through NVIDIA’s pricing dominance and geopolitical vulnerability through Taiwan dependence. The capital and expertise required to build competing foundries or design competing GPUs reinforces incumbent market power (Miller, 2022).

Market participants recognize this concentration as a binding constraint. In October 2025, Anthropic announced a diversified compute strategy across three platforms: Google’s TPUs, Amazon’s Trainium, and NVIDIA’s GPUs, with access to up to one million TPUs. This strategic diversification by a major player holding 31.4% market share signals industry awareness of compute vulnerability. However, success remains uncertain. If alternative architectures fail to match NVIDIA performance at scale, concentration will persist or intensify. The announcement represents market search for solutions rather than solution achievement.

High-quality training data has become a contested resource. Villalobos and colleagues (2024) estimate frontier models are approaching limits of available public text data, with continuous improvement facing diminishing marginal returns (Maslej et al., 2025). Model autophagy, training on AI-generated content, raises quality concerns (Xing et al., 2025). In response, foundation model labs have adopted hybrid strategies combining publicly available information, proprietary third-party data, data-labeling services, user-contributed data, and internally generated data (Anthropic, 2025). This reduces fresh data constraints but introduces new vulnerabilities through legal challenges to web scraping, dependency on expensive specialist labor for reinforcement learning with human feedback, and potential quality degradation from synthetic data.

Foundation models are designed by a small elite group of researchers. Exact talent availability data are scarce; Maslej and colleagues (2025) rely on LinkedIn profiles as proxies, suffering selection bias. However, indirect evidence suggests acute scarcity. Reported salaries reaching multi-million dollar levels for top engineers signal superstar dynamics (Rosen, 1981). Geographic concentration is extreme: MacroPolo tracking indicates 57% of top-tier AI researchers work in the United States (Maslej et al., 2025), with declining international mobility creating national bottlenecks. Demand for AI talent extends beyond foundation model labs to downstream application developers, reducing availability for core research.

Frontier foundation model production represents one of the most capital-intensive R&D processes in history. OpenAI reportedly spent over \$100 million training GPT-4, while burning \$2.5 billion in the first half of 2025 against \$4.3 billion in revenue (Maslej et al., 2025). This capital intensity exceeds traditional venture capital capacity, creating dependency on strategic partners, particularly cloud hyperscalers. Microsoft invested over \$10 billion in

OpenAI; Google and Amazon heavily funded Anthropic. These strategic dependencies create vulnerability as cloud providers gain leverage over foundation model labs' strategic decisions, potentially creating conflicts of interest. The financial sustainability of current business models remains uncertain, evidenced by valuation multiples approaching dotcom-bubble levels.

De Vries (2023) projects that maintaining current AI capacity expansion requires approximately 1.5 million AI server units annually by 2027, demanding 85 to 134 terawatt-hours of electricity. Maslej and colleagues (2025) confirm that while energy efficiency improves, total consumption and CO₂ emissions grow steadily. Industry analysis indicates US datacenter energy consumption grows faster than power availability, creating a binding physical constraint. This energy vulnerability operates on multiple timescales: short-term grid capacity constraints, medium-term permitting and construction delays, and long-term climate policy and carbon regulation. Unlike other inputs, energy infrastructure cannot be rapidly scaled. Anthropic's planned infrastructure deployments will bring well over a gigawatt of capacity online in 2026, highlighting the magnitude of energy requirements extending beyond processors to encompass cooling systems and datacenter facilities requiring substantial power infrastructure as operations scale.

International production networks typically exhibit some substitutability through production splintering and geographic arbitrage (Baldwin and Venables, 2013; Gereffi et al., 2005). Foundation model production permits no such arbitrage. Taiwan's 92% share of leading-edge fabrication capacity cannot be substituted on timescales relevant to business strategy or industrial policy, as foundries require over \$20 billion and five years to construct. NVIDIA's 86% GPU market share reflects not just first-mover advantage but extreme technical barriers. Designing competitive AI accelerators requires years of architectural innovation and extensive software ecosystem development that cannot be purchased or imitated quickly.

This creates novel vulnerabilities. Traditional supply chain risk management employs geographic diversification, multi-sourcing, and strategic stockpiling (Sheffi, 2005). These strategies assume either substitutability or storability. Foundation model inputs violate both assumptions. Semiconductors are not substitutable across fabrication facilities; TSMC's 3nm process cannot be replicated by Samsung's 3nm process due to different manufacturing approaches. Computational infrastructure is not storable; datacenter capacity cannot be stockpiled like rare earth minerals.

The policy implications are profound. If inputs are complements, not substitutes, then industrial policy must address all constraints simultaneously. The US CHIPS Act addresses semiconductor vulnerability through \$52 billion in fabrication subsidies. However, if energy infrastructure cannot support expanded datacenter operations, domestic chip production yields minimal gains in AI production capacity. Conversely, China's strategy of massive grid investment to enable larger, less-efficient compute clusters addresses energy constraints but cannot overcome semiconductor limitations imposed by US export controls. Both strategies implicitly treat inputs as substitutes when the underlying production function requires complementarity.

Foundation models exhibit canonical two-sided platform characteristics (Rochet and Tirole, 2003): they connect developers and end-users, exhibit indirect network effects, and demonstrate increasing returns to scale. However, unlike earlier digital platforms that scaled on venture capital and advertising revenue, foundation model platforms require capital in-

tensity comparable to semiconductor manufacturing or pharmaceutical R&D. This capital intensity creates strategic dependencies absent from Facebook or Google’s development. Foundation model labs cannot self-finance from operating revenue. Our empirical results show they spend 48% of four-year projected revenue on annual infrastructure investment. Traditional venture capital cannot provide the over \$10 billion required for frontier model development and deployment.

This has produced an unprecedented industrial structure: strategic investments from cloud infrastructure providers (Microsoft \$13.8 billion into OpenAI, Amazon \$8.0 billion into Anthropic) that simultaneously supply computational resources and compete in downstream application markets. Economic theory would predict vertical integration to resolve such conflicts of interest (Teece et al., 1997). Instead, we observe incomplete integration creating bilateral dependencies. AI labs depend on cloud providers for capital and compute infrastructure, with switching costs requiring massive re-engineering given proprietary hardware and multi-year service contracts. Cloud providers depend on AI labs for competitive differentiation, with Microsoft’s Azure growth substantially driven by OpenAI integration and Amazon’s Bedrock platform depending on Anthropic’s models.

The foundation model industry emerges amid deliberate technological decoupling between the United States and China. Export controls on advanced semiconductors, restrictions on AI talent mobility, and competing technical standards fragment what was initially a global industry. Farrell and Newman (2019) term this weaponized interdependence, where the United States leverages chokepoints in global technology networks to deny competitors access to critical capabilities. This fragmentation interacts with O-Ring production structure in theoretically novel ways. If successful AI production requires five complementary inputs, and geopolitical competition creates two separate technology ecosystems, then competition becomes a race to achieve simultaneous capability across all five inputs within each ecosystem.

The velocity of technological change in foundation models creates fundamental challenges for empirical research. All current state-of-the-art models were released in a three-month window from September to November 2025. Traditional data sources operate on annual or quarterly cycles. The AI Index Report (Maslej et al., 2025), the most comprehensive industry assessment, uses data typically six to twelve months old at publication. Regulatory filings provide minimal operational detail. OpenAI, Anthropic, and xAI are private companies with no disclosure requirements.

Relevant information exists but is dispersed across grey literature: company technical reports, academic preprints, industry analyst briefs, conference presentations. Manual collection is infeasible at the scale required for systematic industry analysis. Our pilot required screening over 60 sources across multiple languages and formats, with information changing weekly. Traditional scholarly approaches yield analysis that is rigorous but obsolete before publication.

Recent research documents both opportunities and risks in AI-assisted scholarship. Hao and colleagues (2025) demonstrate AI tools expand researcher productivity by automating literature review and data extraction, but may narrow topical focus by biasing researchers toward well-documented areas. Loru and colleagues (2025) warn of epistemia, the substitution of human judgment by AI outputs, particularly concerning when AI systems exhibit overconfidence in uncertain domains. Our human-in-the-loop methodology addresses these

concerns through complete verification protocols where every AI output receives human verification against original sources at a rate of 100%, not a sample. AI serves as tool for information retrieval and preliminary organization; all substantive judgments about source credibility, data interpretation, and theoretical framing remain human-controlled.

3 The Artificial Intelligence Industrial Vulnerability Index

The qualitative analysis in Section 2 clarifies that the five inputs cannot be modeled as perfect substitutes. While marginal trade-offs exist (for example, China’s strategy of using abundant energy to partially compensate for compute constraints through larger, less efficient clusters), these elasticities of substitution are local, not global. Crucially, if supply of any single input falls below a minimum viability threshold, aggregate output collapses regardless of other inputs’ abundance. This weakest link dynamic is precisely the phenomenon O-Ring production functions capture (Kremer, 1993). We therefore adopt the multiplicative formulation:

$$AIIVI = 1 - \prod_{i=1}^5 F_i^{\alpha_i} \quad \text{with} \quad \sum_{i=1}^5 \alpha_i = 1 \quad (2)$$

where F_i denotes the anticipated capacity of input i ’s supply chain to sustain production, bounded in $[0, 1]$. If $F_i = 0$ for any single input, AIIVI reaches maximum vulnerability even if all other inputs are abundant. From the complementarity between force and vulnerability, where $F_i = 1 - V_i$, we can rewrite as:

$$AIIVI = 1 - \prod_{i=1}^5 (1 - V_i)^{\alpha_i} \quad (3)$$

where V_i is the vulnerability sub-index for input i , and α_i represents that input’s criticality weight.

Each V_i aggregates multiple raw indicators. We adopt weighted arithmetic averages at this level for two reasons. Components within each sub-index exhibit lower complementarity than across inputs; data quality and training costs are substitutable in ways that compute and energy are not. Arithmetic averages are less sensitive to weight specification errors than geometric means, a practical advantage given measurement uncertainty. Normalized components ensure $V_i \in [0, 1]$ and allow combining indicators with different units.

Compute vulnerability captures three concentration dimensions in semiconductor supply chains:

$$V_C = w_{C1} \cdot SC_{comp} + w_{C2} \cdot GEO_{comp} + w_{C3} \cdot PR_{comp} \quad (4)$$

where $SC_{comp} = \sqrt{HHI_{foundries}}$ for the leading-edge semiconductor fabrication market, GEO_{comp} is the percentage of leading-edge capacity in the most concentrated geographic region, and $PR_{comp} = \sqrt{HHI_{GPU/ASIC}}$ for the AI accelerator design market. All components are naturally bounded in $[0, 1]$.

Data vulnerability proxies data scarcity through foundation model reliability and training costs:

$$V_D = w_{D1} \cdot R_{data} + w_{D2} \cdot C_{data} \quad (5)$$

where $R_{data} = (100 - \text{AAII})/100$ with AAII being the Artificial Analysis Intelligence Index aggregating ten standard benchmarks, and C_{data} is the weighted average of training cost relative to AI revenue across state-of-the-art players, capped at 1.0 if the ratio exceeds 1 or revenue is missing. The reliability score is inverted because higher reliability indicates lower data vulnerability.

Talent vulnerability captures concentration in human capital and research output:

$$V_T = w_{T1} \cdot E_{talent} + w_{T2} \cdot RC_{talent} + w_{T3} \cdot IP_{talent} \quad (6)$$

where $E_{talent} = \sqrt{HHI_{elite}}$ of elite AI researchers by corporate affiliation, RC_{talent} is the percentage of top-tier conference publications from the top five corporate labs, and $IP_{talent} = \sqrt{HHI_{patents}}$ of granted AI patents.

Capital vulnerability aggregates financial and strategic dependencies:

$$V_K = w_{K1} \cdot DEP_{capital} + w_{K2} \cdot VS_{capital} + w_{K3} \cdot SU_{capital} + w_{K4} \cdot RI_{bubble} \quad (7)$$

where $DEP_{capital} = \sqrt{HHI_{cloud}}$ of strategic capital from cloud providers, $VS_{capital}$ is the share of global venture capital devoted to state-of-the-art foundation model players, $SU_{capital}$ is the weighted average of CAPEX divided by four times annual revenue (capped at 1.0), and RI_{bubble} is the weighted average valuation-to-revenue multiple normalized by the Yahoo! 1999 benchmark of 80x. The four-year denominator in $SU_{capital}$ reflects typical compute infrastructure lifespan.

Energy vulnerability captures consumption growth and efficiency dynamics:

$$V_E = w_{E1} \cdot GR_{energy} + w_{E2} \cdot GR_{co2} + w_{E3} \cdot EB_{energy} \quad (8)$$

where GR_{energy} is the ratio of AI energy consumption growth to global energy production growth, normalized per the following rules: if the ratio exceeds 1 and AI growth is positive, the value is set to 1.0; if the ratio is negative and AI growth is positive, the value is set to 1.0; if the ratio is negative and AI growth is negative, the value is set to 0; otherwise the value equals the ratio. These rules reflect that growth rates exceeding production capacity signal maximum vulnerability. Similarly, GR_{co2} is the ratio of AI CO₂ emissions growth to global emissions growth with analogous normalization rules. The parameter EB_{energy} measures deceleration in energy efficiency improvements, calculated as $1 - \text{Improvement}_{t+1}/\text{Improvement}_t$. This parameter is set to 0 if efficiency accelerates (negative result) and set to 1 if efficiency improvement becomes negative.

4 Methodology

Constructing AIIVI requires 15 distinct parameters, each demanding multiple data sources. Manual collection faces two fundamental constraints. Volume is substantial, requiring screen-

ing over 50 sources across company reports, academic papers, industry analyst briefs, government databases, and news outlets. Obsolescence is rapid, with information changing weekly such that manual collection becomes outdated before completion. Traditional solutions are inadequate. Annual data collection cannot track an industry where all state-of-the-art models were released in a three-month window. Expert surveys suffer from lag and subjective bias. Official statistics do not exist.

We develop a three-stage protocol combining AI capacity with rigorous human verification. In the first stage, AI-assisted information extraction, a large language model executes structured web searches based on pre-specified search protocols detailed in Supplementary Material 3. For each parameter, we require a minimum of three searches with cross-validation from independent sources. The AI extracts relevant data points, performs calculations, and documents reasoning via chain-of-thought analysis. In the second stage, complete human verification, every single AI output is human-verified against original sources. Verification checks include source authenticity (real URLs, reputable publishers), data accuracy (numbers match source documents), calculation correctness (formulas applied properly), and reasoning validity (chain-of-thought logic is sound). Discrepancies trigger AI re-query or manual override. In the third stage, validation and documentation, all sources are catalogued with full citations in Supplementary Material 2. We maintain a complete audit trail including original prompts, AI responses, human feedback, and final values in Supplementary Material 1. Confidence ratings are assigned based on data quality and verification certainty.

The validity of human-in-the-loop methodology rests on three principles. AI serves as tool, not authority, performing information retrieval and preliminary organization while all substantive judgments about source credibility, data interpretation, and weight selection remain human-controlled. Complete verification eliminates hallucination risk; unlike unsupervised AI usage, our protocol requires human verification of every output at a rate of 100%. Transparency enables replication; full documentation in supplementary materials allows independent verification of sources, calculations, and AI reasoning quality.

We tested three current state-of-the-art models: Gemini 3 Pro (highest benchmark performance), Claude Sonnet 4.5 (lowest hallucination rate), and DeepSeek V3.2 (open-source reference). Testing revealed distinct patterns. Gemini 3 Pro showed excellent reasoning but frequent temporal hallucinations, citing future or non-existent sources. DeepSeek V3.2 demonstrated good source retrieval but occasional calculation errors. Claude Sonnet 4.5 proved most reliable in source finding and instruction following with minimal hallucinations. Based on validation testing, Claude Sonnet 4.5 was selected using verification reliability, not reasoning sophistication, as the choice criterion.

Loru and colleagues (2025) warn of epistemia, the substitution of human judgment by AI systems. Our methodology explicitly guards against this through several mechanisms. There is no black-box delegation; all AI reasoning is exposed via chain-of-thought and verified. Researchers maintain human override authority and can reject AI outputs regardless of AI confidence. Substantive decisions remain human-controlled; weight selection, framework design, and interpretation all remain human decisions. Cross-validation is mandatory; AI cannot simply confirm its own outputs but requires independent source corroboration.

To assess methodology robustness, we implement three checks. Monte Carlo sensitivity analysis uses 10,000 simulations with plus or minus 20% uniform distribution on all parameters, assessing AIIVI stability under measurement uncertainty. Weighting sensitivity com-

compares AI-suggested weights based on substitutability logic against equal weights, assessing structural sensitivity. Confidence-weighted analysis excludes parameters with medium-low confidence ratings, assessing robustness to data quality concerns.

5 Results

We focus on foundation models announced since September 2025, ensuring currency. Independent observers including Artificial Analysis and LifeArchitect identified seven state-of-the-art models from six developers. All are reasoning models integrating self-generated data and reinforcement learning with human feedback procedures. The models are Claude Opus 4.5 and Claude Sonnet 4.5 from Anthropic, Gemini 3 Pro from Google DeepMind, GPT-5.1 from OpenAI, Kimi K2 Thinking from Moonshot AI, DeepSeek-V3.2 from DeepSeek-AI, and Grok 4 from xAI. Notably absent are Meta’s Llama 4 (reliability below current state-of-the-art), Mistral AI models (similar reasons), and Generalist’s GEN-0 (insufficient public data). Data collection concluded November 30, 2025.

Information sources span multiple categories: academic journals including Nature, Frontiers, arXiv, and Stanford HAI; research institutions including Epoch AI and Pew Research; agencies including IEA, BCG, and SIA; financial institutions including KPMG, Deloitte, and EY; industry analysts including TrendForce, Jon Peddie Research, and Gartner; news outlets; and company announcements. Total sources verified across 15 parameters: 66.

Table 1 presents parameter values with human-assessed confidence levels. All 15 parameters proved computable: 11 based on empirical data, 2 on mixed estimation, 1 on estimated data, and 1 on 2024 proxy data.

Table 1: Raw Parameter Estimation with Confidence Assessment

Parameter	Value	Nature	Confidence
SC_{comp}	0.708	Empirical	High
GEO_{comp}	0.920	Empirical	High
PR_{comp}	0.868	Empirical	High
R_{data}	0.354	Empirical	High
C_{data}	0.049	Mixed	Medium-High
E_{talent}	0.458	Estimated	Medium-Low
RC_{talent}	0.720	2024 Proxy	High
IP_{talent}	0.376	Empirical	High
$DEP_{capital}$	0.650	Empirical	Medium-High
$VS_{capital}$	0.190	Empirical	Medium-High
$SU_{capital}$	0.480	Mixed	Medium
RI_{bubble}	0.520	Empirical	Medium
GR_{energy}	1.000	Empirical	High
GR_{co2}	1.000	Empirical	High
EB_{energy}	0.000	Empirical	Low

Key observations emerge from these results. Compute and energy parameters show critical vulnerability, all exceeding 0.85 with high confidence. The value of $EB_{energy} = 0$ reflects

the Blackwell architecture’s 3.75-fold efficiency breakthrough over previous generation H100 chips; this parameter may oscillate given innovation cycles in processor development, but currently indicates accelerating rather than decelerating efficiency improvements. Talent concentration measured by E_{talent} has lower confidence due to reliance on indirect indicators such as salaries and geographic distribution rather than direct researcher counts by corporate affiliation. Financial parameters including $SU_{capital}$ and RI_{bubble} suffer from private company opacity but remain calculable with reasonable confidence based on disclosed funding rounds and revenue estimates.

Weight selection is critical for O-Ring indices. We employ a two-tier approach. Sub-index weights reflect substitutability among components within each vulnerability dimension. We estimate substitutability rates under human-in-the-loop supervision, with explicit instruction to justify choices based on economic logic. Criticality weights reflect the relative importance of each input in the production function. Following Kremer (1993), we frame this as: what is the impact on foundation model production if input i is reduced by 50%? This thought experiment yields rankings coherent with O-Ring theory.

Table 2 presents the complete weight structure. For compute vulnerability, geographic concentration receives the highest weight (0.50) because zero short-term geographic substitutes exist, while semiconductor manufacturing (0.25) and GPU design (0.25) have limited alternatives. For data vulnerability, model quality (0.80) dominates because it cannot be fixed by spending alone, while costs (0.20) can be reduced via efficiency. For talent vulnerability, elite researchers (0.50) take years to train, research concentration (0.35) allows knowledge diffusion, and patents (0.15) matter less in fast-moving AI. For capital vulnerability, strategic capital (0.40) creates lock-in effects, operating sustainability (0.35) is fundamental, valuations (0.15) are adjustable, and venture capital (0.10) is most substitutable. For energy vulnerability, grid capacity (0.50) cannot be substituted short-term, regulatory constraints (0.40) are hard to substitute, and efficiency (0.10) can substitute for quantity.

Applying equations (3) through (7) with weights from Table 2 yields sub-index vulnerabilities. Table 3 presents final AIIVI under two scenarios. Scenario A uses substitutability and criticality weights from Table 2. Scenario B employs uniform weights within and across sub-indices.

Weight sensitivity analysis shows that Scenario A (0.823) and Scenario B (0.549) differ significantly, confirming weight choice matters. However, both indicate high vulnerability. Scenario B at 0.55 indicates high vulnerability, while Scenario A at 0.82 indicates extreme vulnerability. The qualitative conclusion that the foundation model industry faces critical supply chain risks holds across specifications.

Monte Carlo robustness with plus or minus 20% parameter uncertainty yields informative results. For Scenario A, the 95% confidence interval [0.763, 0.868] falls entirely within the very high to extreme range. For Scenario B, the 95% confidence interval [0.467, 0.626] spans the moderate-high to high range. The narrow confidence interval width for Scenario A (0.105) indicates stability under measurement uncertainty.

Confidence-weighted analysis provides a critical robustness check. Excluding the two parameters with below medium confidence (E_{talent} with medium-low confidence and EB_{energy} with low confidence) and renormalizing weights reveals the severity of energy constraints. For talent vulnerability, excluding E_{talent} and renormalizing remaining components yields adjusted weights of 0.70 for RC_{talent} and 0.30 for IP_{talent} , producing $V_T^{adj} = 0.617$. For

Table 2: Weight Structure: Substitutability and Criticality

Component	Weight	Rationale
<i>Compute Vulnerability (V_C):</i>		
SC_{comp}	0.25	Alternative foundries exist
GEO_{comp}	0.50	Zero geographic substitutes
PR_{comp}	0.25	Limited GPU alternatives
<i>Data Vulnerability (V_D):</i>		
R_{data}	0.80	Quality fundamental
C_{data}	0.20	Costs reducible
<i>Talent Vulnerability (V_T):</i>		
E_{talent}	0.50	Long training time
RC_{talent}	0.35	Knowledge diffuses
IP_{talent}	0.15	Patents less critical
<i>Capital Vulnerability (V_K):</i>		
$DEP_{capital}$	0.40	Strategic lock-in
$VS_{capital}$	0.10	Most substitutable
$SU_{capital}$	0.35	Cash flow fundamental
RI_{bubble}	0.15	Valuations adjustable
<i>Energy Vulnerability (V_E):</i>		
GR_{energy}	0.50	Grid capacity binding
GR_{co2}	0.40	Regulatory constraints
EB_{energy}	0.10	Efficiency substitutes
<i>Input Criticality (α_i):</i>		
Compute	0.40	No scaling without chips
Energy	0.30	Cannot run clusters
Capital	0.15	Consolidation possible
Talent	0.10	Knowledge diffuses
Data	0.05	Currently sufficient

Table 3: AIIVI Results with Monte Carlo Sensitivity Analysis

Metric	Scenario A Weighted	Scenario B Equal
<i>Sub-Index Vulnerabilities:</i>		
V_C (Compute)	0.854	0.832
V_D (Data)	0.293	0.202
V_T (Talent)	0.537	0.518
V_K (Capital)	0.525	0.460
V_E (Energy)	0.900	0.667
AIIVI	0.823	0.549
<i>Monte Carlo (10,000 simulations):</i>		
Mean	0.819	0.547
Standard Deviation	0.028	0.042
95% CI Lower	0.763	0.467
95% CI Upper	0.868	0.626

energy vulnerability, excluding EB_{energy} and renormalizing yields adjusted weights of 0.556 for GR_{energy} and 0.444 for GR_{co2} . Since both parameters equal 1.000 (maximum vulnerability, capped), this produces $V_E^{adj} = 1.000$. The corresponding force value $F_E^{adj} = 0.000$ represents absolute energy constraint. Under the O-Ring multiplicative structure, any force component equal to zero drives AIIVI to maximum vulnerability regardless of other inputs. The confidence-weighted specification therefore yields $AIIVI = 1.000$, demonstrating that when restricted to highest-quality parameters, the energy bottleneck alone is sufficient to create extreme systemic vulnerability. This strengthens rather than weakens our findings, as the low-confidence efficiency parameter ($EB_{energy} = 0.000$) was the only factor moderating energy vulnerability.

6 Discussion

The foundation model industry exhibits extreme vulnerability with a weighted AIIVI of 0.823, driven primarily by compute (0.854) and energy (0.900) constraints. Compute vulnerability stems from three compounding concentrations: geographic concentration where Taiwan hosts 92% of leading-edge capacity, market concentration where NVIDIA holds 86% of AI accelerators, and fabrication concentration where TSMC controls 70% of leading-edge semiconductor manufacturing. These vulnerabilities are structural; new foundries require over \$20 billion investment and three to five years construction (Miller, 2022).

Energy vulnerability reflects AI consumption growing at 15% annually while global production grows at 4%, a 3.75-fold differential. CO₂ emissions from AI grow at 12% annually versus 0.8% for global emissions, a 15-fold differential. Both ratios exceed unity and trigger normalization rules, capping GR_{energy} and GR_{co2} at 1.000 to indicate maximum vulnerability. Grid capacity cannot scale at AI industry rates on relevant policy timescales. Crucially,

these vulnerabilities are complements per the O-Ring structure in equation (2). Abundant chips without power, or abundant power without chips, both yield near-zero output. Current industrial policy treats them as substitutes, which represents a fundamental error.

Market response to compute concentration provides both encouragement and caution. Anthropic’s October 2025 announcement of a diversified compute strategy across Google TPUs, Amazon Trainium, and NVIDIA GPUs, with planned deployment of million-chip footprints across both AWS and Google Cloud, demonstrates industry awareness and active search for solutions. This symmetrical hyperscale deployment across two clouds and two accelerator families is extraordinarily rare and signals that leading players recognize compute concentration as a binding constraint requiring strategic mitigation. However, success remains uncertain. The ability of alternative architectures to match NVIDIA performance at frontier scale is unproven. If TPU and Trainium deployments fail to deliver competitive training efficiency or encounter unforeseen technical limitations, the industry may revert to deeper NVIDIA dependence, potentially increasing rather than decreasing concentration. The announcement represents market search for alternatives rather than achievement of viable substitutes.

In contrast, energy vulnerability lacks comparable market adjustment mechanisms. Grid capacity expansion requires government coordination, regulatory approvals, and multi-year construction timelines that private actors cannot control. Anthropic’s planned gigawatt-scale deployments in 2026 underscore energy requirements that far exceed current infrastructure, yet no market mechanism exists to accelerate grid build-out at the required pace. While compute concentration may diminish through architectural diversification if market efforts succeed, energy constraints are structurally persistent absent policy intervention. This asymmetry in market adjustment capacity justifies prioritizing energy infrastructure in policy design despite compute receiving higher criticality weight (0.40 versus 0.30) in our framework. The market is attempting to solve compute; only policy can address energy.

Talent at 0.537 and capital at 0.525 show moderate-high vulnerability. Research concentration with RC_{talent} at 0.72 indicates 72% of publications from the top five corporate labs, which is high but less severe than compute and energy. Elite talent concentration with E_{talent} at 0.458 is moderate, consistent with knowledge diffusion through publication and education pipelines. Capital vulnerability reflects strategic dependencies with $DEP_{capital}$ at 0.65 on cloud providers and high burn rates with $SU_{capital}$ at 0.48, but financial markets remain liquid. Valuation multiples with RI_{bubble} at 0.52 approach concerning levels at 52% of the dotcom peak but have not reached extremes.

Data shows lowest vulnerability at 0.293. Model reliability, measured by the Artificial Analysis Intelligence Index version 3.0, averages 64.6 out of 100 across state-of-the-art models, adequate for current deployment though leaving substantial room for improvement. Training costs average 4.9% of revenue, manageable for well-funded players. Synthetic data strategies, while raising long-term quality concerns (Loru et al., 2025; Xing et al., 2025), currently enable continued improvement. However, this may reflect measurement limitations rather than genuine abundance. The intensive use of synthetic data and reinforcement learning with human feedback is recent, dating from after 2024; long-run effects on model quality remain uncertain.

Current industrial policy exhibits asymmetric focus. US and EU semiconductor investment (CHIPS Act \$52 billion, Chips for Europe 43 billion euros) addresses compute vulner-

ability while neglecting energy infrastructure. Chinese strategy emphasizes grid investment, accepting higher compute vulnerability. Both strategies are incomplete. The O-Ring structure implies balanced investment is optimal. With compute criticality at 0.40 and energy at 0.30, and given market mechanisms addressing compute while lacking for energy, marginal returns to reducing energy vulnerability currently exceed returns to reducing compute vulnerability for western economies.

Practical recommendations emerge from this analysis. For energy infrastructure, immediate priorities include fast-tracking datacenter power projects, accelerating renewable deployment specifically for AI infrastructure, and considering small modular reactors for dedicated AI energy supply. For compute diversification over the medium term, priorities include supporting Samsung and Intel foundry expansion to reduce geographic concentration, incentivizing geographic distribution of fabrication capacity beyond Taiwan, and funding AI accelerator diversity through support for AMD and Intel alternatives to reduce NVIDIA dominance. Ongoing efficiency mandates should require datacenters to achieve power usage effectiveness (the ratio of total facility energy to IT equipment energy) below 1.2 and mandate public reporting of floating-point operations per watt and CO₂ per token. For long-term talent development, priorities include expanding PhD programs and streamlining immigration for AI researchers, coupled with industrial capacity to avoid exacerbating brain drain.

Beyond substantive findings, this work demonstrates human-in-the-loop methodology’s viability for constructing composite indices in data-scarce environments. The approach generalizes to other fast-evolving, opaque industries including cryptocurrency, quantum computing, and synthetic biology. Key lessons include the necessity of complete verification at 100% of AI outputs rather than sampling, the value of chain-of-thought documentation for validation transparency, the importance of conservative model selection prioritizing reliability over sophistication, and the continued need for human domain expertise to recognize plausible errors. The methodology’s efficiency gains are substantial. Screening 66 sources across 15 parameters required approximately 40 human-hours with human-in-the-loop methodology versus an estimated 200-plus hours manually, a fivefold efficiency gain enabling near-real-time updating while maintaining verification rigor.

7 Conclusion

This paper makes three contributions. Substantively, we demonstrate extreme supply chain vulnerability (AIIVI equals 0.82) driven by compute and energy constraints. Current industrial policy focused solely on semiconductors misses the emerging energy bottleneck, while market mechanisms show potential to address compute concentration through architectural diversification if ongoing efforts by major players succeed. Methodologically, human-in-the-loop AI assistance enables rigorous composite index construction in opaque, fast-moving industries through complete verification protocols. Theoretically, applying O-Ring production functions to vulnerability measurement formalizes why balanced policy intervention matters when inputs are complements.

Several limitations point toward future research directions. Data quality remains a concern; several parameters rely on estimated rather than reported data. As foundation model

labs mature and disclosure improves, measurement precision will increase. Future iterations should incorporate more granular data as it becomes available. Weight specification, while employing economic logic for substitutability in sub-index weights and criticality in input weights, would benefit from expert panel validation. We view current weights as preliminary; future work should convene industry experts for formal Delphi-style weighting procedures.

Temporal dynamics present another challenge. The index represents a snapshot from November 2025. Vulnerability patterns may shift as technologies evolve, such as if energy efficiency breakthroughs change energy vulnerability or if alternative GPU architectures reduce compute vulnerability. Quarterly updates would track these dynamics effectively. Geographic disaggregation offers another research direction. Current analysis treats western and Chinese strategies asymmetrically in discussion but not in measurement. Future work should construct region-specific AIIVI variants to formalize these differences.

Downstream effects remain unexplored. We analyze upstream foundation model production; downstream application vulnerability dependent on foundation model availability requires investigation. Extension to the full value chain would provide comprehensive risk assessment. Methodological extensions should test the human-in-the-loop protocol on other industries including quantum computing and synthetic biology to assess generalizability. Comparative analysis across different large language models would establish robustness to model choice.

The AI industry’s vulnerability profile resembles historical parallels: oil dependence created 1970s energy shocks, semiconductor concentration enabled Japanese industrial policy in the 1980s, and internet infrastructure vulnerability shaped dotcom dynamics. In each case, concentrated, complementary inputs created systemic fragility. The foundation model industry differs in one critical aspect: time scale. Previous transformations evolved over decades; foundation model capabilities improve monthly. This compression of industrial evolution timescales makes traditional policy responses too slow. AIIVI provides a framework for near-real-time vulnerability assessment, enabling adaptive policy responses.

More fundamentally, this work demonstrates how AI can enable research in environments where traditional methods fail, not by replacing human judgment but by augmenting human capacity. As Hao and colleagues (2025) document, this augmentation comes with risks of narrowed focus. Our human-in-the-loop protocol offers one approach to mitigating these risks while capturing benefits. The broader challenge remains: as industries evolve faster than traditional data systems can track, researchers face a choice between timely but unverified analysis or rigorous but obsolete analysis. Methodologies like human-in-the-loop offer a third path that is rigorous, verified, and timely, but require institutional acceptance of AI-assisted research with proper validation protocols. The foundation model industry itself provides both the motivation and the means for this methodological evolution.

References

- Anthropic (2025). System card: Claude opus 4.5. Technical report.
- Baldwin, R. and Venables, A. J. (2013). Splintering and the unbundling of production and tasks. *Journal of Economic Perspectives*, 27(3):99–118.

- Competition and Markets Authority (2023). Ai foundation models: Initial report. Technical report.
- de Vries, A. (2023). The growing energy footprint of artificial intelligence. *Joule*, 7(10):2191–2194.
- Fang, X., Tao, L., and Li, Z. (2025). Anchoring ai capabilities in market valuations: The capability realization rate model and valuation misalignment risk. arXiv:2505.10590.
- Farrell, H. and Newman, A. L. (2019). Weaponized interdependence: How global economic networks shape state coercion. *International Security*, 44(1):42–79.
- Gereffi, G., Humphrey, J., and Sturgeon, T. (2005). The governance of global value chains. *Review of International Political Economy*, 12(1):78–104.
- Hao, Y., Cao, H., Li, Z., et al. (2025). Artificial intelligence tools expand scientists’ impact but contract science’s focus. *Nature*.
- Kremer, M. (1993). The o-ring theory of economic development. *The Quarterly Journal of Economics*, 108(3):551–575.
- Loru, E., Nudo, J., Di Marco, N., et al. (2025). The simulation of judgment in llms. *Proceedings of the National Academy of Sciences*, 122(42):e2518443122.
- Maslej, N., Fattorini, L., Perrault, R., et al. (2025). The ai index 2025 annual report. Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.
- Miller, C. (2022). *Chip War: The Fight for the World’s Most Critical Technology*. Simon and Schuster.
- Papadopoulos, G. D. and Magafas, L. (2025). The global arms and integrated circuits trade through complex network analysis. *Social Network Analysis and Mining*, 15(1):69.
- Rochet, J.-C. and Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4):990–1029.
- Rosen, S. (1981). The economics of superstars. *The American Economic Review*, 71(5):845–858.
- Schneider, J., Meske, C., and Kuss, P. (2024). Foundation models. *Business & Information Systems Engineering*, 66(2):221–231.
- Sheffi, Y. (2005). *The Resilient Enterprise*. MIT Press.
- Teece, D. J., Pisano, G., and Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, 18(7):509–533.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., and Hobbhahn, M. (2024). Will we run out of data? limits of llm scaling based on human-generated data. arXiv:2211.04325.

We Are Social and Meltwater (2025). Digital 2026: Global overview report. Technical report.

Xing, X., Shi, F., Huang, J., Wu, Y., Nan, Y., Zhang, S., Fang, Y., Roberts, M., Schönlieb, C.-B., Del Ser, J., and Yang, G. (2025). On the caveats of ai autophagy. *Nature Machine Intelligence*, 7(2):172–180.