The Economics of Knowledge in the Age of LLMs

Alec Litowitz* Nick Polson† Vadim Sokolov‡

First Draft: February 2026
This Draft: February 25, 2026

Abstract

This essay examines the economics of knowledge through the lens of large language models—the most consequential knowledge technology since the printing press. It synthesizes contributions from Kenneth Arrow, Friedrich Hayek, Sherwin Rosen, Ronald Coase, George Stigler, Claude Shannon, and John von Neumann into a unified framework for understanding what has changed, what the numbers reveal, and what the physics ultimately permits. The central findings are six: (1) knowledge and information are economically distinct; (2) Hayek's knowledge problem remains unsolved for tacit knowledge; (3) the Rosen superstar mechanism operates with unusual force in AI markets; (4) Coasian firm structure is being bifurcated; (5) capital structure follows knowledge structure; and (6) information processing requires energy as a matter of physical law.

Keywords: large language models, economics of information, knowledge economy, Arrow information paradox, Hayek knowledge problem, superstar economics, scaling laws, Landauer's principle

1 Introduction

Large language models are the most consequential knowledge technology since the printing press. Understanding their economic implications requires synthesizing frameworks from Arrow, Hayek, Rosen, Coase, Stigler, Shannon, and von Neumann—each of whom identified a distinct mechanism by which knowledge resists efficient market allocation. This essay constructs that synthesis. A companion paper [Litowitz et al., 2026] studies the physics of AI through the observation that photons are to energy what tokens are to information, developing the supply-and-demand balance sheet for global token production. Here the focus is economics. Six findings emerge.

*QStar Capital. https://www.qstarcap.com
†University of Chicago. ngp@chicagobooth.edu
‡George Mason University. vsokolov@gmu.edu

I. Knowledge is not information. Shannon's information theory measures uncertainty reduction; Arrow's economics of knowledge concerns the market failures that arise when information is the commodity being traded. LLMs operate at the intersection: they compress Shannon information at unprecedented scale and simultaneously exhibit every Arrow market failure in amplified form.

II. Hayek's knowledge problem remains unsolved. The explicit, documentable portion of human knowledge has been successfully centralized by LLMs. The tacit, local, time-sensitive knowledge that actually coordinates economic activity has not been captured and cannot be, without losing the properties that make it valuable. This is not an engineering limitation—it is an epistemological one, supported by Kolmogorov complexity arguments and Hurwicz's mechanism design theorems.

III. The Rosen superstar mechanism operates with unusual force. Small capability differences produce enormous revenue differences because the joint consumption technology is near-perfect. This creates winner-take-most dynamics in consumer markets and a quality premium in enterprise markets: Claude's 19–30 million monthly users generate roughly one-quarter of ChatGPT's revenue despite a fraction of the user base, implying a revenue-per-user multiple of roughly three to six times—explained by market positioning on the demand curve.

IV. The Coasian theory of the firm is being bifurcated. As LLMs reduce costs for codifiable tasks, the optimal firm boundary shifts outward—producing "one-person unicorns" at one extreme. But the knowledge-based theory predicts that tacit, relational knowledge becomes *more* concentrated inside large organizations at the other extreme. The firms under maximum pressure are those in the middle, whose primary value was coordinating codifiable work.

V. Capital structure follows knowledge structure. Modigliani–Miller does not apply to firms whose primary assets are inalienable human knowledge and model weights that depreciate through competition rather than use. The equity-heavy, cash-burning, perpetual-reinvestment structure of OpenAI and Anthropic is not irrational speculation—it is the only rational capital structure for this kind of firm, derivable directly from Arrow's information asymmetry results applied to the balance sheet.

VI. Information processing requires energy as a matter of physical law. Landauer's principle establishes the thermodynamic floor; Chinchilla scaling laws establish the empirical trajectory. Global LLM token consumption is doubling roughly every eight months. Even with aggressive hardware efficiency improvements, Jevons' paradox means LLM energy demand will likely reach tens to hundreds of terawatt-hours annually by 2030.

The unified message: knowledge is not free at any level of analysis—not economically, not institutionally, not thermodynamically. Every framework examined here arrives at the same conclusion through a different route.

2 From Information to Knowledge

2.1 Three Levels of a Single Problem

To analyze the economics of knowledge, it is necessary first to distinguish three concepts that are often conflated: data, information, and knowledge. They exist at different levels of abstraction and have radically different economic properties.

Data is the raw substrate—text tokens, sensor readings, transaction records. It has the economic character of a commodity: storable, replicable, transmissible. The cost of storing a gigabyte has fallen by a factor of tens of millions since 1980, making data effectively free to hold.

Information, in Shannon's formalization, is the reduction of uncertainty. It is measured in bits, defined by surprise—a message that tells you what you already knew has zero information content, regardless of how many characters it contains. Shannon's framework is about transmission, not meaning.

Knowledge is information that has been absorbed into a structure—that can be recalled, recombined, applied across contexts, and connected to other knowledge. A database contains data. A weather model contains information about atmospheric states. A meteorologist who integrates model output with local geographical knowledge possesses knowledge. The difference is not merely quantitative. Knowledge has a productive character that data and information lack: its value is inseparable from the context in which it is applied.

2.2 Shannon, Entropy, and the Measure of Ignorance

Shannon [1948] gave information theory its quantitative foundation. The key quantity is entropy:

$$H = -\sum_i p_i \log_2 p_i, \tag{1}$$

where p_i is the probability of the i-th possible message. Entropy measures the average uncertainty in a probability distribution. A fair coin has entropy of 1 bit. A biased coin that comes up heads 99% of the time has entropy close to zero. Shannon entropy has exactly the same mathematical form as Boltzmann's thermodynamic entropy—a connection that Jaynes [1957] argued reflects a deep identity: both are instances of the same maximum-entropy inference principle. Large Language Models are trained to calculate those p_i.

2.3 Arrow's Information Paradox

Kenneth Arrow's insight is that information markets are structurally defective in a way that standard goods markets are not. The deeper problem is that exchange is paradoxical: the buyer wants to know the value of information before paying, but the only way to determine value is to receive the information—at which point the buyer has no reason to pay. The market fails not because of externalities but because of a logical impossibility in the transaction structure.

This paradox manifests in the market for expertise, for research, for education, and most acutely in the market for AI model capabilities—you cannot know what a model can do in your specific context without deploying it. Every market responds through institutional workarounds—credentials, reputation, trial periods—none of which fully resolves the underlying problem.

2.4 LLMs as Knowledge Infrastructure

The printing press reduced the cost of copying knowledge. The telegraph and internet reduced the cost of transmitting it. Libraries and search engines reduced the cost of retrieving it—dramatically lowering the search costs that Stigler [1961] identified as a central friction in information markets. **LLMs reduce the cost of synthesizing it**—of taking knowledge from multiple domains and recombining it into a response tailored to a specific query. Stigler showed that buyers and sellers expend real resources discovering prices and evaluating alternatives; LLMs compress these search costs toward zero for any query whose answer exists in the training corpus.

This synthesis function was previously performed only by expert human knowledge workers, at costs reflecting years of training and high ongoing salaries. The LLM performs a similar function at a cost of fractions of a cent per query. The supply curve for knowledge synthesis has shifted so dramatically that the new equilibrium price approaches zero for a vast range of tasks.

2.5 LLMs as Compression Engines

A frontier LLM trained on a substantial fraction of the internet fits in a few hundred gigabytes—roughly the storage required for three days of global weather data. The disproportion is so extreme that it demands explanation, and the explanation reveals something fundamental about the nature of human knowledge itself: most of it is redundant. The same customer relationship described in a Salesforce record reappears in a Slack thread, a Google Doc, an email chain, and a slide deck. The same love story that Euripides staged has been restaged in every literature since. The same analytical frameworks are restated—with variations in notation but not in substance—across thousands of textbooks. An LLM compresses all of this redundancy away. If the model already encodes a pattern, encountering it again barely moves the weights. What remains after training is the essential structure—the *principles*. Knowledge of principles frees one from the necessity of knowing many facts, and the LLM literalizes this ancient insight.

In mathematics, a *basis* is the minimal set of independent elements from which every element in a space can be constructed by linear combination. LLM parameters encode something analogous: a learned basis of linguistic, logical, and factual patterns from which an enormous range of outputs can be generated through recombination. The Greeks wrote many of the basis functions of Western thought—tragedy, comedy, syllogistic reasoning, axiomatic geometry, democratic governance—and subsequent civilizations have largely been writing linear combinations. This is Romer's growth theory in a different language: the knowledge stock A in equation (16) expands by recombining existing designs, and the LLM's learned basis is a compressed representation of that stock. Most "new" ideas are not new dimensions; they are new coordinates in an existing space.

But compression sacrifices precision. Ask an LLM for the opening hours of a specific Starbucks and it will flounder, because that fact either never appeared in training or appeared too rarely to survive compression. General principles compress beautifully (low Kolmogorov complexity); specific, time-local facts do not (high Kolmogorov complexity). The engineering response—retrieval-augmented generation, pioneered commercially by Perplexity and now standard in ChatGPT, Claude, and Gemini—is to combine the compressed basis with a real-time factual index. The LLM provides structure; search provides specifics. Neither alone suffices. Together they approximate the combination of deep understanding and current awareness that previously required an expensive human expert.

The deeper implication is about human knowledge itself. The compression ratio is so favorable—petabytes of text into hundreds of gigabytes of parameters—because genuinely novel ideas are scarce. If human knowledge were mostly original, compression would fail catastrophically. That it succeeds so spectacularly is itself evidence for the Romer worldview: growth is driven by recombination, and the space of truly independent basis functions is far smaller than the volume of text that elaborates them.

3 Arrow's Economics of Knowledge

3.1 The 1962 Paper: Three Fundamental Failures

Arrow's "Economic Welfare and the Allocation of Resources for Invention" [Arrow, 1962b] established three reasons why the market for knowledge production is inherently imperfect. They interact and compound each other, and no institutional arrangement resolves all three simultaneously.

Indivisibility. Knowledge comes in discrete units—an invention, a discovery, a proved theorem. Unlike wheat or labor, you cannot continuously vary the quantity of a specific piece of knowledge. The choice is not how much of a discovery to produce but whether to produce it at all, requiring a comparison of total costs and total benefits rather than marginal analysis.

Inappropriability. Even when knowledge is produced, the producer cannot capture all the value it generates. Knowledge spills over to competitors, to adjacent industries, to future researchers who build on it. Social rates of return to R&D investment consistently run at two to four times private rates of return [Jones and Williams, 1998]—a structural feature, not a rounding error.

Uncertainty. Research outcomes are genuinely unknown before resources are committed. Risk-averse firms underinvest relative to the expected value calculation; investors cannot price research assets without knowing their outcomes; adverse selection—the mechanism Akerlof [1970] formalized for quality uncertainty—drives up the cost of research capital.

These three failures explain the institutional landscape of knowledge production: patent systems, government research funding, research universities, and open-source development. None is optimal; each is a second-best response to one or more of the three failures, typically at the cost of the others.

3.2 Arrow–Debreu: Complete Markets and Their Failure for Knowledge

Arrow and Debreu [1954] proved the existence of competitive equilibrium under general conditions. The fundamental welfare theorems—established separately—state that if markets are complete and competitive, any equilibrium allocation is Pareto efficient. Knowledge markets fail the completeness requirement in multiple ways. Future knowledge goods cannot be traded in advance because their character is unknown before production. The Arrow information paradox means spot markets for knowledge goods fail the basic transaction structure required for market clearing.

Arrow's [1964] paper introduced Arrow–Debreu securities—assets paying off in exactly one state of the world and zero in all others. A complete set of S states requires S independent securities to span the full risk space. Research risk is not merely uncertain but **unspannable**: you cannot write an Arrow security paying off in the state "the cure for Alzheimer's is discovered" because this state is not specifiable before the discovery. This unspannability explains why venture capital takes the form it does—an open-ended claim on an unspecifiable outcome space structured as equity.

3.3 Learning by Doing: Formal Specification

Arrow's [1962a] learning-by-doing model specifies formally how knowledge accumulates as a byproduct of production. Let $A(t)$ denote the cumulative knowledge stock, driven by cumulative gross investment $G(t)$:

$$A(t) = G(t)^{\varphi}, \quad 0 < \varphi < 1. \tag{2}$$

Each firm's labor-augmenting productivity is proportional to $A(t)$. Since $A(t)$ is a public good, social and private returns diverge.[1]

$$r^{\text{social}} = \varphi \cdot A^{-1} \cdot \frac{\partial G}{\partial t} \cdot \sum_i \frac{\partial y_i}{\partial A} > r^{\text{private}} = \varphi \cdot A^{-1} \cdot \frac{\partial G}{\partial t} \cdot \frac{\partial y_j}{\partial A}. \tag{3}$$

Since the sum over all firms exceeds the individual firm's marginal product, private investment falls below the social optimum by the full magnitude of the knowledge spillover externality. For LLMs, this learning curve operates simultaneously at three levels: inference costs falling with hardware scale, model capability improving with training compute, and organizational productivity improving with deployment experience.

[1] The following notation synthesizes Arrow's argument; the specific rate-of-return expressions are not in the original paper.

4 Hayek and the Dispersal of Knowledge

4.1 The 1945 Paper and the Central Claim

Hayek's "The Use of Knowledge in Society" [Hayek, 1945] is one of the most consequential papers in economics—a point he sharpened in his Nobel lecture on "The Pretence of Knowledge" [Hayek, 1974]. It is more precisely an argument about epistemology—about the nature and location of the knowledge that economic coordination requires.

Hayek's central claim: the real economic problem facing any society is the utilization of knowledge that is *not* given to anyone in its totality. The knowledge relevant to economic coordination is dispersed across millions of individuals, much of it inarticulate, and no central authority can aggregate it without catastrophic loss—not because of computational limits but because the knowledge does not exist in a form that can be transmitted to a center at all.

4.2 Scientific vs. Local Knowledge

Hayek drew a sharp distinction between scientific knowledge—general principles, theoretical frameworks, specifications that can be written down and transmitted—and local knowledge—knowledge of particular circumstances of time and place that is ephemeral, contextual, and often not conscious. This distinction anticipates Polanyi's [1966] concept of tacit knowledge: "we can know more than we can tell."

The price system aggregates local knowledge through decentralized action with stakes. When drought reduces wheat supply in Kansas, the resulting price increase transmits this information to millers in New York who have never heard of Kansas droughts—all without anyone having to understand or articulate the underlying knowledge. The price is a sufficient statistic for the local knowledge it encodes.

LLMs are the most ambitious attempt in history to centralize dispersed knowledge. The achievement is genuine for scientific knowledge. The hard wall of Hayek's epistemology remains intact for local, tacit knowledge.

4.3 The Grossman–Stiglitz Paradox

The most direct mathematical treatment of Hayek's price-as-information-aggregator comes from Grossman and Stiglitz [1980]. Their result: if prices fully reflect all available information, there is no incentive for anyone to acquire information to trade on. But if no one acquires information, prices cannot aggregate it. The equilibrium requires prices to be *imperfectly* informative. Let λ denote the fraction of informed traders and c the cost of becoming informed:

$$\lambda^* : E[\text{profit from private information}] = c, \quad \text{Var}(\theta \mid P) > 0. \tag{4}$$

As $c \to 0$, λ^* rises and prices become more informative—but never fully so. The paradox: perfect informational efficiency is unattainable because reaching it would destroy the incentive to acquire the very signals that make prices efficient. For LLMs, the implication is indirect but important: if AI-generated analysis substitutes for costly private research, and if traders treat it as a free signal correlated with the true state, the equilibrium resembles a reduction in effective c. More traders become "informed" in a superficial sense, yet genuine private signals—the ones that correct mispricings—may become scarcer. The price system's aggregation capacity depends on the diversity and independence of private signals, which widely available AI analysis may reduce even as it increases the average level of information.

4.4 Mechanism Design: The Hurwicz Formalization

Hurwicz's mechanism design framework [Hurwicz, 1960] provides the most direct mathematical formalization of Hayek's question: what institutional arrangements can achieve efficient outcomes when relevant information is dispersed?

The key result on informational efficiency, proved by Mount and Reiter [1974] and sharpened by Jordan [1982]: for economies with more than two agents and goods, the competitive price mechanism requires message spaces of dimension L (the number of goods). Any other mechanism achieving Pareto efficiency requires *at least* as large a message space:

$$\dim(M_{\text{price}}) = L, \quad \dim(M_{\text{alternative}}) \geq L \quad \text{for any Pareto-efficient mechanism.} \tag{5}$$

This is a formal proof of Hayek's claim that the price system achieves efficient coordination with minimum communication burden. The price vector is a minimal sufficient statistic for efficient coordination. Any alternative mechanism—including LLM-mediated allocation—requires at least as much communication.

5 Rosen and the Market for Superstars

5.1 The 1981 Mechanism

Rosen's "The Economics of Superstars" [Rosen, 1981] explained why income distributions in certain professions are so extreme. The answer combined two elements: consumers strongly prefer the best to the second-best (imperfect substitution), and the best can reach an unlimited audience at low marginal cost (joint consumption technology).

If consumers prefer the best and the best can reach an unlimited audience, global demand concentrates on a small number of providers. The superstar captures a disproportionate share of total market revenue not because of disproportionate talent but because distribution technology allows talent differences to be amplified into audience differences without limit. Rosen's striking prediction: improvements in distribution technology *increase* income concentration at the top.

5.2 Rosen's Mechanism in AI Markets

The revenue-per-user data for ChatGPT and Claude are a Rosen phenomenon in operation. LLMs represent the limiting case of Rosen's joint consumption technology: a single model can serve a billion users simultaneously at near-zero marginal cost. The result is a market in which capability differences between models produce revenue differences enormous relative to the capability differences themselves.

Table 1: LLM market structure: ChatGPT vs. Claude (2025). Sources in footnotes; both firms' revenues were growing rapidly throughout the year.

Metric	ChatGPT / OpenAI	Claude / Anthropic
Weekly active users (Oct 2025)[1]	∼800 million	—
Monthly active users[2]	∼355 million	∼19–30 million
Daily queries (Jul 2025)[3]	∼2.5 billion	—
Annualized revenue[4]	∼$13 billion	∼$3–4 billion
Revenue per monthly user	∼$37/yr	∼$100–$211/yr
Valuation (late 2025)[5]	∼$500–750 billion	∼$183 billion
Inference cost (Azure, Q1–Q3)[6]	∼$8.7 billion	—
Enterprise revenue share[7]	∼15–20%	∼80%

The monthly-active-user gap is roughly twelve- to eighteen-fold. The revenue gap is three- to four-fold. The implied revenue-per-user gap ranges from roughly three- to sixfold depending on which endpoint of the user and revenue ranges is used. This is the Rosen mechanism in operation: Claude's smaller base consists disproportionately of developers and enterprise teams who pay a substantial premium for marginal quality advantages in high-value domains.

6 The Structure of the Firm and Knowledge

6.1 Coase: Why Firms Exist

Coase's "The Nature of the Firm" [Coase, 1937] answered why firms exist in a world of markets: transaction costs. Discovering relevant prices, negotiating contracts, and enforcing them all consume resources. When these costs exceed the costs of organizing the same activity through hierarchical direction within a firm, the firm supplants the market.

Asset specificity—identified by Williamson [1975]—is the primary driver. The worker whose tacit knowledge is specific to this firm's systems and processes is locked in, and the firm is locked in to them. The firm is, at its core, an institution for managing the transaction costs generated by knowledge that is specific, tacit, and hard to verify.

6.2 The Bifurcation of Firm Structure

LLMs change the parameters of Coase's analysis. As they reduce the cost of specifying, monitoring, and executing codifiable tasks, the optimal firm boundary shifts outward—more activities can be outsourced. The logical extreme is the "one-person unicorn"—a software company built by a single founder using AI tools that substitute for teams of specialized employees. While no verified case has yet reached a billion-dollar valuation on this model, the trajectory is visible in the rapid growth of solo-founder AI-augmented startups.

But the knowledge-based theory of the firm [Kogut and Zander, 1992, Grant, 1996] identifies the incompleteness of this analysis. The activities that LLMs make cheap to outsource are those involving codifiable, propositional knowledge. The activities that remain inside the firm are those requiring tacit, relational, organizationally embedded knowledge: the accumulated understanding of specific customer relationships, the organizational knowledge of how to execute complex multi-team projects, the strategic intuition built from years of operating in a specific competitive environment.

The prediction: a proliferation of small AI-augmented firms executing codifiable tasks at one end; a consolidation of large knowledge-fortress organizations around tacit, relational knowledge at the other. The firms under maximum pressure are those in the middle.

6.3 Capital Structure Follows Knowledge Structure

The optimal capital structure is conventionally analyzed through the Modigliani and Miller [1958] framework and its extensions [Jensen and Meckling, 1976]. This implicitly assumes that primary assets are physical or financial—whose value can be pledged as collateral and whose liquidation value is well-defined.

Knowledge-intensive firms violate every assumption. Their primary assets are non-pledgeable (tacit knowledge cannot be used as loan collateral), illiquid (there is no market for organizational routines), and subject to catastrophic depreciation upon key employee departures. The Stiglitz and Weiss [1981] analysis of credit rationing further reinforces this: creditors who cannot assess the quality of knowledge assets demand a risk premium that makes debt expensive relative to equity from informed investors.

This explains the capital structure of frontier AI firms. OpenAI and Anthropic are running losses of billions annually, financed through equity at valuations that discount cash flows decades into the future. No rational creditor would lend against assets whose value rests entirely on the performance of a small number of key researchers. The equity investors—venture firms, sovereign wealth funds, hyperscalers—have the specialized knowledge to assess these assets. This is a bet that only equity, with its unlimited upside and absence of fixed claims, can appropriately finance.

7 The Physics and Mathematics of Knowledge

7.1 Landauer's Principle: The Thermodynamic Floor

The connection between information and physics is not metaphorical. Boltzmann's thermodynamic entropy and Shannon's information entropy have exactly the same mathematical form—they are the same object viewed from different frames [Jaynes, 1957].

Landauer [1961] established the physical consequence: any logically irreversible computation—any operation that erases a bit—must dissipate a minimum energy as heat. Bennett [1973] proved the converse: logically reversible computations can in principle avoid this cost, confirming that erasure, not computation itself, is the thermodynamic bottleneck. The Landauer limit:

$$E_{\text{min}} = k_B \cdot T \cdot \ln 2 \approx 2.85 \times 10^{-21} \text{ joules (at room temperature).} \tag{6}$$

Modern transistors dissipate roughly 10^6 to 10^7 times this theoretical minimum. Information processing requires energy as a matter of physical law, not engineering limitation. A civilization that continuously increases its knowledge accumulation rate must continuously increase its energy expenditure.

7.2 Chinchilla Scaling Laws

Building on the earlier scaling laws of Kaplan et al. [2020], the Hoffmann et al. [2022] Chinchilla scaling laws established the optimal relationship between model size N, training data D, and compute budget C:

$$\begin{aligned} &N_{\text{optimal}} \propto C^{0.5}, && D_{\text{optimal}} \propto C^{0.5}, \\ &L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}, && (\alpha \approx 0.34,\ \beta \approx 0.28). \end{aligned} \tag{7}$$

The critical implication: because the exponents are less than one, each unit of capability improvement requires superlinearly more compute.[2] This is a property of the statistical learning problem, ultimately rooted in the Shannon entropy structure of the data distribution.

7.3 Token Growth and Energy Projections

OpenAI reported processing approximately 6 billion tokens per minute as of late 2025. Annualized: approximately 3.15×10^{15} tokens/year. Global LLM token consumption across all providers was approximately 10^{16}–10^{17} tokens/year, doubling roughly every eight months:

$$T(2030) \approx T(2025) \times 2^{60/8} \approx 181 \times T(2025). \tag{8}$$

Estimating the current energy baseline requires specifying scope. Back-calculating from OpenAI's reported Azure spending and GPU fleet yields approximately 3 TWh for OpenAI's inference alone.[3] Global LLM inference—including Google, Anthropic, Meta, xAI, Mistral, and the open-source ecosystem—is considerably larger, plausibly 5–15 TWh. A companion paper [Litowitz et al., 2026] constructs the full supply-and-demand balance sheet for global token production using MacKay's arithmetic methodology. Table 2 presents projections under both a conservative (3 TWh) and central (10 TWh) baseline to make this uncertainty explicit.

[2] A replication by Besiroglu et al. (2024) found the original Chinchilla exponent estimates may be unreliable due to data-fitting issues, with corrected values closer to $\alpha \approx 0.35$, $\beta \approx 0.37$. The qualitative conclusion—sublinear scaling, hence superlinear cost—is robust to these revisions.

[3] Based on an estimated 500K–1M inference GPUs at ∼1 kW average draw and ∼70% utilization, consistent with leaked Azure spending figures of ∼\$8.7B in 2025.

Table 2: Projected LLM energy consumption under alternative scenarios. Token growth of $\sim$180$\times$ follows from equation (8). The two baselines bracket the range from a single large provider ($\sim$3 TWh) to a central estimate of global LLM inference ($\sim$10 TWh).

Scenario	Token Growth	Efficiency Gain	2030 Energy (TWh) 3 TWh base	2030 Energy (TWh) 10 TWh base
No efficiency gain	$\sim$180$\times$	1$\times$	$\sim$540	$\sim$1,800
Moderate improvement	$\sim$180$\times$	10$\times$	$\sim$54	$\sim$180
Aggressive improvement	$\sim$180$\times$	32$\times$	$\sim$17	$\sim$56
2025 baseline (est.)			$\sim$3	$\sim$10

7.4 Jevons' Paradox and the Rebound Effect

Jevons [1865] observed that improvements in steam engine efficiency increased, not decreased, total coal consumption. Applied to AI, hardware efficiency improvements reduce cost per token, which increases quantity demanded, which increases total energy consumption. If total energy is $E = S/\eta$ where S is token demand and η is hardware efficiency, then in growth rates:

$$\hat{E} = \hat{S} - \hat{\eta}, \quad \text{where } \hat{S} \approx \varepsilon \hat{\eta} \implies \hat{E} = (\varepsilon - 1)\,\hat{\eta}, \tag{9}$$

with hats denoting proportional growth rates and ε the price elasticity of token demand. Total energy consumption falls only if $\varepsilon < 1$. Given the enormous latent demand for AI services globally, the elasticity almost certainly exceeds 1. The empirical confirmation: combined hardware and algorithmic efficiency for AI inference improved by several orders of magnitude from 2012 to 2025, yet total AI energy consumption grew by several orders of magnitude as well.

8 Mathematical Foundations: Arrow and Hayek Formalized

8.1 Arrow's Impossibility Theorem

Arrow's [1951] impossibility theorem proves that no voting system can simultaneously satisfy four conditions when aggregating individual preferences into a social ranking: Unrestricted domain (U), Pareto efficiency (P), Independence of irrelevant alternatives (IIA), and Non-dictatorship (D).

Theorem 1 (Arrow, 1951). *For $|X| \geq 3$ and $n \geq 2$, any social welfare function $f : \mathcal{P}^n \to \mathcal{P}$ satisfying U, P, and IIA is a dictatorship.*

The price system sidesteps this impossibility not by aggregating preferences into a ranking but by compressing them into a scalar—a price—which loses information but provides a workable sufficient statistic for coordination. The price system does not solve

Arrow's impossibility problem; it circumvents it by operating in a different mathematical space.

For LLMs, Arrow's theorem raises a direct challenge. When a model is asked to produce a recommendation reflecting "what most people would want," it is attempting a social aggregation. The condition most likely violated is IIA: the model's output for A vs. B is affected by how A and B relate to thousands of other alternatives in the training data—precisely the IIA violation Arrow identified as fundamental.

8.2 The Myerson–Satterthwaite Theorem

Myerson and Satterthwaite [1983] proved that with two agents having private information about their valuations, no mechanism can simultaneously achieve efficiency, budget balance, incentive compatibility, and individual rationality.

The proof shows that expected gains from efficient trade exactly equal expected information rents required for truthful revelation—leaving no surplus to cover mechanism costs. For knowledge markets, this formalizes why bilateral trades of information are so persistently difficult. Arrow showed the impossibility from the buyer's inability to evaluate the good; Myerson–Satterthwaite shows additional impossibility from bilateral private information.

8.3 Mutual Information and the Hayek–Shannon Bridge

Let X represent the dispersed private knowledge of all agents in an economy, and Y represent the information encoded in the equilibrium price vector. Mutual information $I(X;Y)$ measures how much knowledge of Y reduces uncertainty about X:

$$I(X;Y) = H(X) - H(X \mid Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}. \tag{10}$$

Perfect price efficiency would require $I(X;Y) = H(X)$, but Grossman and Stiglitz [1980] show this is an unstable equilibrium. The actual equilibrium achieves $I(X;Y) = H(X) - H(X \mid P) > 0$ for a *noisy equilibrium price* P that balances information aggregation against incentives to acquire private signals.

For LLMs, the training process maximizes something approximating $I(\hat{X}; X_{\text{corpus}})$—the mutual information between the model's representation and the training corpus. This is large for well-documented domains. It is small for the local, tacit, time-sensitive knowledge Hayek cared about, because that knowledge was never textualized.

8.4 Kolmogorov Complexity and the Incompressibility of Tacit Knowledge

The Kolmogorov complexity $K(x)$ of a string x is the length of the shortest program that generates x:

$$K(x) = \min_{\{p:\, U(p)=x\}} |p|. \tag{11}$$

The economic knowledge relevant to Hayek has high Kolmogorov complexity: the specific, contextual, time-sensitive knowledge of particular circumstances cannot be compressed into a description shorter than the circumstances themselves. This gives the formal statement of why Hayek's knowledge problem is permanent rather than temporary. It is not a computational limitation—it is informational incompressibility. More compute does not solve this problem.

By contrast, the explicit, general, documentable knowledge that LLMs capture has low Kolmogorov complexity. The division between what LLMs do well and what they do poorly maps almost exactly onto the division between low-complexity (compressible, general) and high-complexity (incompressible, contextual) knowledge.

8.5 Blackwell's Theorem: Comparing Information Structures

Blackwell's [1953] theorem provides a precise way to say when one signal is unambiguously more informative than another. Experiment $\mathcal{E}_1$ Blackwell-dominates $\mathcal{E}_2$ if and only if $\mathcal{E}_2$ can be obtained from $\mathcal{E}_1$ by adding noise—if there exists a stochastic matrix M (a "garbling") such that:

$$P_2(s_2 \mid \theta) = \sum_{s_1} M(s_2 \mid s_1) \cdot P_1(s_1 \mid \theta). \tag{12}$$

Theorem 2 (Blackwell, 1953)**.** *$\mathcal{E}_1 \succ_B \mathcal{E}_2$ if and only if for every decision problem, the maximum expected utility achievable with $\mathcal{E}_1$ is at least as large as with $\mathcal{E}_2$.*

The Blackwell order is a *partial* order—not all information structures are comparable. A price signal and an LLM output are typically not Blackwell-comparable: prices are more informative about real-time supply-and-demand conditions; LLMs are more informative about documented causal relationships. Neither garbles the other. They provide genuinely different kinds of information, useful for different decision problems.

8.6 The Lucas Islands Model

Lucas's [1972] Islands model is the most mathematically precise formalization of the problem Hayek described in 1945: agents observing only local price signals cannot distinguish local shocks from aggregate shocks. Local price:

$$p(z) = P + z, \quad \text{where } P \sim \mathcal{N}(\bar{P}, \sigma_P^2) \text{ and } z \sim \mathcal{N}(0, \sigma_z^2). \tag{13}$$

Optimal Bayesian inference of the relative demand shock z from observed price $p(z)$:

$$E[z \mid p(z)] = \frac{\sigma_z^2}{\sigma_P^2 + \sigma_z^2} \cdot (p(z) - \bar{P}). \tag{14}$$

The signal-to-noise ratio $\sigma_z^2/(\sigma_P^2 + \sigma_z^2)$ determines how much agents respond to price signals. For LLMs: the model knows the distribution P very well—the typical cross-sectional pattern. What it cannot estimate is the island-specific shock z at *this* moment, on *this* island, for *this* producer.

8.7 Romer's Endogenous Growth: The Formal Engine

Romer's [1990] model formalizes what Arrow's learning-by-doing implies for long-run growth. The production function for final goods:[4]

$$Y = H_Y^{\alpha} \int_0^A x_i^{1-\alpha}\, di, \tag{15}$$

where A is the total number of intermediate goods designs (the knowledge stock). The knowledge accumulation equation:

$$\dot{A} = \delta \cdot H_A \cdot A, \quad \text{equilibrium growth rate: } g^* = \delta H_A - \rho, \tag{16}$$

where H_A is human capital devoted to research, δ is research productivity, and ρ is the subjective discount rate. New knowledge builds on existing knowledge linearly—the "standing on shoulders" property. If LLMs raise the effective research productivity δ, the equilibrium growth rate rises. If LLMs reduce the diversity of the research process by concentrating inquiry around modal outputs, the effective δ may fall even as individual researcher productivity rises.

8.8 Evolutionary Game Theory: Competition as Discovery

Hayek described competition as a "discovery procedure." This can be formalized through the replicator dynamics of evolutionary game theory. In a population with strategy frequencies $x_1, \ldots, x_n$:

$$\dot{x}_i = x_i\big[f_i(x) - \bar{f}(x)\big]. \tag{17}$$

Market competition is a replicator dynamic over business strategies: firms that discover better strategies earn above-average profits and expand; inferior strategies contract. The critical Hayekian point is that the fitness landscape—which strategies are profitable given which others are present—is *not known before the process runs*. It is revealed by competition itself.

For LLMs, the structural risk: if LLMs homogenize strategy selection—recommending the same "best practices" to all firms—the population-level diversity that makes the discovery procedure work is reduced. A population of firms all implementing the same LLM-recommended strategy cannot discover through competitive selection which strategies are better in specific contexts, because there is no variation to select on.

9 Von Neumann: The Architect of Computing

John von Neumann died in 1957, having spent the last decade of his life at the precise intersection of computation, thermodynamics, and the theory of mind. His contributions—the stored-program computer architecture, foundational results in game theory, the the-

[4] Romer's original specification includes unskilled labor L as a separate factor: $Y = H_Y^{\alpha} L^{\beta} \int_0^A x_i^{1-\alpha-\beta}\, di$. The two-factor simplification used here follows standard textbook treatments. Note: H_Y and H_A denote human capital (Romer's notation), distinct from the information-theoretic $H(\cdot)$ used earlier.

ory of self-replicating automata, and a thermodynamic theory of information processing that anticipated Landauer—provide a uniquely integrated framework for analyzing LLMs.

9.1 The Architecture and Its Descendants

Von Neumann's [1945] First Draft Report on the EDVAC introduced the stored-program architecture that nearly all modern computers still follow. The transformer architecture underlying modern LLMs extends this architecture in a direction von Neumann himself explored. The attention mechanism—which allows each token to attend to all others in a sequence—resembles the content-addressable memory systems he discussed in his analysis of neural organization [von Neumann, 1958], fundamentally different from address-based retrieval.

The critical departure: the transformer's memory is not stored in a von Neumann memory bank but distributed across the weights of the network. In a standard von Neumann machine, knowledge is stored as explicit bit patterns at specific addresses and retrieved by address. In an LLM, knowledge is stored as a high-dimensional manifold of parameter configurations and accessed by forward propagation during inference. Levin [2026] argues that this geometric representation constitutes a distinct epistemological regime—"navigational knowledge" produced by traversing learned manifolds, irreducible to either symbolic reasoning or statistical recombination.

9.2 The Thermodynamics of Retrieval

In his lectures at the University of Illinois and in his unfinished manuscript *The Computer and the Brain* [von Neumann, 1958], von Neumann was working toward a thermodynamic theory of information processing that anticipated Landauer. He arrived at an expression for the minimum energy per logical operation equivalent to equation (6): $E_{\min} = k_B T \ln 2$.[5] He reportedly presented this result to Shannon, who recognized it as identical to the expression in his communication theory. Extending this logic by analogy, the thermodynamic efficiency of knowledge retrieval can be defined as:

$$\eta_{\text{retrieval}} = \frac{I_{\text{useful}}}{W_{\text{dissipated}}} = \frac{\text{Shannon bits resolved per query}}{\text{joules per query}}. \tag{18}$$

Landauer's principle suggests a Carnot-like bound on this ratio: $\eta_{\text{retrieval}}$ cannot exceed the inverse of $k_B T \ln 2$ per bit resolved (cf. equation 6). This analogy is heuristic—it equates useful information with erasable bits, which is not formally established—but the bound is directionally informative. Current LLMs operate orders of magnitude below it.

[5] The extent to which von Neumann's derivation was independent of earlier thermodynamic arguments is debated; see the historical discussion in Landauer [1961].

9.3 Three Regimes of Knowledge Retrieval

Von Neumann's architectural framework and brain studies suggest three fundamentally different retrieval regimes:

Table 3: Three regimes of knowledge retrieval.

Regime	Mechanism	Cost Structure	Example
Address-based retrieval	Specify memory location $\rightarrow$ content returned	Low, constant; bottleneck is finding the address	Library catalog, database
Content-based retrieval	Specify query $\rightarrow$ most similar content returned	Scales with similarity computation complexity	Human memory, transformer attention
Generative retrieval	Specify prompt $\rightarrow$ content generated consistent with training distribution	Cost of computation, not storage access	LLM inference

Whether generative retrieval is retrieval at all is a substantive question, not a semantic one. Litowitz et al. [2026] argue that it is: LLMs compress a training corpus into parametric storage and inference is retrieval—reconstruction of a targeted answer from compressed representations at machine speed. On this view, what has changed is not the operation but its cost; value lies in retrieval speed per unit energy. Levin [2026] argues the opposite: generative output is not retrieved but *navigated*—produced by traversing learned manifolds in high-dimensional space, constituting a third mode of knowledge production irreducible to either symbolic reasoning or statistical recombination. The distinction matters economically. If LLM inference is retrieval, its value is governed by the economics of speed and compression. If it is navigation, its value is governed by the economics of the manifold—the richness and structure of the learned representation, which can generate outputs that were never stored. In either case, making generative retrieval cheap shifts the binding constraint from retrieval cost to the quality of the generative process—and to the tacit component that was never documented and therefore cannot be generated or retrieved.

9.4 Game Theory and the Common Knowledge Set

Von Neumann's *Theory of Games and Economic Behavior* [von Neumann and Morgenstern, 1944] established that in games with incomplete information, equilibrium strategies depend critically on the information structure. LLMs transform the information structure of essentially every strategic interaction they touch.

The effect can be stated precisely in von Neumann's framework: LLMs move the boundary of the common knowledge set outward, and strategic competition concentrates on whatever remains outside it. In zero-sum information games, the minimax value declines as information asymmetry falls—there is less rent available from pure information advantage. Competition shifts toward judgment, relationships, and execution capability—dimensions where information symmetry does not eliminate advantage.

9.5 Automata Theory and Model Collapse

Von Neumann's theory of self-replicating automata [von Neumann, 1966] proved that above a certain complexity threshold, self-replication can be error-correcting—the automaton is complex enough to detect and correct errors in its own reproduction. Below this threshold, error accumulation dominates and complexity declines.

The model collapse dynamic—LLMs trained on AI-generated text losing diversity and amplifying modal patterns—maps precisely onto sub-threshold automata replication. The central question in von Neumann's framework: are current LLMs above or below the complexity threshold at which self-replication becomes error-correcting? The empirical evidence suggests they are currently below it, which is why model collapse requires human-generated data injection to correct.

In von Neumann's framework, LLMs are the first computational systems large enough that the question of self-elaborating intelligence is not obviously answered in the negative—arguably the most consequential fact about them.

10 Influencers, Attention, and the Superstar Economy

10.1 The Attention Market as Knowledge Market

The influencer economy is best understood not as a deviation from standard knowledge markets but as their purest expression. An influencer is a filtering and certification mechanism in an environment of information overabundance. In a world of infinite content generation, the marginal value of an additional piece of content approaches zero, while the value of a reliable filter that selects high-quality content from the flood rises. The influencer does not primarily produce information—they reduce uncertainty about what information to consume. This reduction of uncertainty is, precisely, Shannon information.

Arrow's paradox reappears in the trust dimension. The influencer's economic value rests on the perception of authentic recommendation—genuine signal, not paid noise. But the economic incentive is to monetize that trust through sponsorship, which introduces noise. The follower cannot reliably distinguish authentic enthusiasm from paid performance. This is Arrow's asymmetric information problem applied to the market for attention.

10.2 LLMs and the Bifurcation of Creator Value

Creator value comes from two sources that LLMs affect very differently. The first source is informational value: knowledge, analysis, explanation, synthesis. This is precisely what LLMs produce cheaply and at scale—facing anyone whose primary value lies in transferring codifiable knowledge with direct commodity competition.

The second source is relational value: the parasocial relationship, the specific human presence and personality that followers invest in. This is what LLMs cannot replicate without replicating the person. The creator who has built a genuine parasocial relationship with an audience has something no AI can straightforwardly substitute.

The prediction: the creator economy bifurcates similarly to the firm economy. Creators whose value is primarily informational face commoditization pressure. Creators whose value is primarily relational face a market in which their scarcity becomes *more* rather than less pronounced as AI floods the informational content space.

11 Open Questions and Further Directions

11.1 The Romer Growth Engine Under Threat

Romer's endogenous growth model predicts that growth is sustained by the continuous production of new ideas, with the rate of idea production depending on the stock of existing knowledge and the number of people working at the frontier. Bloom et al. [2020] document that research productivity has been declining for decades—ideas are getting harder to find. Model collapse—the progressive homogenization of generated content as models trained on AI-generated data amplify common patterns [Shumailov et al., 2024]—poses a direct additional threat to this engine. Romer's mechanism requires a varied landscape of independent inquiries with different priors, approaches, and hypotheses. A world in which most text is generated by a small number of models may be one in which the rate of genuine idea production declines even as the volume of generated content explodes.

11.2 The Data Commons: Sustainability and Compensation

The models were trained on text produced by writers, academics, coders, and ordinary people who received no compensation. Jones and Tonetti [2020] formalize data's defining economic property—nonrivalry—and show that socially optimal data policy diverges sharply from the private equilibrium. If AI systems capture the value created by this commons without contributing to it, and if this reduces the incentive to produce original human text, the quality of future training data deteriorates—a self-undermining dynamic. No satisfactory compensation mechanism has been established; the legal framework for copyright in AI training is unsettled in every major jurisdiction. This is a market failure in the Arrow sense—the social value of original human text generation exceeds its private return, now more than ever.

11.3 Labor Market Displacement: An Inverted Pattern

Previous waves of automation displaced routine, middle-skill work—the hollowing-out hypothesis of Autor et al. [2003]. AI displacement inverts this pattern: it affects high-skill cognitive work most intensely and physical, relational work least. The radiologist, junior lawyer, financial analyst, and entry-level programmer are more exposed than the plumber, home health aide, or kindergarten teacher. This inversion has profound distributional consequences that existing social insurance frameworks—calibrated to the hollowing-out pattern—were not designed to address.

11.4 The Measurement Problem

Standard macroeconomic statistics are poorly suited to capturing the value being created and destroyed. If a lawyer who previously billed 20 hours for document review now completes it in 2 hours with AI assistance and charges less accordingly, measured GDP falls—even though the client receives equivalent value faster and at lower cost. We cannot reliably know whether the AI transition is increasing or decreasing aggregate welfare as measured by existing statistics, which has profound implications for policy evaluation.

11.5 The Geopolitical Economy of Knowledge

The concentration of frontier AI capability in a handful of firms in two countries creates a geopolitical economy of knowledge without precedent in peacetime history. Export controls on advanced semiconductors, restrictions on academic collaboration, sovereign wealth fund investments in AI firms, and the use of AI systems in military and intelligence applications are all aspects of a competition whose ultimate stakes concern which nations and which civilizational values will shape the knowledge infrastructure of the coming century.

12 Conclusion

Six economists—Arrow, Hayek, Rosen, Coase, Shannon, and von Neumann—built the theoretical apparatus that explains the economics of knowledge. Large language models do not require new theory; they require careful application of existing theory to a technology whose scale and speed are genuinely new.

The central lesson is a distinction. Explicit, general, documentable knowledge—low Kolmogorov complexity—is now subject to commodity economics. LLMs produce it cheaply, distribute it globally, and compress the premium that accrued to its human possessors. Tacit, contextual, time-sensitive knowledge—high Kolmogorov complexity—remains scarce, and the frameworks of Hayek, Polanyi, and Grossman–Stiglitz explain why it will remain so. The boundary between these two domains is the most consequential economic frontier of the coming decade.

The physics is equally clear. Information processing requires energy as a matter of thermodynamic law, and Jevons' paradox ensures that efficiency improvements will increase, not decrease, total energy consumption. The economics of knowledge and the physics of computation jointly determine the trajectory: superlinear scaling of capability costs, winner-take-most market structure, bifurcation of firms and labor markets, and an energy footprint bounded from below by Landauer's principle and from above by institutional willingness to pay.

What remains genuinely open is whether LLMs enhance or impair the processes that generate new knowledge—the Romer growth engine, the Hayekian discovery procedure, the diversity of independent inquiry on which both depend. The answer will determine whether the age of large language models is remembered as an acceleration of human knowledge or as the period in which we learned to retrieve what we already knew, ever faster, while the rate of discovering what we did not know quietly declined.

References

George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500, 1970.

Kenneth J. Arrow. *Social Choice and Individual Values*. Wiley, 1951.

Kenneth J. Arrow. The economic implications of learning by doing. *Review of Economic Studies*, 29(3):155–173, 1962a.

Kenneth J. Arrow. Economic welfare and the allocation of resources for invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, pages 609–626. Princeton University Press, 1962b.

Kenneth J. Arrow. The role of securities in the optimal allocation of risk-bearing. *Review of Economic Studies*, 31(2):91–96, 1964.

Kenneth J. Arrow and Gerard Debreu. Existence of an equilibrium for a competitive economy. *Econometrica*, 22(3):265–290, 1954.

David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4):1279–1333, 2003.

Charles H. Bennett. Logical reversibility of computation. *IBM Journal of Research and Development*, 17(6):525–532, 1973.

David Blackwell. Equivalent comparisons of experiments. *Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Nicholas Bloom, Charles I. Jones, John Van Reenen, and Michael Webb. Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144, 2020.

Ronald H. Coase. The nature of the firm. *Economica*, 4(16):386–405, 1937.

Robert M. Grant. Toward a knowledge-based theory of the firm. *Strategic Management Journal*, 17(S2):109–122, 1996.

Sanford J. Grossman and Joseph E. Stiglitz. On the impossibility of informationally efficient markets. *American Economic Review*, 70(3):393–408, 1980.

Friedrich A. Hayek. The use of knowledge in society. *American Economic Review*, 35(4): 519–530, 1945.

Friedrich A. Hayek. The pretence of knowledge. In *Nobel Lectures, Economics 1969–1980*. World Scientific, 1974. Nobel Prize Lecture, December 11, 1974; reprinted in World Scientific, 1992.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Leonid Hurwicz. Optimality and informational efficiency in resource allocation processes. In *Mathematical Methods in the Social Sciences*, pages 27–46. Stanford University Press, 1960.

Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4): 620–630, 1957.

Michael C. Jensen and William H. Meckling. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4):305–360, 1976.

William Stanley Jevons. *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-Mines*. Macmillan, 1865.

Charles I. Jones and Christopher Tonetti. Nonrivalry and the economics of data. *American Economic Review*, 110(9):2819–2858, 2020.

Charles I. Jones and John C. Williams. Measuring the social return to R&D. *Quarterly Journal of Economics*, 113(4):1119–1135, 1998.

James S. Jordan. The competitive allocation process is informationally efficient uniquely. *Journal of Economic Theory*, 28(1):1–18, 1982.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Bruce Kogut and Udo Zander. Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science*, 3(3):383–397, 1992.

Rolf Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961.

Ilya Levin. Epistemology of generative AI: The geometry of knowing. *arXiv preprint arXiv:2602.17116*, 2026.

Alec Litowitz, Nick Polson, and Vadim Sokolov. Photons = tokens: The physics of AI and the economics of knowledge. Technical Report 6265418, SSRN, 2026. Available at `https://papers.ssrn.com/abstract=6265418`.

Robert E. Lucas. Expectations and the neutrality of money. *Journal of Economic Theory*, 4 (2):103–124, 1972.

Franco Modigliani and Merton H. Miller. The cost of capital, corporation finance and the theory of investment. *American Economic Review*, 48(3):261–297, 1958.

Kenneth Mount and Stanley Reiter. The informational size of message spaces. *Journal of Economic Theory*, 8(2):161–192, 1974.

Roger B. Myerson and Mark A. Satterthwaite. Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29(2):265–281, 1983.

Michael Polanyi. *The Tacit Dimension*. Doubleday, 1966.

Paul M. Romer. Endogenous technological change. *Journal of Political Economy*, 98(5): S71–S102, 1990.

Sherwin Rosen. The economics of superstars. *American Economic Review*, 71(5):845–858, 1981.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Ilia Shumailov, Zakhar Shumilo, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

George J. Stigler. The economics of information. *Journal of Political Economy*, 69(3):213–225, 1961.

Joseph E. Stiglitz and Andrew Weiss. Credit rationing in markets with imperfect information. *American Economic Review*, 71(3):393–410, 1981.

John von Neumann. First draft of a report on the EDVAC. Technical report, Moore School of Electrical Engineering, University of Pennsylvania, 1945.

John von Neumann. *The Computer and the Brain*. Yale University Press, 1958.

John von Neumann. *Theory of Self-Reproducing Automata*. University of Illinois Press, 1966. Edited by Arthur W. Burks.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

Oliver E. Williamson. *Markets and Hierarchies: Analysis and Antitrust Implications*. Free Press, 1975.