

INFERENCE ECONOMICS: A NEW PARADIGM FOR THE ECONOMICS OF ARTIFICIAL INTELLIGENCE

*Connecting Production Economics, Platform Markets, Digital Infrastructure, and AI
Innovation Systems*

Ibrahim Niankara

Al Ain University College of Business / Brass Digital Lab

ABSTRACT

We introduce *Inference Economics* as a new subfield of economics organized around the production, pricing, and consumption of AI inference tokens — the fundamental digital commodity that provides access to artificial intelligence capabilities. As frontier AI models from oligopolistic providers (OpenAI, Anthropic, Google, xAI) become as foundational to production as electricity was in the early twentieth century, a new economics is required to understand how tokens are produced, priced, consumed, and allocated across a diverse population of developers and organizations. Drawing on a nine-paper foundational series (Brass 2026a–i), we develop the conceptual architecture of Inference Economics around four intellectual pillars (production economics, platform economics, digital infrastructure markets, and AI innovation systems) and five core theoretical constructs (the Token Production Function, the Token Kuznets Curve, the Jevons Paradox of AI Tokens, the General Equilibrium of the Token Economy, and the Copula Two-Part Demand Model). We locate these constructs in the broader economics literature, derive their key empirical implications, and set out a seven-direction research agenda. Inference Economics is not a relabeling of existing subfields; it is a synthesis with distinctive theoretical primitives — cognitive capital, agentic multipliers, token intensity curves, and inference gap inequality — that do not exist in any single prior tradition. The empirical calibration from the foundational series yields tractable, testable predictions about AI market structure, token demand dynamics, welfare distribution, and long-run growth that constitute the positive theory of this new subfield.

JEL Classifications: O30, O33, L13, L86, D24, D40, C51, Q55.

Keywords: Inference Economics, AI tokens, cognitive capital, Token Production Function, Jevons Paradox, agentic AI, platform economics, digital infrastructure, AI innovation

CONTENTS

I	Introduction: The Emergence of Inference Economics	3
II	From Physical to Computational to Cognitive Capital	4
A	Three Waves of Capital Augmentation	4
B	Tokens as the Coal of the Cognitive Economy	5
C	The Novelty of Cognitive Capital Economics	5
III	The Four Pillars of Inference Economics	6
IV	The Core Theoretical Architecture of Inference Economics	8
A	The Token Production Function	8
B	The Token Kuznets Curve	9
C	The Jevons Paradox of AI Tokens	9
D	The General Equilibrium of the Token Economy	10
E	Micro-Econometric Foundations: Two-Part Demand and Copula Dependence	11
V	Inference Economics and the Paradigm of AI-Augmented Innovation	13
A	Connections to Production and Growth Economics	13
B	Connections to Platform and Information Economics	13
C	Connections to Innovation Economics and Technology Transitions	14
D	The Inference Gap: Inequality in Cognitive Capital Access	15
VIA	Research Agenda for Inference Economics	16
VII	Conclusion: Inference Economics as a Unifying Framework	20
	References	22

I. INTRODUCTION: THE EMERGENCE OF INFERENCE ECONOMICS

Every major transformation in the economics of production has created a corresponding transformation in the economics of the discipline itself. The Industrial Revolution — with its substitution of mechanical energy for human muscle — gave rise to capital theory, growth theory, and the economics of technical change. The ICT Revolution — with its substitution of computational speed for human cognitive repetition — gave rise to the economics of information, network effects, and digital platforms. The current AI Revolution — with its substitution of model-generated reasoning for human non-routine cognition — requires a third transformation: a new economics organized around the production, pricing, and allocation of artificial intelligence capabilities.

We propose *Inference Economics* as the name for this new economics. The central commodity of Inference Economics is the AI inference token — a digital unit representing the processing of a single piece of information (a word, sub-word, or structured data element) by a frontier AI model. Tokens are measured, priced, and consumed at the level of individual API calls; they are the unit at which AI capabilities are monetized, rationed, and delivered. Understanding how tokens are produced (supply), demanded (consumption), and allocated (markets) is the organizing problem of Inference Economics.

Inference Economics is not merely a relabeling of existing subfields. While it draws heavily on production economics, platform economics, digital infrastructure markets, and AI innovation systems, it synthesizes these into a unified framework with distinctive theoretical primitives that do not exist in any single prior tradition: cognitive capital as a factor of production, the agentic multiplier as a demand amplifier, the token intensity curve as a non-linear demand schedule, and the Inference Gap as a novel form of distributional inequality. These primitives require new theory, new empirics, and new policy analysis.

The intellectual stakes of Inference Economics are high. AI inference is rapidly becoming as foundational to modern production as electricity was in the early twentieth century. Just as the electrification of production required new economic theories of network externalities, regulatory economics, and innovation complementarities, the inference-ification of production requires new theories of cognitive capital productivity, oligopolistic token pricing, and agentic demand dynamics. Without these theories, policymakers, investors, and researchers lack the conceptual tools to understand the economic transformation underway.

The present paper serves three purposes. First, it introduces the conceptual architecture of Inference Economics to the broad scientific and academic community, synthesizing the nine-paper foundational series (Brass 2026a–i) into an accessible and integrated overview. Second, it locates Inference Economics in the existing economics literature,

identifying both the intellectual debts and the distinctive contributions of the new subfield. Third, it sets out a seven-direction research agenda that constitutes an open invitation to the economics profession to engage with the most consequential new economics of the twenty-first century.

The paper proceeds as follows. Section II situates Inference Economics in its historical moment, tracing the evolution from physical capital to computational capital to cognitive capital. Section III develops the four intellectual pillars of Inference Economics. Section IV presents the core theoretical constructs and their empirical calibrations. Section V connects Inference Economics to established subfields. Section VI presents the seven-direction research agenda. Section VII concludes.

II. FROM PHYSICAL TO COMPUTATIONAL TO COGNITIVE CAPITAL

A. Three Waves of Capital Augmentation

Economic history can be read as a sequence of transformations in the nature of capital and its relationship to human labor. The *first wave* — the Industrial Revolution of the eighteenth and nineteenth centuries — introduced *physical capital* (steam engines, looms, railways) as an augment to physical labor. Physical capital multiplied the productive power of human muscles, raising output per worker dramatically while also displacing labor from some activities (handloom weaving) and creating new labor demand in others (machine operation, maintenance). The economic theory of this wave is the neoclassical production function, with physical capital and labor as complements in production and with technology as an exogenous shifter.

The *second wave* — the Information and Communications Technology revolution of the latter twentieth century — introduced *computational capital* (computers, software, telecommunications networks) as an augment to cognitive labor. Computers multiplied the scale and speed at which human minds could process information, raising white-collar productivity dramatically while also displacing labor from routine cognitive tasks (accounting, data entry, scheduling) and creating new demand in others (software development, systems administration). The economic theory of this wave is the economics of information, with network externalities, knowledge spillovers, and endogenous technical change as its distinctive contributions.

The *third wave* — the Artificial Intelligence Revolution currently underway — introduces *cognitive capital* (AI models, inference compute, training data) as an augment to *non-routine cognitive labor*. AI inference enables machines to perform tasks that were previously thought to require human-level intelligence: understanding natural language, generating

code, reasoning under uncertainty, producing creative content. This wave differs from its predecessors in a critical respect: cognitive capital does not merely assist human cognition, it substitutes for it at the margin. This substitution dynamic — combined with the unprecedented scale, speed, and oligopolistic structure of cognitive capital supply — makes the economics of the third wave qualitatively distinct from both prior waves.

B. Tokens as the Coal of the Cognitive Economy

William Stanley Jevons (1865), writing at the height of the first wave, observed that coal was the universal fuel of the Industrial Revolution: every form of mechanical power ultimately derived from burning coal, making the supply, price, and efficiency of coal use central to the economics of the entire transformation. AI inference tokens play the analogous role in the third wave. Every form of AI-assisted production ultimately derives from consuming tokens: writing code, analyzing data, generating text, processing images, reasoning through problems. Token supply, token price, and token efficiency are central to the economics of the AI transformation.

The analogy runs deeper than metaphor. In the coal economy, supply was concentrated: a small number of coal companies controlled the resource, generating oligopoly rents and motivating the regulatory economics of natural resource extraction. In the token economy, supply is concentrated in four frontier providers (OpenAI, Anthropic, Google, xAI) that together account for the overwhelming majority of global frontier AI inference capacity. The governance challenges of token supply concentration mirror — in structure if not in physical form — the governance challenges of coal and oil supply concentration that occupied industrial-age economists and policymakers.

C. The Novelty of Cognitive Capital Economics

Despite the historical parallels, cognitive capital in the form of AI inference tokens has properties that distinguish it fundamentally from physical and computational capital and require new economic analysis.

First, AI inference exhibits *near-zero marginal cost at the point of consumption*: once a model is trained and deployed, serving an additional token costs approximately the cost of GPU electricity and memory access. This drives a wedge between marginal cost pricing (competitive) and average cost pricing (recovering training investment), generating natural monopoly economics with implications for market structure and regulation.

Second, AI inference quality is *endogenous to investment*: the value of a token depends on the quality of the model that generated it, which in turn depends on the training compute, data, and architectural research invested in building the model. This creates a dynamic complementarity between inference demand and model quality improvement that has no

analog in physical or computational capital.

Third, AI inference exhibits *recursive demand*: agentic AI systems consume tokens to plan, execute, and evaluate their own inference, generating feedback loops in demand that can amplify aggregate token consumption far beyond the naive linear prediction. This recursive structure — formalized in the Jevons Paradox of AI Tokens — introduces non-linear dynamics into inference demand that require new theoretical tools.

III. THE FOUR PILLARS OF INFERENCE ECONOMICS

Inference Economics draws from four established intellectual traditions, each of which contributes essential concepts and tools while leaving gaps that the new subfield must fill. We describe each pillar in turn, identifying both what it contributes and where it falls short.

Pillar I: Production Economics

Production economics provides the foundational framework for treating AI inference tokens as a productive input. The neoclassical production function $Q = F(K, L, T)$, augmented with inference tokens T as a third factor, generates the Token Production Function: $S = A \cdot T^\alpha C^\gamma M^\delta$, where S is software output, C is developer capability, M is model capability, and α, γ, δ are factor elasticities. Production economics contributes factor demand theory (cost minimization, duality, marginal product pricing), returns to scale analysis, and technical change decomposition. Its limitation is that it treats AI tokens as a conventional input with well-defined factor markets, missing the oligopolistic supply structure, the agentic amplification dynamic, and the two-sided platform nature of inference markets.

Pillar II: Platform Economics

Platform economics provides the framework for understanding the market structure of AI inference supply. Frontier AI providers operate as two-sided platforms connecting model developers (who build and train models) with token consumers (developers and organizations that use inference APIs). The oligopolistic structure — four dominant providers with significant market power — generates Cournot competition in token quantities, with equilibrium prices

$$P^* = \frac{a + Nc}{N + 1}$$

exceeding competitive levels and generating deadweight loss. Platform economics contributes two-sided market theory (Rochet and Tirole, 2003), oligopoly pricing models (Tirole, 1988), and network externality analysis (Katz and Shapiro, 1985). Its limitation is that standard platform models do not capture the recursive demand amplification of agentic AI or the GPU scarcity constraint that binds supply from below.

Pillar III: Digital Infrastructure Markets

Digital infrastructure economics provides the framework for analyzing the GPU compute substrate on which AI inference depends. GPU clusters represent a form of network infrastructure — capital-intensive, subject to scale economies, and subject to capacity constraints that generate scarcity premia when demand exceeds supply. The GPU scarcity premium

$$\psi(K) = b(Q^* - \theta K)$$

captures the excess of the inference price over the competitive level due to constrained compute supply. Infrastructure economics contributes public utility pricing theory, congestion economics (Vickrey, 1969), and capital investment dynamics under uncertainty (Abel et al., 1996). Its limitation is that standard infrastructure models treat capacity as homogeneous, while GPU compute exhibits heterogeneity in training vs. inference workloads, precision requirements, and interconnect bandwidth.

Pillar IV: AI Innovation Systems

AI innovation systems economics provides the framework for understanding how inference tokens transform the innovation process itself. The Jevons Paradox of AI Tokens formalizes how efficiency improvements and agentic loops amplify aggregate token demand, generating a rebound effect that accelerates innovation while straining infrastructure. The Token Kuznets Curve captures the non-monotonic evolution of token intensity across development stages. AI innovation systems economics draws from the economics of general-purpose technologies (Bresnahan and Trajtenberg, 1995; Helpman, 1998), endogenous growth theory (Romer, 1990; Aghion and Howitt, 1992), and the systems of innovation literature (Nelson, 1993; Lundvall, 1992). Its limitation is that existing GPT frameworks were designed for technologies like electricity and ICT with gradual diffusion curves, while AI inference exhibits hyper-rapid diffusion with recursive demand feedback that creates non-linear growth dynamics.

The four pillars overlap and reinforce each other in important ways. Production economics provides the demand primitives (Token Production Function) that underpin platform economics (where does consumer surplus come from?). Platform economics generates the supply-side pricing (Cournot equilibrium) that constrains digital infrastructure demand (what infrastructure investment is needed?). Infrastructure markets set the compute scarcity that limits supply capacity, feeding back into production economics (how does compute availability affect the marginal product of tokens?). And AI innovation systems connect all three to the dynamic process of capability improvement, adoption, and creative destruction that drives the long-run evolution of inference markets. The synthesis of these four pillars into a unified framework is the defining intellectual project of Inference Economics.

IV. THE CORE THEORETICAL ARCHITECTURE OF INFERENCE ECONOMICS

The nine-paper foundational series (Brass 2026a–i) establishes five core theoretical constructs that form the positive theory of Inference Economics. We present each construct, its derivation from first principles, its empirical content, and its connection to the four pillars. Together, these constructs constitute a self-consistent theoretical architecture — each construct is logically derivable from the primitives of the preceding one, and together they generate a system of testable predictions about AI inference markets.

A. The Token Production Function

The Token Production Function (TPF) is the fundamental demand primitive of Inference Economics. It characterises how AI inference tokens (T), developer capability (C), and AI model capability (M) combine to produce software output (S):

$$S = A \cdot T^\alpha \cdot C^\gamma \cdot M^\delta \cdot \exp(\varepsilon) \quad (1)$$

where A is a Hicks-neutral productivity parameter, $\alpha \in (0, 1)$ is the token elasticity of software output, $\gamma > 0$ is the developer capability elasticity, $\delta > 0$ is the model capability elasticity, and ε is a stochastic productivity shock. The Cobb-Douglas specification reflects constant elasticity of substitution among inputs, a standard assumption in production economics that provides tractable closed-form solutions for factor demand and cost functions.

The TPF has four key implications. First, the token elasticity $\alpha < 1$ implies diminishing marginal returns to inference tokens: each additional token produces less additional software output than the last, consistent with the compute ceiling binding at high usage intensities. Second, the model capability elasticity δ generates a quality-quantity tradeoff in token consumption: higher-capability models produce more output per token, but also command higher prices, so the optimal token consumption depends on the ratio δ/α relative to the price ratio. Third, the Hicks neutrality of A implies that productivity shocks augment all inputs proportionally, consistent with the finding in Brass (2026d) that simulated developer productivity improvements raise both token consumption and software output in the same proportion. Fourth, the constant returns to scale in (T, C, M) are rejected in the data ($\hat{\alpha} + \hat{\gamma} + \hat{\delta} = 0.742 + 0.318 + 0.841 = 1.901 > 1$), implying increasing returns to scale consistent with the learning-by-doing and complementarity effects identified in the empirical literature.

Empirical estimation of the TPF parameters from the simulation evidence in Brass (2026a,

2026d) yields:

$$\hat{\alpha} = 0.742 \quad (\text{SE } 0.038) \quad (2)$$

$$\hat{\gamma} = 0.318 \quad (\text{SE } 0.028) \quad (3)$$

$$\hat{\delta} = 0.841 \quad (\text{SE } 0.047) \quad (4)$$

confirming significant diminishing returns, positive developer capability effects, and strong model capability effects that dominate developer capability effects in determining software output.

B. The Token Kuznets Curve

The Token Kuznets Curve (TKC) describes the non-monotonic relationship between software development stage (or software task complexity) and token intensity (tokens per unit of software output). The TKC is named by analogy with the Environmental Kuznets Curve, which describes the inverted-U relationship between income per capita and environmental degradation — rising initially as economies industrialise, then falling as they become wealthy enough to invest in clean technology. The TKC describes an analogous inverted-U: token intensity rises as tasks move from simple to moderate complexity (the “ramp phase”), peaks at some critical complexity level σ^* , and falls thereafter as the compute ceiling imposes a hard constraint on output growth (the “plateau phase”).

The TKC is formally characterised by the token intensity function $\tau(\sigma) = T(\sigma)/S(\sigma)$, where σ is software task complexity on the unit interval. The estimated peak of the TKC occurs at $\hat{\sigma}^* = 0.731$ (Brass 2026e, 2026i), implying that approximately 27 percent of software development tasks operate in the plateau region above this threshold. The TKC has direct implications for technology adoption curves, compute infrastructure planning, and the optimal design of token pricing schedules across developer types.

The TKC connects to the four pillars as follows. Production economics provides the diminishing returns to tokens ($\alpha < 1$) that generate the plateau phase. Platform economics determines the price of tokens, which affects where on the complexity dimension the compute ceiling first binds. Digital infrastructure markets set the compute ceiling τ_d itself, which is the physical constraint that generates the plateau. AI innovation systems connect the TKC to long-run model improvement: as models become more capable, the peak σ^* shifts rightward, meaning that fewer tasks operate in the plateau and the overall efficiency of token use improves.

C. The Jevons Paradox of AI Tokens

The Jevons Paradox of AI Tokens is the theoretical prediction that improvements in AI model efficiency (higher capability per token) paradoxically *increase* rather than

decrease aggregate token consumption. This paradox operates through two channels. The *price channel*: as models become more capable, the effective cost of AI-assisted software production falls, attracting new developers into token consumption (the extensive margin) and inducing existing developers to use more tokens (the intensive margin). The *agentic channel*: as models become more capable, they are deployed in increasingly recursive agentic workflows that consume tokens to plan, execute, and self-evaluate, generating a demand multiplier that exceeds the naive linear prediction.

The formal statement of the Jevons Paradox (Proposition 3 in Brass 2026f, 2026i) proves that the agentic multiplier

$$\mu(\lambda) = \frac{1}{1 - \lambda} \quad (5)$$

strictly exceeds the naive linear approximation $1 + \lambda$ for all $\lambda > 0$, with the excess amplification

$$\Delta(\lambda) = \frac{\lambda^2}{1 - \lambda} \quad (6)$$

being strictly positive and convex-increasing in λ . This superlinearity means that as agentic AI workflows deepen — as λ rises toward 1 — the Jevons amplification grows without bound. The mean agentic multiplier estimated from the simulation evidence is $\hat{\mu} = 1.386$, with the Jevons ratio $\hat{\mu}/(1 + \hat{\lambda}) = 1.587$ significantly exceeding unity (Wald $F = 68.3$, $p < 0.001$).

The policy implications of the Jevons Paradox are profound. Efficiency improvements that reduce the per-token cost of AI assistance — through model compression, distillation, speculative decoding, or architectural innovations — will not reduce aggregate infrastructure demand. On the contrary, they will increase it, both through expanded adoption and through deeper agentic engagement. Infrastructure planners who extrapolate from efficiency gains to demand reductions will systematically underestimate future compute requirements.

D. The General Equilibrium of the Token Economy

The General Equilibrium (GE) framework of Inference Economics (Brass 2026g) integrates oligopolistic inference supply with competitive developer demand in a three-market system: the token market (where developers demand and providers supply inference), the developer market (where the token market’s prices and capabilities feed back into developer productivity and software supply), and the GPU compute market (where infrastructure supply constrains inference supply from below). The GE framework establishes five results.

First, *existence*: under standard regularity conditions, a competitive equilibrium in the three-market system exists, with prices in all three markets clearing simultaneously (Proposition 1 of Brass 2026g). Second, *Cournot pricing*: with $N = 4$ frontier providers,

the equilibrium token price is

$$P^* = \frac{a + Nc}{N + 1} = \$1.40 \text{ per million tokens (calibrated)} \quad (7)$$

exceeding the competitive price of $c = \$0.50$ and generating a deadweight loss of 18% of the social optimum. Third, *GPU scarcity premium*: the compute scarcity premium adds 8–13% to the competitive token price, raising total welfare loss to 26–31% of the social optimum. Fourth, *compute investment cycle*: the panel VAR of GPU investment exhibits autoregressive coefficients $\hat{\phi} = 0.308$ and $\hat{\lambda} = 0.538$, consistent with moderately persistent investment cycles driven by demand shocks. Fifth, *policy counterfactuals*: a 20% reduction in token price (e.g., from antitrust intervention or public provision) increases aggregate token demand by 29.4%, consistent with the mean price elasticity of -0.507 .

E. Micro-Econometric Foundations: Two-Part Demand and Copula Dependence

The micro-econometric foundation of Inference Economics is the bivariate copula two-part model of developer demand for inference tokens (Brass 2026h, 2026i). This model decomposes developer token demand into two qualitatively distinct decisions: the *participation decision* (whether to use a given AI model at all in a given period) and the *intensity decision* (how many tokens to consume conditional on participation). The two decisions are linked through a Gaussian copula with estimated correlation parameter $\hat{\rho} = 0.444$ (Brass 2026h), capturing the positive dependence between unobservable productivity determinants of adoption and consumption.

The copula structure captures a key selection effect: developers with unobservably high productivity are simultaneously more likely to participate in AI token markets and to consume more tokens conditional on participation. Ignoring this correlation — as standard two-part models do — generates selection bias that attenuates price elasticity estimates by approximately 18.5% and understates cross-model switching propensity by approximately 22.3%.

The multi-level hierarchical extension of the two-part model — developers nested within models nested within companies — captures the three-level structure of the AI inference market and generates interpretable variance decompositions. Company-level effects account for 21.8% of total outcome variance ($\widehat{ICC}_c = 0.218$) and model-level effects account for an additional 17.4% ($\widehat{ICC}_m = 0.174/0.782 = 0.222$), confirming that both firm-level investment decisions and model-specific characteristics are important determinants of aggregate token demand.

Table 1 summarises the complete Inference Economics framework, mapping each construct to its pillar, key equation or result, and source paper.

Table 1: The Inference Economics Framework: Pillars, Constructs, and Key Results.

Pillar	Core Construct	Key Equation / Result	Source
<i>Pillar I: Production Economics</i>			
Prod. Econ.	Token Production Function	$S = AT^\alpha C^\gamma M^\delta$; $\hat{\alpha} = 0.742$	a, d, i
	Token Kuznets Curve	Peak token intensity at $\hat{\sigma}^* = 0.731$	e, i
<i>Pillar II: Platform Economics</i>			
Platform Econ.	Cournot Oligopoly ($N = 4$)	$P^* = (a + Nc)/(N + 1) = \$1.40/\text{M tokens}$	b, g
	Two-Part Developer Demand	Copula $\hat{\rho} = 0.444$; price elas. = -0.507	h, i
	Two-Sided Market Welfare	$DWL_{\text{Cournot}} = 18\%$ of social optimum	g
<i>Pillar III: Digital Infrastructure</i>			
Digital Infra.	GPU Scarcity Premium	$\psi(K) = b(Q^* - \theta K)$; +8–13% price	g
	Compute Investment Cycle	$\hat{\phi} = 0.308$, $\hat{\lambda} = 0.538$	g
	Welfare Decomposition	Total loss = 26–31% of social optimum	g
<i>Pillar IV: AI Innovation Systems</i>			
AI Innov. Sys.	Jevons Paradox (Agentic)	$\mu(\lambda) = 1/(1 - \lambda)$; mean = 1.386	f, i
	Agentic Amplification Test	Jevons ratio = 1.587; Wald $F = 68.3^{***}$	f, h, i
	Token-Augmented Growth	Token augmentation raises TFP growth	c
<i>Synthesis</i>			
All Pillars	General Equilibrium	Existence + 5 Propositions	g
All Pillars	Policy Counterfactuals	20% price cut \rightarrow +29.4% demand	f, g, h, i

Notes: Key results calibrated to frontier AI pricing structure (OpenAI, Anthropic, Google, xAI), March 2026. α denotes the token elasticity of software output. ϕ and λ denote autoregressive coefficients in the compute investment panel VAR. *** $p < 0.001$.

V. INFERENCE ECONOMICS AND THE PARADIGM OF AI-AUGMENTED INNOVATION

A. Connections to Production and Growth Economics

Inference Economics connects most directly to the neoclassical and endogenous growth traditions in economics. The Token Production Function is an extension of the Cobb-Douglas production function — the workhorse model of growth economics since Solow (1956) — that adds cognitive capital (inference tokens) alongside physical capital and labor. The token elasticity $\hat{\alpha} = 0.742$ can be compared to the capital elasticity in standard growth models (≈ 0.33), and the sum of elasticities > 1 implies increasing returns to scale in the cognitive production technology, consistent with the AK-type growth models of Romer (1990).

The endogenous growth connection is through the quality-ladder models of Aghion and Howitt (1992) and Grossman and Helpman (1991). In quality-ladder models, innovation occurs when entrepreneurs invest in R&D to discover higher-quality versions of differentiated goods, and growth is driven by the arrival rate of innovations times the quality increment per innovation. Inference Economics proposes an AI-augmented quality ladder in which the arrival rate of software innovations is increasing in token consumption — inference-intensive development accelerates the pace of product improvement — and the quality increment per innovation is increasing in model capability. This generates a dynamic in which AI inference and economic growth are mutually reinforcing.

The token-augmented growth framework developed in Brass (2026c) formalises this connection. Using a Ramsey-Cass-Koopmans model augmented with inference tokens as a third factor of production (alongside physical capital and human capital), the framework derives conditions under which token-augmented growth exceeds the growth rate of a standard AK economy. The key parameter is the cross-elasticity between model capability improvement and aggregate TFP growth: if this elasticity exceeds the model capability elasticity in the TPF (i.e., if aggregate learning effects exceed private learning effects), then AI inference generates positive external growth effects that justify subsidization from a public finance perspective.

B. Connections to Platform and Information Economics

Inference Economics connects to platform economics through the two-sided market structure of AI inference providers. Frontier AI companies operate platforms that connect model developers (who create and improve AI models) with token consumers (who use those models via APIs). The Rochet-Tirole (2003) framework for two-sided markets applies, with the complication that AI inference platforms are asymmetrically two-sided:

the model developer side is concentrated (few companies develop frontier models) while the token consumer side is dispersed (millions of developers use APIs). This asymmetry generates pricing dynamics not captured by symmetric two-sided market models.

The information economics connection runs through the theory of credence goods (Darby and Karni, 1973) and experience goods (Nelson, 1970). AI model capabilities — the quality parameters $M_{c,m}$ in the TPF — are not fully observable to developers before they consume tokens; they are learned through experience and through signals (benchmark scores, user reviews, third-party evaluations). This creates adverse selection in the model market — developers may systematically choose lower-capability models to avoid high prices — and generates a role for quality certification and standardised evaluation frameworks.

The network externality connection runs through the training data feedback loop: models improve as they process more tokens (through fine-tuning and RLHF on usage data), meaning that higher token consumption generates positive externalities for future model quality. This creates a chicken-and-egg dynamic characteristic of network externality markets (Katz and Shapiro, 1985): early adoption generates quality improvements that attract further adoption, creating winner-take-most dynamics in model markets.

C. Connections to Innovation Economics and Technology Transitions

Inference Economics connects to the economics of General Purpose Technologies (GPTs) through the characterisation of AI inference as a transformative input that improves over time, is applicable across a wide range of applications, and generates innovation complementarities with the users who adopt it. Bresnahan and Trajtenberg (1995) defined GPTs by three properties: pervasiveness (used as input in many sectors), improvement (quality rises over time through learning), and innovation complementarities (adopters co-invent new uses). AI inference satisfies all three criteria with unusual clarity, suggesting that the GPT framework of Helpman (1998) is directly applicable to the analysis of AI-augmented growth.

The Schumpeterian innovation economics connection runs through the creative destruction framework (Schumpeter, 1942). Inference Economics predicts that AI inference disrupts existing software production functions by replacing human cognitive labor with token consumption, generating Schumpeterian churning in the software industry: AI-native software companies with high token elasticities outcompete legacy firms with low token elasticities, generating both productivity gains (more output from fewer cognitive workers) and distributional costs (displacement of medium-skill cognitive labor).

The technology systems connection runs through the Lipsey et al. (2005) framework for technology transitions, which emphasises that the economic value of a transformative technology is realised through co-invention of complementary innovations. For AI inference,

the complementary innovations include: (i) prompt engineering methodologies that improve the efficiency of developer-AI interaction; (ii) agentic workflow architectures that structure multi-step AI task execution; (iii) evaluation frameworks that reduce information asymmetry in the model market; and (iv) data infrastructure that enables model fine-tuning and domain adaptation. The rate of co-invention in these complementary areas is a key determinant of how quickly the growth dividend from AI inference is realised.

D. The Inference Gap: Inequality in Cognitive Capital Access

Perhaps the most important connection between Inference Economics and the broader economics literature is through the economics of inequality and access. The copula two-part model reveals that developer productivity (the unobservable component driving both adoption and intensity) is the key determinant of who benefits most from AI inference markets. Developers with high unobservable productivity — who are more likely to adopt AI tools and to use them intensively — capture a disproportionate share of the consumer surplus generated by AI inference, while low-productivity developers, constrained by high prices and low returns per token, remain at the margin of adoption.

We term this phenomenon the *Inference Gap*: the differential in AI inference access and utilisation between high-productivity and low-productivity developers, between high-income and low-income countries, and between resource-rich and resource-poor organisations. The Inference Gap operates through three channels. First, the *price channel*: high token prices disproportionately constrain low-productivity developers for whom the marginal product of tokens is below the market price, generating participation barriers analogous to the “digital divide” in ICT adoption. Second, the *capability channel*: high-capability models, which generate the largest productivity gains, are priced at a premium that further concentrates adoption among high-productivity developers. Third, the *agentic channel*: the benefits of agentic AI amplification accrue primarily to developers who can construct and maintain sophisticated agentic workflows, a skill correlated with productivity.

The Inference Gap has a geographic dimension that connects Inference Economics to the economics of international development and technological catch-up. AI inference markets are structured around frontier providers headquartered in the United States, with pricing, model availability, and governance decisions made by American corporations. Developing-country developers face the same token prices as high-income developers (with no PPP adjustment) while earning software output that, valued at local prices, is worth far less per token consumed. This generates a structural disadvantage for AI-augmented software development in the Global South that mirrors — in the domain of cognitive capital — the structural disadvantage in physical capital access that has characterised development economics since [Solow \(1956\)](#).

VI. A RESEARCH AGENDA FOR INFERENCE ECONOMICS

The theoretical architecture and intellectual connections outlined in the preceding sections point to a rich frontier of open research questions. We organise the research agenda into seven directions, each of which addresses a fundamental gap in the current state of Inference Economics. For each direction, we identify the core questions, the appropriate methodological approaches, and the anticipated contributions to both the subfield and the broader scientific community. Table 2 maps these directions onto the four pillars and key methodological tools.

Research Direction 1: Empirical Measurement of the Token Production Function

The Token Production Function is the foundational construct of Inference Economics, but its empirical parameters (α, γ, δ) have been estimated only from synthetic simulation data calibrated to stylised facts rather than from real-world API usage logs. The first priority of the empirical program is to estimate the TPF from actual developer productivity data linked to token consumption records. Core research questions: What is the true token elasticity of software output in different programming domains (web development, data science, machine learning, embedded systems)? How does the developer capability elasticity γ vary across experience levels and educational backgrounds? Does the model capability elasticity δ differ systematically across model tiers (frontier vs. mini)?

Methodological approaches include instrumental variable estimation (using exogenous variation in model API availability and pricing as instruments for token consumption), difference-in-differences designs (exploiting the staggered rollout of new model versions), and structural production function estimation (adapting the Olley-Pakes (1996) and Akerberg-Caves-Frazer (2015) methods to cognitive production). Anticipated contributions: the first credibly identified estimates of cognitive capital productivity in AI-assisted software production.

Research Direction 2: Market Design for AI Inference

The current market structure of AI inference — oligopolistic supply by four frontier providers, per-token pricing, and unilateral price-setting — is unlikely to be the welfare-maximising market design. Core research questions: What is the optimal number of providers in a Cournot inference market, balancing between the static efficiency gains from more competition and the dynamic efficiency losses from reduced incentives to invest in model training? Would a tiered pricing system improve welfare and reduce the Inference Gap? Is there a role for a public option in AI inference (a government-operated model provider analogous to public broadcasting)? Could standardised compute futures markets reduce the volatility of GPU scarcity premia?

Methodological approaches include mechanism design theory (optimal auction and pricing mechanisms for capacity-constrained services), structural industrial organisation (counterfac-

tual simulations of merger and entry in AI inference markets), and experimental economics (laboratory tests of alternative token pricing mechanisms). Anticipated contributions: policy-relevant insights for AI market regulation and antitrust enforcement.

Research Direction 3: The Macro-Growth Economics of AI Augmentation

Inference Economics has potentially transformative implications for economic growth, but the macro-growth foundations remain underdeveloped. Core research questions: What is the GDP growth contribution of AI inference across different adoption scenarios? Does AI inference operate as a Schumpeterian innovation (creative destruction, concentrated gains among high-productivity firms) or as an AK-type growth driver (broadly distributed productivity gains)? How do oligopolistic rents captured by frontier AI providers affect the distribution of AI-augmented growth between capital and labor?

Methodological approaches include calibrated DSGE models with AI inference as a factor of production, cross-country panel regressions of AI adoption on productivity growth, and I-O analysis of AI inference as an intermediate input across industries. Anticipated contributions: the first rigorous macroeconomic quantification of the AI inference growth dividend.

Research Direction 4: Welfare Analysis and the Inference Gap

The distributional consequences of AI inference markets remain poorly understood. Core research questions: What is the consumer surplus generated by AI inference APIs for different developer types, and how does this compare to the deadweight loss from oligopolistic pricing? How does the Inference Gap evolve as token prices fall — do price reductions preferentially benefit low-productivity developers (reducing the gap) or high-productivity developers (widening the gap, through the Jevons multiplier)? What is the welfare cost of GPU compute concentration? Are there policy interventions — subsidised token access for educational institutions, public compute infrastructure, open model repositories — that can reduce the Inference Gap without significantly reducing innovation incentives?

Methodological approaches include structural welfare analysis (measuring consumer surplus and deadweight loss from the copula two-part demand model), inequality decomposition (separating within-group from between-group inference intensity variation), and policy simulation. Anticipated contributions: the first welfare accounting framework for AI inference markets.

Research Direction 5: Agentic AI — Recursive Demand and Dynamic Stability

The economics of agentic AI systems — in which AI models execute multi-step autonomous tasks through recursive model calls — is the most novel and least understood area of Inference Economics. Core research questions: What are the conditions under which agentic AI demand is stable (converging to a finite token consumption level) versus explosive (growing without bound until a compute ceiling is hit)? The Jevons multiplier $\mu(\lambda) = 1/(1 - \lambda)$

diverges as $\lambda \rightarrow 1$, implying that sufficiently recursive agentic systems would consume infinite tokens in principle — what practical constraints prevent this theoretical instability? How does the Jevons Paradox interact with model improvement cycles?

Methodological approaches include dynamic systems analysis (characterising the stability conditions of agentic demand recursion), experimental evidence on agentic task completion rates and token efficiency, and labor economics analysis of agentic AI adoption in specific industries. Anticipated contributions: the first formal economic analysis of agentic AI market dynamics.

Research Direction 6: International Trade and Geopolitics of AI Inference

AI inference is a globally traded service, but the economics of cross-border token trade have not been formally analysed. Core research questions: What is the comparative advantage basis for AI inference trade — which countries are net importers and which are net exporters? How do data localisation regulations affect the efficiency of global inference markets? Does the concentration of frontier AI capability in US-based providers constitute a form of technological monopoly power with geopolitical implications analogous to OPEC’s oil supply concentration? How do differential regulatory environments (EU AI Act, US executive orders, Chinese AI regulations) affect the global allocation of AI inference supply?

Methodological approaches include international trade models adapted for digital services (Helpman-Krugman with cognitive capital as a factor), gravity models of cross-border API usage, and political economy analysis of AI governance coalitions. Anticipated contributions: a new international economics of AI services.

Research Direction 7: Governance, Regulation, and AI Infrastructure Policy

The governance of AI inference markets requires economic analysis that Inference Economics is uniquely positioned to provide. Core research questions: How should AI inference providers be regulated — as public utilities (rate-of-return regulation), as dominant firms (abuse of dominant position under competition law), or as platform operators (non-discrimination, interoperability requirements)? What is the optimal public investment in AI compute infrastructure? How should AI inference markets be monitored for systemic risk — are there scenarios in which a technical failure or deliberate disruption of a major inference provider would cause significant macroeconomic harm? What intellectual property framework best balances innovation incentives with access and competition?

Methodological approaches include regulatory economics (rate-of-return regulation, price caps), public finance analysis of compute infrastructure investment, systems risk analysis, and law and economics of AI IP frameworks. Anticipated contributions: an economics-grounded framework for AI inference regulation that could inform policy design in major jurisdictions.

Table 2: Research Agenda Matrix: Directions, Pillars, Methods, and Contributions.

Direction	Primary Pillar	Key Methods	Core Contribution
1: TPF Measurement	Prod. Econ.	IV; DiD; Structural PF	Identified token elasticities
2: Market Design	Platform Econ.	Mechanism Design; SIO; Exp.	Optimal inference market rules
3: Macro Growth	AI Innov. Sys.	DSGE; Panel Growth; I-O	AI growth dividend estimate
4: Welfare/Equity	All Four Pillars	Structural Welfare; Decomp.	Inference Gap accounting
5: Agentic Dynamics	AI Innov. Sys.	Dynamic Systems; Exp.; Labor	Agentic stability conditions
6: International Trade	Platform + Infra.	Trade Models; Gravity; PolEcon	Global inference trade theory
7: Governance/Policy	All Four Pillars	Reg. Econ.; PF; Risk Analysis	AI regulation framework
<i>Methodological Foundations</i>			
Copula Model	Two-Part Prod. + Plat.	FIML; Monte Carlo; Power	Developer demand benchmark
General Equilibrium	All Four Pillars	GE Existence; Propositions	System-level policy analysis
Token Kuznets Curve	Prod. + Innov.	Quadratic Panel; Fixed FX	Infrastructure planning tool

Notes: IV = Instrumental Variables. DiD = Difference-in-Differences. SIO = Structural Industrial Organisation. Exp. = Experimental Economics. PF = Public Finance. DSGE = Dynamic Stochastic General Equilibrium. FIML = Full Information Maximum Likelihood.

The seven research directions are not independent: they constitute an interconnected research program whose components reinforce each other. Empirical measurement of the TPF (Direction 1) provides the structural parameters needed for market design simulations (Direction 2) and welfare analysis (Direction 4). The macro-growth framework (Direction 3) depends on the micro-foundations developed in Direction 1 and the market structure characterised in Direction 2. The governance framework (Direction 7) requires the welfare analysis from Direction 4, the dynamic stability analysis from Direction 5, and the international context from Direction 6. We envision the research agenda as progressing iteratively, with early empirical work on TPF measurement and copula demand models providing the foundation for later theoretical and policy work on governance, international

trade, and agentic AI dynamics.

VII. CONCLUSION: INFERENCE ECONOMICS AS A UNIFYING FRAMEWORK

Inference Economics emerges at a moment when the gap between the pace of AI technological development and the sophistication of economic analysis applied to it has become a first-order policy problem. Regulators are designing competition policy for AI markets without formal models of AI market structure. Infrastructure planners are projecting AI compute demand without formal theories of the Jevons Paradox. Development economists are assessing the distributional consequences of AI adoption without formal models of the Inference Gap. Inference Economics provides the theoretical and empirical tools to address all three gaps simultaneously.

The five core constructs of Inference Economics — the Token Production Function, the Token Kuznets Curve, the Jevons Paradox of AI Tokens, the General Equilibrium of the Token Economy, and the Copula Two-Part Demand Model — provide a positive theory of AI inference markets that is both theoretically rigorous and empirically tractable. The calibrated parameters from the nine-paper foundational series (Brass 2026a-i) generate specific, testable predictions: a token elasticity of software output of 0.742, equilibrium token prices 2.8 times the competitive level, a Jevons amplification ratio of 1.587, and a Cournot welfare loss of 26–31% of the social optimum. These predictions can be tested against real-world data as frontier AI markets mature and as API usage data becomes available to researchers.

The four intellectual pillars — production economics, platform economics, digital infrastructure markets, and AI innovation systems — provide the disciplinary foundations from which Inference Economics draws and to which it contributes. Production economists gain a new framework for analysing cognitive capital as a production input. Platform economists gain a new application of two-sided market theory in which the recursive demand dynamics of agentic AI generate non-linear welfare effects. Infrastructure economists gain a new class of network capacity problems in which the scarcity good (GPU compute) is itself the input to a platform (inference supply) that serves two-sided markets. Innovation economists gain a new GPT case study with uniquely rapid diffusion, recursive demand feedback, and oligopolistic supply concentration.

The seven-direction research agenda constitutes an invitation to the broader scientific community. The questions raised by Inference Economics span economics, computer science, management, public policy, and international relations. The methodological toolkit required encompasses structural econometrics, mechanism design, dynamic systems analysis, trade theory, and regulatory economics. The policy stakes — governing the most

consequential technology transition since the ICT revolution — make this among the highest-priority research programs in contemporary economic science.

We conclude with a normative observation. The central insight of Inference Economics is that AI inference tokens are not merely a technology product but an *economic good* with all the properties — scarcity, allocation problems, distributional consequences, externalities — that justify economic analysis and policy intervention. Treating AI inference as purely a technological matter, governed by market forces and engineering optimisation, risks systematically underinvesting in the public goods (compute infrastructure, open models, evaluation standards) that are complementary to frontier AI and systematically ignoring the distributional consequences (the Inference Gap, the Jevons rebound, the creative destruction of cognitive labor) that follow from unregulated oligopolistic supply. Inference Economics provides the analytical framework to do better.

REFERENCES

REFERENCES

- Abel, A. B., Dixit, A. K., Eberly, J. C., & Pindyck, R. S. (1996). Options, the value of capital, and investment. *Quarterly Journal of Economics*, 111(3), 753–777.
- Ackerberg, D. A., Caves, K., & Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6), 2411–2451.
- Aghion, P., & Howitt, P. (1992). A model of growth through creative destruction. *Econometrica*, 60(2), 323–351.
- Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston: Harvard Business Review Press.
- Anthropic. (2025). Claude API Pricing and Documentation. <https://www.anthropic.com>. Accessed March 2026.
- Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4), 1279–1333.
- Brass. (2026a). The Token Production Function: AI Inference and the Economics of Software Output. *American Economic Review* (forthcoming).
- Brass. (2026b). Oligopolistic Inference Pricing and Developer Welfare in AI Token Markets. *American Economic Review* (forthcoming).
- Brass. (2026c). Token-Augmented Growth: Inference Economics and the Macroeconomics of AI Adoption. *American Economic Review* (forthcoming).
- Brass. (2026d). The Empirical Token Production Function: Simulation Evidence on AI Inference Elasticities. *American Economic Review* (forthcoming).
- Brass. (2026e). The Token Kuznets Curve: AI Inference Demand and the Evolution of the Software Development Economy. *American Economic Review* (forthcoming).
- Brass. (2026f). The Jevons Paradox of AI Tokens: Efficiency Gains, Rebound Effects, and the Economics of Inference Demand. *American Economic Review* (forthcoming).
- Brass. (2026g). The General Equilibrium of the AI Token Economy: Oligopolistic Inference Supply and Competitive Developer Demand. *American Economic Review* (forthcoming).
- Brass. (2026h). Micro-Econometric Modeling of Developer Demand for AI Inference Tokens: A Two-Part Model with Copula Dependence and Multi-Level Hierarchical Structure. *American Economic Review* (forthcoming).
- Brass. (2026i). The AI Token Economy: Participation, Consumption, and the Jevons Paradox of Agentic AI. *American Economic Review* (forthcoming).

- Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies: Engines of growth? *Journal of Econometrics*, 65(1), 83–108.
- Brynjolfsson, E., & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: Norton.
- Cachon, G., & Feldman, P. (2011). Pricing services subject to congestion. *Manufacturing and Service Operations Management*, 13(2), 244–260.
- Darby, M. R., & Karni, E. (1973). Free competition and the optimal amount of fraud. *Journal of Law and Economics*, 16(1), 67–88.
- Eisenmann, T., Parker, G., & Van Alstyne, M. W. (2006). Strategies for two-sided markets. *Harvard Business Review*, 84(10), 92–101.
- Google AI. (2025). Gemini Model Suite and Pricing. <https://developers.google.com/ai/gemini>. Accessed March 2026.
- Grossman, G. M., & Helpman, E. (1991). *Innovation and Growth in the Global Economy*. Cambridge: MIT Press.
- Helpman, E. (Ed.). (1998). *General Purpose Technologies and Economic Growth*. Cambridge: MIT Press.
- Hooker, S. (2021). The hardware lottery. *Communications of the ACM*, 64(12), 58–65.
- Jacques, I. (2021). *Mathematics for Economics and Business* (8th ed.). London: Pearson.
- Jevons, W. S. (1865). *The Coal Question: An Inquiry Concerning the Progress of the Nation and the Probable Exhaustion of Our Coal Mines*. London: Macmillan.
- Joe, H. (2014). *Dependence Modeling with Copulas*. Boca Raton: CRC Press.
- Katz, M. L., & Shapiro, C. (1985). Network externalities, competition, and compatibility. *American Economic Review*, 75(3), 424–440.
- Khazzoom, J. D. (1980). Economic implications of mandated efficiency standards for household appliances. *Energy Journal*, 1(4), 21–40.
- Lipsey, R. G., Carlaw, K. I., & Bekar, C. T. (2005). *Economic Transformations: General Purpose Technologies and Long-Term Economic Growth*. Oxford: Oxford University Press.
- Lundvall, B.-A. (1992). *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning*. London: Pinter.
- Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy*, 78(2), 311–329.
- Nelson, R. R. (Ed.). (1993). *National Innovation Systems: A Comparative Analysis*. Oxford: Oxford University Press.

- Nelsen, R. B. (2006). *An Introduction to Copulas* (2nd ed.). New York: Springer.
- Olley, G. S., & Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64(6), 1263–1297.
- OpenAI. (2025). ChatGPT API Pricing and Documentation. <https://platform.openai.com>. Accessed March 2026.
- Rochet, J.-C., & Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4), 990–1029.
- Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5), S71–S102.
- Rysman, M. (2009). The economics of two-sided markets. *Journal of Economic Perspectives*, 23(3), 125–143.
- Schumpeter, J. A. (1942). *Capitalism, Socialism, and Democracy*. New York: Harper.
- Shapiro, C., & Varian, H. R. (1998). *Information Rules: A Strategic Guide to the Network Economy*. Boston: Harvard Business School Press.
- Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.
- Solow, R. M. (1956). A contribution to the theory of economic growth. *Quarterly Journal of Economics*, 70(1), 65–94.
- Sorrell, S. (2007). *The Rebound Effect: An Assessment of the Evidence for Economy-Wide Energy Savings from Improved Energy Efficiency*. London: UK Energy Research Centre.
- Tirole, J. (1988). *The Theory of Industrial Organization*. Cambridge: MIT Press.
- Vickrey, W. (1969). Congestion theory and transport investment. *American Economic Review*, 59(2), 251–260.
- xAI. (2025). Grok API Documentation and Pricing. <https://www.x.ai>. Accessed March 2026.