Measuring AI Capital Services Through Compute: A Framework for Productivity Analysis and Monetary Policy

Joel Christoph*

March 2026

Abstract

Recent advances in artificial intelligence are beginning to affect production, task allocation, and the composition of investment, but standard macroeconomic statistics remain poorly suited to measuring these effects in real time. This paper develops a theoretical framework for measuring AI capital services through compute and for mapping these services into productivity analysis relevant to monetary policy.

The core idea is that compute is not merely an intermediate input used by technology firms, but an economically meaningful carrier of AI capital services whose productive contribution depends on model quality, utilisation, task substitutability, and complementarities with labour and organisational capital. I formalise these relationships in a growth-accounting framework that distinguishes between raw compute inputs, effective AI capital services, and realised output gains. This helps clarify when measured productivity may lag behind underlying technological change and when official statistics may overstate or understate inflationary pressure.

The paper makes three contributions. First, it provides a tractable way to analyse AI as a form of capital deepening rather than as a diffuse general-purpose technology shock. Second, it identifies channels through which AI adoption can generate temporary wedges between observed prices, quality-adjusted output, and underlying marginal costs. Third, it shows how heterogeneous diffusion across sectors and occupations can produce mixed signals for labour market slack, trend productivity, and the appropriate stance of monetary policy.

The paper is theoretical, but it is designed to support empirical implementation using data on compute deployment, model capabilities, occupational exposure, sectoral

*Harvard Kennedy School and Economic Research Institute for ASEAN and East Asia (ERIA). Email: joel.christoph@proton.me. *Disclosure:* The author has no financial relationships or other potential conflicts of interest related to this research. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

adoption, and emerging evidence on AI-related productivity, employment, and inflation dynamics. I discuss how the framework can be brought to data and how it can improve interpretation of productivity dynamics during periods of rapid AI diffusion. For central banks, the main implication is that unmeasured AI capital can distort real-time estimates of trend productivity, potential output, sectoral inflation, and labour-market slack, even when aggregate demand conditions are unchanged.

By linking AI adoption, productivity measurement, and monetary policy analysis, the paper offers a framework for interpreting productivity, inflation, labour-market slack, and potential output during rapid AI diffusion.

JEL Codes: E01, E22, E31, E52, O33, O47

Keywords: artificial intelligence, productivity measurement, capital services, compute, growth accounting, monetary policy, inflation, labour markets

1 Introduction

Artificial intelligence is reshaping production across a growing set of sectors in the modern economy. Firms deploy large language models and other AI systems for tasks ranging from customer service and legal research to coding, forecasting, and scientific discovery. Emerging evidence already points to meaningful productivity gains from generative AI at the worker and firm level, although effects are heterogeneous across tasks, firms, and industries (Noy and Zhang, 2023; Peng et al., 2023; Misch et al., 2025; Aldasoro et al., 2026). Yet despite the growing economic relevance of these systems, the measurement infrastructure economists rely on to track capital, productivity, and growth still offers no systematic way to quantify AI's contribution to output.

This gap matters for monetary policy. Growth accounting, the primary tool for decomposing the sources of economic expansion, requires that all economically significant inputs be identified, measured, and assigned appropriate factor shares. When an important input is omitted, its contribution is absorbed into the residual, inflating measured total factor productivity (TFP) and obscuring the true sources of growth. If central banks cannot distinguish genuine productivity acceleration from measurement artefacts driven by unmeasured AI capital, they risk misinterpreting the output gap, misjudging inflationary pressure, and setting policy on the basis of distorted signals. The rise of information technology in the 1990s created a similar measurement challenge, which was addressed through hedonic price indices for semiconductors and the incorporation of IT capital into national accounts (Jorgenson, 2001; Triplett, 2006; Nordhaus, 2007). What remains missing for AI is a framework that maps compute and capability improvements into capital services in a way that is usable for productivity measurement and monetary policy analysis (Hartmann and Maver, 2025; Melina and Villa, 2025; Gambacorta et al., 2025).

The measurement of AI capital is complicated by features that distinguish it from traditional capital, from IT hardware, and even from conventional software. Three properties are particularly consequential. First, trained AI models exhibit non-rivalry within the firm. Once a model has been trained at substantial cost, deploying it to additional tasks or users requires only the marginal cost of inference compute, not retraining from scratch. Second, AI capabilities are improving at rates that far exceed those of earlier general-purpose technologies. Frontier models become obsolete within months rather than years, and the effective capability per unit of compute has been growing by roughly an order of magnitude every one to two years (Hernandez and Brown, 2020). Third, AI capital services are deeply entangled with complementary human capital and organisational restructuring, making it difficult to attribute output gains to AI investment alone. Recent evidence suggests that productiv-

ity gains from AI adoption are real but uneven across firms and sectors, and often largest where complementary capabilities are strongest (Brynjolfsson et al., 2021; Misch et al., 2025; Aldasoro et al., 2026).

This paper develops a theoretical framework for measuring AI capital services and incorporating them into growth accounting, with particular attention to the implications for monetary policy analysis. The central idea is that compute, measured in standardised physical units (such as floating-point operations or GPU-hours) and adjusted for quality, can serve as a sufficient statistic for the contribution of AI to output under standard assumptions on market structure and firm behaviour. Compute is the common physical substrate through which AI systems generate economic value. Whether a firm uses AI for text generation, image recognition, or scientific simulation, the productive input it purchases from cloud providers or deploys on proprietary hardware is ultimately compute. This makes compute the natural analogue of "machine-hours" for conventional equipment or "kilowatt-hours" for electricity as an input to production.

The paper makes four main contributions. The first is to establish conditions under which the compute expenditure share in total revenue equals the output elasticity of AI capital services (Proposition 1). Under competitive factor markets, cost minimisation, and constant-returns-to-scale production technology, the statistician needs only data on compute prices and quantities, together with total revenue, to recover AI's share in output. Neither direct observation of the stock of trained models nor measurement of "AI services" in some abstract sense is required.

The second contribution is to develop a hedonic framework for constructing a quality index for compute (Propositions 2 and 3). Just as hedonic methods have been used to quality-adjust prices of semiconductors, personal computers, and other IT goods (Griliches, 1961; Triplett, 2006), I propose a quality index for AI compute based on model capability benchmarks. The structural relationships documented in the scaling laws literature (Kaplan et al., 2020; Hoffmann et al., 2022) provide an empirical foundation for this index, linking compute inputs to measurable capability characteristics.

The third contribution is to derive a decomposition of AI's contribution to output growth (Proposition 4). This decomposition separates AI's effect into three parts, namely compute deepening (growth in the physical quantity of compute per worker), compute quality improvement (growth in AI capability per unit of compute), and AI-augmenting technical change (improvements in how effectively firms convert AI services into output). This three-way decomposition provides a richer picture than the simple capital-deepening narrative that dominates current discussions.

The fourth contribution is to characterise the measurement bias that arises when AI

capital is omitted from growth accounting (Proposition 5). I show that the direct omission bias is the product of the compute expenditure share and the growth rate of quality-adjusted compute, and that total bias can be larger once compute's income share is misattributed to traditional capital and labour. As AI becomes a more important input to production, this bias will grow, potentially leading to substantial overestimation of TFP growth and distorting policy conclusions about the sources of economic performance. For central banks, this bias can generate misleading estimates of potential output and trend productivity, with direct consequences for the calibration of monetary policy.

The framework is designed with implementation in mind. Cloud computing providers record detailed usage data on compute consumption by customer and workload type. Hardware manufacturers track shipments of AI-capable processors. National statistical agencies are beginning to collect data on AI adoption and IT spending at the firm and industry level. Recent empirical work already points in this direction, using cross-country, firm-level, and experimental data to quantify AI-related productivity gains and their heterogeneity across tasks, firms, and regulatory environments (Noy and Zhang, 2023; Peng et al., 2023; Misch et al., 2025; Aldasoro et al., 2026; Borowski et al., 2025; Gambacorta et al., 2025). The theoretical results in this paper identify exactly which data are needed, how they should be combined, and what assumptions underpin the resulting estimates.

This paper relates to several strands of the literature. It builds on the tradition of growth accounting initiated by Solow (1957) and extended by Jorgenson and Griliches (1967), adapting the framework to handle an input whose economic importance is rising rapidly but whose services are poorly measured. It draws on the hedonic measurement literature (Griliches, 1961; Triplett, 2006) for the quality-adjustment component. It connects to recent evidence on generative AI and worker productivity (Noy and Zhang, 2023; Peng et al., 2023), task-based macroeconomic analyses of AI (Acemoglu, 2024), new evidence on AI adoption, productivity, employment, inflation, and short-run growth effects (Misch et al., 2025; Aldasoro et al., 2026; Borowski et al., 2025; Gambacorta et al., 2025), and early work on the implications of AI for monetary policy and the natural rate (Hartmann and Maver, 2025; Melina and Villa, 2025). It also contributes to the literature on intangible capital measurement (Corrado et al., 2009), by proposing a concrete approach for measuring one particularly important and elusive form of intangible investment.

The rest of the paper is organised as follows. Section 2 discusses the measurement problem posed by AI capital in greater detail. Section 3 presents the production framework and defines AI capital services. Section 4 establishes the main sufficiency result. Section 5 develops the hedonic quality adjustment methodology. Section 6 derives the growth accounting decomposition and characterises the measurement bias. Section 7 considers extensions, in-

cluding international comparisons, welfare measurement, and distributional considerations. Section 8 concludes.

2 AI Capital and the Measurement Problem

To appreciate why AI capital poses a distinctive measurement challenge, it is useful to review how earlier forms of capital and technology have been incorporated into national accounts and growth accounting.

2.1 Lessons from IT Capital Measurement

The measurement of information technology capital was transformed in the 1990s by the adoption of hedonic price indices for semiconductors and computers. Griliches (1961) introduced the hedonic approach, which decomposes the price of a good into the implicit prices of its characteristics. Triplett (2006) provided a comprehensive guide to hedonic methods for quality adjustment. Applied to IT, these methods allowed statisticians to separate the enormous improvements in computing power per dollar from pure price changes, yielding quality-adjusted investment series that dramatically revised the picture of IT capital accumulation (Jorgenson, 2001).

The IT capital measurement framework rests on a chain of well-defined steps. Hardware investment is observed through expenditure data. Hedonic price indices convert nominal investment into real (quality-adjusted) terms. The perpetual inventory method accumulates real investment into capital stocks, applying depreciation rates estimated from used-equipment markets or engineering data. Capital service flows are then computed by multiplying the capital stock by an estimated service price (the Jorgensonian user cost), which incorporates the interest rate, depreciation, and expected capital gains.

This approach has been enormously successful for IT hardware, but it does not translate directly to AI. The fundamental reason is that AI capital takes a very different form from IT hardware. Where IT capital is embodied in physical equipment (servers, routers, personal computers), AI capital is embodied in trained models, which are weightless, non-rival, and rapidly depreciating bundles of parameters produced by computationally intensive training processes.

2.2 Three Distinctive Properties of AI Capital

Three properties of AI capital create measurement challenges that go beyond those encountered with IT hardware or conventional software.

The first property is non-rivalry within the firm. A trained neural network, once produced, can be deployed simultaneously across an arbitrary number of tasks, users, and applications within the organisation, at a marginal cost that consists only of the inference compute required to run the model. This contrasts with physical capital, where each machine can only be used for one task at a time, and with most software, which is typically licensed per seat or per server. The non-rivalry of trained models means that the "stock" of AI capital and the "flow" of AI capital services are partially decoupled. A firm may hold a fixed set of models but vary the intensity of their use by adjusting inference compute.

The second property is rapid capability improvement and obsolescence. Frontier AI models are replaced by substantially more capable successors on timescales of months. The effective capability per unit of training compute has been growing at extraordinary rates, driven by a combination of hardware improvements (faster processors, more memory bandwidth), algorithmic innovations (better architectures, training procedures, and data strategies), and scale (larger training runs). Hernandez and Brown (2020) estimate that the compute required to reach a given level of performance on standard benchmarks has been halving roughly every sixteen months due to algorithmic progress alone. Combined with hardware improvements tracked by Moore's Law and its successors, the effective quality of AI compute has been improving far faster than that of IT hardware in previous decades.

The third property is entanglement with complementary inputs. The productivity of AI systems depends heavily on the availability of complementary human capital (workers who can design prompts, interpret outputs, and integrate AI into workflows), organisational restructuring (changes in business processes, management practices, and task allocation), and data assets (training data, fine-tuning data, and feedback loops that improve model performance over time). Brynjolfsson et al. (2021) argue that these complementary investments create a "productivity J-curve" in which the full benefits of AI are delayed because the complementary restructuring takes time and is not well measured in standard statistics.

2.3 Why Compute Is the Right Measurement Basis

Despite these complications, there is a natural measurement basis for AI capital services, namely computational resources. Every AI system, regardless of its architecture, application domain, or deployment mode, requires compute to function. Training a model requires a large, one-time expenditure of compute. Deploying a model for inference requires a smaller but ongoing flow of compute for each query or task. Fine-tuning and adaptation require intermediate amounts of compute.

Compute has several attractive properties as a measurement basis. It is physically well-

defined, measurable in standardised units (floating-point operations, GPU-hours, or similar metrics), and traded in increasingly liquid markets (cloud computing services). Expenditure data on compute are available from cloud providers' pricing and usage records, from hardware shipment data, and increasingly from firm-level IT expenditure surveys.

The analogy to earlier measurement approaches is instructive. In national accounts, electricity is measured as a flow of energy services (kilowatt-hours) consumed by each sector, priced at market rates. Capital services are measured as a flow of machine-hours or equipment-hours, priced at the Jorgensonian user cost. The proposal advanced in this paper is to measure AI capital services as a flow of compute-hours, priced at market rates and adjusted for quality using a hedonic index based on capability benchmarks. The quality adjustment is essential because a GPU-hour in 2026 produces far more AI capability than a GPU-hour in 2020, just as a transistor in 2000 was far more powerful than a transistor in 1970.

3 A Production Framework with AI Capital Services

This section presents the formal production framework. I begin with the basic setup, define AI capital services in terms of compute and quality, and then state the assumptions under which the main results are derived.

3.1 Setup

Consider an economy in continuous time t. There are J sectors indexed by $j = 1, \ldots, J$. In each sector, a representative firm produces output $Y_j(t)$ using three aggregate inputs: traditional capital $K_j(t)$, labour $L_j(t)$, and AI capital services $S_j(t)$. The production function is

$$Y_j(t) = A_j(t)\, F_j(K_j(t),\, L_j(t),\, S_j(t))\,, \tag{1}$$

where $A_j(t)$ is Hicks-neutral total factor productivity and F_j is the sector-specific production function.

Assumption 1 (Production Technology)**.** *For each sector j, the production function F_j is (i) homogeneous of degree one in (K_j, L_j, S_j), (ii) twice continuously differentiable with $F_{j,x} > 0$ and $F_{j,xx} < 0$ for each input $x \in \{K_j, L_j, S_j\}$, and (iii) satisfies the Inada conditions.*

Constant returns to scale (CRS) is a standard assumption in growth accounting that ensures the exhaustion of product across factor payments. The concavity and Inada conditions

are technical regularity assumptions that guarantee interior solutions.

Traditional capital $K_j(t)$ encompasses all non-AI physical and intangible capital, including structures, equipment, vehicles, and conventional (non-AI) software. It evolves according to the standard accumulation equation

$$\dot{K}_j(t) = I_j^K(t) - \delta_K K_j(t), \tag{2}$$

where I_j^K is gross investment and δ_K is the depreciation rate.

Labour $L_j(t)$ is measured in quality-adjusted hours, following the convention in the growth accounting literature (Jorgenson, 2001).

3.2 AI Capital Services

I now define AI capital services in terms of observable quantities. The key definitional step is to write AI capital services as the product of a quality index and a compute quantity.

Definition 1 (AI Capital Services)**.** *AI capital services in sector j at time t are defined as*

$$S_j(t) = Q(t) \cdot C_j(t), \tag{3}$$

where $C_j(t)$ is the quantity of compute used by sector j at time t, measured in standardised physical units (e.g., petaFLOP-days or GPU-hours of a reference processor), and $Q(t)$ is an economy-wide quality index that captures the effective capability per unit of compute at time t.

The quality index $Q(t)$ is common across sectors, reflecting the fact that improvements in AI models and algorithms are largely non-rival and diffuse across the economy through shared model releases and cloud platforms. Sector-specific differences in how effectively AI is used are captured by the TFP term $A_j(t)$ and by the functional form of F_j, not by the quality index.

The decomposition in equation (3) parallels the decomposition of IT capital services in Jorgenson (2001), where IT capital services equal a quality-adjusted capital stock times a utilisation rate. The key difference is that for AI, I propose to measure the flow of compute directly (as the "utilisation" variable) rather than constructing a capital stock through the perpetual inventory method. This is appropriate because the extremely rapid depreciation of AI models (often within months) makes stock-based measurement unreliable. The flow-based approach avoids the need to estimate depreciation rates for trained models, which are both difficult to observe and likely heterogeneous across model types.

The quality index $Q(t)$ aggregates two sources of improvement over time. Specifically,

$$Q(t) = Q_H(t) \cdot Q_A(t), \tag{4}$$

where $Q_H(t)$ captures hardware efficiency gains (more floating-point operations per unit of compute expenditure, driven by advances in processor design and manufacturing) and $Q_A(t)$ captures algorithmic efficiency gains (more capability per floating-point operation, driven by improvements in model architectures, training methods, and data strategies). Section 5 develops a hedonic methodology for constructing $Q(t)$ from observable data.

3.3 Market Structure

Assumption 2 (Competitive Factor Markets)**.** *Goods and factor markets are perfectly competitive. Each firm takes the output price $P_j(t)$, the rental rate of traditional capital $r(t)$, the wage $w(t)$, and the price of compute $p^C(t)$ as given.*

Assumption 3 (Cost minimisation)**.** *Each firm chooses inputs to minimize the total cost of producing a given level of output. That is, firm j solves*

$$\min_{K_j, L_j, C_j} \quad r(t)K_j + w(t)L_j + p^C(t)C_j \quad \textit{subject to} \quad Y_j = A_j(t)\, F_j(K_j, L_j, Q(t)C_j)\,. \tag{5}$$

The price of compute $p^C(t)$ is the per-unit cost of raw (non-quality-adjusted) compute, observable from cloud computing prices or imputed from hardware costs and utilisation rates. Note that the firm purchases compute at price p^C per physical unit, but receives AI capital services equal to $Q \cdot C$ per physical unit of compute, because the quality index converts physical compute into effective AI services.

4 Compute as a Sufficient Statistic

This section establishes the main result, that under the assumptions stated above, the expenditure share of compute in total revenue is a sufficient statistic for the output elasticity of AI capital services, and that compute expenditure data are sufficient to measure AI's contribution to output growth.

4.1 The Revenue Share Equivalence

Proposition 1 (Revenue Share Equivalence)**.** *Under Assumptions 1–3, the share of compute expenditure in total revenue of sector j equals the output elasticity of AI capital services.*

Formally,

$$\alpha_j^C(t) \;\equiv\; \frac{p^C(t)\,C_j(t)}{P_j(t)\,Y_j(t)} \;=\; \frac{\partial \ln Y_j(t)}{\partial \ln S_j(t)}, \tag{6}$$

where $P_j(t)$ is the output price in sector j.

Proof. The Lagrangian for the cost minimisation problem in equation (5) is

$$\mathcal{L} = rK_j + wL_j + p^C C_j + \lambda\left[Y_j - A_j F_j(K_j, L_j, QC_j)\right].$$

The first-order condition with respect to C_j gives

$$p^C = \lambda\, A_j\, F_{j,S}\, Q, \tag{7}$$

where $F_{j,S} \equiv \partial F_j / \partial S_j$ is the marginal product of AI capital services. Similarly, the first-order conditions with respect to K_j and L_j give

$$r = \lambda\, A_j\, F_{j,K} \quad \text{and} \quad w = \lambda\, A_j\, F_{j,L}.$$

Under perfect competition in the output market, $\lambda = P_j$ (the marginal cost equals the output price). From equation (7),

$$p^C C_j = P_j A_j F_{j,S}\, Q\, C_j = P_j A_j F_{j,S}\, S_j. \tag{8}$$

Dividing both sides by $P_j Y_j = P_j A_j F_j$,

$$\frac{p^C C_j}{P_j Y_j} = \frac{F_{j,S}\, S_j}{F_j} = \frac{\partial \ln F_j}{\partial \ln S_j} = \frac{\partial \ln Y_j}{\partial \ln S_j},$$

where the second equality uses the definition of the elasticity and the third equality follows because $Y_j = A_j F_j$ with A_j not depending on S_j. □

Proposition 1 is the fundamental measurement result of the paper. It states that the statistician does not need to observe AI capital services S_j directly. All that is needed to recover the output elasticity of AI services (and hence AI's share of output) is (i) compute expenditure $p^C C_j$, and (ii) total revenue $P_j Y_j$. Both quantities are, in principle, observable from existing economic data.

Remark 1. *The result has the same structure as the standard result in growth accounting that equates the capital income share with the output elasticity of capital. What is new here is the identification of compute as the relevant "physical" input whose expenditure share*

recovers the AI elasticity. The quality index Q does not appear in the expenditure share formula, because improvements in quality are already reflected in the higher effective services $S_j = QC_j$ that the firm receives per dollar of compute, and hence in the higher willingness to pay for compute.

Remark 2. *The result extends naturally to imperfect competition with constant markups. If firms charge a constant markup $\mu_j > 1$ over marginal cost, the revenue share of compute expenditure equals the output elasticity of AI services divided by μ_j. Since markups can be estimated from firm-level data (De Loecker and Warzynski, 2012), the framework remains implementable under imperfect competition.*

4.2 Factor Share Exhaustion

Under CRS, Euler's theorem implies that output is exhausted by factor payments,

$$P_j Y_j = rK_j + wL_j + p^C C_j. \tag{9}$$

The factor shares therefore sum to unity,

$$\alpha_j^K + \alpha_j^L + \alpha_j^C = 1, \tag{10}$$

where $\alpha_j^K = rK_j/(P_j Y_j)$ and $\alpha_j^L = wL_j/(P_j Y_j)$.

This exhaustion property is important for practical measurement, because it means that the compute share can also be recovered residually as $\alpha_j^C = 1 - \alpha_j^K - \alpha_j^L$, provided that capital and labour shares are correctly measured. In practice, measuring the compute share directly (from expenditure data) provides a useful cross-check.

4.3 Separating Training and Inference Compute

In practice, firms allocate compute to two distinct activities, namely training (producing new models or fine-tuning existing ones) and inference (running models to generate outputs). It is natural to ask whether the framework requires separate measurement of each.

Let C_j^T denote training compute and C_j^I denote inference compute, with $C_j = C_j^T + C_j^I$. Training compute is an investment that produces a durable asset (a trained model), while inference compute generates an immediate flow of services. Under the flow-based measurement approach proposed here, both types of compute are included in the total compute flow C_j.

This inclusive approach is justified by two considerations. First, from a growth accounting perspective, what matters is the total resource cost of AI services, which includes both

training and inference. Second, the extremely rapid depreciation of trained models means that the distinction between "investment" and "operating expenditure" is less meaningful for AI than for traditional capital, since a model trained today may be obsolete within months, so the training cost is effectively a short-lived expenditure rather than a long-lived investment.

Formally, if we write the production function as

$$Y_j = A_j F_j\big(K_j, L_j, G_j(M_j, Q_I C_j^I)\big)\,, \tag{11}$$

where M_j is the stock of trained models and G_j aggregates model quality and inference compute into AI services, and if models depreciate at rate δ_M, then total compute expenditure (training and inference combined) captures the full flow cost of AI. In the rapid-depreciation limit ($\delta_M \to \infty$), there is no meaningful model stock, and total compute expenditure reduces exactly to the flow measure in the baseline framework.

5 Hedonic Quality Adjustment for Compute

The framework in Section 4 shows that the compute expenditure share α_j^C can be recovered from data on compute spending and revenue. However, to decompose AI's contribution to growth into quantity and quality components, we need the quality index $Q(t)$. This section develops a hedonic methodology for constructing this index.

5.1 Characteristics and the Hedonic Approach

The hedonic approach to quality measurement, introduced by Griliches (1961) and formalized by Rosen (1974), treats a differentiated good as a bundle of measurable characteristics. The price of the good is a function of these characteristics, and the quality index tracks changes in the characteristic bundle over time, holding prices constant.

For AI compute, the relevant characteristics describe the capability of the AI models that a unit of compute can run or produce. I define these characteristics in terms of performance on standardised evaluation benchmarks.

Definition 2 (Capability Characteristics)**.** *Let $\mathbf{z}(t) = (z_1(t), \ldots, z_M(t))$ be a vector of M capability characteristics at time t, where each $z_m(t)$ measures the performance of the frontier AI model on a standardised evaluation benchmark m. Benchmarks may include tasks such as mathematical reasoning, natural language understanding, code generation, scientific question-answering, and multimodal perception.*

Definition 3 (Hedonic Price Function)**.** *The hedonic price function for compute relates the effective price of a unit of compute to its capability characteristics,*

$$\ln p^S(t) = \beta_0 + \sum_{m=1}^{M} \beta_m z_m(t) + \gamma t + \epsilon(t), \tag{12}$$

where $p^S(t)$ is the price per unit of effective AI services, β_m is the implicit price of characteristic m, γ captures the pure time trend in prices holding quality constant, and $\epsilon(t)$ is a residual.

The quality index is then constructed by isolating the contribution of characteristic improvements to the effective price of compute. Following standard practice in the hedonic literature (Triplett, 2006), the quality-adjusted price index satisfies

$$\ln p^{QA}(t) - \ln p^{QA}(t-1) = \hat{\gamma}, \tag{13}$$

and the quality index satisfies

$$\ln Q(t) - \ln Q(t-1) = \sum_{m=1}^{M} \hat{\beta}_m \left[z_m(t) - z_m(t-1) \right]. \tag{14}$$

Proposition 2 (Hedonic Quality Index)**.** *If the hedonic price function in equation (12) is correctly specified, then the quality index $Q(t)$ defined by equation (14) satisfies $S_j(t) = Q(t)C_j(t)$, and the quality-adjusted compute price $p^{QA}(t) = p^C(t)/Q(t)$ isolates the pure price change holding AI capability constant.*

The proof follows directly from the definitions. The hedonic decomposition separates the observed price into a quality component (captured by characteristics) and a pure price component (the time trend). The quality index $Q(t)$ converts physical compute into effective AI services, just as a hedonic computer price index converts nominal hardware expenditure into quality-adjusted real investment.

5.2 Practical Construction of the Quality Index

Implementing the hedonic quality index requires data on model capability benchmarks over time. Several features of the current AI landscape make this feasible.

First, there is a well-established and expanding ecosystem of AI evaluation benchmarks. Benchmarks such as MMLU (measuring broad knowledge), GSM8K (mathematical reasoning), HumanEval (code generation), and many others provide standardised, reproducible measures of model capability. New benchmarks are regularly introduced as AI capabilities

advance, and cross-temporal comparisons are possible for benchmarks that remain in use over multiple model generations.

Second, the prices and specifications of AI compute are increasingly transparent. Cloud computing providers publish pricing for GPU instances. Hardware manufacturers report the computational throughput of each processor generation. Third-party organisations track and compare the price-performance ratios of different compute offerings.

Third, the open-source AI ecosystem provides a rich set of models spanning a wide range of capability levels and training compute budgets, enabling the estimation of hedonic regressions with substantial cross-sectional variation.

The practical challenge is that benchmarks may saturate (ceiling effects), that new capabilities emerge for which no benchmark exists, and that the implicit prices β_m may shift over time as the composition of AI applications changes. These are familiar issues in hedonic measurement (Triplett, 2006) and can be addressed through standard techniques such as adjacent-year hedonic regressions, chained indices, and careful benchmark selection.

5.3 Connection to Scaling Laws

The hedonic quality index gains additional structure from the empirical scaling-laws literature. Foundational papers show that model capability on a wide range of benchmarks follows a predictable relationship with training compute, model size, and data (Kaplan et al., 2020; Hoffmann et al., 2022). This micro-level relationship is consistent with newer macro evidence linking AI adoption to firm productivity, employment, inflation, and short-run growth differentials across countries and sectors (Misch et al., 2025; Aldasoro et al., 2026; Borowski et al., 2025; Gambacorta et al., 2025). Specifically, for a given benchmark m,

$$z_m(t) = a_m + b_m \ln C^{train}(t) + d_m(t), \tag{15}$$

where $C^{train}(t)$ is the training compute used for the frontier model at time t, $b_m > 0$ is the scaling exponent for benchmark m, and $d_m(t)$ captures algorithmic progress (improvements in capability holding compute constant).

Proposition 3 (Scaling-Law-Based Quality Growth)**.** *If model capabilities follow the log-linear scaling law in equation (15), the growth rate of the compute quality index is*

$$g_Q = \sum_{m=1}^{M} \hat{\beta}_m \left(b_m \, g_{C^{train}} + g_{d_m} \right), \tag{16}$$

where $g_{C^{train}}$ is the growth rate of training compute per frontier model and g_{d_m} is the rate of

algorithmic improvement for benchmark m.

Proof. Differentiating equation (15) with respect to time gives $\dot{z}_m = b_m \, g_{C^{train}} + g_{d_m}$. Substituting into equation (14) yields the result. □

Proposition 3 provides a structural decomposition of quality growth into a component driven by scale (larger training runs) and a component driven by algorithmic progress. The scaling exponents b_m are estimable from cross-model data, and the rate of algorithmic improvement g_{d_m} can be estimated by comparing the compute required to achieve a given performance level across model generations (Hernandez and Brown, 2020).

To give a sense of magnitudes, if the effective algorithmic efficiency doubles every 16 months (Hernandez and Brown, 2020) and frontier training compute doubles every 6 to 10 months (Sevilla et al., 2022), a rough calibration yields quality growth rates on the order of 50 to 100 percent per year. These rates are an order of magnitude larger than those for IT hardware, where the hedonic quality adjustment for semiconductors has historically been in the range of 30 to 50 percent per year.

6 Growth Accounting with AI Capital

This section derives the growth accounting decomposition that incorporates AI capital services, and characterises the measurement bias that arises when AI capital is omitted from the standard framework.

6.1 The Extended Solow Decomposition

Differentiating the log of the production function in equation (1) with respect to time and using the definitions of factor shares gives the following decomposition of output growth.

Proposition 4 (Growth Decomposition with AI Capital)**.** *Under Assumptions 1–3, the rate of output growth in sector j decomposes as*

$$g_{Y_j} = g_{A_j} + \alpha_j^K \, g_{K_j} + \alpha_j^L \, g_{L_j} + \alpha_j^C \left(g_Q + g_{C_j} \right), \tag{17}$$

where $g_X \equiv \dot{X}/X$ denotes the growth rate of variable X, and α_j^K, α_j^L, and α_j^C are the revenue shares of traditional capital, labour, and compute, respectively.

Proof. Taking the total time derivative of $\ln Y_j = \ln A_j + \ln F_j(K_j, L_j, S_j)$ and using the fact

that F_j is homogeneous of degree one, we have

$$g_{Y_j} = g_{A_j} + \frac{F_{j,K}K_j}{F_j} g_{K_j} + \frac{F_{j,L}L_j}{F_j} g_{L_j} + \frac{F_{j,S}S_j}{F_j} g_{S_j}.$$

By the first-order conditions and the definition of revenue shares, $F_{j,x}x_j/F_j = \alpha_j^x$ for each input. Since $S_j = QC_j$, we have $g_{S_j} = g_Q + g_{C_j}$. The result follows from Proposition 1, which establishes that $\alpha_j^S = \alpha_j^C$. □

The decomposition in equation (17) extends the standard Solow decomposition by adding a term for the contribution of AI capital services. This AI contribution, $\alpha_j^C(g_Q + g_{C_j})$, can be further decomposed into two subcomponents.

The first subcomponent, $\alpha_j^C g_{C_j}$, represents *compute deepening*, that is, growth in the physical quantity of compute deployed, holding quality constant. This captures the expansion of AI infrastructure through purchases of additional GPUs, expansion of cloud capacity, and similar investments.

The second subcomponent, $\alpha_j^C g_Q$, represents *compute quality improvement*, that is, growth in AI capability per unit of compute. As shown in Section 5, this component can be further decomposed into hardware efficiency gains and algorithmic efficiency gains using the scaling-law framework.

The three-way decomposition of AI's contribution (compute deepening from hardware expansion, quality improvement from better models, and TFP growth from organisational adaptation) provides a much richer picture than treating AI as a single undifferentiated input. It allows policymakers and researchers to distinguish between countries or sectors that are growing through compute accumulation (building more data centres) versus those growing through quality improvement (adopting better models and algorithms) versus those growing through complementary innovation (reorganizing production to exploit AI more effectively).

6.2 Aggregate Growth Accounting

For the aggregate economy, the growth decomposition becomes

$$g_Y = g_A + \bar{\alpha}^K g_K + \bar{\alpha}^L g_L + \bar{\alpha}^C \left(g_Q + g_C\right), \tag{18}$$

where the bars denote Domar-weighted aggregate factor shares and the aggregate compute quantity C is the sum of sectoral compute quantities.

Corollary 1. *True aggregate TFP growth is the residual from the three-factor decomposition,*

$$g_A = g_Y - \bar{\alpha}^K g_K - \bar{\alpha}^L g_L - \bar{\alpha}^C \left(g_Q + g_C\right). \tag{19}$$

6.3 Measurement Bias from Omitting AI Capital

The central empirical concern motivating this paper is that current growth accounting omits AI capital, attributing its contribution to the Solow residual. The following proposition characterises the resulting bias.

Proposition 5 (Omission Bias)**.** *Suppose a statistician performs growth accounting using only traditional capital and labour, computing the "measured" Solow residual as*

$$\tilde{g}_A = g_Y - \tilde{\alpha}^K g_K - \tilde{\alpha}^L g_L, \tag{20}$$

where $\tilde{\alpha}^K + \tilde{\alpha}^L = 1$*. Then the measured Solow residual exceeds the true TFP growth rate by*

$$\tilde{g}_A - g_A = \alpha^C \left(g_Q + g_C\right) + \left(\tilde{\alpha}^K - \alpha^K\right) g_K + \left(\tilde{\alpha}^L - \alpha^L\right) g_L. \tag{21}$$

The first term is the direct omission bias, equal to the AI contribution to growth. The second and third terms capture the indirect bias from misattributing compute's income share to traditional capital and labour.

Proof. From equation (18),

$$g_A = g_Y - \alpha^K g_K - \alpha^L g_L - \alpha^C (g_Q + g_C).$$

Subtracting this from equation (20) gives

$$\tilde{g}_A - g_A = (\tilde{\alpha}^K - \alpha^K) g_K + (\tilde{\alpha}^L - \alpha^L) g_L + \alpha^C (g_Q + g_C).$$

□

The bias has a clear structure. The dominant term, $\alpha^C(g_Q+g_C)$, is the direct contribution of AI to growth, which is erroneously attributed to TFP. This term is increasing in both the compute expenditure share α^C (how important AI is as an input) and the growth rate of quality-adjusted compute $g_Q + g_C$ (how fast AI is improving and expanding).

The indirect terms reflect the distortion of factor shares. When the statistician imposes $\tilde{\alpha}^K + \tilde{\alpha}^L = 1$, the compute share is implicitly reallocated to capital and labour. If compute expenditure is classified as a capital expenditure (because it involves purchasing or renting hardware), it will inflate the measured capital share. If it is classified as an intermediate input or operating expense, it may partly reduce measured value added.

Corollary 2. *If the compute expenditure share* α^C *and the growth rates of quality-adjusted compute are both positive and increasing over time, the omission bias is positive and growing.*

In particular, the bias accelerates when AI enters a period of rapid capability improvement and adoption.

This result is quantitatively relevant. Even a modest compute share of 2 to 3 percent of GDP, combined with quality-adjusted compute growth of 50 to 100 percent per year, would produce an omission bias on the order of 1 to 3 percentage points of annual TFP growth. By comparison, measured TFP growth in the United States has averaged roughly 1 percent per year over recent decades. The bias from omitting AI capital could therefore be substantial relative to measured productivity growth.

6.4 The Productivity J-Curve Through the Measurement Lens

Brynjolfsson et al. (2021) argue that the adoption of general-purpose technologies like AI follows a "productivity J-curve" in which measured productivity initially declines or stagnates as firms invest in intangible complements (reorganisation, training, new processes) that are not captured in output statistics. The present framework offers a complementary perspective on this phenomenon.

During the investment phase, firms allocate resources to building AI infrastructure and acquiring models. If these investments are classified as intermediate consumption rather than capital formation, they reduce measured value added and hence measured productivity. Meanwhile, the benefits of AI may not yet be fully realized because complementary reorganisation is incomplete. The result is a period in which true TFP growth (net of AI capital deepening) may be positive, but measured TFP understates it because the AI capital input is not subtracted from the residual.

In the mature adoption phase, the AI contribution term $\alpha^C(g_Q + g_C)$ becomes large and positive. If AI capital remains omitted from growth accounting, measured TFP surges, not because of a sudden acceleration in true technical change but because the return on accumulated AI capital is now showing up in output. The growth accounting framework developed here can in principle separate these effects, attributing the J-curve dynamics to the time profile of AI capital accumulation and complementary investment rather than to fluctuations in true TFP.

6.5 Implications for Monetary Policy

The measurement distortions characterised above matter for monetary policy because they affect the real-time interpretation of productivity, inflation, and slack rather than the underlying technology itself. Recent work increasingly emphasizes that AI can affect inflation

dynamics, the natural rate, and policy transmission through productivity, labour-market, and investment channels (Hartmann and Maver, 2025; Borowski et al., 2025; Melina and Villa, 2025). AI capital deepening is a genuine supply-side development. The problem for central banks is that, if it is omitted or badly measured, standard statistics may misattribute part of the resulting output growth to TFP, miss quality-adjusted price declines in AI-intensive sectors, and understate cross-sector heterogeneity in adoption.

First, mismeasurement can distort estimates of trend productivity and therefore of potential output growth. A central bank observing faster output growth alongside weak measured input accumulation may infer a broad-based acceleration in TFP when part of the change instead reflects unmeasured AI capital services. In real time, this complicates the separation of temporary disinflation from a persistent improvement in supply conditions and can bias estimates of the neutral policy stance.

Second, AI diffusion can distort inflation measurement through quality adjustment. If AI improves service quality or lowers effective marginal cost in ways not captured by official deflators, measured inflation in AI-intensive sectors may overstate true quality-adjusted price growth. This matters for both sectoral monitoring and the interpretation of core inflation when adoption is uneven across sectors.

Third, heterogeneous diffusion across firms and occupations can weaken historical relationships between labour-market indicators and price pressure. If AI substitutes for some tasks and complements others, unemployment, vacancies, wage growth, and the labour share may no longer map into inflation in the same way as before. For monetary policy, the practical implication is not that central banks need a separate "AI rule", but that they need better real-time measurement of compute use, AI adoption, and quality-adjusted prices to interpret standard macro signals correctly.

In this sense, the framework developed here is best viewed as a measurement layer for monetary policy analysis. It helps central banks distinguish AI-driven supply improvements, relative-price changes, and mismeasured productivity from aggregate-demand fluctuations that call for a policy response.

7 Extensions

7.1 International Comparisons and Compute Sovereignty

The framework extends naturally to international comparisons of AI capital and its contribution to growth. Cross-country differences in AI-driven productivity growth can be decomposed into differences in (i) compute intensity (compute per worker), (ii) compute quality

(which may differ if countries have differential access to frontier models or hardware), and (iii) TFP in the use of AI.

This decomposition is particularly relevant for understanding the economic implications of compute governance regimes. Export controls on AI-capable hardware, such as those imposed by the United States on advanced semiconductor sales to certain countries, directly affect the compute quantity C_j available to firms in affected economies. If these controls also limit access to frontier models (by constraining the compute available for training), they additionally affect the effective quality index $Q(t)$ in those economies.

The framework allows quantification of these effects. If country A faces a binding constraint on compute quantity $C_A < C_A^*$ (where C_A^* is the unconstrained optimum), the output loss can be approximated as

$$\Delta \ln Y_A \approx \alpha_A^C \cdot \ln\left(\frac{C_A^*}{C_A}\right). \tag{22}$$

Similarly, if country B has access only to models with quality index $Q_B < Q$ (because export controls prevent access to frontier models), the output loss is

$$\Delta \ln Y_B \approx \alpha_B^C \cdot \ln\left(\frac{Q}{Q_B}\right). \tag{23}$$

These expressions provide a measurement-based approach to quantifying the economic effects of compute governance policies, complementing the strategic and geopolitical analyses that dominate current policy discussions.

7.2 Welfare Measurement

The growth accounting framework measures AI's contribution to output, but it does not directly capture welfare effects. AI may generate consumer surplus through new goods and services that are poorly measured in GDP, a concern emphasized by Brynjolfsson et al. (2019).

The framework can be extended to welfare measurement by augmenting the production-side analysis with a demand-side model. If AI-enabled goods and services enter the consumer's utility function as new varieties (in the spirit of Feenstra (1994)), the welfare gain from AI includes both the productivity gains captured in growth accounting and the consumer surplus from new products.

Formally, if utility over differentiated goods is CES with elasticity of substitution $\sigma > 1$, and AI introduces $N(t)$ new varieties at time t, the welfare gain from variety expansion is approximately

$$\Delta W \approx \frac{1}{\sigma - 1} \ln \frac{N(t)}{N(0)}. \tag{24}$$

The compute framework provides a natural way to proxy $N(t)$, since new AI-enabled products and services require inference compute to operate. Growth in inference compute may therefore serve as an indicator of new variety creation, enabling the construction of welfare-augmented growth measures.

7.3 Distributional Considerations

The aggregate framework can be extended to analyse the distribution of AI's benefits across workers, firms, and sectors. If AI substitutes for some types of labour while complementing others (Acemoglu and Restrepo, 2019; Acemoglu, 2024), the distributional effects depend on the cross-partial derivatives F_{LS} and on the heterogeneity of compute adoption across firms.

The compute-based measurement approach is particularly useful for distributional analysis because compute usage data can potentially be disaggregated by firm, sector, region, and even task type. Cloud providers record which customers use how much compute and for what workloads. This granularity, if made available to researchers and statistical agencies, would enable distributional analyses that are impossible with aggregate capital or investment data.

8 Conclusion

This paper has proposed a theoretical framework for measuring the contribution of artificial intelligence to economic output, using compute as the measurement basis. The framework rests on a simple but powerful observation: under standard assumptions on market structure and technology, the expenditure share of compute in total revenue is a sufficient statistic for the output elasticity of AI capital services. This means that existing data on compute prices, quantities, and firm revenues can, in principle, support a complete accounting of AI's contribution to growth.

The framework yields several results with practical implications for economic measurement and for the conduct of monetary policy. First, the growth accounting decomposition separates AI's contribution into compute deepening and compute quality improvement, providing a richer picture than treating AI as a monolithic input. Second, the hedonic quality index, constructed from model capability benchmarks and grounded in empirical scaling laws, provides a concrete methodology for quality-adjusting compute that parallels established methods for IT hardware. Third, the characterization of the omission bias demon-

strates that failing to account for AI capital in growth accounting will produce increasingly large overestimates of TFP growth as AI becomes more important in production.

For central banks, the implications are practical rather than doctrinal. Better measurement of AI capital services would improve real-time estimates of trend productivity and potential output growth, sharpen interpretation of quality-adjusted inflation in AI-intensive sectors, and help policymakers distinguish AI-driven relative-price changes from broader demand conditions.

Several important tasks remain for future work. On the theoretical side, the framework should be extended to handle imperfectly competitive AI markets (where a small number of cloud providers and model developers exercise market power), to incorporate network effects and data externalities, and to analyse the dynamics of AI capital accumulation when training and inference compute have different investment characteristics. On the empirical side, implementing the framework requires assembling data on compute expenditures by sector and firm, estimating hedonic regressions from cross-model benchmark data, and calibrating the scaling-law parameters that determine quality growth.

The most immediate practical recommendation is that national statistical agencies should begin systematically collecting data on compute usage, disaggregated by sector, firm size, and compute type (training versus inference). Cloud computing providers, who already record this information for billing purposes, could be required to report aggregate usage statistics, analogous to the energy consumption data that utilities report for the construction of energy accounts. Without such data, the measurement gap identified in this paper will persist, and central banks will lack the informational foundation needed to interpret macroeconomic dynamics in an economy increasingly shaped by AI.

References

Acemoglu, D. and Restrepo, P. (2019). Automation and New Tasks: How Technology Displaces and Reinstates Labour. *Journal of Economic Perspectives*, 33(2), 3–30.

Acemoglu, D. (2024). The Simple Macroeconomics of AI. *NBER Working Paper* No. 32487.

Aghion, P., Jones, B.F. and Jones, C.I. (2018). Artificial Intelligence and Economic Growth. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Aldasoro, I., Gambacorta, L., Pal, R., Revoltella, D., Weiss, C. and Wolski, M. (2026). AI Adoption, Productivity and Employment: Evidence from European Firms. *BIS Working Papers* No. 1325.

Borowski, J., Fidrmuc, J. and Jaworski, K. (2025). Artificial Intelligence and Inflation in the EU. *Economics Letters*, 257, 112638.

Brynjolfsson, E., Collis, A., Diewert, W.E., Eggers, F. and Fox, K.J. (2019). GDP-B: Accounting for the Value of New and Free Goods in the Digital Economy. NBER Working Paper No. 25695.

Brynjolfsson, E., Rock, D. and Syverson, C. (2021). The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. *American Economic Journal: Macroeconomics*, 13(1), 333–372.

Corrado, C., Hulten, C.R. and Sichel, D.E. (2009). Intangible Capital and U.S. Economic Growth. *Review of Income and Wealth*, 55(3), 661–685.

De Loecker, J. and Warzynski, F. (2012). Markups and Firm-Level Export Status. *American Economic Review*, 102(6), 2437–2471.

Feenstra, R.C. (1994). New Product Varieties and the Measurement of International Prices. *American Economic Review*, 84(1), 157–177.

Gambacorta, L., Kharroubi, E., Mehrotra, A. and Oliviero, T. (2025). Artificial Intelligence and Growth in Advanced and Emerging Economies: Short-Run Impact. *BIS Working Papers* No. 1321.

Griliches, Z. (1961). Hedonic Price Indexes for Automobiles: An Econometric Analysis of Quality Change. In *The Price Statistics of the Federal Government*. NBER.

Hartmann, P. and Maver, V. (2025). Implications of Artificial Intelligence for Monetary Policy: A First Conceptual Assessment. *SUERF Policy Brief*, No. 1080.

Hernandez, D. and Brown, T.B. (2020). Measuring the Algorithmic Efficiency of Neural Networks. arXiv preprint arXiv:2005.04305.

Hoffmann, J. et al. (2022). Training Compute-Optimal Large Language Models. arXiv preprint arXiv:2203.15556.

Jorgenson, D.W. and Griliches, Z. (1967). The Explanation of Productivity Change. *Review of Economic Studies*, 34(3), 249–283.

Jorgenson, D.W. (2001). Information Technology and the U.S. Economy. *American Economic Review*, 91(1), 1–32.

Kaplan, J. et al. (2020). Scaling Laws for Neural Language Models. arXiv preprint arXiv:2001.08361.

Melina, G. and Villa, S. (2025). From Servers to Rates: AI, ICT Capital, and the Natural Rate. *IMF Working Paper* No. 25/224.

Misch, F., Park, B., Pizzinelli, C. and Sher, G. (2025). Artificial Intelligence and Productivity in Europe. *IMF Working Paper* WP/25/67.

Nordhaus, W.D. (2007). Two Centuries of Productivity Growth in Computing. *Journal of Economic History*, 67(1), 128–159.

Noy, S. and Zhang, W. (2023). Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science*, 381(6654), 187–192.

Peng, S., Kalliamvakou, E., Cihon, P. and Demirer, M. (2023). The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. arXiv preprint arXiv:2302.06590.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1), 34–55.

Sevilla, J. et al. (2022). Compute Trends Across Three Eras of Machine Learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*.

Solow, R.M. (1957). Technical Change and the Aggregate Production Function. *Review of Economics and Statistics*, 39(3), 312–320.

Triplett, J.E. (2006). *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products*. OECD Publishing.