BANK OF CANADA
BANQUE DU CANADA

Staff working paper / Document de travail du personnel—2026-20

Last updated: June 18, 2026

Measuring the AI Economy

Anton Korinek
PIIE and University of Virginia
akorinek@virginia.edu

Patrick McKelvey
Data and Digital Services
Bank of Canada
pmckelvey@bank-banque-canada.ca

DOI: https://doi.org/10.34989/swp-2026-20 | ISSN 1701-9397

Measuring the AI Economy

Anton Korinek* Patrick McKelvey†

May 8, 2026

Abstract

We construct a macroeconomic estimate of total AI production for the United States, combining inference and R&D/training activities and applying quality adjustments based on the evolution of API prices at fixed performance levels and the pace of algorithmic progress. We estimate that nominal AI compute spending grew over 140% per year each in 2024 and 2025, raw compute capacity grew over 200% per year, and quality-adjusted AI output grew over 2000% per year. These growth rates reflect three compounding forces: expanding data-center capacity, continued improvements in chip efficiency, and rapid algorithmic progress. We then employ our estimates to develop a nascent framework for "AI GDP" that tracks the AI economy as a coherent whole rather than dispersed across standard industry classifications. Quality-adjusted AI GDP grew by more than 2500% each in 2024 and 2025. Our measures complement traditional national accounts by providing visibility into a fast-moving sector whose activity is difficult to isolate in existing statistics, and they may serve as building blocks for satellite accounts that track AI's growing role in the economy.

Acknowledgements. We thank Martin Chorzempa, Cullen Hendrix, Patrick Honohan, Adam Posen, and David Wilcox for excellent comments. Kody Karmody and Dylan Ryfe provided reliable research assistance. We thank Leopold Brown and Yuval Rhymon for their contributions to early stage research and data collection, Future Impact Group for their support, and Andrey Fradkin for generously sharing data on inference prices. Korinek also works at the Anthropic Institute. This work was conducted in his capacity as a nonresident senior fellow at PIIE and professor at the University of Virginia. The views expressed herein are solely those of the authors and do not necessarily represent the views of the Bank of Canada, of the Anthropic Institute, or of the Peterson Institute for International Economics.

*PIIE and University of Virginia: akorinek@virginia.edu
†Bank of Canada: pmckelvey@bank-banque-canada.ca

Contents

1 Introduction

Among artificial intelligence (AI) researchers and leading technology companies, there is broad agreement that AI capabilities are advancing at a remarkable pace—with some arguing that artificial general intelligence (AGI) may be achieved soon. Yet when we look at traditional economic statistics, we see only upstream investment in data centers, while downstream impacts from this revolution remain nearly invisible. GDP growth in the United States and other advanced economies has remained moderate, and productivity statistics have barely ticked up. The question "when will we see AI in the GDP statistics?" has become a recurring theme in economic commentary. One natural response is patience: AI adoption takes time, and transformative economic effects may simply lie ahead. This is almost certainly part of the story.

But we believe there is an additional, complementary issue worth taking seriously. National accounts were designed for an economy in which all production is ultimately organized around humans as the central point of value creation. This was an entirely appropriate design for most of economic history, and it continues to serve its core purpose well. However, the rapid growth of the AI sector introduces measurement challenges that existing statistical categories were not built to address. The difficulty is that AI activity is hard to see through the lens of traditional national accounts and in the ways we typically measure GDP.

The challenge operates through several channels. First, AI activity is scattered: spending on AI compute, model development, and AI-powered services is spread across dozens of industry categories—data processing, cloud computing, software publishing, professional services—making it difficult to track the AI economy as a coherent whole. Second, AI quality improvements are unusually rapid: the pace of improvement in AI capabilities is far faster than in most sectors for which statistical agencies have developed quality adjustment methods, raising questions about whether standard hedonic techniques capture what is happening. Third, AI's role in the economy is evolving: as AI systems become more capable, they may transition from being one among many intermediate inputs to playing a more central role in production, potentially straining categories that were designed for a world in which machines are passive capital rather than active contributors to economic output.

AI is the latest in a series of new rapidly-growing technologies to introduce new measurement challenges, such as semiconductors and the internet. But in contrast with previous episodes, AI won't necessarily be constrained by the supply of complementary human labor, which could open the door to much larger mismeasurement.

These measurement challenges matter today, and they will matter much more in the near future. If AI capabilities continue to advance rapidly, policymakers and researchers will need tools that can track the AI economy's growth alongside traditional statistics.

The case for building such tools now—while the AI sector is still relatively small—rests on the simple observation that statistical infrastructure takes time to develop, and waiting until measurement gaps become acute means arriving too late.

This paper makes two related but separable contributions. First, we construct production-side estimates of U.S. AI output combining inference and training activities, quality-adjusted using API prices at fixed performance levels and estimates of algorithmic progress. These estimates reveal that real AI output has grown at rates vastly exceeding nominal spending—a finding that can be incorporated into conventional GDP statistics as a hedonic correction, without requiring any reorganization of existing national accounts. Second, we propose a framework for "AI GDP" that treats the AI sector as a coherent economic entity trading with the human economy, and use it to produce an integrated picture of the AI economy's size and growth.

2 Approach and Contributions

In this paper, we construct a macroeconomic estimate of total AI production for the United States and a corresponding quality-adjusted AI production index. Our measures aggregate through several layers. First, we begin with estimates of raw compute capacity, anchored in electricity usage projections (Patel et al., 2024) and data-center characteristics (Epoch AI, 2026b), and cross-checked against chip sales data (Epoch AI, 2026a). Second, we map raw compute to nominal spending using a collected dataset of GPU-hour rental prices. Third, we construct quality adjustments: for inference, using the evolution of API prices[1] at fixed benchmark performance (Demirer et al., 2025); for training, using the pace of algorithmic progress (Ho et al., 2024) as measured by the compute required to reach a given level of model performance.

We use these estimates to develop a nascent broader framework for calculating AI GDP—the portion of economic value creation more closely associated with AI computation rather than human computation.This framework provides a potential conceptual foundation for tracking AI's contribution to the economy on its own terms rather than as an incidental byproduct of human-centered accounting categories, offering a novel and complementary perspective on the forces shaping our economy.

Relationship to Traditional National Accounting

It is useful to spell out concretely how our approach relates to standard national accounting practice and where it provides additional information. Our measurements of AI

[1] API (Application Programming Interface) prices refer to the per-unit fees charged by AI providers—such as OpenAI, Anthropic, and Google—for access to large language models (LLMs) via software interfaces. Prices are typically quoted per million tokens processed, where a token is roughly a word fragment.

production represent refinements that follow the spirit of traditional national accounting rules but adapt them to provide a clearer picture of AI:

Tracking the AI sector as a coherent whole. National accounts organize economic activity by institutional unit and industry. AI-related activity is therefore distributed across many categories, including cloud computing, software and professional services. Our approach instead tracks the AI economy as a coherent whole—all compute production, model development, and inference output—regardless of which industry classification it falls under. This is analogous to how trade economists sometimes construct accounts for the "tradable sector" or how energy economists track the energy sector across standard industry boundaries. Several of the measures we construct, particularly our estimates of nominal AI compute spending and raw compute capacity, could serve as building blocks for an AI satellite account within the existing national accounting framework, providing a structured view of AI activity without requiring changes to headline GDP methodology. This is important because AI may soon become one of the primary drivers of value creation, making it necessary to have an integrated picture of its impact.

More granular quality adjustment. Standard statistical agencies apply hedonic price adjustments conservatively. Quality-adjusted price declines of 20–30% per year in fast-moving technology sectors like semiconductors count as outliers. Our inference deflator declines approximately 94% per year—roughly a 16-fold increase in quality-adjusted output for a given dollar spent. This reflects the compounding of two forces: continued chip efficiency improvements and rapid algorithmic progress. This challenge is not without precedent: Hausman (1999) showed that the BLS telecommunications CPI was biased upward by roughly 2.3 percentage points per year simply by failing to account for the introduction of cellular telephone service, a new good that official statistics were slow to incorporate. The extent to which this aggressive adjustment is appropriate depends on the degree to which benchmark-based performance gains translate into economic value—a question we acknowledge as an important caveat. But we believe the exercise is informative precisely because it highlights how much the choice of deflator methodology matters for our understanding of AI's economic trajectory. If AI crosses a threshold such as AGI, behind which it becomes broadly useful across the economy, our estimates may shed important light on the resulting macroeconomic implications. In that case, the deflator for AI GDP may no longer decline at its current pace, and the rapid increases in AI production may translate more directly into traditional GDP growth.

Treating model development as capital formation. Our framework treats AI training as investment in "model capital"—intangible assets that improve the quality of future inference output. This extends the logic already embedded in national accounts

since the 2008 SNA revision, which capitalized R&D as intellectual property products. As AI model development grows in scale, statistical agencies may find it useful to identify it as a distinct asset category within the existing IPP framework.

Our proposal for developing a nascent framework for "AI GDP" represents a more fundamental reorientation of national accounting statistics but may become increasingly useful as the role of AI in our world grows:

A Nascent Framework for AI GDP. In standard national accounting, most of AI production counts as an intermediate good that nets out in the calculation of final GDP. This obscures the rapid changes occurring in the AI sector. We propose a new framework that separates out human GDP and AI GDP, where the latter consists of all economic activity that is driven by AI-based computation. Inputs such as electricity, compute, and maintenance that AI systems require to operate are classified as imports to the AI economy, whereas inference outputs that are sold to the human economy are counted as expenses and thus imports. Model training counts as investment in model capital. This departure is more consequential than the others: it reflects a view—developed further in our AI GDP framework in Section 5—that as AI systems become more capable, the boundary between "capital being used up" and "an agent producing output" may become increasingly blurred. We present this as an alternative lens to current practices of compiling GDP that will become more and more useful as AI capabilities grow.

Our framework of AI GDP is designed to remain informative across a range of scenarios for AI's future economic role. Standard accounts handle AI adequately for today's purposes because AI is a relatively small intermediate input. But if AI's role expands substantially—whether gradually or through a more rapid transition—then tracking the AI economy on its own terms provides an early signal that complements what traditional statistics can offer.

We emphasize that our approach is intended to complement, not replace, traditional GDP measurement. Standard accounts will largely continue to serve their core purpose of tracking economic welfare from a human perspective, although they may have to be adjusted on the margins. Our measures provide additional visibility into this fast-moving sector.

Data Limitations and Uncertainty. Our estimates represent an initial attempt at taking a macroeconomic view of the AI economy, and data limitations mean that strong assumptions were required. The allocation of compute between training and inference is largely unobserved, forcing us to lean on anecdotal evidence that there is a roughly equal split. Our visibility into AI companies' gross margins is limited. And the relationship between benchmark performance—which anchors our quality adjustments—and actual economic value remains uncertain. These gaps point toward valuable collaborations be-

tween statistical agencies, AI companies, and researchers. As the AI economy grows, so does the case for systematic measurement infrastructure.

3 Methodology

The central focus of our measurement efforts is AI production, which encompasses the creation of inference outputs (token generation for the rest of the economy) and the formation of model capital. We provide an overview of our measurement methodology and data sources here. The in-depth methodology is described in the technical appendix. Below, we will also use AI production to develop a nascent conceptual framework for measuring AI GDP.

3.1 Measuring AI Production

Figure 1 shows a conceptual outline of AI production. The central driver of AI production is the generation of AI compute, which requires data center capital (i.e., GPUs) and the electricity to run them. Compute is then allocated between inference and training. Inference compute is the compute used to produce AI outputs for use in other tasks. Inference compute is combined with AI models (intangible capital) to produce inference outputs. Training and R&D compute includes all compute used in the AI R&D process including pretraining and post-training of AI models, experiments, and synthetic data generation. Training compute is combined with R&D algorithms to produce AI R&D, which leads to the accumulation of intangible model capital. In this way, training output improves with the improvement of R&D algorithms, and inference outputs improve with the development of new model capital.

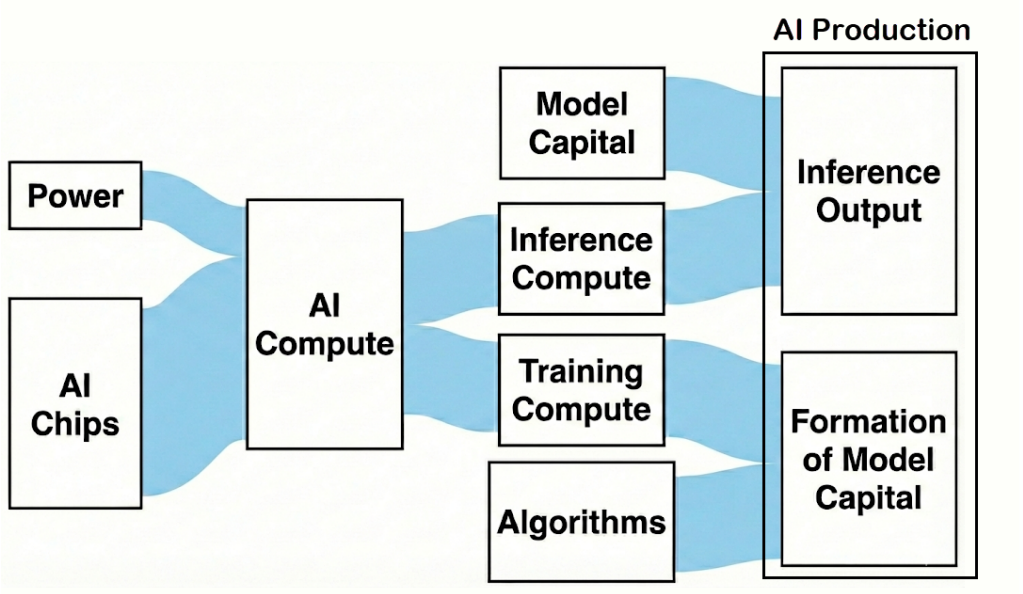

Figure 1: Conceptual outline of AI production. Compute is the central input, allocated between inference and training activities, each of which benefits from quality improvements over time.

Power-based estimates of U.S. AI compute. Our primary methodology for calculating AI compute production for the United States stems from the insight that, for a given composition of the stock of AI chips, compute output scales with power usage. Starting from electricity usage projections from Patel et al. (2024), we apply chip-level characteristics from the Epoch AI ML Hardware dataset (Epoch AI, 2024)—notably Thermal Design Power (TDP) and bit-level processing performance—together with GPU capital-stock shares from the Epoch AI GPU Clusters dataset (Epoch AI, 2026b), to convert total facility energy into the associated working time in GPU-hours for each chip type. We then apply hourly rental rates for GPUs, collected independently, to obtain estimates of total nominal spending on AI compute, on an imputed rental price basis [2]. The same characteristics let us compute total physical compute production (in FLOPs or H100-equivalents) [3]. Derivations and assumptions behind these calculations are provided in the online technical appendix.

As a cross-check, for an alternative estimate of compute production, we leverage recently released data from Epoch AI on global chip sales from leading producers (Epoch AI, 2026a). By assuming a constant usage intensity and accumulating quarterly deployments into a cumulative stock, we obtain a bottom-up estimate of AI compute spending with

[2] As simplifying assumptions, we hold rental prices constant over time for a given GPU type, and we apply the lowest-available contract rate for each given GPU type, as much large-scale compute is purchased through long-term private contracts.

[3] It should also be noted that FLOP-based compute measures likely understate improvements in chip effectiveness as they fail to capture improved memory capacity and connection bandwidth in newer chip generations.

global scope, complementing the U.S.-only power-based method. The two approaches draw on largely independent data, and their agreement—once adjusted for the difference in geographic coverage—provides an informal validation of the power-based estimates.

Training vs. inference allocation. Data on how to attribute compute between training and inference is extremely limited. We therefore impose the assumption, based on narrative accounts, that physical compute is split roughly equally between inference serving and R&D activities. More data on this split would be very valuable for researchers aiming to understand AI activity at a macro level.

Inference output and its price deflator. Raw inference spending does not capture the rapid quality gains in AI output. To build a quality-adjusted inference deflator, we use data generously shared by Demirer et al. (2025), which records the minimum prompt price (per million tokens) for the cheapest available model within each intelligence performance tier on OpenRouter, observed weekly. We compute a chained Fisher price index across consecutive monthly pairs, including only tiers already present in the prior month—so a newly introduced frontier tier enters the index the month *after* its introduction, rather than artificially inflating it. In this way, we capture price movements within performance tiers, averaged across all available intelligence levels in a given period. This yields a trend decline in per-token prices of approximately 97% per year (a 35× efficiency gain), as illustrated in Figure 2. We adjust this index by the approximately 2.2× annual trend increase in benchmark response lengths (Emberson et al., 2025), leading to a net inference deflator that declines 94% per year (roughly a 16× fall in the effective price of AI inference output). As documented by Demirer et al. (2025), this decline is broad-based across intelligence tiers and model types, consistent with broad-based algorithmic improvements over time allowing smaller models to match the performance of older larger models.

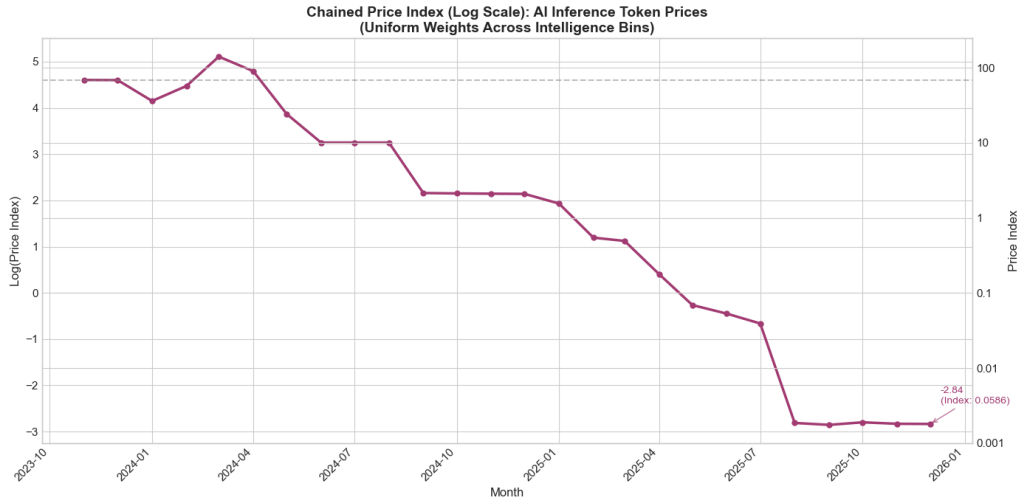

Figure 2: Chained Fisher Price Index for AI inference tokens at constant capability. The index tracks the minimum price per million tokens within each intelligence tier and chains them into a single quality-adjusted series. The rapid decline reflects falling inference costs at fixed model capability, not a shift toward lower-capability models.

Training output and its price deflator. To price-adjust training production, we apply the pace of algorithmic improvement estimated by Ho et al. (2024), who find that the compute required to train a model at fixed performance falls by roughly two thirds each year—an annualized efficiency gain of approximately 65%, or a 3× increase in effective compute per unit of physical compute.

One caveat is that we use a mixed methodology to accommodate limitations in the available data: we estimate nominal spending on the production side, and combine it with a price deflator based on the price and characteristics of inference outputs. This assumes that inference providers' gross margins did not materially change over time. Sevilla et al. (2026) find that AI companies have positive but modest gross margins on inference; since this is consistent with relatively stable margins, we believe our estimates capture the appropriate first-order dynamics determining total growth rates.

4 Results

This section presents our main empirical findings for the United States AI economy from 2023 to 2025.

Table 1 reports annual nominal compute spending and physical compute output for the US AI economy, based on the power-based methodology described in Section 3.1. [4]

[4]Global estimates based on the chip-sales methodology show more rapid growth, with nominal spending growing 291% in 2024 and 166% in 2025. For further details, see the online appendix.

Table 1: US AI Production Estimates (Annual)

Year	Spending ($B)	Spending Growth	Compute (H100e)	Compute Growth
2023	36.92	—	1.09×10^6	—
2024	90.46	145.0%	3.41×10^6	211.9%
2025	219.17	142.3%	1.07×10^7	213.9%

Nominal compute spending grew at roughly 144% per year in 2024 and 2025. Physical compute output—measured in H100-equivalent units—grew even faster at roughly 213% per year, reflecting both the deployment of more chips and the transition to higher-performance hardware. Continued improvements in AI chip efficiency from Moore's Law meant that a given dollar could buy more FLOPs. As such, physical compute production outpaced nominal spending.

4.1 Quality-Adjusted AI Production

Rapid progress on AI algorithms means that each unit of compute could be used to produce drastically more AI output. Table 2 reports quality-adjusted growth rates for inference, training, and aggregate AI output. Applying the inference and training deflators described in Section 3.1, the real volume of AI inference output grew roughly 39× per year. Quality-adjusted AI production growth combines inference and training growth weighted by their nominal shares, with growth in our final quality-adjusted AI Production Index surpassing 2000% per year.

Table 2: Quality-Adjusted AI Production Growth

Year	QA Inference Growth	QA Training Growth	QA AI Production Growth
2024	3,798%	782%	2,290%
2025	3,754%	788%	2,271%

Note: QA Inference represents the quality-adjusted amount of inference tokens (model outputs) and is a measure of production by AI models. QA Training is a measure of the amount of quality-adjusted investment in new model capital produced by compute used in the AI R&D process. QA AI Production is the aggregate of the previous two quantities.

5 From AI Production to AI GDP

In this section, we go beyond the traditional national accounting concepts and use our measure of AI production to construct a novel framework for AI GDP, which we define

as all economic value creation associated with AI computation rather than human brainpower. To illustrate, imagine that all AI activities (data centers and R&D labs) were located on a separate island, with all human labor and other inputs like electricity being imported from off the island. If this AI island was its own country, then it is straightforward to calculate the island's GDP by adding up total production of final product on the island (i.e., excluding intermediate inputs) and subtracting any imported inputs from off the island, i.e., from the human economy. This quantity, which we call "AI GDP," conceptually corresponds to the contribution of AI to the GDP of the overall economy.

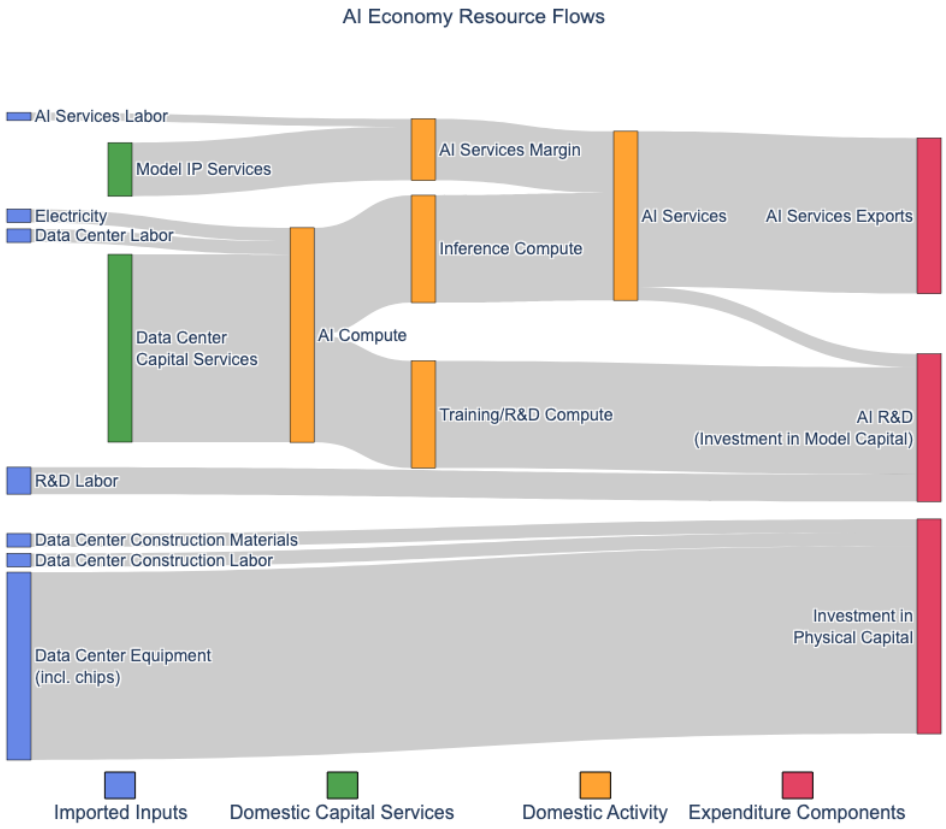

Figure 3: Economic flows in the AI economy. Imported inputs (left, blue) combine to produce final products (right, red). AI GDP equals final products minus imports.

Figure 3 outlines the key economic flows within the AI economy. AI Production — the process of combining imported electricity with AI chips to produce AI compute which is then applied to either inference or AI R&D — forms the core of the economy. Along the top of the figure, AI production is augmented by including consideration of the margin accruing to companies selling inference-based AI services. The AI services margin represents the return to intangible capital embodied in AI model intellectual property (IP) ownership. We also note "imported" labor inputs which net out from AI GDP.

In our framework, most inference-based AI services are "exported" from the AI econ-

omy, with the exception of the use of AI tools in the AI R&D process itself, which remains on the island. AI R&D and training represent investment in intellectual property (IPP) and are considered the formation of "model capital." This is conceptually similar to the treatment of R&D in current National Accounts [5]. These contributions directly increase AI GDP, though current methods are not directly designed to capture AI effects, and the intermediate nature of AI makes it less visible in expenditure accounts.

Physical capital formation consists of the installation of AI chips and data centers. As seen in the bottom section of Figure 3, these are currently created in the human economy and are "imported" into the AI economy, leading to a net-zero impact on final AI GDP. Capital formation of course increases the stock of AI chips and thus the compute available for production in future periods.

Moving from AI production to headline AI GDP requires estimates for the additional flows laid out above:

- **Electricity** consumption—already used to derive compute spending above—is converted to nominal cost using industrial electricity prices from the U.S. Energy Information Administration (Energy Information Administration, 2024).
- **Data center labor** is estimated from a constant staffing-density assumption based on physical rack space, yielding an order-of-magnitude figure that is distributed across months using the same power-weighted shares as compute spending.
- **AI services revenue and labor.** We model AI services revenue as a 1.5× markup on inference compute spending, anchored to gross-margin data from the Epoch AI Companies Dataset (Epoch AI, 2025). The markup reflects the return to intangible model-IP capital embodied in trained models, plus the value added by software scaffolding delivering services to users. Building and running those services requires engineering labor, which is imported in our framework. We provisionally set AI services labor at 10% of inference compute spending, anchored loosely to labor-cost shares at consumer web companies; this is a first-pass figure that can be refined as better data become available.

Real AI GDP. We construct real AI GDP using a chained Fisher quantity index—the same methodology employed by the U.S. Bureau of Economic Analysis—with component-specific deflators for training compute, AI services revenue, datacenter labor, electricity, and AI services labor.

[5] Our methodology is aligned with current national accounts for nominal R&D investment spending, but differs in the treatment of productivity improvements from AI R&D compute.

Nominal and Real AI GDP

Table 3 decomposes nominal AI GDP into its output and imported-input components. Note that these are rough calculations, relying on several simplifying assumptions, as outlined in Section 5. Our estimates yield overall nominal AI GDP in 2025 of $250B, which is similar in magnitude to the revenues of the U.S. scheduled passenger airline industry (Bureau of Transportation Statistics, 2025).

Table 3: Nominal AI GDP by Component ($B)

Component	**2023**	**2024**	**2025**
+ Training compute	18.46	45.23	109.58
+ AI services revenue	27.69	67.84	164.38
− Datacenter labor	0.07	0.16	0.38
− Power costs	2.31	5.89	11.90
− AI services labor	1.85	4.52	10.96
= Nominal AI GDP	**41.93**	**102.49**	**250.72**
Nominal Growth	—	144.5%	144.6%

Table 4 shows the corresponding real AI GDP index, which similarly relies on simplifying assumptions for the evolution of labor input prices.

Table 4: Real AI GDP (Chained Fisher Quantity Index, 2023 = 100)

Metric	**2023**	**2024**	**2025**
Fisher Index (2023 = 100)	100	2,700	74,445
Real AI GDP ($B, 2023 $)	41.93	1,131.83	31,213.18
Real Growth Rate	—	2,600%	2,658%

We estimate approximately 2,600% growth in real AI GDP per year in 2024 and 2025—faster even than our AI output measure. The gap between nominal growth (≈145% per year) and real growth (≈2,600% per year) reflects the rapid decline of inference token prices: users receive far more AI capability per dollar than nominal spending figures capture. This divergence echoes the experience of the semiconductor industry, where hedonic price indices revealed real output growth far exceeding nominal figures, and suggests that conventional measures may substantially understate the true pace of AI economic expansion.

This measure takes an AI-centric approach by quality-adjusting AI outputs that are considered intermediates in standard GDP accounting. This leads to a substantial degree

of uncertainty when assessing implications for overall real GDP for the US as a whole, as the production function for final goods based on inference outputs is not well understood. As an upper bound, if the quality improvements in intermediates we measure were to pass directly into final output, this would imply a boost of +2 and +4 percentage points to real GDP in 2024 and 2025, respectively.

Real AI GDP grows faster than AI production because imported inputs are not growing as fast as quality-adjusted outputs and have much more stable prices. Efficiency improvements mean electricity consumption is growing less rapidly than physical compute production. Labor inputs are also growing more slowly than physical compute production due to scale economies and due to AI becoming increasingly useful in creating future versions of AI. Since these negatively contributing imported inputs are growing more slowly than AI output, headline AI GDP is growing at a faster rate than AI production. That said, our visibility of AI companies' gross margins is quite limited, so we cannot rule out the possibility that margins are shrinking in a way that AI GDP growth lags AI production.

6 Discussion

Our estimates reveal AI production growing at extraordinary rates. This pace of growth reflects three compounding exponentials: expanding data center capacity, continued Moore's Law improvements in compute efficiency, and algorithmic progress that makes each FLOP more productive. Of particular importance are algorithmic AI model improvements which enable vastly more quality-adjusted inference outputs for the same compute input. These three forces operating simultaneously result in dramatic growth in quality-adjusted output.

The gap between benchmarks and economic value. Our quality-adjusted figures are based on benchmark performance, but we acknowledge that the link between benchmarks and economic usefulness remains uncertain. Current AI models score impressively on standardized tests, but are only starting to become broadly useful across the macroeconomy. Our numbers capture the *potential* for rapid economic impact, but production statistics alone cannot tell us whether or when AI usage will become widespread enough to realize that potential. Furthermore, inefficiencies in model and task allocation may attenuate the ultimate economic impact of AI production growth, as users use over-powered models for simple tasks, or use AI tools to complete tasks with only marginal benefits.

Consumer surplus and the relationship between economic impact and consumer utility. Our measure carries two-sided uncertainty depending on the definition used when quantifying economic impact. On the one hand, consumer surplus likely ex-

ceeds recorded revenues substantially, as with many digital services (Brynjolfsson et al., 2025; Hausman, 1999). On the other hand, the improved capabilities we capture as quality improvements could also pass primarily to consumers, and might not "count" if the final arbiter of economic impact is a traditional GDP measure. This issue is sharpened by the fact that a substantial share of consumer-facing AI is currently offered at zero overt price. Standard revenue-based measures record nothing for these services. Our production-side methodology is partly robust to this gap by construction: because we measure output from compute rather than observed revenue, free-tier inference enters our estimates regardless of how it is monetized — a practical argument in favor of a production-based AI satellite account. Similar issues exist with other free/ad-funded digital services, which has led to the creation of a Digital Economy Satellite Account (U.S. Bureau of Economic Analysis, 2023).

Why rapid AI growth is invisible in headline GDP. Part of the divergence between our AI production estimates and headline GDP growth has a straightforward accounting explanation. Quality-adjusted AI output is growing at extraordinary rates, but our current GDP measures do not accurately reflect these quality adjustments, and prices for a given quality level are falling nearly as fast, so nominal revenue grows only moderately. That pattern is familiar from decades of Moore's Law in semiconductors: enormous quality-adjusted output growth coexisted with a modest and relatively stable semiconductor share of GDP, precisely because each generation of chips was dramatically cheaper per unit of performance than the last. Indeed, even within that familiar pattern, work by Byrne et al. (2018) showed that official semiconductor deflators substantially understated the true pace of quality-adjusted price declines, suggesting that conventional statistics underrecorded semiconductor output during a period when Moore's Law was well understood. In effect, the AI sector's terms of trade are rapidly deteriorating—massive increases in the quantity and quality of its output are offset by collapsing prices. This is analogous to what Bhagwati (1958) termed "immiserizing growth," in which a country's expansion in output is offset by adverse terms-of-trade movements, leaving gains invisible in terms of market value. However with rising prices for compute, and rapid growth in enterprise usage of AI tools like Claude Code, this dynamic might be less pronounced going forward.

This is not the first time observers have worried that GDP is missing something important about a fast-moving technology sector. A substantial literature has asked whether the productivity slowdown of the past two decades reflects, in part, the failure of national accounts to capture the value created by the internet and digital services (see, e.g., Brynjolfsson et al., 2025; Aghion et al., 2023; Byrne et al., 2018). Byrne et al. (2016) and Syverson (2017) concluded that mismeasurement is real but too small—and too uncorrelated with the slowdown patterns across countries—to explain the puzzle. We

take that conclusion seriously, and we do not claim that AI mismeasurement today is large enough to move headline numbers materially. Our point is forward-looking rather than retrospective.

The reason we think the AI case may eventually prove different from the semiconductor and internet cases comes down to the economic role the mismeasured sector plays. In the prior episodes, the mismeasured technology was ultimately a *complement* to human labor and to the rest of the economy: better semiconductors made workers and devices more productive; free digital services raised consumer welfare in ways that flowed through human time and attention. Because the gains had to pass through a human bottleneck, their macroeconomic footprint was bounded by the size of the activities they enabled, which is part of why quantitative estimates of mismeasurement remained modest even when researchers looked hard for them. AI is the first plausible candidate for large-scale mismeasurement in which the rapidly improving sector is a potential *substitute* for labor itself, rather than a complement to it. If AI capabilities continue to broaden—first across cognitive work, and eventually, with more capable robotics, across physical work—the human bottleneck that disciplined the magnitude of prior mismeasurement episodes is precisely what begins to give way. That is why we believe the lessons of the internet and semiconductor debates, while genuinely cautionary, do not straightforwardly carry over.

As AI becomes broadly applicable across the economy, the relevant price comparison may shift from AI-versus-prior-generation-AI to AI-versus-prior-human-wages. At that point, the decline in the price of AI services may slow or even reverse, and the rapid growth in AI productive capacity could translate much more directly into broad GDP growth, if measured correctly—a potential regime shift that is precisely why building measurement infrastructure now matters.

The policy gap. The case for closing the measurement gap is sharpest when one considers how policymakers must plan for the medium term. In fiscal policy, income and payroll taxes are the backbone of advanced-economy revenue, and the central fiscal question raised by AI (Trammell and Korinek, 2023) is how quickly the wage tax base will erode and what, if anything, will replace it. Answering that question requires tracking the AI sector's productive capacity, not just its current revenue, and our measures show a sector whose underlying capacity is more than doubling annually. A finance ministry running ten-year revenue projections off conventional data will dramatically underweight the probability of a labor-tax-base shock, and will be correspondingly unprepared to design policy responses such as sovereign wealth funds, compute dividends, or other benefit-sharing schemes that such a shock would call for. A windfall that cannot be seen cannot be shared. The point generalizes beyond revenue forecasting: any proposal to tax AI directly, whether through a compute tax, an excise on AI services, or a windfall levy on frontier labs, presupposes knowing the size and growth trajectory of the base being

taxed. Right now, official statistics cannot tell us what a given AI tax instrument would raise.

Monetary policy faces an analogous challenge, with an additional twist. Central banks set policy on the basis of measured output gaps, productivity, and the natural rate of interest, all of which are inferred from national accounts that struggle to see the AI sector. As AI investment absorbs an increasing share of real resources—electricity, capex, and skilled labor—the natural rate of interest is likely to rise, but the signal in conventional statistics will be muted and lagged. More importantly, the policy challenge sharpens dramatically at the phase transition we have flagged above: if and when AI becomes a close enough substitute for labor that its output prices stop falling at the current pace, the rapid growth in productive capacity that is currently invisible in nominal terms may begin to register in headline statistics, potentially abruptly. A monetary authority that has been reading a low-growth, disinflationary economy may find itself behind a regime change it had no statistical apparatus to anticipate. Building the measurement infrastructure now is, among other things, an insurance policy against being caught flat-footed at exactly the moment when getting policy right matters the most.

7 Conclusion

We have constructed a first macroeconomic estimate of aggregate AI production for the United States, combining measures of raw compute capacity, nominal spending, and quality adjustment for both inference and training. The resulting picture is striking: quality-adjusted AI production grew at rates exceeding 2000% per year over the past two years—growth rates far outside the range of historical experience in economic measurement.

These growth rates are dramatic, and we have tried to be transparent about the assumptions and data limitations that underlie them. The gap between benchmark-based quality adjustment and realized economic value is real and important. But we believe the core signal in our estimates—that something is growing very fast in the AI economy, far faster than traditional statistics suggest—is robust to reasonable alternative assumptions. Even if our quality adjustments were to overstate the improvement in AI capabilities somewhat, the resulting growth rates would still be extraordinary.

Two economies, one set of statistics. Our estimates make concrete that the AI economy can look fundamentally different depending on how you measure it. From the vantage point of traditional GDP statistics, recent years show an economy growing at a moderate pace, with AI playing just a minor role as one input among many. From the vantage point of the AI economy—measured in compute, capability, and quality-adjusted output—the picture is one of rapid expansion.

Both perspectives capture real features of the same underlying economy. Traditional GDP accurately reflects the pace of improvement in human-experienced economic welfare—which, so far, AI has affected only modestly. Our new measures capture the pace of growth in AI productive capacity—which may or may not ultimately translate into proportional welfare gains for humans. The divergence between these two views is itself informative: it provides a signal for how rapidly the AI sector is expanding relative to the economy it operates within, and it highlights the potential for a future in which these two perspectives either converge (as the economic impact of AI becomes broadly felt) or remain decoupled (if AI capability growth does not translate into broadly shared economic benefits for humans).

Preparing the measurement infrastructure. We have argued that standard national accounts, while well-suited to their core purpose, provide limited visibility into the AI economy as a coherent sector. This limitation may become more consequential as AI grows, especially if AI crosses a threshold beyond which it becomes broadly useful across the economy. Several of our measures—nominal compute spending, raw compute capacity, and quality-adjusted output indices—could serve as components of an AI satellite account that provides structured visibility into AI activity without requiring changes to headline GDP methodology. More ambitiously, our AI GDP framework offers a way to track AI's net contribution to the economy that could prove useful if AI's economic role grows to the point where it strains existing categories.

Building such measurement tools takes time. The revisions to national accounts that capitalized R&D as investment took decades from initial proposal to implementation. The growth rates we estimate suggest that the window for building AI measurement infrastructure may be shorter. We view this article as an early contribution to that effort—one that will need to be refined as better data becomes available and as the AI economy's contours become clearer.

From measurement to policy. Accurate measurement of the AI economy is not merely a theoretical exercise. The question of how to allocate the gains from AI—across workers, firms, and society—will be one of the defining policy challenges of the coming years. Answering it requires knowing, at a minimum, how large the AI economy actually is, how fast it is growing, and where the value is being created and captured. Our estimates provide a first approximation to some of these questions. We hope they also make the case for the systematic collection of data—on compute allocation, margins, and the economic value of AI outputs—that will be needed to sharpen our estimates as the stakes grow higher.

References

Aghion, P., Bergeaud, A., Boppart, T., Klenow, P. J., and Li, H. (2023). A theory of falling growth and rising rents. *Review of Economic Studies*, 90(6):2714–2746. Working paper version: 2019.

Bhagwati, J. (1958). Immiserizing growth: A geometrical note. *The Review of Economic Studies*, 25(3):201–205.

Brynjolfsson, E., Collis, A., Diewert, W. E., Eggers, F., and Fox, K. J. (2025). GDP-B: Accounting for the value of new and free goods. *American Economic Journal: Macroeconomics*, 17(4):312–344.

Bureau of Transportation Statistics (2025). U.s. airlines net profit was $1.6 billion in third quarter 2025, a decrease in profit over third quarter 2024. https://www.bts.gov/newsroom/us-airlines-net-profit-was-16-billion-third-quarter-2025-decrease-profit-over-thir Accessed: 2026-04-27.

Byrne, D. M., Fernald, J. G., and Reinsdorf, M. B. (2016). Does the United States have a productivity slowdown or a measurement problem? *Brookings Papers on Economic Activity*, 2016(1):109–182.

Byrne, D. M., Oliner, S. D., and Sichel, D. E. (2018). How fast are semiconductor prices falling? *Review of Income and Wealth*, 64(3):679–702.

Demirer, M., Fradkin, A., Ifrach, B., Tadelis, N., and Peng, S. (2025). The emerging market for intelligence: Pricing, supply, and demand for LLMs. Working Paper 34608, National Bureau of Economic Research.

Emberson, L., Cottier, B., You, J., Adamczewski, T., and Denain, J.-S. (2025). LLM responses to benchmark questions are getting longer over time. https://epoch.ai/data-insights/output-length. Accessed: 2026-02-20.

Energy Information Administration (2024). Electricity data browser. https://www.eia.gov/electricity/data/browser. Industrial electricity prices.

Epoch AI (2024). Data on machine learning hardware. https://epoch.ai/data/machine-learning-hardware. Accessed: 2026-02-20.

Epoch AI (2025). Data on AI companies. https://epoch.ai/data/ai-companies. Accessed: 2026-02-20.

Epoch AI (2026a). Data on AI chip sales. https://epoch.ai/data/ai-chip-sales. Accessed: 2026-02-20.

Epoch AI (2026b). Data on gpu clusters. Accessed: 20 Mar 2026.

Hausman, J. (1999). Cellular telephone, new products, and the CPI. *Journal of Business and Economic Statistics*, 17(2):188–194.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., and Sevilla, J. (2024). Algorithmic progress in language models.

Patel, D., Nishball, D., and Eliahou Ontiveros, J. (2024). AI datacenter energy dilemma—race for AI datacenter space. *SemiAnalysis*.

Sevilla, J., Petrovic, H., and Ho, A. (2026). Can AI companies become profitable? https://epochai.substack.com/p/can-ai-companies-become-profitable. Epoch AI Gradient Updates.

Syverson, C. (2017). Challenges to mismeasurement explanations for the US productivity slowdown. *Journal of Economic Perspectives*, 31(2):165–186.

Trammell, P. and Korinek, A. (2023). Economic growth under transformative AI. Working Paper 31815, National Bureau of Economic Research.

U.S. Bureau of Economic Analysis (2023). Digital economy. https://www.bea.gov/data/special-topics/digital-economy. Last update to the Digital Economy Satellite Account: December 6, 2023. Accessed April 27, 2026.