Technical Appendix

Measuring the AI Economy

Conceptual Framework and Implementation

Patrick McKelvey* Leopold Brown Yuval Rhymon
Anton Korinek†

Technical Appendix to PIIE Working Paper 26-9 *Measuring the AI Economy* by Anton Korinek and Patrick McKelvey
May 11, 2026

Acknowledgements. We thank Kody Karmody and Dylan Ryfe for excellent research assistance, and Andrey Fradkin for generously sharing data on inference prices. We also thank Future Impact Group for enabling this collaboration. 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.

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

Contents

1 Overview

This technical appendix details the conceptual framework and implementation behind our piece on measuring the AI economy. We develop methods for tracking economic activity attributable to artificial intelligence. By defining an "AI economy" boundary encompassing all economic value creation involving AI computations rather than human neural computations, we construct an analog to Gross Domestic Product for the AI sector. Flows between human-attributable and AI-attributable activities are treated as trade between the two partitions. Our measurement approach centers on AI compute production, 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.

Our central contribution is to define and estimate an analog to Gross Domestic Product for the AI sector. Just as GDP measures the total value added within a national boundary, we define an "AI economy" boundary and measure the value added within it. The boundary encompasses all economic value creation involving foundational AI computations—that is, the portion of economic activity orchestrated by AI cognition rather than human cognition. In effect, we construct GDP from the perspective of an imaginary "AI Statistical Agency" tasked with producing national accounts for the AI economy.

Partitioning the economy into human-attributable and AI-attributable segments requires us to account for flows between the two, much as international trade accounts track flows between countries. In our framework, AI services sold to humans—such as inference API calls or chatbot subscriptions—are exports from the AI economy, adding to its GDP. Conversely, human labor, electricity, and other inputs purchased by the AI sector are imports, which subtract from gross product just as imported intermediate goods do in national accounting. These flows net out when the two partitions are recombined, ensuring consistency with aggregate GDP. Where humans and AI systems collaborate closely, the result is substantial bilateral trade in intermediate goods, analogous to the integrated cross-border supply chains observed in international trade.

This trade-based perspective clarifies several conceptual issues. Physical capital such as chips and datacenters is attributed to the AI economy because that is where its capital services are deployed, even though the capital goods themselves are currently produced entirely by human labor. The situation is analogous to homebuilding in a country that imports all of its construction materials: the new construction counts as domestic residential investment, offset by imports of materials, yielding minimal contribution to GDP but an increase in the capital stock. Similarly, the construction of AI datacenters represents investment within the AI economy, but is fuelled entirely by imports from the human economy.

In contrast, the production of intangible model capital through R&D spending does contribute directly to AI GDP. While researchers' labor is imported in this framework, training compute is domestically produced, and thus represents investment in model capital, the AI equivalent of intellectual property products in conventional national accounts.

Our measurement strategy centers on AI compute as the fundamental driver of the AI economy. Measuring AI output directly—through the revenues of all firms selling AI services—would be ideal but is not yet feasible, given limited disclosure by AI companies and the difficulty of isolating AI-attributable revenue within diversified technology firms. We therefore take an input-based approach, estimating total AI compute production and converting it to nominal spending using market prices for GPU usage. The key empirical insight is that total AI compute can be estimated from datacenter power consumption, which provides a top-down comprehensive signal of macroeconomic activity. We combine power consumption estimates with detailed data on GPU capital stocks and their characteristics to produce monthly time series of compute production, which we then decompose into inference (current output) and training (investment in model capital).

Applying this framework to the United States for the period 2023–2025, we estimate that nominal AI compute spending grew from approximately $37 billion in 2023 to $90 billion in 2024 and $219 billion in 2025, representing annual growth of roughly 144%. These nominal figures, however, substantially understate the growth in real AI output. Inference—the component of compute that produces current AI services—has seen dramatic declines in the cost per unit of capability-equivalent output, driven by improvements in model architecture, hardware efficiency, and inference optimization. After adjusting for these quality improvements using a chained Fisher quantity index, real AI GDP grew at approximately 2,600% per year—a more than 26-fold annual expansion in real output. The divergence between nominal growth ($\approx$144% per year) and real growth ($\approx$2,600% per year) 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.

2 Defining the Boundary of the AI Economy

2.1 Core Definition

In this work, we define and measure an analog to Gross Domestic Product for the AI segment of the economy. To do this, we segment overall economic activity into human- and AI-attributable GDP. We define "AI-attributable" economic activities as:

Definition 1 (AI-Attributable Economic Activity)**.** *All economic value creation involving foundational AI computations rather than human computations.*

Partitioning the economy in this way requires us to account for "trade" between the partitions. In our framework, any AI services sold to humans are treated as **exports**, adding to AI GDP. Any human-attributable inputs to production in the AI economy are treated as **imports**, subtracting from AI GDP. These imports include all human labor inputs, as well as material inputs such as electricity and GPU chips.

As in international trade, these flows net out in the calculation of the gross product of the combined entity. In many cases, there is substantial trade in intermediate goods, particularly where humans and AI systems work closely together in integrated workflows.

2.2 Capital Attribution

We attribute capital based on where its capital services are used:

- The chips and datacenters that support AI calculations are part of the AI economy
- This holds even though their production and installation is currently fully human-attributable
- In our framework, chips are produced in the human-attributable economy, then exported to the AI-attributable economy, where they are invested, resulting in a net-zero impact on AI GDP from physical capital formation.

Housing Analogy. A useful analogy for the role of physical capital (i.e., chips) in the AI economy is to imagine the treatment of housing if it were built entirely with imported materials. Imagine further that the labor input for building the housing could also be imported. Then:

- The construction of the housing would count as residential investment
- The imported materials and labor would count negatively against GDP in equal measure
- The result is no net gain in GDP, but nonetheless an increase in the housing stock

2.3 Scope and Limitations

Edge AI Exclusion. In practice, our measure focuses on compute produced in AI datacenters and excludes AI workloads run on consumer hardware. Models that are small enough to run locally are currently not a major contributor to AI output, but as smaller models become more capable, this "edge AI" could become a large enough factor to warrant incorporation into our methodology.

What Counts as "AI Computations"? In order to fully operationalize this framework, a cut-off needs to be defined on what constitutes an AI computation. Most observers would not consider a 3 layer neural network to be "AI," but such a model exists on a continuum that reaches frontier AI models, with many intermediate-size models in between. At this stage of development, our definition is functionally limited by the available data. As such, in the estimates shared below, an AI Computation is defined as a computational workload running in an AI-accelerated datacenter, as defined by Patel et al. (2024).

3 Overview of Relevant Flows in the AI Economy

This section describes the key quantities at a conceptual level, their relationships to each other, and their contributions to the final AI GDP measure.

3.1 Conceptual Flow Diagram

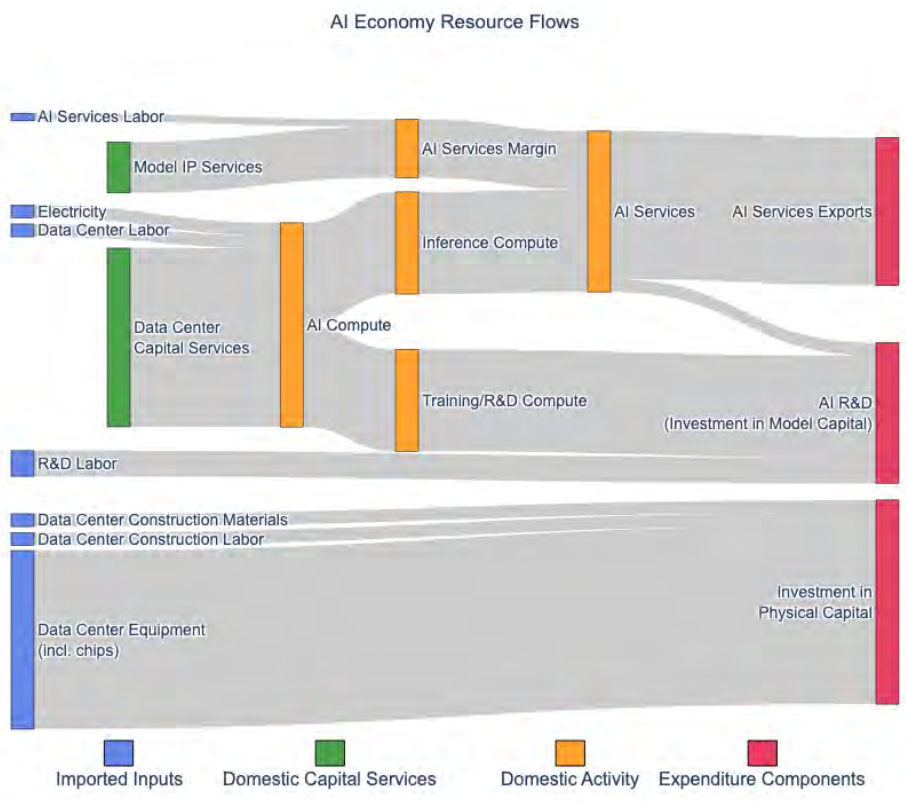

Figure 1: Economic flows in the AI economy. Imported inputs (left, blue) combine to produce final products (right, red). AI GDP equals final products minus imports.

3.2 Central Role of Total AI Compute

The central quantity in AI value production is **total AI compute**. This is produced using:

- Domestic **physical capital services** from AI datacenters
- Imported **electricity**
- Imported **labor inputs** needed to run datacenters

In our methodology, the quantity of electricity consumed is a key input to total compute. The growth rate of AI electricity consumption is a key metric for understanding the growth rate of the AI economy.

3.3 Decomposition: Inference vs. Training

Total compute is divided into two categories:

Inference Compute. This is the compute used to produce AI outputs for use in other tasks. This combines with intangible **model capital services** to produce **AI services**.

Training and R&D Compute. Compute used in training processes involving back-propagation. This includes:

- Final pretraining and post-training
- Compute used for experiments
- Compute used for generating synthetic data

3.4 AI Services Production

Inference compute combines with model capital to produce AI services. The production process also incorporates:

- **Labor inputs to AI services**: engineers building chatbot platforms, software scaffolding, etc.
- **AI services margin**: the difference between revenues (less labor costs) and inference compute costs, which accrues to services providers as gross income.

The AI services margin represents the return to intangible capital embodied in AI model IP ownership. Model capital is treated analogously to software or other Intellectual Property Products.

Exports of AI Services. Most AI services are sold back to the human economy and are thus **exports** in our framework. A small portion of AI services are used internally in the AI R&D process and would count as R&D investment (in practice we are not yet able to measure this).

3.5 Investment Activities

Investment in Intangible Capital. Training/R&D compute, combined with R&D labor inputs, results in the formation of new model capital. This is counted toward investment in intangible capital.

Investment in Physical Capital. This consists of all creation of physical capital within the AI economy, including:

- Construction of datacenter structures
- Installation of chips, servers, cooling, and interconnect
- Both materials and associated labor costs

All of these flows come from outside the boundary of the AI economy and are therefore treated as imports. As a result, this activity makes **no net contribution to headline AI GDP**. However, physical capital investment serves as a leading indicator of the magnitude of future capital services (i.e., compute generation).

3.6 Future Extensions

Model Capital. Explicit tracking of the quality-adjusted rate of investment in model research and development implies the existence of a model capital stock. This points to potential work on quantifying the model capital stock and returns to model capital investment.

Autonomous Robots. Once their deployment is more widespread, autonomous robots will be another type of physical capital within the AI economy. This opens up the possibility that physical investment could become a positive net contributor to AI GDP. More work, both conceptual and empirical, will be needed to incorporate robots into our framework.

4 Estimation Methodology: Compute

The production of AI compute is the central economic activity within our AI economy boundary. Compute is generated using domestic physical capital services (datacenter

GPUs) combined with imported inputs (electricity and labor), and represents the fundamental productive capacity from which all AI economic value flows.

This section details how we measure total compute production and partition it between inference and training activities. Quality adjustments that account for rapid improvements in what a unit of compute can deliver are described in Section 6.

4.1 Total Compute

Total compute production is the foundational quantity in our measurement framework. Because compute generation is directly tied to electricity consumption and GPU characteristics, we can measure it by working backward from observable power usage data.

4.1.1 Data Sources and Relevant Quantities

Electricity Consumption. Total annual electricity consumption by AI-accelerated datacenters is estimated from SemiAnalysis projections of AI datacenter critical IT power, utilization rates, and power usage effectiveness (Patel et al., 2024). We assume a Power Usage Effectiveness (PUE) of 1.3, in line with modern hyperscale facilities. The resulting annual totals are shown in Table 1.

Table 1: Annual AI Datacenter Power Consumption

Year	Power (TWh)
2023	28.70
2024	72.30
2025	138.00

GPU Usage Prices. Per GPU-hour prices by GPU type are collected from various public sources online. Our calculations use the lowest available market price for each GPU type, as much AI compute is provided through long-term negotiated contracts at rates below published consumer cloud prices. We hold these prices constant over time.

GPU Power Fraction. Total IT power measures the power used by a whole server, including GPUs, CPUs, memory, and other components. To convert power usage to GPU-hours requires estimates of the GPU power share in total IT power. These vary across GPU types and server setups. Collected independently from chip producer specifications on recommended hardware setups. When data weren't available, assumptions were used in line with the closest comparable GPU type with available data.

GPU-Specific Estimates. Sourced from the Epoch AI ML Hardware dataset (Epoch AI, 2024a):

- **Thermal Design Power (TDP)**: maximum power input a chip could require (watts)
- **Total processing performance** (bit-OP/s): proportional to Tensor FP-16 performance when available

GPU Capital Stock Characteristics. Sourced from the Epoch AI GPU Clusters dataset (Epoch AI, 2026b):

- Share of each GPU type in the capital stock (US only)
- Coverage share by provider type, used to map the dataset to the macroeconomy (Epoch AI, 2024b)

4.1.2 Key Assumptions

Table 2: Model Parameters and Assumptions

Parameter	Value	Description
PUE	1.3	Power Usage Effectiveness (industry standard for modern hyperscale)
Sellable Fraction	0.9	Allocation efficiency—accounts for GPU time that cannot be sold
TDP Adjustment Factor	0.9	Chips not always at full engineered capacity

Sellable Fraction. Our aim is to measure the total amount of money spent on AI compute. However, some power usage occurs when GPUs are running but not being paid for by customers. The "sellable fraction" proportionally reduces the number of billable hours.

TDP Adjustment Factor. GPU chips are not run at their maximum engineered capacity for extended durations. This factor accounts for average utilization below TDP.

Both parameters directly affect the level of output, and more work is needed to ground these numbers better. However, if the values are constant over time (as we assume), growth rates are unaffected.

4.1.3 Coverage Factors

Coverage factors estimate what fraction of total deployed chips (by type) are captured in the Epoch clusters dataset.

Table 3: Coverage Factors by Chip Type

Chip Type	**Coverage**	**Notes**
Hopper (H100/H200/GH200)	0.30	From Epoch dataset analysis
A100	0.12	From Epoch dataset analysis
V100	0.10	Assumption, slightly less than A100
MI300	0.18	From Epoch dataset analysis
Google TPU	0.04	From Epoch dataset analysis
Other	0.10	Assumption, in line with average

Source: Coverage factors derived from Epoch AI (2024b).

4.1.4 Calculation Methodology

Step 1: Temporal Disaggregation. Annual power usage is converted to a monthly time series using a constant monthly growth rate assumption. Given annual totals T_1, T_2, the monthly growth rate r satisfies:

$$\frac{T_2}{T_1} = (1 + r)^{12} \tag{1}$$

Step 2: GPU Share Calculation. Shares of each GPU type are calculated from the GPU clusters dataset, and scaled to total IT power. To adjust for sampling bias, we scale GPU shares by the coverage fraction:

$$\text{IT\_power\_adjusted} = \frac{\text{chip\_quantity} \times \text{TDP} \times \text{PUE}}{\text{gpu\_fraction} \times \text{coverage\_share}} \tag{2}$$

The coverage share changes the relative weighting of clusters to approximate their shares in the total population.

Step 3: Revenue per Watt-hour. To determine the revenue associated with each unit of energy:

$$E_{\text{IT}} = E_{\text{total}}/\text{PUE} \tag{3}$$

$$E_{\text{GPU}} = E_{\text{IT}} \times \text{gpu\_fraction} \tag{4}$$

$$\text{GPU\_hours} = E_{\text{GPU}}/(\text{TDP} \times \text{adjustment\_factor}) \tag{5}$$

$$\text{Effective\_GPU\_hours} = \text{GPU\_hours} \times \text{sellable\_fraction} \tag{6}$$

$$\text{Revenue} = \text{Effective\_GPU\_hours} \times \text{price\_per\_hour} \tag{7}$$

Therefore:

$$\text{revenue\_per\_Wh} = \frac{\text{price} \times \text{gpu\_fraction} \times \text{sellable\_fraction}}{\text{PUE} \times \text{TDP} \times \text{TDP\_adjustment}} \tag{8}$$

Step 4: Monthly Allocation. For each month, total energy is allocated across GPU types proportionally to their IT_power_adjusted values. Spending is then calculated as:

$$\text{spending}_i = \text{power\_allocated}_i \times \text{revenue\_per\_Wh}_i \tag{9}$$

Further exposition on these calculations is provided in Section 9.

4.2 Inference Share

Distinguishing between inference and training compute is essential because these activities play fundamentally different roles in the AI economy:

- **Inference compute** combines with model capital to produce AI services—the current output and "exports" of the AI sector
- **Training compute** represents investment activity that creates new model capital for future use

As the inference share evolves over time, it affects both the composition of AI GDP (exports versus investment) and our interpretation of productivity trends.

Current Data Limitations. Available data on this partition is limited. A common assumption is that inference is approximately half of AI compute. Analysts expect the share of compute going to inference to rise, so developing a way to track this is a priority.

4.3 Inference Output

Inference output—the economic output of models used to produce AI services—does not scale directly with inference compute:

- For larger models, more compute is needed per token
- For reasoning models, more inference compute is used for generating reasoning tokens
- There has been a general trend of increasing response token count per response over time, on top of reasoning tokens.

To account for the trend in token counts per response, we adjust our quality adjustment measure by the growth in output response lengths, which has been approximately 2.2× per year (Emberson et al., 2025).

5 Alternative Methodology: Chip Sales Approach

The power-based methodology described in Section 4 works top-down from aggregate electricity consumption to compute and spending estimates. As a complement, we develop a bottom-up approach that starts from chip deployment data.

This alternative methodology serves two purposes:

1. **Validation**: Provides an independent cross-check on the power-based estimates.
2. **Global scope**: While the power-based method is restricted to US datacenters (due to the geographic scope of its electricity inputs), the chip sales method produces **global** estimates of AI compute capacity and spending, since chip deployment data from Epoch AI covers worldwide shipments.

5.1 Overview

The chip sales method estimates compute spending by:

1. Starting with quarterly chip deployment estimates (from Epoch AI)
2. Accumulating deployments into a cumulative stock with no depreciation
3. Converting stock to spending using cloud rental prices
4. Converting stock to implied power consumption using chip TDP
5. Converting stock to physical compute capacity using H100-equivalent factors

The key methodological difference from the power-based method is the direction of estimation. Where the power-based method starts from an observable macroeconomic aggregate (total electricity) and distributes it across GPU types, the chip sales method builds up from microeconomic data (individual chip deployments) to produce aggregate estimates.

5.2 Data Sources

Epoch AI Chip Sales Timelines. Quarterly chip deployment estimates by chip type. Each observation includes the chip type, quarter, median unit estimate with 5th and 95th percentile bounds, and an incomplete flag for partial quarters. (Epoch AI, 2026a)

Epoch AI Chip Types. Chip specifications and performance mappings, including TDP, 16-bit FLOP/s, 8-bit OP/s, and a normalized H100-equivalent (H100e) performance factor. These H100e factors allow conversion of a heterogeneous chip stock into a single comparable unit.(Epoch AI, 2024a)

GPU Cloud Pricing. The same cloud rental pricing data used in the power-based method (Section 4.1). The conservative `price_low` values are used throughout.

5.3 Model Parameters

Table 4: Chip Sales Method: Model Parameters

Parameter	Value	Description
Operational hours/year	8,000	Approximately 91% uptime, typical for hyperscale operations
PUE	1.3	Power Usage Effectiveness (for implied power calculations)
H100 price factor	\$1.80/hr	Fallback pricing for chips without market data

5.4 Calculation Pipeline

5.4.1 Cumulative Chip Stock

Quarterly chip sales are converted to a cumulative stock per chip type:

$$S_{c,q} = \sum_{q' \leq q} \text{units}_{c,q'} \tag{10}$$

where $S_{c,q}$ is the cumulative stock of chip type c through quarter q.

No depreciation is applied—chips remain in the active stock indefinitely once deployed. This reflects the reality that AI accelerators are rarely decommissioned given ongoing demand. If a chip type has no reported sales in later quarters, its cumulative stock is carried forward.

5.4.2 Price Assignment

Each chip type is assigned an hourly rental price through one of two mechanisms:

1. **Market pricing** (preferred): The chip's name is looked up in the cloud pricing dataset to obtain `price_low`.

2. **H100e fallback**: For chips without market pricing data, the price is estimated as:

$$\hat{p}_c = h_c \times p_{\text{H100}} \tag{11}$$

   where h_c is the chip's H100-equivalent factor and $p_{\text{H100}} = \$1.80$/hr is the H100's market price. This assumes rental prices scale proportionally with compute capability.

Each chip is tagged with its price source (`market` or `estimated`) for transparency.

5.4.3 Quarterly Spending

For each chip type c in each quarter q:

$$\text{spending}_{c,q} = S_{c,q} \times h_q \times p_c \tag{12}$$

where $h_q = 2{,}000$ hours per quarter (8,000/4) and p_c is the hourly price. This represents the revenue that would be generated if the entire installed base were rented at `price_low` rates with 91% uptime.

Quarterly totals are computed by summing across all chip types:

$$\text{spending}_q = \sum_c \text{spending}_{c,q} \tag{13}$$

5.4.4 Implied Power Consumption

For each chip type in each quarter:

$$E_{c,q} = S_{c,q} \times \text{TDP}_c \times h_{\text{year}} \times \text{PUE} \tag{14}$$

where $h_{\text{year}} = 8{,}000$ hours. This gives the total energy consumption implied by the installed base running at rated TDP with PUE overhead.

Unlike the power-based method, the implied power calculation does not use `gpu_fraction` or `TDP_adjustment_factor`. It represents total facility power at the chip level, scaled only by PUE.

5.4.5 Physical Compute (H100 Equivalents)

For each chip type in each quarter:

$$C_{c,q} = S_{c,q} \times h_c \tag{15}$$

where h_c is the normalized H100-equivalent performance factor from the chip types dataset. Selected values:

Table 5: Selected H100-Equivalent Performance Factors

Chip Type	H100e Factor
H100	1.000
A100	0.315
B200	2.527
TPU v4	0.139
MI300X	1.321

Total physical compute capacity in each quarter is the sum across chip types, measured in H100-equivalent chips.

5.5 Growth Rate Estimation

5.5.1 Quarterly Growth

For each metric (spending, power, compute), quarter-over-quarter growth is:

$$g_q = \frac{X_q - X_{q-1}}{X_{q-1}} \tag{16}$$

Annualized growth extrapolates the quarterly rate:

$$g_{\text{annual}} = (1 + g_q)^4 - 1 \tag{17}$$

5.5.2 Adjusted Annual Growth

Raw quarterly growth rates have two known issues:

1. A data artifact in 2024Q1 causes an anomalous growth spike
2. The most recent quarter typically has incomplete data

We correct these by:

- Setting 2024Q1 growth to the average of 2023Q4 and 2024Q2

- Setting the final quarter's growth equal to the penultimate quarter

A cumulative index is then constructed from adjusted quarterly growth rates, and annual growth is calculated from average quarterly index values:

$$g_{\text{year}} = \frac{\bar{I}_{\text{year}}}{\bar{I}_{\text{year}-1}} - 1 \tag{18}$$

We use average-based (rather than point-to-point) annual growth because the underlying quantities represent a *flow* of compute services: the cumulative chip stock in a given quarter represents the capacity available to produce outputs during that quarter, and annual total output is proportional to the average of quarterly capacities.

5.6 Comparison with Power-Based Method

Table 6: Comparison of Estimation Approaches

Aspect	**Power-Based**	**Chip Sales**
Direction	Top-down (power → spending)	Bottom-up (chips → spending)
Geographic scope	**US only**	**Global**
Coverage adjustment	Yes (coverage factors)	No (full market estimates)
Power input	External annual estimates	Implied from chip stock

The differing geographic scope is an important distinction. The power-based method estimates are restricted to the United States because the underlying electricity consumption data covers US AI datacenters only. In contrast, the chip sales estimates cover **global** chip deployments, since Epoch AI's chip sales estimates are based on worldwide manufacturer shipments and supply chain analysis. When comparing results across the two methods, this scope difference must be accounted for—the chip sales method should produce larger estimates, all else equal, because it includes non-US deployments.

5.7 Assumptions and Limitations

No depreciation. Chips remain operational indefinitely once deployed. This likely overestimates stock in later periods as some chips are decommissioned, but given the rapid growth of AI compute, the effect is small relative to new deployments.

Uniform utilization. All chips are assumed to operate at the same uptime (91%). In reality, utilization varies by operator, workload, and chip age.

Price stability. Cloud rental rates are treated as constant over time. Actual prices decline as chips age and newer models become available.

Conservative pricing. The low end of cloud pricing (`price_low`) is used throughout. Actual revenue could be higher (especially for scarce chips) or lower (for volume contracts), adding uncertainty to our estimates.

H100e proportionality. Fallback pricing assumes rental value scales linearly with H100-equivalent compute. This may not hold for chips with very different architectures or market positions.

Data uncertainty. Epoch's chip sales estimates are modeled, not directly observed. The 5th–95th percentile ranges span roughly a factor of two, representing substantial uncertainty in the underlying data.

No deployment lag. Chips are counted as operational from the quarter they appear in the sales data. In practice, there is a lag between sale and deployment that is not modeled.

6 Quality-Adjusted Price Indices for AI Compute

Raw compute spending figures miss a crucial dimension of AI economic growth: the rapid improvement in what a unit of compute can deliver. Both inference and training compute become more productive over time, and measuring real GDP requires price indices that capture these technological improvements.

For inference, the relevant price index tracks the cost of AI output at a constant level of model capability. For training, it tracks the effective cost of building a model of a given quality. We describe each in turn.

6.1 Inference: Chain Fisher Price Index

A central input to our real GDP calculation is a price index for AI inference tokens that holds model capability constant. Because AI models are improving rapidly, a simple average of token prices would conflate two distinct phenomena: (1) prices falling for a given capability tier, and (2) the capability frontier shifting upward. We need to isolate (1).

We construct a Chain Fisher Price Index for AI inference that measures how much cheaper it becomes to purchase inference at a fixed level of model intelligence.

6.2 Data

The index is constructed from 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, observed weekly, collected from OpenRouter.

Table 7: Inference Price Data Structure

Variable	Description
`week`	Observation date (weekly frequency)
`intelligence_bin`	Model capability tier (ranges on a benchmark score; higher = more capable)
`min_prompt_price`	Lowest available price in that tier ($/million tokens)

Intelligence bins represent ranges on a standardized benchmark score. Tracking prices within bins—rather than across all models indiscriminately—holds quality constant.

6.3 Filtering

Two filters are applied before any index computation:

1. **Date filter**: Observations before November 2023 are dropped. Coverage of intelligence bins in earlier months is too sparse to support reliable index construction.

2. **Price ceiling**: Observations with prices above $7.00 per million tokens are dropped. Early observations for high-capability tiers contain extreme outlier prices that would dominate the index without providing useful signal about the evolution of accessible inference costs.

6.4 Index Construction

Weekly prices are averaged within each calendar month for each intelligence bin, with missing months forward-filled from the most recent prior observation. We then compute a chained Fisher price index over consecutive month pairs.

Bin selection. Only bins present in the previous month $t-1$ are included. A newly introduced intelligence tier first enters the index in the month *after* its introduction, preventing new high-priced tiers from artificially inflating the index.

Quantity weights. We currently assume uniform weights: $q_{i,t} = 1$ for all bins and months. Under uniform weights, $P_L = P_P = P_F$, and the Fisher index reduces to the ratio of summed prices across bins. These should ideally be weighted by relative query volume within each capability bin once such data are available. That said, as seen in Demirer et al. (2025), prices are declining rapidly across all capabilities, so we expect the index to be robust to reasonable weighting schemes.

6.5 Results and Deflator Used in GDP Calculation

The chain index implies approximately a 35× annual decline in the price of inference tokens at constant intelligence levels. This price deflator is then adjusted for a partially offsetting trend: response lengths have been increasing at approximately 2.2× per year (Emberson et al., 2025), reflecting that models are generating longer outputs per query.

The net year-over-year inference price ratio used as the deflator in the Fisher quantity index for AI GDP is:

$$\text{Inference Deflator Ratio} = \frac{2.2}{35} \approx 0.063 \tag{19}$$

That is, after accounting for increasing output lengths, the effective price of a unit of AI inference output falls to roughly 6% of its prior-year value—an annual real price decline of approximately 94%. This deflator is applied to AI services revenue in the GDP quantity index (Section 7.5).

Figure 2 shows the evolution of the chained inference price index over our sample period, illustrating the rapid and sustained decline in the cost of inference tokens at constant capability levels.

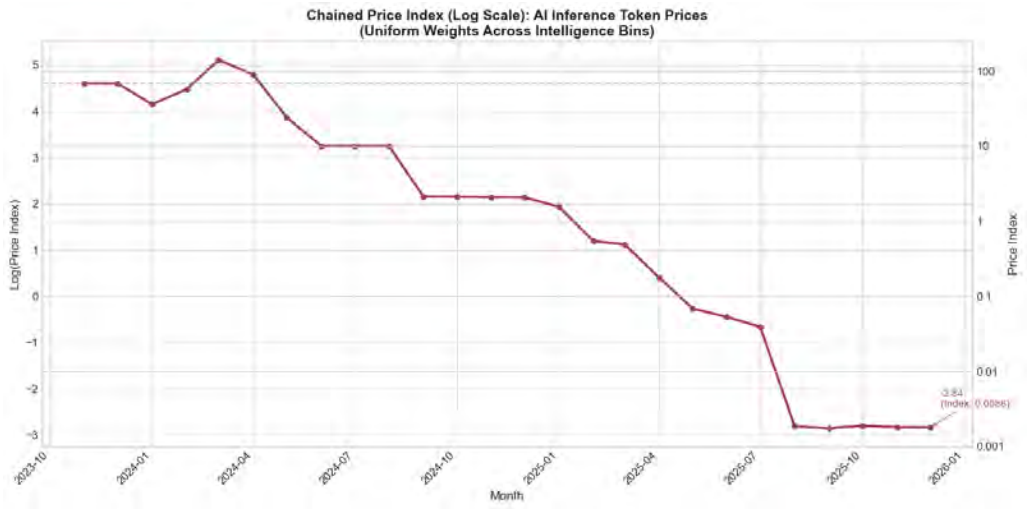

Figure 2: Chained Fisher Price Index for AI Inference Tokens (constant capability). The index tracks the minimum price per million tokens within each intelligence tier and chains them into a single quality-adjusted price series. The rapid decline reflects falling inference costs at constant model capability,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

6.6 Training: Algorithmic Progress Deflator

Just as inference output benefits from model improvements, training compute becomes more productive over time due to algorithmic progress. Advances in training methods, architectures, and techniques mean that achieving a given level of model capability requires steadily less compute.

This algorithmic efficiency gain represents real productivity growth in the "investment goods" sector of the AI economy—each dollar spent on training compute yields more valuable model capital than previously.

Methodology. Ho et al. (2024) find that algorithmic progress in AI training methods has been making compute approximately 3× more effective each year. Equivalently, the compute needed to train a model to a given capability level today is roughly $\frac{1}{3}$ what was needed a year ago.

Based on an 8-month doubling time for effective compute per unit of physical compute:

$$\text{Training Deflator Ratio} = 0.5^{(12/8)} \approx 0.354 \tag{20}$$

The quality-adjusted training growth is then:

$$\text{QA Training Growth} = \frac{\text{Compute Growth Ratio}}{\text{Training Deflator Ratio}} - 1 \tag{21}$$

This deflator is applied to training compute spending in the GDP quantity index (Section 7.5).

Applying this deflator to the underlying physical compute growth yields quality-adjusted training output growth of approximately 782% in 2024 and 788% in 2025—far exceeding nominal training compute spending growth of roughly 144% per year.

7 Other Inputs and Components

Beyond compute, several other economic flows contribute to (or subtract from) AI GDP. This section describes these components and our current approaches to measuring them.

7.1 Datacenter Labor Inputs

Operating AI datacenters requires human labor inputs for maintenance, monitoring, and technical support. Within our framework, these labor services come from outside the AI economy boundary and are treated as imported inputs that reduce AI GDP, similar to how imported intermediate goods reduce national GDP in trade accounting. While likely a small component currently, tracking this quantity is important for completeness of the GDP calculation and for understanding how the labor-to-compute ratio may change over time as datacenters scale and become more automated.

Staffing Model. We estimate datacenter labor costs using a simple staffing density model based on physical rack space. This is a highly stylized model based on rough assumptions, and should be seen as an order-of-magnitude estimate for staffing costs. The key assumptions are:

Table 8: Datacenter Labor Parameters

Parameter	Value	Description
Staffing density	50[a] / 100,000 sqft[b]	Staff per square foot of rack space
Annual salary	$100,000[c]	Average fully-loaded cost per staff member
Hours per year	8,760	24×365

[a] Microsoft has stated that they employ roughly 50 employees per building. [b] Described as the size of a typical datacenter building in media sources. [c] Assumption within the range of posted salaries for datacenter technicians, engineers, and electricians on ZipRecruiter.

Rack geometry (square footage per rack, chips per rack) was collected from suggested specifications for the most prominent chip types, with simple assumptions applied to less popular chips. Labor cost per watt-hour of total facility power is:

$$\text{labor\_cost\_per\_Wh} = \frac{\text{sqft\_per\_rack}}{\text{chips\_per\_rack}} \times \frac{\text{staff\_per\_sqft} \times \text{annual\_salary}}{\text{hours\_per\_year}} \times \frac{\text{gpu\_fraction}}{\text{TDP} \times \text{PUE}} \tag{22}$$

Labor costs are distributed across months using the same `IT_power_adjusted` shares as compute spending, so that labor scales with the power footprint of each chip type.

Price assumption. Wages are held constant across years.

7.2 Electricity Costs

Electricity is the largest imported input into AI compute production. In our framework, power originates outside the AI economy boundary and is purchased from the broader economy. Annual electricity consumption by AI datacenters is estimated from SemiAnalysis projections (see Section 4.1). We convert consumption to nominal spending using industrial electricity prices from the U.S. Energy Information Administration (EIA) (Energy Information Administration, 2024).

Table 9: Industrial Electricity Prices Used

Year	Price ($/kWh)
2023	0.0804
2024	0.0815
2025	0.0862

7.3 AI Services Revenues and Margin

Beyond the direct cost of inference compute, AI services command additional value that reflects:

- The intangible capital embodied in trained models
- The software scaffolding that delivers services to users

The margin—the difference between AI service revenues and the cost of underlying compute—represents the return to model intellectual property and engineering investments.

Inference/Training Split. We assume inference accounts for 50% of total compute spending in each year (2023–2025). The remainder is attributed to training. This is broadly consistent with industry estimates and is treated as constant pending better data.

Revenue Calculation. AI services revenue is modelled as a fixed markup on inference compute spending:

$$\text{AI Services Revenue} = 1.5 \times \text{inference compute spending} \tag{23}$$

The 1.5× markup implies AI service providers earn 50% above the underlying inference compute cost [1]. This premium reflects the intangible capital embodied in trained models, the software scaffolding that delivers services to users, and API infrastructure—all of which add value beyond raw compute.

Deflation. The inference quality-adjusted price deflator (Section 6) is used as the price index for AI services revenue in the Fisher quantity index. As inference prices decline rapidly, the real value of AI services output grows far faster than nominal revenue.

7.4 AI Services Labor

Many AI services require substantial human engineering inputs to:

- Build platforms
- Integrate models into applications
- Create user interfaces

These labor inputs represent imported services in our framework—they enable AI service production but originate from outside the AI economy boundary.

Distinguishing AI services labor from pure compute costs is important for accurately measuring value-added. If we count full service revenue but fail to subtract these imported labor inputs, we would overstate AI GDP by double-counting human contributions that properly belong to the traditional economy.

Current Assumption. We assume AI services labor equals 10% of inference compute spending:

$$\text{AI Services Labor} = 0.1 \times \text{inference compute spending} \tag{24}$$

This covers human workers across the AI services sector—customer support, fine-tuning specialists, prompt engineers, and ML engineers maintaining production systems. The 10% ratio is a provisional assumption anchored loosely to labor cost shares at consumer web companies.

[1] The markup is anchored to OpenAI 2024 margins from the Epoch Companies Dataset. See https://epoch.ai/data/ai-companies

Price assumption. Constant prices are assumed for AI services labor (price ratio = 1.0 in the Fisher index), reflecting that wages in this sector are not declining at the rate of inference compute costs.

7.5 Real AI GDP: Chained Fisher Quantity Index

Measuring real AI GDP requires deflating nominal output and imported inputs by appropriate price indices. The extraordinary pace of price change in AI—particularly the rapid decline of inference token prices—makes the choice of index methodology consequential in a way that is unusual for most sectors of the economy.

Production approach. Nominal AI GDP follows the standard production approach:

$$\text{Nominal AI GDP} = \underbrace{(\text{Training Compute} + \text{AI Services Revenue})}_{\text{Output}} - \underbrace{(\text{DC Labor} + \text{Power Costs} + \text{AI Ser}}_{\text{Imported Inputs}} \tag{25}$$

where each imported input represents a flow originating outside the AI economy boundary—analogous to imports in national accounts.

Price deflators. Each component has a distinct year-over-year price ratio:

Table 10: Price Deflators by AI GDP Component

Component	**Sign**	**Price Ratio (YoY)**
Training compute	+ (output)	≈ 0.354 (algorithmic progress)
AI services revenue	+ (output)	≈ 0.063 (inference price index)
DC labor	− (import)	1.0 (flat wages assumed)
Power costs	− (import)	P_t/P_{t-1} (actual electricity prices)
AI services labor	− (import)	1.0 (constant prices assumed)

Why the Fisher index matters. The inference price deflator declines at approximately 35× per year—an order of magnitude faster than price changes in most industries. This creates a severe index number problem: a fixed-weight index would use one year's relative prices to value all subsequent output, producing results that are highly sensitive to the choice of base year.

The chained Fisher quantity index—the standard methodology used by statistical agencies such as the U.S. Bureau of Economic Analysis—addresses this by re-weighting annually. For each consecutive year pair $(t-1, t)$, we compute:

$$Q_L = \frac{\sum_i \text{sign}_i \cdot v_i(t)/\pi_i}{\text{Nominal GDP}(t-1)}, \quad \text{(Laspeyres: current quantities at prior prices)} \tag{26}$$

$$Q_P = \frac{\text{Nominal GDP}(t)}{\sum_i \text{sign}_i \cdot v_i(t-1) \cdot \pi_i}, \quad \text{(Paasche: prior quantities at current prices)} \tag{27}$$

$$Q_F = \sqrt{Q_L \cdot Q_P}, \quad \text{(Fisher: geometric mean)} \tag{28}$$

where $v_i(t)$ is the nominal value of component i in year t and π_i is its year-over-year price ratio. The chain index is then:

$$I(t) = I(t-1) \times Q_F(t-1, t), \quad I(2023) = 100 \tag{29}$$

Consequence for measured growth. Because inference prices are falling roughly 16× faster than nominal inference spending is growing, the Fisher index reveals real output growth that is dramatically larger than nominal figures suggest. The methodology choice is not a technical detail—it is central to our core finding that AI economic output is growing far faster than conventional price-adjusted measures would indicate.

8 Results

This section presents our main empirical findings for the United States AI economy from 2023 to 2025. We report nominal spending and physical compute, quality-adjusted output growth, and real AI GDP.

8.1 AI Production Estimates

Table 11 reports annual nominal compute spending and physical compute output for the US AI economy. All AI GDP calculations in this section are based on the electricity-based methodology described in Section 4, which works top-down from aggregate AI datacenter power consumption to estimate compute production and spending.

Table 11: 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. 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.

Figure 3 illustrates how successive quality adjustments compound to produce the overall real output growth estimate. Starting from nominal spending growth, each layer—hardware performance improvements, inference price declines, and algorithmic progress in training—adds a further dimension of real growth that is invisible in the raw spending figures.

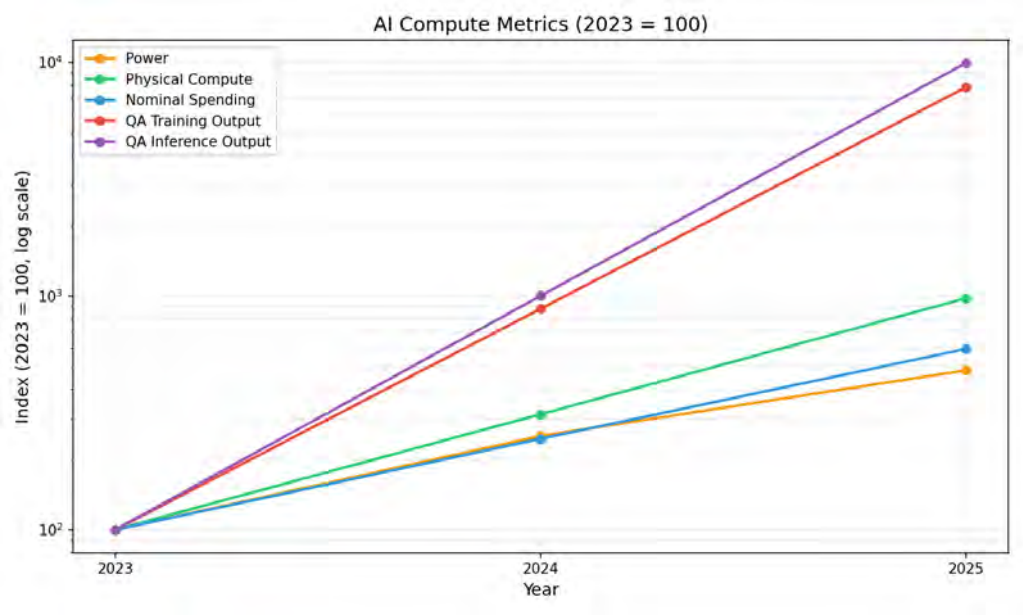

Figure 3: Layers of AI output growth. The figure shows how nominal spending growth (bottom layer) is amplified by hardware efficiency gains, inference price declines at constant capability, and algorithmic progress in training, yielding quality-adjusted output growth that far exceeds headline spending figures.

8.2 Quality-Adjusted Output Growth

Table 12 reports quality-adjusted growth rates for inference, training, and aggregate AI output. These measures correct for the rapid improvements in what a unit of compute can deliver.

Table 12: Quality-Adjusted AI Production Growth

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

Quality-adjusted inference growth deflates nominal inference spending by the inference price index (deflator ratio ≈ 0.063 per year; Section 6.1). The approximately 35× annual decline in per-token prices at constant capability, adjusted for a 2.2× increase in

response lengths, implies that the real volume of AI inference output grew roughly 39× per year.

Quality-adjusted training growth deflates compute growth by the algorithmic progress deflator (≈ 0.354 per year; Section 6.6), reflecting an approximate 3× improvement per year in the capability achievable per unit of training compute.

Quality-adjusted AI output growth combines inference and training growth weighted by their nominal shares. This measures quality-adjusted *production* growth—the expansion in real output before netting out imported inputs. The corresponding real AI *GDP* growth (output minus imports), computed via the chained Fisher quantity index (Section 7.5), is reported in Table 14 below.

8.3 Nominal and Real AI GDP

Table 13 decomposes nominal AI GDP into its output and imported-input components, and Table 14 shows the resulting real AI GDP index.

Table 13: 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 14: 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%

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 capabil-

ity per dollar than nominal spending figures capture. This is not primarily a story about spending—it is a story about dramatically falling prices at a constant quality standard.

8.4 Global Growth Rates (Chip Sales)

Table 15 reports global annual growth rates derived from the chip-sales methodology (Section 5), which covers worldwide shipments and is therefore not restricted to the United States.

Table 15: Global AI Growth Rates (Chip Sales Method)

Year	Nominal Spending	Physical Compute
2023	442%	466%
2024	291%	362%
2025	166%	235%

Note: Epoch chip sales data were adjusted in two quarters: quarterly growth for 2024Q1 was interpolated to correct for a structural break, and 2025Q4 was assumed to grow at the same rate as 2025Q3 to adjust for incomplete data.

Comparison with power-based estimates. The chip-sales method yields higher growth rates—particularly in 2023 and 2024—than the power-based approach, which records 145% spending growth in 2024. The divergence likely reflects: (i) the chip-sales method is global while the power-based method covers US datacenters only; and (ii) the two approaches differ in how they handle the rapid build-out of chip stock during this period. The chip-sales method captures spending on chips *purchased* (i.e., additions to the stock), whereas the power-based method captures spending on compute *consumed* (hours billed to customers). Given the rapid expansion of installed capacity relative to utilization, spending growth based on stock additions outpaces spending growth based on usage.

9 Supplementary Details

9.1 Data Sources

9.1.1 Epoch AI Datasets

GPU Clusters Dataset.

- **URL**: https://epoch.ai/data/generated/gpu_clusters.zip
- **Contents**: Known GPU clusters worldwide with:

  - Cluster name, location, operator
  - Chip type (primary and secondary)
  - Chip quantity
  - First operational date
  - Decommissioned date (if applicable)
  - Status (Confirmed, Likely, Planned)
  - Certainty indicators

ML Hardware Dataset.

- **URL**: https://epoch.ai/data/generated/ml_hardware.zip
- **Contents**: Hardware specifications including:
  - TDP (Thermal Design Power) in watts
  - Total processing performance (bit-OP/s)
  - Memory bandwidth
  - Release dates

9.1.2 GPU Cloud Pricing Data

- **Source**: Various public cloud provider listings
- **Contents**: Hourly rental rates for approximately 30 GPU types
- **Format**: Price range (low, high) for each GPU type

Table 16: Example GPU Cloud Pricing

GPU Type	Price Low ($/hr)	Price High ($/hr)
NVIDIA H100 SXM5 80GB	1.80	6.16
NVIDIA A100 80GB SXM4	0.50	4.22
Google TPU v4	1.45	3.22
AMD Instinct MI300X	1.50	7.86

9.1.3 GPU Power Fractions

- **Source**: Technical specifications by GPU type
- **Contents**: Fraction of node IT power consumed by GPU accelerators
- **Purpose**: Convert between total IT power and GPU power

Table 17: Example GPU Power Fractions

GPU Type	GPU Fraction
NVIDIA H100 SXM5	0.55
NVIDIA A100 SXM4	0.492
Google TPU v4	0.88
AMD MI300X	0.60

9.2 Detailed Formula Derivations

9.2.1 Revenue per Watt-hour

The `revenue_per_Wh` formula converts total facility power (in Wh) to revenue ($).

Formula:

$$\text{revenue\_per\_Wh} = \frac{\text{price\_low} \times \text{gpu\_fraction} \times \text{sellable\_fraction}}{\text{PUE} \times \text{TDP} \times \text{TDP\_adjustment\_factor}} \tag{30}$$

Unit Verification:

$$\begin{aligned}
&\frac{[\$/\text{hr}] \times [\text{unitless}] \times [\text{unitless}]}{[\text{unitless}] \times [\text{W}] \times [\text{unitless}]} \\
&= \frac{[\$/\text{hr}]}{[\text{W}]} \\
&= \frac{[\$/\text{hr}]}{[\text{J/s}]} \\
&= \frac{[\$ \times \text{s}]}{[\text{J} \times \text{hr}]} \\
&= \frac{[\$]}{[\text{J} \times 3600]} \\
&= \frac{[\$]}{[\text{Wh}]} \quad \checkmark
\end{aligned}$$

Conceptual Chain (working through 1 Wh of total facility power):

1. Start with 1 Wh of total facility power
2. IT power = 1 Wh / PUE = 0.769 Wh (for PUE = 1.3)
3. GPU power = IT power × gpu_fraction
4. GPU-hours = GPU power / (TDP × TDP_adjustment) = GPU energy / actual GPU power draw
5. Sellable GPU-hours = GPU-hours × sellable_fraction
6. Revenue = Sellable GPU-hours × price_low

9.2.2 Operations per Watt-hour

The `OP_per_Wh` formula converts total facility power to bit-operations.

Formula:

$$\text{OP\_per\_Wh} = \frac{\text{bit\_OP\_per\_s} \times 3600 \times \text{gpu\_fraction}}{\text{PUE} \times \text{TDP}} \tag{31}$$

Unit Verification:

$$\begin{aligned}
&\frac{[\text{bit-OP/s}] \times [\text{s/hr}] \times [\text{unitless}]}{[\text{unitless}] \times [\text{W}]} \\
&= \frac{[\text{bit-OP/hr}]}{[\text{W}]} \\
&= \frac{[\text{bit-OP/hr}]}{[\text{J/s}]} \\
&= \frac{[\text{bit-OP} \times \text{s}]}{[\text{hr} \times \text{J}]} \\
&= \frac{[\text{bit-OP}]}{[3600 \times \text{J}]} \\
&= \frac{[\text{bit-OP}]}{[\text{Wh}]} \quad \checkmark
\end{aligned}$$

Note: Unlike revenue, OP_per_Wh doesn't include sellable_fraction or TDP_adjustment because:

- Operations scale linearly with actual power draw
- We're measuring total operations performed, not "sold" operations
- The bit-OP/s spec is for maximum performance at maximum power

9.3 Worked Calculation Example

Given Inputs:

- Annual power 2023: 28.7 TWh, 2024: 72.3 TWh
- A cluster with 1000 H100 chips
- H100 specs: TDP = 700W, gpu_fraction = 0.55, price_low = \$1.80/hr, coverage = 0.30

Step 1: Calculate Derived Columns for This Cluster

$$\begin{aligned} \text{IT\_power} &= \text{chip\_quantity} \times \text{TDP} \times \text{PUE}/\text{gpu\_fraction} && (32)\\ &= 1000 \times 700 \times 1.3/0.55 && (33)\\ &= 1{,}654{,}545 \text{ W} && (34) \end{aligned}$$

$$\begin{aligned} \text{IT\_power\_adjusted} &= \text{IT\_power}/\text{coverage} && (35)\\ &= 1{,}654{,}545/0.30 && (36)\\ &= 5{,}515{,}152 \text{ W} && (37) \end{aligned}$$

$$\begin{aligned} \text{revenue\_per\_Wh} &= \frac{\text{price\_low} \times \text{gpu\_fraction} \times \text{sellable\_fraction}}{\text{PUE} \times \text{TDP} \times \text{TDP\_adj}} && (38)\\ &= \frac{1.80 \times 0.55 \times 0.9}{1.3 \times 700 \times 0.9} && (39)\\ &= \frac{0.891}{819} && (40)\\ &= 0.00109 \text{ \$/Wh} && (41) \end{aligned}$$

Step 2: Monthly Distribution Assuming this cluster is 10% of total IT_power_adjusted:

$$\text{If monthly\_power} = 2.87 \times 10^{12} \text{ Wh (Jan 2023, approximately)} \tag{42}$$

$$\text{proportion} = 0.10 \tag{43}$$

$$\text{power\_allocated} = \text{monthly\_power} \times \text{proportion} \tag{44}$$

$$= 2.87 \times 10^{12} \times 0.10 \tag{45}$$

$$= 2.87 \times 10^{11} \text{ Wh} \tag{46}$$

$$\text{spending} = \text{power\_allocated} \times \text{revenue\_per\_Wh} \tag{47}$$

$$= 2.87 \times 10^{11} \times 0.00109 \tag{48}$$

$$= \$312.8 \text{ million (for this cluster, this month)} \tag{49}$$

Key Insight: The power is distributed proportionally to IT_power_adjusted (not raw IT_power). This means clusters with low coverage factors (representing rare chips in Epoch's dataset) receive a larger share of total power, correcting for the incomplete capture of deployments.

References

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

Epoch AI (2024b). GPU clusters coverage analysis. https://arxiv.org/html/2504.16026v2. Coverage factors by chip type.

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.

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*.