Tokens Are Not Value:
Economic Accounting for AI Inference

Matthew Long
The YonedaAI Collaboration
YonedaAI Research Collective
Chicago, IL
matthew@yonedaai.com · https://yonedaai.com

23 July 2026

Abstract

Token counts are useful for billing, capacity planning, context management, and reproducible operation under a fixed tokenizer. They are not task outcomes or economic value. This paper develops a causal accounting chain for AI inference:

$$E_{\text{wall}} \longrightarrow C \longrightarrow T_\kappa \longrightarrow O \longrightarrow A \longrightarrow \Delta W.$$

Here T_κ is a token count under named tokenizer κ, O is a verified outcome under a task and service protocol, A is adoption, and ΔW is incremental net value relative to a counterfactual. We prove that tokens per joule is not invariant to tokenizer choice or verbosity. A ranking can reverse even when wall energy and task outcomes remain fixed. This does not make tokens useless. It limits the questions they can answer.

For field evaluation, we define a causal estimand as the ratio of the average treatment effect on net value to the average treatment effect on wall energy. That ratio is distinct from the expectation of unit-level value-to-energy ratios. We derive its difference-in-means estimator, variance and covariance terms, Fieller confidence set, and a stratified percentile bootstrap. Weak incremental-energy denominators can require unbounded confidence sets. The framework records quality, latency, retries, rework, terminal failures, adoption, shared idle energy, and interference between workers or tasks.

Existing workplace studies show that generative AI effects vary by task, worker, tool, and deployment. They provide important evidence about outcome stages, but they generally do not measure wall energy at the same experimental unit. Observational revenue per joule is therefore not causal productivity. A standard-library Python implementation supplies schemas, counterexamples, experimental estimators, and interval routines. The mathematical non-invariance results are PROVED; the software fixtures are COMPUTATIONAL; realized-value claims remain CONDITIONAL on a research design and declared boundary.

Keywords: AI inference; tokenization; energy measurement; causal inference; productivity; ratio estimation; Fieller interval; evaluation

Nonclaims. Tokens are not useless. A fixed-tokenizer token count can be the right operational unit. A benchmark score is not realized economic value. Observational revenue per joule is not causal productivity. This paper does not estimate an economy-wide return to AI energy.

1 Introduction

Inference services expose token counts because tokens are close to the computational interface. Providers meter them. Engineers use them to estimate memory, latency, throughput, and cost. Users see them in context limits and bills. This operational success makes tokens tempting as an economic output unit.

The temptation should be resisted. A tokenizer is a versioned segmentation rule. The same sentence can have different token counts under two tokenizers. A longer answer can use more tokens without being more correct or useful. A short answer can avert a costly mistake. A long answer can create rework. The token count is real, but its economic interpretation is not built into the count.

The accounting problem has two parts. First, a study needs a chain from wall energy to a verified service outcome:

$$E_{\text{wall}} \to \text{computation} \to \text{tokens} \to \text{quality-constrained task completion.}$$

Second, it needs a causal chain from the service to realized value:

$$\text{verified outcome} \to \text{adoption} \to \text{downstream effect} \to \Delta W.$$

Each arrow can fail. More output can increase review time. A correct output can arrive after a deadline. A useful recommendation can be ignored. An adopted recommendation can displace another worker's task or change team behavior.

This paper formalizes both parts. It follows the measurement tuple

$$\mathcal{M} = (b, c, W, \tau, e, a, \pi),$$

where b is boundary, c counterfactual, W value functional, τ time horizon, e energy convention, a attribution, and π uncertainty model. The tuple is not decoration. Changing one element can change the reported value-per-joule result.

The main claims are narrow.

1. Tokens per joule is not tokenizer invariant.
2. Tokens per joule is not verbosity invariant.
3. A causal incremental-value-per-joule estimand can be defined as a ratio of two average treatment effects.
4. A ratio of expectations is generally not an expectation of ratios.
5. Confidence sets for a ratio can be unbounded when incremental energy is weakly identified.

The paper also gives reporting rules for quality, latency, retries, rework, failures, spillovers, and uncertainty. A reference implementation turns those rules into testable schemas and estimators.

1.1 Why this is an economic accounting paper

Hardware efficiency and benchmark quality matter, but neither identifies economic value. Economic value requires a counterfactual. The relevant question is not how much revenue

appeared next to an AI system. It is how outcomes changed relative to what would have happened without the intervention, net of declared costs and harms.

This distinction is familiar in program evaluation [20, 12]. It becomes easy to lose in AI measurement because providers observe enormous streams of tokens and relatively few well-identified outcomes. A large intermediate data set does not compensate for a missing counterfactual.

1.2 Contribution and novelty boundary

Subword tokenization, model evaluation, energy reporting, causal inference, cost-benefit analysis, and ratio inference have mature literatures. This paper does not replace them. Its contribution is a shared accounting contract that keeps their outputs typed and forces a study to state where evidence ends.

2 Related work

2.1 Tokenization

Modern language models commonly operate on subword units. Byte-pair encoding for neural machine translation was developed to handle rare and open-vocabulary forms [22]. SentencePiece provides language-independent subword tokenization from raw text [13]. These methods were designed for modeling and engineering goals, not as natural units of economic output.

Tokenizer choice can affect sequence length, multilingual representation, performance, and compute. Rust et al. [21] show that tokenizer quality helps explain differences between multilingual and monolingual models. Ali et al. [1] report that tokenizer choices can alter downstream performance and training cost. This empirical literature studies more than the formal counting issue proved here. Our theorem needs only the simpler fact that different deterministic segmentations can assign different counts to the same string.

2.2 Evaluation and TEVV

HELM argues for multi-metric, scenario-based evaluation rather than one aggregate score [14]. NIST's AI Risk Management Framework and TEVV program emphasize context, documented metrics, validity, reliability, robustness, safety, and ongoing evaluation [17, 18]. These sources support the separation between model output and validated system performance.

A task metric still does not equal realized value. Accuracy on a frozen data set is evidence about one stage. Workflow adoption and downstream outcomes require field evidence.

2.3 Inference energy

MLCommons measures full-system AC power at the wall for the accompanying benchmark configuration and warns that thermal design power or power-supply ratings are not validated measurements [16]. Henderson et al. [9] propose systematic reporting of machine-learning energy and carbon. Luccioni et al. [15] compare inference energy across model and task classes.

These studies make the denominator more credible. They do not choose the economic numerator. Carbon also differs from energy because location, time, and generation mix intervene.

2.4 AI productivity evidence

Noy and Zhang [19] ran an incentivized experiment on professional writing tasks and found faster completion and higher evaluated quality with generative AI in that setting. Brynjolfsson, Li, and Raymond [4] studied a staggered deployment to 5,172 customer support agents and estimated a 15 percent increase in issues resolved per hour, with substantial heterogeneity.

Dell'Acqua et al. [5] report field-experimental evidence from consulting tasks. Performance improved on tasks within the model's capability frontier, while participants using AI were less likely to reach correct solutions on a selected task outside that frontier. Becker et al. [3]

randomized AI access across 246 tasks completed by 16 experienced open-source developers and estimated a slowdown in that narrow setting. Humlum and Vestergaard [11] use survey and administrative data to study broader workplace and labor-market adjustment.

The mixed results are informative. They do not imply that AI is generally helpful or harmful. They show why task, worker, deployment, and horizon belong in the estimand. None of these studies jointly meters experimental-unit wall energy and counterfactual net value, so none directly estimates the ratio proposed here.

2.5 Ratios and causal inference

Fieller [7] developed confidence sets for ratios whose denominator is estimated. Gleser and Hwang [8] show why bounded confidence procedures can fail badly in related weak-denominator problems. Efron and Tibshirani [6] provide the standard bootstrap reference. Interference and spillovers require potential outcomes indexed by more than a unit's own assignment [10, 2].

3 The inference accounting chain

3.1 Typed stages

For an experimental unit i, define:

- E_i: operational wall energy under a named boundary;
- C_i: named computation events under a hardware and software stack;
- $T_{i,\kappa}$: tokens under tokenizer κ;
- O_i: verified task outcome under protocol P;
- A_i: adopted or realized use;
- ΔW_i: counterfactual net value under functional W.

The path is

$$E_i \to C_i \to T_{i,\kappa} \to O_i \to A_i \to \Delta W_i.$$

The token node is explicitly indexed by κ. The outcome node is indexed by task distribution, scoring rule, quality threshold, latency budget, and version. The value node is indexed by counterfactual and horizon.

3.2 Why stages do not collapse

An output token can be part of a correct answer, an unnecessary restatement, or a harmful fabrication. Verification changes the type from emitted symbol to scored outcome. Adoption changes the type again. A correct answer that nobody uses has no realized workflow effect, although it remains a correct answer.

The final stage is causal. If $Y_i(1)$ and $Y_i(0)$ are potential outcomes with and without AI access, then

$$\Delta W_i = W(Y_i(1)) - W(Y_i(0)) .$$

Only one potential outcome is observed. Randomization or another identification strategy is needed.

3.3 Net value

Let $G_i(z)$ be gross realized benefit, $R_i(z)$ rework cost, $F_i(z)$ failure cost, and $K_i(z)$ other nonenergy opportunity cost under assignment z. One declared net-value functional is

$$N_i(z) = G_i(z) - R_i(z) - F_i(z) - K_i(z).$$

Energy purchase cost can be included or excluded, but the choice must be explicit. When the denominator is physical joules, excluding monetary energy cost from N often avoids mixing an electricity price into both the numerator and interpretation. Environmental externalities can be reported as a separate vector component or monetized under published weights.

3.4 Boundary manifest

The denominator records model, precision, hardware, software, prompt distribution, output rule, batching, latency target, retries, idle allocation, host and memory, networking, facility overhead, and time interval. Embodied energy is separate unless a lifecycle study allocates it.

The numerator records task version, affected population, quality rule, adoption rule, rework, failure costs, counterfactual, price year, discounting, and spillover exposure. A study that measures one side at the request level and the other at the firm-quarter level needs a bridge.

4 Tokenizer non-invariance

Definition 4.1 (Tokenizer)**.** A tokenizer κ is a deterministic map from strings to finite sequences of token identifiers. Its count is

$$T_\kappa(s) = |\kappa(s)|.$$

For fixed wall energy $E(s) > 0$, define the operational metric

$$\eta_\kappa(s) = \frac{T_\kappa(s)}{E(s)}.$$

Theorem 4.2 (Tokenizer non-invariance)**.** *There exist a string s, positive energy E, and tokenizers κ_1, κ_2 such that*

$$\eta_{\kappa_1}(s) \neq \eta_{\kappa_2}(s),$$

while the emitted string and declared task outcome are identical.

Proof. Let $s =$ `abcdefghij`. Let κ_1 map the complete string to one token, and let κ_2 map each character to one token. Then $T_{\kappa_1}(s) = 1$ and $T_{\kappa_2}(s) = 10$. Dividing by the same positive energy preserves the inequality. The string and any outcome determined solely by that string remain unchanged. □

Theorem 4.3 (Tokenizer rank reversal)**.** *There exist outputs s_A, s_B, equal positive energies, and tokenizers κ_1, κ_2 for which the tokens-per-joule ordering reverses.*

Proof. Let $s_A =$ `abcdefghij` and $s_B =$ `a b c d e`. Under whitespace tokenization, the counts are one and five, so B ranks higher. Under a space-preserving character tokenizer, the counts are ten and nine, so A ranks higher. Equal energy leaves the rankings unchanged within each tokenizer and reversed across tokenizers. □

The theorem is PROVED. It does not claim that real tokenizers are arbitrary or that their compute costs are equal. It shows that the token numerator is representation-dependent.

4.1 Multilingual consequences

Different languages and scripts can have different token fertility under the same multilingual tokenizer. A token-priced service can therefore charge different token counts for semantically comparable content. This is an engineering and distributional issue, not proof that every language has a different economic value. Comparisons should report language mix and tokenizer version.

4.2 The legitimate fixed-tokenizer use

When tokenizer, model, precision, decoding rule, and workload are fixed, tokens per joule can diagnose serving efficiency. It can compare batching or kernel changes for the same service contract. The conclusion is not "stop measuring tokens." It is "do not relabel the operational measure as value."

5 Verbosity non-invariance

Definition 5.1 (Outcome equivalence)**.** Two outputs are outcome-equivalent under protocol P if the protocol assigns them the same task score, service status, adoption state, and downstream action.

Theorem 5.2 (Verbosity non-invariance)**.** *For a tokenizer that assigns a positive count to appended text, there can be outcome-equivalent outputs with different token counts.*

Proof. Consider a task whose correct answer is 42. Let $s_1 =$ `42` and $s_2 =$ `The answer is 42`. A protocol that checks the extracted answer can assign both the same correct outcome. A whitespace tokenizer assigns one token to the first and four to the second. Thus equal outcome does not imply equal token count. □

If energy also rises with verbosity, tokens per joule may rise, fall, or remain constant. None of those directions determines task value. Longer reasoning can improve a difficult answer; it can also add delay and review burden. The relationship is empirical and task-specific.

5.1 Billing is not welfare accounting

Token billing allocates provider costs and manages demand. A billing unit need not equal a welfare unit. Postal services charge by weight even though a letter's value is not its mass. The analogy is limited but useful: an operational meter can work well without measuring benefit.

5.2 Goodhart pressure

If teams are rewarded for tokens per joule, they can raise the numerator by changing tokenizer or verbosity. A fixed protocol reduces this gaming but does not connect tokens to usefulness. Outcome metrics should sit beside operational metrics, and economic claims should use downstream evidence.

6 Quality, latency, reliability, and failure

6.1 Task schema

A task specification is

$$P = (\mathcal{D}, S, q^\star, L^\star, d, v),$$

where $\mathcal{D}$ is task distribution, S scoring rule, $q^\star$ quality threshold, $L^\star$ latency budget, d difficulty weight, and v version. For unit i,

$$O_i = d_i \mathbf{1}\{S_i \geq q_i^\star\} \mathbf{1}\{L_i \leq L_i^\star\} \mathbf{1}\{\text{no terminal failure}\}.$$

Quality and latency are joint constraints. An accurate response delivered after a contractual deadline can fail the service. A fast incorrect response also fails.

6.2 Retries

Retries consume energy and time. Let attempts $j = 1, \ldots, J_i$ have energy E_{ij}. The task denominator includes

$$E_i^{\text{attempt}} = \sum_{j=1}^{J_i} E_{ij}$$

when retries fall inside the boundary. Reporting only the successful attempt conditions on success and understates energy per delivered outcome.

Retries can be automatic, user-initiated, or caused by validation failure. Their causes belong in diagnostics. Their energy belongs in the total once.

6.3 Rework

Rework includes human review, correction, additional inference, testing, and downstream repair. Monetary rework cost enters net value. Measured compute used for rework enters energy if the boundary covers the full workflow. Human labor is not converted into metabolic joules. It remains labor time or cost.

6.4 Terminal failures and missing outcomes

Timeouts, refusals, malformed responses, abandoned sessions, and unmeasured downstream results need explicit branches. Dropping them can make quality and energy metrics look better. Missingness assumptions should be stated, and bounded analyses should replace point estimates when outcomes cannot be recovered.

6.5 Reliability

One successful run does not establish a reliable service. Repeated tasks, prompt variation, temporal drift, and rare high-cost errors matter. A study can report a service frontier over quality, latency, failure probability, and wall energy rather than collapsing every dimension.

7 Operational wall-energy accounting

7.1 Non-overlapping components

For unit i, define

$$E_i^{\text{wall}} = E_i^{\text{initial}} + E_i^{\text{retry}} + E_i^{\text{host}} + E_i^{\text{network}} + E_i^{\text{facility}} + E_i^{\text{idle,alloc}} + E_i^{\text{rework}}.$$

The components must be non-overlapping. If a wall meter already captures host and network energy, those terms are not added again. The equation is a schema, not a command to sum duplicate telemetry.

7.2 Idle allocation

Serving systems reserve capacity. Idle energy can be allocated by request time, peak capacity, tokens, or another rule. Rankings can change with the rule. Publish both active-only and allocated operational results when possible.

7.3 Wall measurement

Wall measurement integrates power:

$$E_i = \int_{t_{i,0}}^{t_{i,1}} P(t)\, dt.$$

Clock alignment matters when task logs and power logs use different systems. Sampling must capture bursts. Calibration, uncertainty, warm-up, cooldown, and background load need documentation.

MLCommons provides a strong benchmark precedent because its measured power is valid for the accompanying system and benchmark, not every deployment [16]. A field study needs the same humility.

7.4 Control energy

The control arm can use energy. A human workflow may use laptops, search, databases, and office infrastructure. The causal denominator is

$$\tau_E = \mathbb{E}[E_i(1) - E_i(0)],$$

not treatment energy alone. A system that replaces more energy-intensive work can have $\tau_E < 0$. The sign changes interpretation.

7.5 Operational is not lifecycle

Operational wall energy excludes embodied equipment and construction unless allocated separately. A lifecycle study can add non-overlapping embodied components under a lifetime and utilization model. It should not call a wall meter lifecycle energy.

8 Causal estimands

8.1 Potential outcomes

Let $Z_i \in \{0, 1\}$ assign AI access. Define potential net value $N_i(z)$ and operational wall energy $E_i(z)$. The average effects are

$$\tau_N = \mathbb{E}[N_i(1) - N_i(0)], \qquad \tau_E = \mathbb{E}[E_i(1) - E_i(0)].$$

Definition 8.1 (Incremental net value per incremental joule)**.** When $\tau_E \neq 0$, define

$$\theta = \frac{\tau_N}{\tau_E}.$$

For $\tau_E > 0$, θ is expected incremental net value per additional operational joule under the study population and assignment policy. For $\tau_E < 0$, the intervention saves energy; the sign and numerator should be reported separately rather than summarized with a slogan.

8.2 Identification

In a randomized experiment with no relevant interference, arm means identify $\mathbb{E}[N_i(z)]$ and $\mathbb{E}[E_i(z)]$. In observational deployment, selection into AI use can depend on task difficulty, worker skill, expected benefit, and management. Regression adjustment needs assumptions. Revenue divided by observed AI electricity does not identify θ.

8.3 Treatment policy

"AI access" is incomplete. Treatment includes model and version, interface, training, price, usage limits, workflow, and enforcement. An intention-to-treat estimand measures assignment to access. A treatment-on-the-treated estimand needs uptake modeling. Selective usage can make naive user-versus-nonuser comparisons misleading.

Corollary 8.2 (Partial compliance under exclusion)**.** *Let assignment Z be an instrument for adoption A. Under random assignment, monotonicity, a nonzero first stage $\tau_A = \mathbb{E}[A(1) - A(0)]$, and exclusion of any route from assignment to net value or energy other than adoption, the ratio of complier effects equals the intention-to-treat ratio:*

$$\frac{\tau_N/\tau_A}{\tau_E/\tau_A} = \frac{\tau_N}{\tau_E},$$

provided $\tau_E \neq 0$.

Proof. The Wald estimands for complier net value and energy are respectively τ_N/τ_A and τ_E/τ_A. Their ratio cancels the common, nonzero adoption first stage. □

The result is conditional, not automatic. Exclusion fails if mere access changes training time, reserved capacity, idle energy, or behavior among nonadopters. In that case assignment and adoption define different policies, and the numerator and denominator need a design that represents those routes.

8.4 Net value and transfers

Revenue is not always social value. Payments can transfer surplus. A firm perspective can use profit or contribution margin. A social perspective can include consumer surplus, labor effects, externalities, and distributional weights. The paper permits several W functions but requires one to be named.

8.5 Horizon

Short-run completion time can improve while long-run learning, skill, quality, or market equilibrium changes in another direction. The estimand includes a horizon τ. A one-day experiment should not be silently projected over years.

9 Ratio of effects versus effect of ratios

Two summaries are often confused:

$$\mathrm{RoE} = \frac{\mathbb{E}[N]}{\mathbb{E}[E]}, \qquad \mathrm{EoR} = \mathbb{E}\left[\frac{N}{E}\right].$$

Proposition 9.1 (Non-equivalence)**.** *In general,* $\mathrm{RoE} \neq \mathrm{EoR}$.

Proof. Let (N, E) equal $(1, 1)$ or $(9, 3)$ with equal probability. Then

$$\mathrm{RoE} = \frac{(1+9)/2}{(1+3)/2} = \frac{5}{2},$$

while

$$\mathrm{EoR} = \frac{1/1 + 9/3}{2} = 2.$$

□

The ratio of means weights units in proportion to energy when rewritten as a weighted average of unit ratios. The mean of ratios gives every defined unit equal weight. Both can answer legitimate questions, but they answer different ones.

9.1 Causal version

Our target is

$$\theta = \frac{\mathbb{E}[N(1) - N(0)]}{\mathbb{E}[E(1) - E(0)]}.$$

It is not

$$\mathbb{E}\left[\frac{N(1) - N(0)}{E(1) - E(0)}\right].$$

The unit-level expression may be undefined when an individual energy effect is zero, and individual potential-outcome differences are not jointly observed.

9.2 Aggregation level

Pure regrouping does not change a ratio of totals. If net value and energy are summed over the same units with the same weights and boundary, then

$$\frac{\sum_i N_i}{\sum_i E_i} = \frac{\sum_g \sum_{i \in g} N_i}{\sum_g \sum_{i \in g} E_i}.$$

The ratio of expectations inherits this volume-aggregation invariance under consistent weighting. The expectation of unit ratios does not: changing from requests to workers changes which ratios receive equal weight and can also change which zero denominators are discarded.

Aggregation can still change the causal ratio when it changes weights, captures shared idle energy, or incorporates team spillovers and interference missing at request level. Equal weighting of unequal-sized clusters is not pure regrouping. A study should define its unit, boundary, and weighting rule before estimation.

10 Estimation and interval uncertainty

10.1 Difference-in-means estimator

For treated and control sample means,

$$\widehat{\tau}_N = \bar{N}_1 - \bar{N}_0, \qquad \widehat{\tau}_E = \bar{E}_1 - \bar{E}_0, \qquad \widehat{\theta} = \frac{\widehat{\tau}_N}{\widehat{\tau}_E}.$$

Under random assignment, the first two estimators have standard randomization-based interpretations. Ratio bias can remain in finite samples.

Let

$$\widehat{V}_N = \widehat{\mathrm{Var}}(\widehat{\tau}_N), \qquad \widehat{V}_E = \widehat{\mathrm{Var}}(\widehat{\tau}_E), \qquad \widehat{C}_{NE} = \widehat{\mathrm{Cov}}(\widehat{\tau}_N, \widehat{\tau}_E).$$

These are estimated sampling variances and covariance of the two effect estimators, not raw sample variances of N_i and E_i. The covariance matters because high-value tasks may also consume more energy.

10.2 Delta interval

When $\widehat{\tau}_E$ is far from zero, the delta-method variance is

$$\widehat{\mathrm{Var}}(\widehat{\theta}) \approx \frac{\widehat{V}_N}{\widehat{\tau}_E^2} + \frac{\widehat{\tau}_N^2 \widehat{V}_E}{\widehat{\tau}_E^4} - 2\frac{\widehat{\tau}_N \widehat{C}_{NE}}{\widehat{\tau}_E^3}.$$

This approximation can fail near a zero denominator and can hide the correct unbounded geometry.

10.3 Fieller confidence set

For normal critical value q, candidate ratio r belongs to the Fieller set when

$$(\widehat{\tau}_N - r\widehat{\tau}_E)^2 \le q^2(\widehat{V}_N - 2r\widehat{C}_{NE} + r^2\widehat{V}_E).$$

Rearranging gives

$$Ar^2 + Br + C \le 0,$$

with

$$A = \widehat{\tau}_E^2 - q^2\widehat{V}_E, \quad B = -2(\widehat{\tau}_E\widehat{\tau}_N - q^2\widehat{C}_{NE}), \quad C = \widehat{\tau}_N^2 - q^2\widehat{V}_N.$$

Depending on A and the discriminant, the confidence set can be bounded, disjoint and unbounded, all real numbers, or empty under the approximation.

An unbounded result is not a software defect. It says the data do not exclude a denominator near zero strongly enough to support a bounded ratio.

10.4 Bootstrap

A stratified bootstrap resamples treated and control units separately and recomputes both effects and their ratio. It preserves arm sizes and the within-unit value-energy relation. Cluster assignment requires cluster resampling.

Percentile intervals can behave poorly near zero denominators. The software counts resamples whose incremental energy is too close to zero and refuses to report an interval when too many are invalid. Fieller output remains the primary weak-denominator diagnostic.

10.5 Multiple outcomes

Quality, latency, failures, rework, and value create multiple testing and selection risks. The analysis plan should identify primary outcomes and report the rest as a vector. Tuning the value functional after seeing results defeats its role as a declared estimand.

11 Spillovers and interference

11.1 Why SUTVA can fail

One worker's AI use can affect colleagues through shared queues, copied outputs, training, review load, and changed norms. One task can populate a cache or knowledge base used by later tasks. Then

$$N_i(\mathbf{Z}) \neq N_i(Z_i)$$

because outcomes depend on the assignment vector.

11.2 Exposure mappings

An exposure mapping $g_i(\mathbf{Z})$ can classify a unit by own treatment and neighbors' treatment [2]. Direct, indirect, total, and overall effects are different estimands. Energy must be allocated at the same exposure level.

11.3 Cluster randomization

Randomizing teams or time blocks can reduce contamination. It lowers the effective sample size and requires cluster-aware inference. Shared facility energy may naturally live at the cluster level.

11.4 General equilibrium

A firm deployment can change prices, hiring, task volume, and demand. A local randomized effect does not equal an economy-wide equilibrium effect. Humlum and Vestergaard [11] illustrate the value of studying broader labor outcomes separately from controlled task productivity.

11.5 Rebound

Cheaper inference can increase demand. Even if value per inference joule rises, total energy may rise. The causal ratio in this paper is policy and horizon specific. It is not a forecast of aggregate energy demand.

12 Reading the field evidence

12.1 Writing tasks

Noy and Zhang's experiment [19] offers credible evidence that generative AI access changed completion time and evaluated quality for a specified set of professional writing tasks. In our chain, it informs the verified-outcome stage. To estimate incremental value per joule, the same units would need metered treatment and control energy, adoption and rework records, and a value functional.

12.2 Customer support

Brynjolfsson, Li, and Raymond [4] analyze deployment at scale and find heterogeneous productivity changes. Issues resolved per hour is closer to a workflow outcome than token volume. It still requires quality checks, customer effects, costs, and wall-energy measurement for our target ratio.

12.3 Consulting

The jagged-frontier study [5] is especially relevant because effects changed with task location relative to model capability. More output was not uniformly better. A task distribution that omits outside-frontier work can overstate deployment value.

12.4 Software development

Becker et al. [3] study experienced developers working in familiar repositories and estimate slower completion with the tested early 2025 tools. The sample and tool era limit generalization. The study still demonstrates that review, prompting, and context costs can outweigh generation speed in a real workflow.

12.5 What the studies do not say

The studies do not show that tokens are useless. They do not establish a universal effect of AI. They also do not measure causal value per wall joule. Their strength is design-specific evidence about intermediate and outcome stages. Combining their productivity estimates with unrelated provider energy numbers would create a synthetic ratio with mismatched units and populations.

13 Experimental protocol

13.1 Unit and assignment

Choose the unit at which treatment and spillovers are credible: request, task, worker, team, or time block. Record assignment before uptake. Preserve failed and abandoned units.

13.2 Treatment definition

Pin model, endpoint, date, system prompt, tools, tokenizer, decoding parameters, context policy, user interface, training, and quotas. Provider model aliases can drift and should not stand alone as version identifiers.

13.3 Control definition

The control can be usual workflow, alternative software, another model, or no AI. Record its compute and labor inputs under a compatible boundary. A zero-energy control is rarely appropriate for workplace tasks.

13.4 Energy collection

Use synchronized wall meters where feasible. Record idle, warm-up, networking, retries, and rework. If facility overhead is allocated, publish the rule and a sensitivity analysis. Report device telemetry separately.

13.5 Outcome collection

Blind evaluators to assignment when possible. Record quality, latency, reliability, adoption, rework, and harms. Predefine the task distribution and failure policy. Link downstream outcomes without dropping unmatched cases.

13.6 Value collection

Choose firm, worker, customer, or social perspective. State price year, discount rate, transfers, externalities, and horizon. Estimate numerator components before aggregation so readers can reweight them.

13.7 Analysis

Report $\widehat{\tau}_N$ and $\widehat{\tau}_E$ separately with intervals. Then report the ratio, Fieller set, bootstrap diagnostic, covariance, and weak-denominator warning. Show task-stratum and allocation sensitivity without turning exploratory subgroups into confirmatory findings.

14 Constructed examples

14.1 Tokenizer rank reversal

The Python fixture compares `abcdefghij` with `a b c d e` at one joule each. Whitespace tokenization assigns counts one and five. Space-preserving character tokenization assigns counts ten and nine. The ordering reverses. No claim about semantic quality is needed.

14.2 Verbosity

At two joules, `42` has one whitespace token and `The answer is 42` has four. If the task protocol extracts the same correct answer, tokens per joule changes from 0.5 to 2.0 while the declared outcome is unchanged.

14.3 Energy ledger

A constructed task uses 10 joules for initial inference, 3 for retries, 2 for host and memory, 1 for network, 4 for facility overhead, 5 allocated idle, and 6 for downstream compute rework. The non-overlapping total is 31 joules. If the wall meter already includes several components, the ledger must not add them again.

14.4 Randomized ratio fixture

Four treated units have net values $13, 15, 17, 19$ and wall energies $16, 17, 18, 19$. Four controls have values $8, 10, 12, 14$ and energies $10, 11, 12, 13$. Therefore

$$\widehat{\tau}_N = 5, \qquad \widehat{\tau}_E = 6, \qquad \widehat{\theta} = \frac{5}{6} \approx 0.8333.$$

The software retains variance 3.3333, denominator variance 0.8333, and covariance 1.6667 for the effect estimators.

Under its normal approximation, the 95 percent Fieller set is approximately $[0.338, 1.101]$. A 20,000-draw stratified percentile bootstrap with seed 9 returns approximately $[0.444, 1.067]$. These are COMPUTATIONAL fixture results, not an empirical AI estimate.

15 Reference implementation

The module `src/joule_standard/inference_accounting.py` uses only the Python standard library.

15.1 Schemas

`TaskSpec` records task identifier, version, quality threshold, latency budget, and difficulty weight. `TaskOutcome` records score, latency, adoption, failure, retries, gross realized value, rework, and other nonenergy cost. `EnergyAccount` records non-overlapping operational components.

15.2 Counterexamples

Whitespace, character, and UTF-8 byte tokenizers make the representation choice inspectable. `has_rank_reversal` checks whether two named tokenizers reverse an ordering.

15.3 Estimators

`estimate_incremental_ratio` computes arm differences, sampling variances, covariance, and their ratio. `fieller_confidence_set` returns bounded, unbounded, all-real, or empty approximate sets. `bootstrap_ratio_interval` resamples within assignment arm and counts invalid weak-denominator draws.

15.4 Tests

Tests cover tokenizer counts, rank reversal, verbosity, task success, latency, adoption, retries, energy aggregation, ratio non-equivalence, difference in means, covariance, bounded and unbounded Fieller behavior, deterministic bootstrap, weak-denominator rejection, and schema validation.

Passing tests support COMPUTATIONAL claims about the code. They do not validate an external study's randomization, meter, value functional, or spillover model.

16 Claim status and falsifiers

Status	Claim	Falsifier or limit
PROVED	Tokens per joule is not invariant to tokenizer choice.	Comparison fixes tokenizer and claims only an operational metric.
PROVED	Outcome-equivalent answers can have different token counts.	Protocol defines verbosity itself as part of the outcome.
PROVED	A ratio of expectations can differ from an expectation of ratios.	Special distributions make them equal.
COMPUTATIONAL	Reference estimators reproduce declared fixtures.	A test fails or code accepts an invalid record.
CONDITIONAL	Randomized arm differences identify incremental net value and wall energy.	Randomization, measurement, no-interference, or attrition assumptions fail.
OBSTRUCTED	Observational revenue per joule identifies causal productivity without additional assumptions.	A credible design identifies the counterfactual numerator and denominator.
OPEN	A field study can estimate the full causal ratio at useful scale.	Requires synchronized value and energy evidence.

Table 1: Evidence status for the paper's main claims.

17 Limitations

17.1 No new field estimate

The paper proposes an estimand and implementation but supplies only constructed data. It does not estimate the return to a deployed AI service.

17.2 Value remains normative

Net value depends on perspective, horizon, prices, externalities, and distributional weights. A declared functional makes the choice visible but does not make it neutral.

17.3 Energy measurement can be expensive

Request-level wall metering is difficult in shared cloud infrastructure. Provider telemetry may omit host, network, and facility components. Allocation uncertainty can dominate statistical precision.

17.4 Rapid tool drift

Models, interfaces, user skill, and prices change quickly. A field estimate is tied to its deployment period. Replication over versions is necessary.

17.5 Fieller approximation

The implementation uses a normal approximation. Small samples, clustering, heavy tails, and complex assignment can require randomization inference or a more specialized model.

17.6 Bootstrap limits

The percentile bootstrap is simple and reproducible, but it is not reliable for every ratio problem. Near zero denominators demand unbounded-set diagnostics, not forced bounded intervals.

17.7 Interference

Exposure mappings can miss unknown spillovers. Team and market effects may remain after cluster randomization. The local estimand should not be advertised as an economy-wide effect.

17.8 Human outcomes

Productivity is not the only outcome. Skill, autonomy, satisfaction, error severity, privacy, and distribution matter. They should appear as separate outcomes before monetization.

18 Conclusion

Tokens are useful measurements. They are not natural units of task success or economic value. Their counts depend on tokenizer and verbosity, and their per-joule rankings can reverse without a change in declared outcome.

A defensible economic account follows the full chain from wall energy through computation, named-tokenizer output, verified service, adoption, and counterfactual net value. Quality, latency, retries, rework, failures, idle allocation, and spillovers are part of that chain.

The causal target proposed here is a ratio of average effects, not an average of unit ratios. Its numerator and denominator should be reported separately. Fieller confidence sets make weak incremental-energy evidence visible, even when that produces an unbounded result.

The field literature already shows large heterogeneity across work settings. Those studies are evidence against universal slogans, not against careful use of AI. The next empirical step is to pair credible productivity designs with synchronized wall-energy measurement and a predeclared value functional.

A Proof details

A.1 General rank reversal construction

Let two strings have tokenizer counts

$$(T_{\kappa_1}(A), T_{\kappa_1}(B)) = (a_1, b_1)$$

and

$$(T_{\kappa_2}(A), T_{\kappa_2}(B)) = (a_2, b_2).$$

For equal energy, a reversal occurs when

$$(a_1 - b_1)(a_2 - b_2) < 0.$$

For unequal energies E_A, E_B, replace each count with count divided by its energy. The reference function implements this sign test.

A.2 Ratio weighting identity

For positive E_i,

$$\frac{\sum_i N_i}{\sum_i E_i} = \sum_i \left(\frac{E_i}{\sum_j E_j} \right) \frac{N_i}{E_i}.$$

Thus the aggregate ratio is an energy-weighted mean of unit ratios. The unweighted mean answers another question.

A.3 Fieller geometry

The Fieller inequality is obtained by testing

$$H_0 : \tau_N - r\tau_E = 0$$

with estimated variance

$$\widehat{V}_N - 2r\widehat{C}_{NE} + r^2\widehat{V}_E.$$

Inverting the tests over r gives a confidence set. When the data do not exclude $\tau_E = 0$, arbitrarily large ratios can remain plausible. A bounded interval would conceal that feature.

B Adversarial audit checklist

1. Swap tokenizer version without changing the label.
2. Add redundant text and see whether the metric improves.
3. Drop failed and abandoned requests.
4. Count only final successful attempts.
5. Substitute device telemetry for wall energy.
6. Add facility overhead already captured by the meter.

7. Compare AI users with self-selected nonusers.
8. Treat gross revenue as incremental welfare.
9. Ignore review and correction time.
10. Ignore spillovers to untreated workers.
11. Average unit ratios with zero denominators removed.
12. Force a bounded interval when incremental energy is near zero.

An accounting system that accepts these substitutions without a warning has failed the intended contract.

C Minimal reporting manifest

1. Study, protocol, and preregistration identifiers.
2. Experimental unit, assignment, and exposure mapping.
3. Model, endpoint, version date, tools, and interface.
4. Tokenizer and decoding configuration.
5. Task distribution, quality scorer, threshold, and latency budget.
6. Failure, retry, refusal, and missingness policy.
7. Adoption and rework definitions.
8. Wall meter, sampling, calibration, and time synchronization.
9. Idle, host, network, facility, and lifecycle allocation.
10. Counterfactual and value functional.
11. Currency year, horizon, discounting, and perspective.
12. Variance, covariance, clustering, and spillover assumptions.
13. Separate numerator and denominator effects.
14. Ratio interval and weak-denominator diagnostic.
15. Exclusions and evidence status.

References

[1] M. Ali, M. Fromm, K. Thellmann, A. Ebert, A. Weber, M. L. Arnett, M. T. N. Vu, F. Busch, P. L. T. Schulz, J. Berrospi Ramis, and others, "Tokenizer choice for LLM training: negligible or crucial?" *Findings of NAACL 2024*, pp. 3907 to 3924, 2024. https://doi.org/10.18653/v1/2024.findings-naacl.247

[2] P. M. Aronow and C. Samii, "Estimating average causal effects under general interference, with application to a social network experiment," *Annals of Applied Statistics*, vol. 11, no. 4, pp. 1912 to 1947, 2017. https://doi.org/10.1214/16-AOAS1005

[3] J. Becker, N. Rush, E. Barnes, and D. Rein, "Measuring the impact of early-2025 AI on experienced open-source developer productivity," arXiv:2507.09089, 2025. https://doi.org/10.48550/arXiv.2507.09089

[4] E. Brynjolfsson, D. Li, and L. R. Raymond, "Generative AI at work," *Quarterly Journal of Economics*, vol. 140, no. 2, pp. 889 to 942, 2025. https://doi.org/10.1093/qje/qjae044

[5] F. Dell'Acqua, E. McFowland III, E. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani, "Navigating the jagged technological frontier: field experimental evidence of the effects of AI on knowledge worker productivity and quality," *Organization Science*, vol. 37, no. 2, pp. 403 to 423, 2026. https://doi.org/10.1287/orsc.2025.21838

[6] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993. https://doi.org/10.1201/9780429246593

[7] E. C. Fieller, "Some problems in interval estimation," *Journal of the Royal Statistical Society, Series B*, vol. 16, no. 2, pp. 175 to 185, 1954. https://doi.org/10.1111/j.2517-6161.1954.tb00159.x

[8] L. J. Gleser and J. T. Hwang, "The nonexistence of 100(1-alpha) percent confidence sets of finite expected diameter in errors-in-variables and related models," *Annals of Statistics*, vol. 15, no. 4, pp. 1351 to 1362, 1987. https://doi.org/10.1214/aos/1176350597

[9] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," *Journal of Machine Learning Research*, vol. 21, no. 248, pp. 1 to 43, 2020. https://jmlr.org/papers/v21/20-312.html

[10] M. G. Hudgens and M. E. Halloran, "Toward causal inference with interference," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 832 to 842, 2008. https://doi.org/10.1198/016214508000000292

[11] A. Humlum and E. Vestergaard, "Still waters, rapid currents: early labor market transformation under generative AI," NBER Working Paper 33777, 2025, revised 2025. https://doi.org/10.3386/w33777

[12] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press, 2015.

[13] T. Kudo and J. Richardson, "SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing," *Proceedings of EMNLP: System Demonstrations*, pp. 66 to 71, 2018. https://doi.org/10.18653/v1/D18-2012

[14] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, and others, "Holistic evaluation of language models," *Transactions on Machine Learning Research*, 2023. https://openreview.net/forum?id=iO4LZibEqW

[15] A. S. Luccioni, Y. Jernite, and E. Strubell, "Power hungry processing: watts driving the cost of AI deployment?" *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 85 to 99, 2024. https://doi.org/10.1145/3630106.3658542

[16] MLCommons, "MLPerf inference power measurement," 2026. https://docs.mlcommons.org/inference/power/

[17] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023. https://doi.org/10.6028/NIST.AI.100-1

[18] National Institute of Standards and Technology, "AI test, evaluation, validation and verification," 2026. Available at the NIST TEVV program page.

[19] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, vol. 381, no. 6654, pp. 187 to 192, 2023. https://doi.org/10.1126/science.adh2586

[20] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688 to 701, 1974. https://doi.org/10.1037/h0037350

[21] P. Rust, J. Pfeiffer, I. Vulić, S. Ruder, and I. Gurevych, "How good is your tokenizer? On the monolingual performance of multilingual language models," *Proceedings of ACL-IJCNLP 2021*, pp. 3118 to 3135, 2021. https://doi.org/10.18653/v1/2021.acl-long.243

[22] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *Proceedings of ACL 2016*, pp. 1715 to 1725, 2016. https://doi.org/10.18653/v1/P16-1162