The Cognitive Rebound Effect:
Why Efficient AI May Increase Energy Demand

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

23 July 2026

Abstract

An efficient inference system uses fewer joules for a fixed cognitive service. It does not follow that the system, firm, or economy will use less energy. Lower cost, shorter latency, wider access, new tasks, and changes in task mix can expand demand. This paper states the narrow result that can be proved and separates it from empirical questions that remain open.

Let Q_{cog} denote quality-adjusted cognitive service under a fixed task and outcome protocol, and let η denote units of that service per joule within a declared energy boundary. Since

$$E_{\text{AI}} = \frac{Q_{\text{cog}}}{\eta},$$

logarithmic differentiation gives

$$\frac{d\ln E_{\text{AI}}}{d\ln\eta} = \frac{d\ln Q_{\text{cog}}}{d\ln\eta} - 1.$$

Thus direct AI energy rises after an efficiency improvement if and only if the elasticity of consistently measured service demand with respect to efficiency exceeds one. This threshold is PROVED as an identity. Its empirical size is OPEN. The result is not an economy-wide theorem, and it fails as a comparison if service quality, energy scope, population, or horizon silently changes.

We distinguish direct, indirect, and economy-wide rebound. Within direct service demand, we separate intensive use, extensive adoption, and task composition. We also isolate cost and price pass-through because a hardware gain that is retained as margin differs from one passed to users through lower prices, higher quotas, or shorter queues. A proposed stepped rollout combines randomized serving-stack assignment with synchronized wall-energy, quality, price, latency, and usage records. Cohort-aware event studies, pretrend equivalence bands, shifted-date placebos, unaffected-outcome placebos, and predeclared heterogeneity checks diagnose the design. They do not identify economy-wide rebound.

A dependency-light Python module implements the identity, finite-change rebound, demand and energy decompositions, pass-through, not-yet-treated event studies, pretrend and placebo diagnostics, heterogeneity summaries, and deterministic simulations. Tests establish COMPUTATIONAL agreement between equations and fixtures. Efficient AI need not increase energy demand. It need not reduce it either.

Keywords: rebound effect; artificial intelligence; energy efficiency; demand elasticity; event study; pass-through; data centers

Nonclaims. This paper does not estimate an AI rebound magnitude. It does not infer economy-wide rebound from provider electricity growth. It does not treat requests, tokens, or revenue as quality-adjusted cognitive service. It does not claim that efficiency policy is futile, or that efficient AI necessarily raises total energy demand.

1 Introduction

Two observations about AI energy can both be true. A new accelerator, compiler, model architecture, or serving policy can reduce joules per completed task. At the same time, total electricity used for AI can rise. The first observation holds service fixed. The second reflects realized service volume, task mix, prices, capacity, and adoption.

Arguments about this pattern often skip the middle. One side points to engineering efficiency and predicts falling energy. Another invokes Jevons and predicts explosive demand. Neither move is sufficient. An efficiency gain is a technical fact under a test protocol. A rebound magnitude is a behavioral and economic response relative to a counterfactual. Backfire, in which total energy rises because of the efficiency change, is a stronger causal claim still.

The distinction has a long history. Jevons examined coal use in an expanding industrial economy [9]. Khazzoom showed why appliance standards cannot assume use stays fixed [10]. Borenstein decomposed microeconomic rebound into income and substitution channels and emphasized prices, capital costs, and marginal-cost pricing [1]. Lemoine showed that general-equilibrium channels can amplify or dampen savings depending on sector structure, substitution, and the energy-supply response [12]. Brockway, Sorrell, Semieniuk, Heun, and Court reviewed an economy-wide evidence base whose methods and assumptions remain diverse [2].

AI changes the application, not the logic. A cheaper cognitive service may be used more intensively by existing users. New users may enter. Tasks may shift toward longer contexts, higher reasoning budgets, agents, multimodal inputs, or continuous monitoring. Firms may embed inference in products that previously used no model. Savings may be retained by providers, passed through in prices, or converted into lower latency and higher quotas. Each path has different data requirements.

This paper makes four contributions.

1. It proves the direct elasticity threshold under a fixed declared scope and gives its exact finite-change counterpart.
2. It separates intensive, extensive, composition, pass-through, indirect, and economy-wide mechanisms without adding overlapping energy.
3. It proposes an empirical design that can identify direct rebound for a named serving intervention while stating why broader rebound remains open.
4. It supplies executable reference code and adversarial tests for the identities, decompositions, diagnostics, and simulations.

The contribution is not a new claim that demand responds to cost. It is a boundary-explicit translation of rebound economics into quality-adjusted AI service, paired with an empirical contract that can fail visibly.

2 Literature and novelty boundary

2.1 From Jevons to modern microeconomics

Jevons argued that economical coal use could enlarge the set of profitable uses for coal [9]. The historical claim should not be turned into a universal law. Modern rebound analysis asks a counterfactual question: how much of the engineering energy saving is offset by behavioral and economic responses caused by the efficiency improvement?

Khazzoom focused on mandated appliance efficiency and the error in projecting savings mechanically from engineering efficiency while holding utilization fixed [10]. Saunders connected the broader Khazzoom-Brookes proposition to neoclassical growth models [14]. These foundations motivate the question but do not give a portable rebound number for AI.

Borenstein's framework is especially useful here [1]. An efficiency upgrade changes the effective price of an energy service, may create disposable income, and induces substitution across goods. The energy content of displaced expenditure matters. So do upgrade costs and the gap between price and marginal social cost. A provider discount financed by lower compute cost is not equivalent to a quota increase with a fixed subscription price.

Chan and Gillingham show that common elasticity shortcuts can be biased when complements, substitutes, and welfare channels are not modeled correctly [4]. Gillingham, Rapson, and Wagner likewise stress that direct, indirect, and macroeconomic rebound answer different questions [6]. This paper keeps that separation.

2.2 Direct empirical evidence

Sorrell, Dimitropoulos, and Sommerville review direct rebound estimates for household energy services and explain how endogenous efficiency choice, measurement error, capital cost, and the substitution of energy-price elasticities for efficiency elasticities can bias estimates [16]. Their numerical findings concern transport, heating, and other household services. They are not priors that can simply be assigned to AI inference.

For AI, both the service and its price are difficult to observe. A request can contain a trivial completion or a difficult verified task. Token prices do not fully reflect subscriptions, volume discounts, latency, user time, or quotas. Quality and capability can change at the same moment as efficiency. These features make a design that isolates efficiency more valuable than a large observational panel with an ambiguous treatment.

2.3 General equilibrium

Lemoine develops a multisector framework in which sector prices, non-energy inputs, consumption substitution, and the energy-supply sector transmit an efficiency innovation [12]. The calibrated results show both amplifying and dampening channels. The direction is not fixed by the word "general equilibrium."

Brockway and coauthors review economy-wide estimates from computable general equilibrium models and other approaches [2]. They conclude that large erosion of engineering savings is plausible in many studies, while also documenting wide variation in model structure, mechanisms, assumptions, and estimated magnitudes. Their review is a warning against omitting macroeconomic channels. It is not an empirical estimate of cognitive rebound.

2.4 AI electricity context

The IEA reports that data centers used about 415 TWh globally in 2024 and presents scenarios in which demand grows substantially through 2030 [7]. Its scenarios vary efficiency, adoption, and deployment conditions. The IEA's 2026 update again identifies efficiency, uptake, and new capabilities as distinct uncertain drivers [8]. DOE and Lawrence Berkeley National Laboratory report rapid growth and a wide projected range for United States data-center electricity through 2028 [5, 11].

These sources establish policy relevance and measurement uncertainty. They do not identify rebound. Observed electricity growth combines baseline digital demand, model capability, new investment, prices, macroeconomic conditions, and technical efficiency. A rebound estimate needs the energy path that would have occurred without the efficiency change.

3 Measurement contract

3.1 The tuple

Every result is indexed by

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

where b is system boundary, c counterfactual, W outcome functional, τ horizon, e energy convention, a attribution rule, and π uncertainty model. For rebound, the tuple must also hold the cognitive-service definition fixed.

The minimum manifest is:

1. quality-adjusted service and functional unit;
2. model, software, hardware, precision, and serving policy;
3. physical energy boundary;
4. population and assignment unit;
5. start, end, and adaptation horizon;
6. geographic and grid location;
7. counterfactual serving stack and demand path;
8. shared infrastructure and idle allocation;
9. failed requests, retries, and rework;
10. embodied-energy treatment;
11. price, quota, latency, and pass-through policy;
12. uncertainty and missing-data model.

3.2 Quality-adjusted cognitive service

Let tasks belong to strata k. A service unit can be written

$$Q_{\text{cog}} = \sum_k w_k \sum_{i \in k} \mathbf{1}\{\text{quality}_i \geq q_k^\star\}\mathbf{1}\{\text{latency}_i \leq L_k^\star\}\mathbf{1}\{\text{no terminal failure}_i\}.$$

The weights w_k, thresholds $q_k^\star$, latency limits $L_k^\star$, and task distribution are declared before comparison. More tokens do not automatically create more Q_{cog}. Retries consume energy but add service only if the final task satisfies the protocol.

The metric can be specialized. A coding assistant study might use accepted tests of preregistered repository tasks. A support study might use quality-constrained resolved cases. A medical setting would need much stronger safety and outcome rules. The symbol does not make these services interchangeable.

3.3 Efficiency

Define

$$\eta = \frac{Q_{\text{cog}}}{E_{\text{AI}}},$$

where E_{AI} is non-overlapping energy inside boundary b. For a deployed service, facility operational electricity is often preferable to accelerator telemetry. Device energy can still support engineering analysis if it is named as such.

Efficiency can improve because hardware uses less energy, software performs less work, batching improves utilization, failed work falls, or quality rises at fixed energy. These interventions need not have the same demand response. A faster service can relax a latency constraint even when its posted price does not change.

4 The threshold identity

Assumption 4.1 (Fixed comparison)**.** Across the derivative, Q_{cog} retains the same task protocol, quality weights, population, and accounting rule; E_{AI} retains the same energy boundary and allocation; and $\eta > 0$.

Theorem 4.2 (Cognitive rebound threshold)**.** *Under assumption 4.1, if*

$$E_{\text{AI}} = \frac{Q_{\text{cog}}}{\eta},$$

then

$$\frac{d \ln E_{\text{AI}}}{d \ln \eta} = \epsilon_{Q,\eta} - 1, \qquad \epsilon_{Q,\eta} = \frac{d \ln Q_{\text{cog}}}{d \ln \eta}.$$

For an efficiency improvement, direct AI energy rises locally if and only if $\epsilon_{Q,\eta} > 1$, is locally unchanged if $\epsilon_{Q,\eta} = 1$, and falls locally if $\epsilon_{Q,\eta} < 1$.

Proof. Taking logarithms gives

$$\ln E_{\text{AI}} = \ln Q_{\text{cog}} - \ln \eta.$$

Differentiate both sides with respect to $\ln \eta$. The three sign cases follow immediately. □

The theorem is PROVED. It is an accounting identity under a fixed definition, not a behavioral model. It says where the threshold lies. It does not say which side of the threshold an AI service occupies.

4.1 Why quality adjustment matters

Suppose raw requests double after a serving change, but the added requests are mostly failures or low-value duplicates. A request-count elasticity can exceed one while the Q_{cog} elasticity does not. Conversely, a model may solve harder tasks with fewer requests. Raw volume can understate service expansion.

The protocol must not be rewritten after the efficiency change to make the threshold come out one way. If quality weights change, the comparison is a joint change in measurement and behavior. It can still be studied, but it is not theorem 4.2 under a fixed Q.

4.2 Why scope matters

If E_{AI} initially means accelerator energy and later means facility electricity, the difference includes a boundary change. If the initial sample contains one product and the later sample contains the whole firm, the population changed. If a monthly comparison is placed next to a five-year forecast, the horizon changed. The logarithmic algebra remains true within each state, but the derivative no longer has the claimed empirical meaning.

5 Finite changes and rebound measures

Derivatives are useful for a threshold. Real deployments make discrete changes. Let state 0 precede an efficiency improvement and state 1 follow it.

Proposition 5.1 (Finite-change threshold)**.** *Under the same fixed comparison,*

$$\frac{E_{\mathrm{AI},1}}{E_{\mathrm{AI},0}} = \frac{Q_{\mathrm{cog},1}/Q_{\mathrm{cog},0}}{\eta_1/\eta_0}.$$

When $\eta_1 > \eta_0$, energy rises if and only if

$$\frac{Q_{\mathrm{cog},1}}{Q_{\mathrm{cog},0}} > \frac{\eta_1}{\eta_0}.$$

Proof. Divide $E_{\mathrm{AI},1} = Q_{\mathrm{cog},1}/\eta_1$ by $E_{\mathrm{AI},0} = Q_{\mathrm{cog},0}/\eta_0$, then compare the ratio with one. □

Define the fixed-service engineering counterfactual

$$E^{\mathrm{eng}}_{\mathrm{AI},1} = \frac{Q_{\mathrm{cog},0}}{\eta_1}.$$

Engineering savings are

$$S_{\mathrm{eng}} = E_{\mathrm{AI},0} - E^{\mathrm{eng}}_{\mathrm{AI},1},$$

while actual savings are

$$S_{\mathrm{act}} = E_{\mathrm{AI},0} - E_{\mathrm{AI},1}.$$

For $S_{\mathrm{eng}} > 0$, finite direct rebound is

$$R_{\mathrm{direct}} = 1 - \frac{S_{\mathrm{act}}}{S_{\mathrm{eng}}} = \frac{E_{\mathrm{AI},1} - E^{\mathrm{eng}}_{\mathrm{AI},1}}{E_{\mathrm{AI},0} - E^{\mathrm{eng}}_{\mathrm{AI},1}}.$$

Label	Condition	Meaning within the declared direct scope
Superconservation	$R < 0$	Actual saving exceeds fixed-service engineering saving.
Partial rebound	$0 < R < 1$	Demand erodes some engineering saving.
Full offset	$R = 1$	Actual energy returns to its initial level.
Backfire	$R > 1$	Actual energy exceeds its initial level.

Table 1: Finite rebound labels are statements about a named counterfactual.

A percentage is not meaningful without the engineering baseline. Model capability changes, new products, and population growth can raise both service and energy for reasons unrelated to the efficiency intervention. Assigning all of that growth to rebound exaggerates the causal response.

6 Demand margins

6.1 Intensive use

The intensive margin asks how much existing users consume. Examples include more requests per account, longer operating hours, more agent steps, more frequent monitoring, or higher reasoning budgets. The metric should be quality-adjusted. A rise in retries caused by worse reliability is energy use, not useful intensive service.

6.2 Extensive adoption

The extensive margin counts new users, firms, products, or tasks entering the service. Lower price can expand adoption. So can lower latency, easier integration, higher reliability, or a larger capacity allocation. A free service still has extensive rebound because user time, queueing, rate limits, and integration costs create shadow prices.

6.3 Composition

Efficiency can change the mix of tasks. Let

$$Q_{\text{cog}} = N\bar{x}m,$$

where N is the number of active service units, $\bar{x}$ is raw task volume per active unit, and m is quality-adjusted service per raw task under fixed task weights and outcome rules. Thus $N\bar{x}$ has units of raw tasks and m converts raw tasks into quality-adjusted service. Then

$$\epsilon_{Q,\eta} = \underbrace{\frac{d\ln\bar{x}}{d\ln\eta}}_{\epsilon_{\text{int}}} + \underbrace{\frac{d\ln N}{d\ln\eta}}_{\epsilon_{\text{ext}}} + \underbrace{\frac{d\ln m}{d\ln\eta}}_{\epsilon_{\text{comp}}}.$$

The direct energy elasticity is therefore

$$\epsilon_{E,\eta}^{\text{direct}} = \epsilon_{\text{int}} + \epsilon_{\text{ext}} + \epsilon_{\text{comp}} - 1.$$

This decomposition is exact when the multiplicative index is the declared measurement model. It is not a license to tune m after seeing the data. Report component levels alongside the aggregate.

6.4 Capacity and queueing

Inference demand can be rationed by capacity. An efficiency gain may first appear as lower queue time rather than a lower bill. Shorter queues can make latency-sensitive applications feasible. If the study observes price but not latency, it misses this pass-through channel. If it observes submitted requests but not rejected or throttled attempts, it confuses desired demand with served demand.

7 Cost and price pass-through

Let c be provider marginal cost per quality-adjusted service and p the user-facing generalized price. The generalized price can include money, latency, quotas, integration labor, and expected failure cost. Define

$$\rho_{c,\eta} = -\frac{d\ln c}{d\ln \eta}, \qquad \pi_{p,c} = \frac{d\ln p}{d\ln c}.$$

The efficiency-to-price pass-through is

$$\rho_{p,\eta} = -\frac{d\ln p}{d\ln \eta} = \rho_{c,\eta}\pi_{p,c}$$

when the chain rule applies and other price determinants are fixed.

Using the conventional signed demand-price elasticity

$$\epsilon_{Q,p} = \frac{d\ln Q_{\text{cog}}}{d\ln p},$$

the price-mediated service elasticity is

$$\epsilon_{Q,\eta}^{\text{price}} = -\rho_{p,\eta}\epsilon_{Q,p}.$$

With ordinary downward-sloping demand, $\epsilon_{Q,p} < 0$, so positive pass-through raises service demand.

These scalar derivatives require a scalar p with fixed weights across its money, latency, quota, failure, and integration components. When no defensible fixed-weight index exists, the price remains a vector and each pass-through channel is estimated separately. A changed index is not a price response.

The equations do not imply full pass-through. A provider can retain savings as margin, use them to buy more capacity, improve latency, raise free-tier quotas, or change product quality. Competition, contracts, market power, capacity scarcity, and fixed-cost recovery all matter. Borenstein's analysis of prices that differ from marginal cost is directly relevant [1].

7.1 A factorial test

A clean design can separate technical efficiency from monetary pass-through. Randomize eligible tenant clusters to the serving-stack improvement. Within technical treatment and control, independently randomize a predeclared price credit or quota increase. The technical treatment identifies the effect of the stack under the existing product policy. The price or quota treatment identifies a user-facing pass-through channel. Their interaction tests whether technical capacity changes the response to price.

This design still has limits. A temporary credit may not mimic a permanent price change. Users can learn. Quotas may bind only for some accounts. Spillovers can occur when teams share outputs across clusters. Each limitation belongs in the estimand, not in a footnote after the result.

8 Direct, indirect, and economy-wide rebound

8.1 Direct AI energy

Direct rebound concerns energy used to provide the focal cognitive service inside the declared AI boundary. It includes initial and retry inference, host and memory, network, facility overhead, allocated idle energy, and covered rework only when those components are non-overlapping.

8.2 Indirect energy

Indirect rebound covers energy changes outside the focal service caused by the efficiency improvement. Examples include client devices, additional network traffic, human review tooling, newly induced complementary software, and goods or services purchased with cost savings. A component is not "indirect" merely because it is hard to meter. If the facility wall meter already captures networking or host energy, adding an estimate again is double counting.

8.3 Economy-wide response

Let

$$E_\Omega = E_{\mathrm{AI}} + E_{\mathrm{other}},$$

where the two terms are non-overlapping under an economy-wide energy convention. Let $s_{\mathrm{AI}} = E_{\mathrm{AI}}/E_\Omega$. Then the local elasticity is

$$\frac{d \ln E_\Omega}{d \ln \eta} = s_{\mathrm{AI}}(\epsilon_{Q,\eta} - 1) + (1 - s_{\mathrm{AI}})\frac{d \ln E_{\mathrm{other}}}{d \ln \eta}.$$

Proposition 8.1 (Economy-wide threshold is not the direct threshold)**.** *The condition $\epsilon_{Q,\eta} > 1$ is sufficient and necessary for direct AI backfire under assumption 4.1. It is neither necessary nor sufficient for economy-wide backfire unless the response of E_{other} is restricted.*

Proof. The direct claim follows from theorem 4.2. The economy-wide derivative contains the additional weighted term for E_{other}. A sufficiently positive other-energy response can make total energy rise when direct AI energy falls. A sufficiently negative response can make total energy fall when direct AI energy rises. □

Lemoine's framework supplies concrete reasons for the added term: consumption-good prices, demands for non-energy inputs, and changes in the energy-supply sector [12]. The Brockway et al. review catalogs further macroeconomic mechanisms and shows that model coverage varies [2]. Calling direct compute growth "economy-wide rebound" would erase those mechanisms rather than estimate them.

9 Why AI is a difficult rebound application

9.1 Capability and efficiency move together

A new model can use fewer joules for one benchmark while unlocking tasks the old model could not perform. If rollout changes both capability and efficiency, demand growth cannot be assigned to efficiency alone without a design that separates them. Holding the model, prompt distribution, quality rule, and output constraint fixed is therefore valuable for the primary technical treatment.

9.2 Prices are nonlinear

AI services combine subscriptions, token charges, reserved capacity, free tiers, minimum commitments, discounts, and internal transfer prices. Posted token price is not the generalized price of a verified task. The empirical record should include invoices, credits, quotas, queue time, and failure risk.

9.3 Supply can bind

Observed use is the minimum of desired demand and available service. Capacity constraints, rate limits, regional availability, and power interconnection can mask the demand response. An efficiency gain that expands capacity can reveal previously rationed demand. That is a real channel, but a naive price elasticity will miss it.

9.4 Service is heterogeneous

Inference covers translation, code, search, image generation, control, monitoring, customer support, scientific workloads, and entertainment. A single request count gives each the same weight. A quality-adjusted service index makes weights explicit but remains task-distribution specific.

9.5 Training and inference interact

Cheaper inference can change demand for model training, fine-tuning, evaluation, and synthetic data. Training improvements can change inference demand. A study limited to inference should name these exclusions. A broader study must meter them without treating transferred computation as an energy saving.

10 A credible direct-rebound experiment

10.1 Treatment

The proposed treatment is a serving-stack change that raises verified service per facility joule while holding model weights, quality protocol, user interface, posted price, context policy, and maximum output constant. Examples could include a compiler improvement, kernel change, scheduling policy, or hardware substitution whose output equivalence is tested.

Eligible tenant clusters are assigned a rollout date by lottery within region, baseline demand, and service-tier strata. A persistent stepped rollout is used only if rollback is infeasible. If rollback is safe, a shorter randomized cross-over can strengthen within-cluster comparisons, provided carryover and learning are controlled.

10.2 Unit and horizon

The assignment unit is a tenant cluster large enough to contain team spillovers. Outcomes are recorded at fixed daily or weekly intervals. The study includes a baseline long enough to inspect trends and a post period long enough to capture initial demand response without claiming a permanent equilibrium.

10.3 Primary outcomes

For cluster i, period t, record:

$$\begin{aligned}
Q_{it} &= \text{quality-adjusted completed service,} \\
\eta_{it} &= Q_{it}/E_{it}^{\text{fac,alloc}}, \\
E_{it}^{\text{fac,alloc}} &= \text{facility operational energy allocated to the focal service,} \\
N_{it} &= \text{active users or active workflows,} \\
\bar{x}_{it} &= \text{raw task volume per active unit,} \\
m_{it} &= \text{quality-adjusted service per raw task.}
\end{aligned}$$

Secondary outcomes include generalized price, latency, throttling, failed requests, retries, rework, adoption, and user time. Treatment integrity checks verify model and quality equivalence.

The allocation in $E_{it}^{\text{fac,alloc}}$ is fixed before assignment. It excludes training, legacy stacks, and other tenant workloads except for a declared share of common idle and facility load. A total facility meter can anchor the allocation but cannot be inserted wholesale into the focal denominator when unrelated loads share the site.

10.4 Why randomization matters

Providers often deploy efficiency improvements first where demand is high, hardware is new, or operators expect a benefit. Those same conditions predict future energy. Random rollout timing removes that selection channel in expectation. It does not solve measurement error, interference, attrition, or general-equilibrium identification.

11 Estimands and event-study structure

Let G_i be cluster i's randomized adoption period. For an outcome $Y \in \{\ln Q, \ln \eta, \ln E, \ln p\}$, define a horizon-$h$ cohort effect:

$$\tau_Y(h) = \mathbb{E}\left[Y_{i,G_i+h}(1) - Y_{i,G_i+h}(0)\right].$$

A not-yet-treated comparison estimates cohort-specific changes without using already treated clusters as controls and normalizes the estimator to event time -1. This avoids the contamination that can arise in conventional two-way fixed-effect event studies with staggered timing and heterogeneous effects [17, 3].

The direct service-response ratio is

$$\widehat{\epsilon}_{Q,\eta}(h) = \frac{\widehat{\tau}_{\ln Q}(h)}{\widehat{\tau}_{\ln \eta}(h)}$$

when the efficiency first stage is nonzero. Report both effects and their joint uncertainty. The ratio can be unstable when the denominator is weak. The ratio is secondary to the directly observed energy effect $\widehat{\tau}_{\ln E}(h)$.

Algebraically, this is a Wald ratio with randomized rollout as an instrument for efficiency. A structural demand-elasticity interpretation requires exclusion of rollout effects on Q other than through the declared efficiency change, a stable treatment version, and an appropriate monotonicity condition under heterogeneous first stages. A serving change that also alters latency, capability, or user interface violates exclusion. Without those assumptions, the ratio is a policy-specific rollout response, not a general demand elasticity. The first stage should be reported for a fixed service basket as well as for realized $\eta = Q/E$, whose numerator contains the outcome Q.

11.1 Identity audit

If each observation uses $E = Q/\eta$ with matched scope, then

$$\tau_{\ln E}(h) = \tau_{\ln Q}(h) - \tau_{\ln \eta}(h).$$

Failure of this equality beyond numerical tolerance reveals mismatched aggregation, measurement error, or a boundary change. Passing it confirms accounting consistency, not causal identification.

11.2 Weights

Tenant-weighted, user-weighted, task-weighted, and energy-weighted estimands are different. The primary weight is chosen before rollout. Equal weighting of tenant clusters answers a different question from summing all service and energy. Both can be reported if labeled.

11.3 Inference

Uncertainty should respect randomization strata and assignment clusters. Simultaneous event-time intervals are preferable to a collection of pointwise claims. For staggered observational adoption, use a cohort-aware estimator and state conditional parallel trends. A default two-way fixed-effect regression is not a substitute for examining timing and treatment heterogeneity.

12 Diagnostics

12.1 Pretrends

Plot cohort-aware lead coefficients for service, efficiency, energy, price, latency, and task composition. A useful diagnostic declares equivalence bands before analysis. Ask whether leads are small enough to be substantively compatible with the design rather than relying on whether a low-powered test fails to reject zero.

Pretrend tests do not prove parallel trends. Conditioning publication on a non-significant pretest can distort the subsequent estimate [13]. Randomized timing gives a design basis for identification; lead diagnostics can still reveal implementation leakage, anticipation, data misalignment, or failed randomization.

12.2 Placebos

The first placebo shifts adoption dates earlier by a fixed number of periods. An apparent effect before access suggests anticipation, trend imbalance, or a coding error. The second uses unaffected outcomes, such as energy for an isolated legacy service that does not share the treated stack. The third uses an efficiency change that passed engineering tests but was disabled before traffic reached it.

A placebo need not be exactly zero in a finite sample. It should be interpreted with its assignment-aware uncertainty and a predeclared materiality threshold.

12.3 Heterogeneity

Predeclare estimates by baseline quota binding, price schedule, latency sensitivity, task family, firm size, region, and baseline service intensity. Separate intensive and extensive effects. Heterogeneity discovered after many searches is descriptive until replicated.

12.4 Attrition and migration

Tenant migration across clusters can break assignment. Preserve intention-to-treat records and report migration. If a cluster leaves the platform, its disappearance is an outcome, not a reason to delete its pretreatment observations. Missing energy intervals should be flagged rather than imputed from accelerator TDP.

13 Pass-through and composition diagnostics

13.1 Pass-through table

For each cluster-period, record a price vector rather than a single token price:

$$p_{it} = (p_{it}^{\text{money}}, p_{it}^{\text{latency}}, p_{it}^{\text{quota}}, p_{it}^{\text{failure}}, p_{it}^{\text{integration}}).$$

The elements need not be collapsed into dollars. A vector preserves the mechanism. If a scalar generalized price is constructed, publish its weights and sensitivity.

13.2 Composition stability

Report treatment effects on predeclared task shares. If the treated stack attracts new task types, a fixed-mix efficiency benchmark and the realized-mix efficiency answer different questions. Both are useful:

- fixed mix isolates engineering efficiency for a common service basket;
- realized mix measures the deployed system after demand responds.

The gap is a composition effect. It should not be hidden inside an average joules-per-request statistic.

13.3 Intensive and extensive records

Active accounts, new workflows, and task entry identify the extensive margin. Service per active account, agent steps per workflow, and operating hours identify the intensive margin. A firm may show little request growth but a large extensive effect if many low-volume users enter. Another may show the reverse.

14 Indirect and economy-wide research design

The provider experiment identifies a local direct effect. A broader study needs linked data and a separate estimand.

14.1 Indirect digital energy

Client devices and networks can be sampled for a consented subset. Metering must distinguish energy already captured by the provider boundary. If more efficient inference moves work from server to client, a provider-only saving can be a transfer.

14.2 Expenditure and production channels

Cost savings can change expenditure on other goods. Firms can change labor, capital, and product output. Energy suppliers can change prices and capacity. Borenstein's income and substitution decomposition and Lemoine's sectoral model show why these paths depend on prices, cost shares, and substitution [1, 12].

An input-output extension can trace embodied and purchased energy under fixed coefficients. It cannot identify price, substitution, innovation, or growth responses. A computable general equilibrium model can represent more channels but adds functional-form and calibration assumptions. Neither method turns the direct experiment into an economy-wide natural experiment.

14.3 Economy-wide status

For current AI, the empirical magnitude of economy-wide rebound is OPEN. Scenario analysis can vary demand elasticities, pass-through, sector substitution, and energy supply. It should not attach causal language to a scenario because it reproduces recent electricity growth.

15 Constructed examples

15.1 Threshold paths

Begin with $Q_0 = 100$ quality-adjusted tasks and $\eta_0 = 2$ tasks per joule, so $E_0 = 50$ joules. Efficiency rises by 25%.

Service elasticity	Q_1	η_1	E_1	Regime
0.5	$100(1.25)^{0.5}$	2.5	below 50	energy saving
1.0	125	2.5	50	full offset
1.5	$100(1.25)^{1.5}$	2.5	above 50	backfire

Table 2: Constructed direct responses under a fixed service definition.

The example is COMPUTATIONAL. It demonstrates the identity, not an estimate for a provider.

15.2 Demand decomposition

Suppose the intensive elasticity is 0.25, the extensive elasticity is 0.50, and composition contributes -0.10. The service elasticity is 0.65, so direct energy elasticity is -0.35. Adoption erodes much of the engineering saving, but direct energy still falls locally.

15.3 Economy-wide reversal

Suppose direct AI energy falls by 10 joules. Indirect digital and induced expenditure energy rise by 10, and other equilibrium responses add 5. The economy-wide change is +5 even though direct energy falls. Reversing the signs can produce economy-wide saving even when direct AI energy rises. The example proves that scopes cannot be substituted.

15.4 Pass-through

If efficiency doubles, marginal cost halves, user price halves, and service rises from 100 to 150, cost and price pass-through with respect to efficiency are each one in logs. The signed demand-price elasticity is negative. The price-mediated efficiency elasticity is

$$\frac{\ln(1.5)}{\ln 2}.$$

If user price stays fixed, the two-state data do not identify a demand-price elasticity. Service growth must be attributed, if at all, through capacity, latency, capability, or another channel supported by the design.

16 Reference implementation

The module `src/joule_standard/rebound.py` uses only the Python standard library. Its public components are:

1. a scope manifest and strict numeric validation;
2. $E = Q/\eta$, log changes, elasticity, and regime classification;
3. finite engineering and actual savings with rebound labels;
4. intensive, extensive, and composition elasticities;
5. cost, user-price, and price-mediated demand pass-through;
6. a non-overlapping direct, indirect, and general-equilibrium ledger;
7. panel records and not-yet-treated event-time contrasts;
8. pretrend, shifted-adoption placebo, and heterogeneity diagnostics;
9. deterministic constant-elasticity simulations.

The event-study routine is intentionally transparent. Each treated unit is differenced against its event-time reference and against units untreated at both dates. It is a diagnostic implementation, not a general inference package. A publication analysis still needs assignment-aware standard errors, simultaneous intervals, and a design appropriate to timing.

16.1 Test contract

Tests use generic constructed panels and algebraic invariants. They check:

- energy and logarithmic identities;
- regimes below, at, and above the unit elasticity threshold;
- finite partial rebound, backfire, and superconservation;
- additive demand margins and non-overlapping energy channels;
- pass-through signs and unchanged-price non-identification;
- zero pretrends and shifted-date placebos in a balanced rollout;
- heterogeneous effects by declared segment;
- simulated energy directions for three elasticities;
- rejection of invalid levels and duplicate panel records.

Passing tests support COMPUTATIONAL claims about the reference code. They do not establish that an actual rollout was randomized, that its meter was calibrated, or that its quality index measured welfare.

17 Interpreting official energy scenarios

The IEA's 2025 report gives an observed 2024 global data-center electricity estimate and scenario paths through 2030 and 2035 [7]. Its High Efficiency Case shows lower energy for a stated service path than its Base Case. Its higher-adoption cases show greater electricity use. This is exactly why efficiency and demand must be reported separately.

The IEA's 2026 discussion emphasizes three evolving drivers: efficiency, uptake, and changing capabilities [8]. The categories align with this paper's decomposition, but the report is not a randomized rebound study.

DOE's release of the LBNL United States report gives a wide 2028 range [5, 11]. The width is informative. Hardware shipments, utilization, data-center type, cooling, and AI adoption all contribute. Selecting one point from the range and calling the residual "rebound" would replace a counterfactual with a label.

Official scenarios are valuable for grid planning. They describe possible loads and sensitivities. Causal rebound analysis asks a different question: how would energy have changed in the same population and horizon absent a specific efficiency improvement?

18 Policy interpretation

An energy-saving technical improvement can raise welfare even when rebound is positive. More cognitive service may create useful outcomes. Conversely, energy can fall while welfare worsens if quality, access, or reliability deteriorates. Rebound is an energy quantity relative to a counterfactual, not a welfare verdict.

Efficiency policy and energy policy therefore address different margins. Efficiency reduces energy per fixed service. Prices, quotas, carbon policy, capacity rules, procurement, and product design influence total demand and its external costs. Whether a particular combination is desirable depends on benefits, costs, distribution, market power, and grid conditions.

Borenstein separates quantity accounting from welfare and shows why retail prices, marginal costs, and transfers matter [1]. Chan and Gillingham derive welfare implications under a broader microeconomic model [4]. This paper does not reproduce those welfare models. It prevents the threshold identity from being mistaken for one.

18.1 No policy from a slogan

"Efficiency will solve the energy problem" assumes demand response is small. "Jevons makes efficiency useless" assumes backfire and ignores benefits, prices, and policy. Neither statement follows from the evidence assembled here. A serious policy analysis reports technical efficiency, service demand, energy, and welfare components separately.

19 Claim status and falsifiers

Status	Claim	Falsifier or limit
PROVED	Under fixed scope, direct energy elasticity equals service-demand elasticity minus one.	Q, η, or energy boundary changes definition.
PROVED	Finite energy rises exactly when service grows proportionally more than efficiency.	Initial and final states use incompatible service or energy accounting.
PROVED	Direct and economy-wide thresholds can differ.	Other-energy response is restricted to zero by the declared scope.
COMPUTATIONAL	Reference code reproduces identities, diagnostics, and threshold paths.	A generic invariant or validation test fails.
CONDITIONAL	Random rollout identifies direct dynamic effects for the assigned service.	Randomization, integrity, measurement, or interference assumptions fail.
OPEN	Magnitude of direct cognitive rebound in deployed AI.	Requires a valid field study with synchronized service and wall energy.
OPEN	Magnitude of economy-wide cognitive rebound.	Requires broader causal or structural evidence beyond the provider trial.

Table 3: Claims keep mathematical, computational, and empirical status separate.

The central claim corresponds to registry claim JS-C011. It is proposed PROVED as an identity and leaves empirical size OPEN.

20 Limitations

20.1 No field estimate

The paper provides a design, not a deployment result. Constructed panels show how diagnostics behave under known effects. They cannot support a claim about current providers, models, or total data-center electricity.

20.2 Quality aggregation

A quality-adjusted service index requires task definitions and weights. Fixed weights improve comparability but may miss new capabilities. Changing weights captures a new question but breaks the fixed-service comparison. Publishing the task vector reduces dependence on one index.

20.3 Generalized price

Money, latency, quota, failure risk, privacy, and user time do not have a natural common scale. A scalar price can conceal mechanism changes. The factorial design identifies a particular monetary or quota intervention, not every shadow cost.

20.4 Interference

Tenant clusters can share outputs, workers, caches, and product demand. Capacity freed in one cluster can be reallocated to another. Cluster randomization reduces some contamination but does not eliminate platform-wide effects.

20.5 Long horizons

Short-run demand response can differ from product entry, capital investment, innovation, and economy-wide adaptation. Extrapolating a twelve-week rollout to a decade is a structural assumption. The IEA and DOE scenarios are useful for long-range planning, but they do not supply the missing causal bridge.

20.6 Energy stages

Facility operational electricity excludes embodied hardware and construction unless added through a non-overlapping lifecycle account. Grid emissions are not energy. Marginal and average grid effects answer different questions.

21 Conclusion

Efficiency fixes one side of an accounting identity. Demand determines the other. For a fixed definition of quality-adjusted cognitive service and a fixed energy boundary,

$$\frac{d \ln E_{\mathrm{AI}}}{d \ln \eta} = \epsilon_{Q,\eta} - 1.$$

This equation gives a clean threshold. It does not give the elasticity.

The empirical work begins where the identity ends. Existing users can consume more intensively. New users and tasks can enter. Task composition can change. Providers can pass cost savings through money, quotas, latency, or capacity. Indirect energy and general-equilibrium responses can reverse the sign of the direct effect.

A randomized stepped rollout of a serving-stack improvement can identify a named direct response if quality, price policy, population, and energy scope are held or measured carefully. Cohort-aware event studies, pretrend bands, placebos, heterogeneity, and accounting audits make failures visible. They do not turn a provider experiment into an economy-wide estimate.

The responsible conclusion is conditional. Efficient AI may increase energy demand. It may also reduce it. The direction and magnitude are empirical, scope-dependent questions, and the burden of proof belongs to the claimed scope.

A Proof details

A.1 Finite log elasticity

For two positive states, define

$$\widehat{\epsilon}_{Q,\eta} = \frac{\ln(Q_{\text{cog},1}/Q_{\text{cog},0})}{\ln(\eta_1/\eta_0)}.$$

Using proposition 5.1,

$$\frac{\ln(E_{\text{AI},1}/E_{\text{AI},0})}{\ln(\eta_1/\eta_0)} = \widehat{\epsilon}_{Q,\eta} - 1.$$

This is an exact two-point log-change identity. It is an arc elasticity, not a causal estimate unless the efficiency change is identified and definitions remain fixed.

A.2 Local rebound fraction

For a small positive $d\ln\eta$, fixed-service engineering energy changes by

$$d\ln E^{\text{eng}} = -d\ln\eta.$$

Actual energy changes by

$$d\ln E = (\epsilon_{Q,\eta} - 1)d\ln\eta.$$

The share of the engineering saving offset by service expansion is locally $\epsilon_{Q,\eta}$. This familiar equality is scope-specific. Indirect and economy-wide rebound add other responses and weights.

A.3 Share-weighted economy-wide derivative

For $E_\Omega = E_{\text{AI}} + E_{\text{other}}$,

$$\begin{aligned} d\ln E_\Omega &= \frac{dE_\Omega}{E_\Omega} \\ &= \frac{E_{\text{AI}}}{E_\Omega} d\ln E_{\text{AI}} + \frac{E_{\text{other}}}{E_\Omega} d\ln E_{\text{other}}. \end{aligned}$$

Divide by $d\ln\eta$ and substitute theorem 4.2. The result depends on baseline energy shares and the other-energy elasticity.

B Empirical data dictionary

C Adversarial audit checklist

1. Replace quality-adjusted service with raw requests after treatment.
2. Change from device to facility energy between periods.

Field	Unit	Rule
Assignment	cluster and date	Preserve original randomized date.
Service	quality-adjusted units	Fixed task weights, thresholds, and latency.
Allocated facility energy	joules or kWh	Focal-service allocation, synchronized meter, calibration, missingness flag.
Efficiency	service per joule	Derived only from matched service and energy.
Intensive use	raw tasks per active unit	Include failed demand separately.
Extensive use	active users or workflows	Fixed entry and activity definition.
Composition	task-share vector	Report vector before any index.
Generalized price	vector	Money, quota, latency, failure, integration.
Retries and rework	counts, time, energy	Keep outcome and energy roles separate.
Shared idle	joules	Publish allocation and sensitivity.

Table 4: Minimum records for the proposed direct study.

3. Drop failed, throttled, or abandoned requests.
4. Attribute a simultaneous capability release to efficiency.
5. Treat a posted token price as the complete user price.
6. Ignore queue time and quota pass-through.
7. Use already-treated units as controls in staggered rollout.
8. Interpret insignificant leads as proof of parallel trends.
9. Search heterogeneity until a backfire subgroup appears.
10. Add networking energy already captured by the wall meter.
11. Call direct provider energy economy-wide energy.
12. Infer rebound from a forecast without a no-efficiency counterfactual.

D Reproducibility manifest

- Source: `papers/latex/cognitive-rebound-effect.tex`.
- Module: `src/joule_standard/rebound.py`.
- Tests: `tests/test_rebound.py`.
- Runtime: Python standard library plus `pytest` for tests.
- Randomness: none in reference simulations or diagnostics.
- Numeric policy: positive levels for logs and explicit failure on weak or invalid comparisons.
- Evidence: algebraic proofs are PROVED; test fixtures are COMPUTATIONAL; empirical magnitudes are OPEN.

References

[1] S. Borenstein, "A microeconomic framework for evaluating energy efficiency rebound and some implications," *The Energy Journal*, vol. 36, no. 1, pp. 1 to 22, 2015. doi:10.5547/01956574.36.1.1

[2] P. E. Brockway, S. Sorrell, G. Semieniuk, M. K. Heun, and V. Court, "Energy efficiency and economy-wide rebound effects: A review of the evidence and its implications," *Renewable and Sustainable Energy Reviews*, vol. 141, article 110781, 2021. doi:10.1016/j.rser.2021.110781

[3] B. Callaway and P. H. C. Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of Econometrics*, vol. 225, no. 2, pp. 200 to 230, 2021. Journal DOI record

[4] N. W. Chan and K. Gillingham, "The microeconomic theory of the rebound effect and its welfare implications," *Journal of the Association of Environmental and Resource Economists*, vol. 2, no. 1, pp. 133 to 159, 2015. doi:10.1086/680256

[5] U.S. Department of Energy, "DOE releases new report evaluating increase in electricity demand from data centers," 20 December 2024. DOE release and report link

[6] K. Gillingham, D. Rapson, and G. Wagner, "The rebound effect and energy efficiency policy," *Review of Environmental Economics and Policy*, vol. 10, no. 1, pp. 68 to 88, 2016. doi:10.1093/reep/rev017

[7] International Energy Agency, *Energy and AI*, Paris, 2025. IEA report

[8] International Energy Agency, *Key Questions on Energy and AI*, Paris, 2026. IEA report

[9] W. S. Jevons, *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-Mines*. London: Macmillan, 1865.

[10] J. D. Khazzoom, "Economic implications of mandated efficiency standards for household appliances," *The Energy Journal*, vol. 1, no. 4, pp. 21 to 40, 1980. publisher DOI page

[11] A. Shehabi, S. J. Smith, A. Hubbard, A. Newkirk, N. Lei, M. A. B. Siddik, B. Holecek, J. Koomey, E. Masanet, and D. Sartor, *2024 United States Data Center Energy Usage Report*, Lawrence Berkeley National Laboratory, LBNL-2001637, 2024. LBNL report page

[12] D. Lemoine, "General equilibrium rebound from energy efficiency innovation," *European Economic Review*, vol. 125, article 103431, 2020. doi:10.1016/j.euroecorev.2020.103431

[13] J. Roth, "Pretest with caution: Event-study estimates after testing for parallel trends," *American Economic Review: Insights*, vol. 4, no. 3, pp. 305 to 322, 2022. doi:10.1257/aeri.20210236

[14] H. D. Saunders, "The Khazzoom-Brookes postulate and neoclassical growth," *The Energy Journal*, vol. 13, no. 4, pp. 131 to 148, 1992.

[15] S. Sorrell and J. Dimitropoulos, “The rebound effect: Microeconomic definitions, limitations and extensions,” *Ecological Economics*, vol. 65, no. 3, pp. 636 to 649, 2008. doi:10.1016/j.ecolecon.2007.08.013

[16] S. Sorrell, J. Dimitropoulos, and M. Sommerville, “Empirical estimates of the direct rebound effect: A review,” *Energy Policy*, vol. 37, no. 4, pp. 1356 to 1371, 2009. doi:10.1016/j.enpol.2008.11.026

[17] L. Sun and S. Abraham, “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,” *Journal of Econometrics*, vol. 225, no. 2, pp. 175 to 199, 2021. doi:10.1016/j.jeconom.2020.09.006