

Can We Test Jevons Paradox in AI Inference? A Longitudinal Audit of Token Prices, Demand, and Energy (2022–2026)

Eugen Spivak

publications@eugenspivak.com

Independent researcher <https://orcid.org/0009-0005-7509-8609>

Research Article

Keywords: AI inference, rebound effect, Jevons paradox, token prices, inference volume, energy reporting, measurement boundary, data provenance, web archives

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11425034/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare potential competing interests as follows: The author develops commercial data-compression software. No other competing interests.

Abstract

Claims that AI inference efficiency will limit, or instead raise, energy demand rest on evidence that has not been joined: a dated record of price, volume and energy for the same service. I built that record from public sources for 2022 to 2026: 1,392 list prices of text-output models from archived copies of eleven developers' own price pages at 19 quarter ends, 49 dated volume statements, and 122 energy figures tagged with a measurement-boundary class. I then developed the Inference Rebound Identifiability Framework (IRIF 1.0), which separates descriptive comparability, energy accounting (volume times energy per unit), causal identification of the demand response (an exogenous change or a control), and a quantified causal energy rebound that needs both. Of 12 candidate comparisons, two (Google and OpenAI) align at provider and API scope but not by token type or model mix, four are partly comparable and six are not; none reaches energy accounting, causal identification or a quantified rebound. In the two aligned series, API tokens per minute grew about 4.2 to 5.5 times a year (Google) and at least 5.6 times a year (OpenAI) in 2025 to 2026, while published output-token prices under a price-defined class definition did not fall. This does not contradict falling prices per unit of capability, and the volume growth does not show a price or efficiency response. For the same model at consecutive quarter ends, list prices were unchanged in 698 of 725 observations (266 model series, 10 developers), so class movements come from models entering and leaving. The public record cannot yet show whether AI inference exhibits rebound, so neither the claim that efficiency caps AI's energy demand nor the claim that it raises it can be checked against public data today. I release the ledger, the framework, a minimum reporting schema with a validator, a tested checker and nine recommendations.

1. Introduction

Agencies, investors and developers face two opposing claims about AI's energy future. First, energy per AI task is falling fast: the International Energy Agency writes that software and hardware advances have cut "the energy use per AI task" by at least an order of magnitude a year in recent years [IEA 2026]. Second, use is rising faster: the same report finds that electricity consumed by AI-focused data centres grew 50% in 2025, while all data centres together grew 17% to 485 TWh [IEA 2026]. The question is whether the first trend causes part of the second. This is the rebound or Jevons question [Jevons 1865; Luccioni et al. 2025], and it matters when judging whether efficiency standards can bound AI's energy demand or only slow it.

Three questions are often run together. Is demand rising while prices or energy per token fall? That is a descriptive comparison. Did total energy rise, given volume and energy per unit on one boundary? That is energy accounting. Did the efficiency gain itself cause part of the demand growth, measured against what demand would otherwise have been? Only that is rebound, and only a rebound above 100% is the Jevons paradox. These questions need different evidence. Showing how demand responds to a price change needs an exogenous change or a control, and no energy data. Accounting for energy needs volume and energy per unit on one boundary, and no counterfactual. A quantified energy rebound needs both.

The literature provides the series in pieces. Price declines at fixed capability are well measured [Cottier et al. 2025; Gundlach et al. 2025; Du 2026]. Volume and its response to price have been studied on one marketplace, where short-run price elasticities just above one were found [Demirer et al. 2025]. Energy per request has been measured, but on boundaries that differ from one figure to the next [Elsworth et al. 2025; companion audit, in preparation]. I found no source joining the three for the same scope and dates, or stating which of the questions the public record can answer.

This paper builds the joined record and sets out what it can establish. It does four things. First, a dated, source-linked ledger for 2022 to 2026: list prices from archived copies of developers' own price pages, volume statements coded for scope and counting rule, and energy figures tagged with a boundary class. Second, the Inference Rebound Identifiability Framework (IRIF 1.0): a named, versioned rule set with definitions, pseudocode and a missing-evidence schema, released so that it can be used without this dataset. Third, its application to 12 candidate comparisons, with the outcome per branch and the minimum evidence that would advance each. Fourth, audits of the price series themselves: same-model repricing against composition, the independence of the open price datasets used for cross-checks, and the provenance of each archived capture. A minimum reporting schema with a validator turns the findings into something providers can adopt. Section 2 reviews the existing work. Section 3 explains the method. Section 4 answers five analysis questions from the ledger. Section 5 discusses the findings, Section 6 gives nine recommendations, Section 7 sets out open problems and a roadmap, and Section 8 states the limitations.

2. Background

Rebound has a definition. Jevons argued in 1865 that more efficient use of coal raised its total consumption [Jevons 1865]. The modern literature separates direct rebound (more use of the same service once it is cheaper to use), indirect rebound (spending the saving on other goods and services) and economy-wide effects [Sorrell and Dimitropoulos 2008; Sorrell et al. 2009]. Sorrell and Dimitropoulos give a definition for a single energy service and show that estimating it from price variation depends on assumptions about capital costs, whether efficiency is exogenous, and the cost of time. Sorrell's review of the evidence for backfire, a rebound above 100%, finds it inconclusive [Sorrell 2009]. I use direct rebound for one service. For a change from energy per unit I_0 to I_1 (I_1 below I_0), observed volume V_1 and the volume V_{cf} expected without the change, the rebound share is $R = (V_1 - V_{cf}) / (V_{cf} (I_0 / I_1 - 1))$: the extra energy used by the extra volume, divided by the saving expected at constant volume. The counterfactual V_{cf} is never observed; it needs a control or an exogenous change. A price elasticity can stand in for R only under further assumptions: that the price change passes the efficiency gain through, that efficiency is exogenous, that energy per unit of service is constant, and that the elasticity is for the same service and scope. An elasticity at the model-provider level also includes substitution between providers, which moves volume among providers without raising the total.

Price series exist, mostly at fixed capability. Epoch AI measured the price of reaching a given benchmark score and found declines of 9x to 900x a year depending on the task [Cottier et al. 2025]. Gundlach et al.

found the price for a given level of benchmark performance falling about 5x to 10x a year, while the price of running frontier models rose 3x to 18x a year [Gundlach et al. 2025]. Du reports an approximately 600-fold decline in token prices from marketplace and milestone records, 2020 to 2026, with faster declines in cheaper tiers [Du 2026]. These studies track the price of a level of capability, not the dated list price of a named model on its developer's page. Open historical price datasets begin in 2025: llm-prices.com in February 2025, model-prices (built from a model catalogue's commit history) in June 2025, and the AI Signal dataset's change log in September 2026 [llm-prices.com; model-prices; AI Signal 2026].

Volume is measured on platforms; statements elsewhere are scattered. OpenRouter data underlie a study of over 100 trillion tokens of real-world use [Aubakirova et al. 2026], descriptive evidence on substitution and multihoming [Fradkin 2025], and the market study that estimated short-run elasticities of -1.08 and -1.11 at the model-provider level [Demirer et al. 2025; Demirer et al. 2026]. The Journal of Economic Perspectives study of how firms buy and sell AI is a within-market analysis of one marketplace [Demirer et al. 2026]. This paper's difference is a source-auditable record across providers, joined to boundary and identifiability checks. Outside platforms, volume appears as statements in keynotes, earnings calls and agency releases. One index compiles such disclosures into a modelled global count with archived links [Tokens Per Day 2026]; it publishes no price and gives energy no boundary.

Energy per token, request or task is measured on many boundaries. Google measured its median Gemini Apps text prompt at 0.24 Wh on a boundary that counts accelerators, host machines, idle capacity and facility overhead, and reported a 33x reduction over one year [Elsworth et al. 2025]. A public leaderboard reported energy per output token falling 15–41% for the same models on the same accelerators between two versions [Chung 2026]. A study of inference energy, efficiency pathways and test-time scaling shows that longer reasoning outputs multiply the energy of a query, and that efficiency levers act at model, serving and hardware level [Oviedo et al. 2026]. Task complexity is therefore a driver of energy separate from efficiency per token: energy per token and energy per task can move in different directions, and a per-token price does not line up with a per-task energy figure without a stated conversion. A companion audit of 126 published inference energy claims finds that about half do not state what they count [companion audit, in preparation]; I use its boundary vocabulary unchanged.

The AI rebound literature is conceptual, theoretical, firm-level or planned. Luccioni, Strubell and Crawford explain why efficiency gains may raise consumption and call for joining life-cycle assessment with socio-economic analysis [Luccioni et al. 2025]. Gujral et al. test sustainability narratives for metric, boundary and reinvestment [Gujral et al. 2026]. A theoretical model derives a structural Jevons effect from agent architectures [Zhang and Zhang 2026]. Training impacts have been shown to keep rising despite efficiency gains [Morand et al. 2025]. Yu and Xu study Chinese listed firms and find that AI adoption lowers carbon intensity but raises total carbon emissions, which they call an AI Jevons paradox [Yu and Xu 2026]. That is firm-level adoption and emissions, not inference electricity or token prices. A preregistered study plans to estimate a causal elasticity on marketplace data and map it to national

electricity [Ekeocha 2026]. The missing step is practical: put the three public series side by side, keep their source dates and boundaries, and show which kind of comparison they actually support.

3. Data and method

I fixed the method in a written protocol before collecting any values. Later changes were logged with their reasons and the rows they affect. No change was made after coding began on the rows it affects, except the re-sourcing of one volume statement to its primary page, the three-branch reading of Section 3.5 and the token-type audit of Section 3.3, all made after reviews and logged.

3.1 Series and sources

Table 1 lists the three series and their sources.

Table 1 The three series: unit of record, source classes and what was not used.			
Series	Unit of Record	Source Classes	Not Used
Price	One model's standard pay-as-you-go input and output price per token on one developer's price page at one quarter end	Internet Archive captures of the developer's own price pages; three open price datasets for cross-checking only	Prices of third-party hosts; batch, cached, priority, regional and long-context surcharges; any estimate
Volume	One dated public statement of tokens, requests or messages served	Provider first-party texts; platform statistics; government agencies; press reports only where the primary was unreachable	User counts, revenue, capacity, forecasts, modelled global counts
Energy	One published energy figure for AI inference, or a facility total	Papers, provider pages, benchmarks, statements, agency reports	Any measurement of mine; any figure converted to another boundary

Time grid. Nineteen quarter ends, 2022-03-31 to 2026-09-30. A price cell holds the latest capture on or before the quarter end and at most 92 days older; otherwise it stays empty. No value is interpolated, carried over a longer gap or converted between currencies.

Developers. The frame is every maker in the three open price datasets (31). A developer is included when it sells per-token access to its own text-output models on a public page, that page has an archived copy dated on or before 2024-12-31, and the developer appears in at least two of the datasets. Eleven developers qualify (Alibaba, Amazon, Anthropic, Cohere, DeepSeek, Google, MiniMax, Mistral, Moonshot AI, OpenAI, xAI); Meta failed only the archive condition, and nineteen appear in one dataset only. Developers appear in alphabetical order everywhere.

3.2 Price capture

For each developer, I listed the URLs serving its API price page in a fixed priority order and read the Internet Archive's capture index. For each quarter end, the script takes the highest-priority URL with a capture in the window, newest first. It keeps the first capture whose text contains a currency figure within 120 characters of a per-token unit. Raw bytes and extracted text are hashed (SHA-256). When an archived page renders prices in the browser, the script reads data blocks embedded in the same page or, for empty shells, the page's own route script as archived at the same timestamp. A capture-level provenance table records, for each of 127 capture-bearing cells, where the price text sits, whether a separate script was read and the extraction rule (Section 4.7). Requests were spaced at least three seconds apart; refusals were retried and never recorded as "no capture". A cell before a developer's own dated statement that its API was available is not-offered; a page priced only per character or per request is other unit.

3.3 Coding, agreement and adjudication

Collection and coding were separate. Two coding passes used the fixed codebooks, the second blind to the first. For price captures, they coded model name, listing status, reasoning label, input and output price, unit and a quote of at most 40 words. For volume statements, they coded quantity, period, scope, counting rule, bound and confidence. For energy figures, they coded all fields of the companion boundary vocabulary. Disagreements were adjudicated under written rules: price rules R0-R12 were applied by script and their counts published; every volume and energy field received a written decision with a reason. Agreement between the passes shows consistency under the written rules; Section 4.7 states what was checked independently of the passes and what was not.

Token-type audit. The price series are output-token list prices, so a volume statement is matched to them only if it counts tokens of the same type. I audited the 27 token statements used in the price-volume comparisons for any stated token type (input, output, cached, or an aggregate with its split). Three state one (DeepSeek's input and output counts, and OpenRouter's prompt plus completion tokens); the Google and OpenAI statements say "tokens". The audit is released as a table and the framework reads it (Section 3.5).

Energy figures from the companion audit. For 2024 to 2026, energy point figures are the companion audit's adjudicated primary claims at its adjudication commit (d49a59da), the cut-off for this paper; a later revision of that audit goes into a versioned update of the ledger. The point figures of 2022 to 2023 are coded here by the same rules, because the companion audit's window starts in 2024.

3.4 Derived series

Classes. Per developer and date, over current models with positive prices: flagship = highest output price, economy = lowest, middle = the lower median of the rest. The classes are defined by price, not by capability. A class price changes when models enter, leave or replace each other, so a class comparison is not a price comparison for an equivalent service, and it implies nothing about capability or product mix.

Three price comparisons, kept apart. Same-model repricing follows one model name of one developer across consecutive grid dates in one currency (266 series). Price-defined composition follows the classes and records, for each pair of consecutive dates, whether the models that define the class changed. Fixed-capability price evolution, the price of reaching a given benchmark level, is not computed here; external series appear for context only (Section 2).

Factors. $f = (v_{\text{last}} / v_{\text{first}})^{(365.25 / \text{days})}$: a factor of 0.5 is a halving a year. Factors are descriptive; I fit no regression on few points and estimate no elasticity.

Bounded volume factors. A printed number is an interval. "Over" or "more than" gives a lower bound; a rounded value such as "6B" covers half a unit of its last digit either side (5.5 to 6.5 billion); a vague date such as "early 2024" is a date interval. Each growth factor is reported as the range of factors over these intervals, with a side left open where the sources do not bound it. Of the 49 statements, 23 are lower bounds and 6 approximate.

Volume per day. Calendar arithmetic only (per minute x 1,440; a month divided by its days), on the date the statement refers to, not the date it was made.

3.5 The identifiability framework (IRIF 1.0)

A comparison of price, volume and energy can support four different things (Fig. 1, Table 2). Descriptive comparability (D) comes first. Two branches follow it and do not depend on each other: energy accounting (E) and causal identification (K). A quantified causal energy rebound (Q) needs both, at compatible scope. Passing a stage never implies a later one, and E and K are reported separately.

Table 2

The four stages of IRIF 1.0: question, minimum conditions and the evidence whose absence stops the stage.

Stage	Question	Minimum Conditions	Missing Evidence That Stops It
D Descriptive Comparability	Are price, volume or energy observations matched well enough to compare?	A shared scope; two or more dated observations that count the same thing on one basis; prices within one quarter; for token volume against a token price, the same token type and model mix (without that match the category is provider-and-scope alignment); for energy against price, the same subject, boundary class and per-token unit	No shared scope; volume not in tokens; one date; different bases; token type or model mix not shown; undeterminable boundary
E Energy Accounting	Can energy be followed as volume times energy per unit on one boundary?	An intensity series (per token or request) on one determinable boundary at two or more dates, in the unit and scope of the volume series	No intensity series; boundary, unit or scope mismatch
K Causal Identification	Can the demand response to a price or efficiency change be identified?	Two or more dated demand observations, a stated change, an exogenous change or a control, and other changes in the window ruled out	No stated change; no design; other changes (products, capacity, prices) not ruled out
Q Quantified Causal Energy Rebound	Can the rebound share be estimated?	E and K supported at compatible scope, and an efficiency gain stated	Either branch not supported; no efficiency gain; scopes differ

Definitions

Volume growth is not rebound by itself. A price fall is not evidence of an energy-efficiency gain. Rising energy is not by itself caused by efficiency. Rebound is the share R defined in Section 2, with its counterfactual. It is computed only when Q is supported. The Jevons paradox requires R above 1; rising consumption alone does not show it. A price fall that is not an efficiency gain can have an identified demand response (K) without any energy rebound being defined: the framework reports that case as K supported and Q not. "Q supported" means that the rebound share can be computed. It does not mean that rebound was found: the share may be zero, between zero and one, or above one. Demand response and elasticity are printed whenever the counterfactual values are given, and they are descriptive, not identified, unless K is supported.

Provider-and-scope alignment. When a provider's volume statement and its price series cover one service and are dated alike, but the volume statement does not say which token types it counts or which models, the comparison is aligned at provider and API scope. It supports a statement about the direction

of the two series for that provider. It does not support a comparison of magnitudes, a statement about the price of equivalent capability, or any step toward E, K or Q. Model mix counts as aligned only when the volume statement names the model or set of models counted and the price series covers exactly those models, or the volume is for a single priced model. My own judgement is not enough.

Missing-evidence schema. Each stage below the top of its scale carries codes from a table of 24 (ME-D1 to ME-D10, ME-E1 to ME-E6, ME-K1 to ME-K4, ME-Q1 to ME-Q4). A code says what is missing, what closes it, the kind of evidence (documentation, data extraction, measurement or identification) and who can supply it. The minimum missing evidence of a comparison is the set of codes of its next step: those of D while D is not comparable (for the two aligned comparisons, the token type and the models counted, which are documentation only), otherwise those of the unsupported branch with the fewest codes.

Implementation. The framework is a standard-library reference implementation of under 300 lines with a JSON description of a comparison as input. It is tested on thirteen synthetic gold cases (labelled synthetic: they test the rules, not the world) and on the 12 real comparisons. IRIF does not verify that the efficiency endpoints I0 and I1 of a counterfactual are the energy-intensity series of branch E on the same boundary class; the analyst states them. The framework folder of the release holds the specification, pseudocode, code, schema, cases and tests, and needs none of the data. I wrote and tested it. No other researcher has applied it, and its stage rules have not been checked against an independent coding of the same comparisons, so it is a proposed framework and not an established standard.

3.6 Analysis questions

AQ1 How did list prices change, and how much of the change is repricing, composition or capability?
AQ2 Within an aligned scope, did volume grow while price-defined class prices fell? AQ3 Where energy figures exist for the same subject at two dates on one boundary, does their slope match price? AQ4 Which outcome does each comparison reach, per branch, and what evidence would advance it? AQ5 What do facility totals show, and can they be attributed to inference?

3.7 Use of AI tools

AI assistants supported evidence collection and processing, as well as manuscript preparation. I made the research and publication decisions and am responsible for the content.

4. Findings

4.1 What the public record holds

Price. Of 209 developer-quarter cells, 100 hold priced captures, 32 fall before the developer's own dated availability statement, 34 have no capture in the window, 33 have captures without a usable price and 10 price in another unit. The filled cells give 1,392 model prices, 1,202 of them current. Coverage is thin before 2024: 4 filled cells in 2022 and 9 in 2023, against 34 in 2025 (Table 3, Fig. 2a). In several cases, the archive kept the page layout but lost the prices: client-rendered shells on Moonshot AI's pages in

2025–2026, unrendered price placeholders on Amazon's Bedrock page from 2024-03 on, and a model table that did not load on xAI's documentation in three quarters.

Table 3
Price cells by year (eleven developers, four quarter ends a year; 2026 has three).

Year	Filled	Not Offered	No Capture	No Price in Capture	Other Unit
2022	4	20	16	0	4
2023	9	9	15	6	5
2024	27	3	2	11	1
2025	34	0	0	10	0
2026	26	0	1	6	0

Volume. 49 statements from 14 organisations, made between November 2024 and September 2026: 43 count tokens, 3 requests and 3 messages. None was found for 2022 or 2023. Only 9 state which tokens they count, and 40 of the 43 token statements do not say whether they count input, output or cached tokens. 34 are graded A (the organisation's own text, number verbatim) and 15 C (reachable only through a press report; this includes every statement by a national agency). 23 are lower bounds ("over", "more than") and 6 approximate.

Energy. 122 figures: 11 longitudinal points from eight within-source comparisons, 100 point figures and 11 facility totals. 69 (57%) are undeterminable (class CU): the text does not say whether the core components are counted, or does not define the unit. Of the 53 determinable figures, 5 are per token, and 2 of the 122 are per token or per request on a determinable boundary with dated data.

4.2 AQ1: Prices fell through new models, not repricing

Three different comparisons are often read as one. I keep them apart.

Same-model repricing. The ledger follows 266 model series from 10 developers across consecutive quarter ends. They give 725 consecutive observations, which are repeated observations of the same models and not 725 independent models. Output prices were unchanged in 698 (96.3%), fell in 21 and rose in 6 (Table 4). Only 25 of the 266 series (9.4%) changed price at all. The result does not depend on one developer or one naming rule (Table 5). Alibaba's many models contribute 450 of the 725 pairs; without Alibaba, 257 of 275 pairs (93.5%) are unchanged. When dated snapshots, "latest" and "preview" names are merged into one model family, 515 of 551 pairs (93.5%) are unchanged. Developers differ: the unchanged share is 90% or more for nine of ten developers, and 66.7% for DeepSeek (12 pairs, 3 series). The developer-weighted mean is 93.4%.

Table 4
The same model at consecutive quarter ends.

Window	Consecutive Observations	Output Price Fell	Unchanged	Rose
2022–2023	32	5	27	0
2024	133	6	124	3
2025–2026	560	10	547	3

Table 5
Robustness of the unchanged share of same-model observations to naming rules, developers and resampling.

Rule or Resampling	Pairs	Unchanged	Share	95% Interval
As coded: developer, currency, normalised name	725	698	96.3%	94.6 to 97.4 (Wilson)
Series resampled (2,000 draws, seed fixed)	725	698	96.3%	94.8 to 97.6
Developers resampled (2,000 draws)	725	698	96.3%	90.3 to 97.6
Name without dated, "latest" or "preview" stamps (family key)	551	515	93.5%	91.1 to 95.2 (Wilson)
Series with no stamp in the published name	548	521	95.1%	92.9 to 96.6 (Wilson)
Without Alibaba	275	257	93.5%	89.9 to 95.8 (Wilson)
Developer-weighted mean (10 developers)	725		93.4%	Developers range 66.7% to 100%

Price-defined composition. The cheapest output price listed by any developer fell from \$0.80 per million tokens in March 2022 to \$0.04 at the end of 2024, a twentyfold fall, and rose to \$0.13 by September 2026. Within developers, the class factors are wide in 2024 (economy-class factors per year from 0.02 to 1.67 across seven developers, median 1.00) and close to flat or rising in 2025–2026 (median factors 1.00 for flagship, 1.28 for middle and 1.10 for economy across six to nine developers). Across the 251 developer-class-date cells there are 215 transitions between consecutive dates. The output price of the class moved in 92 of them, and in 82 of those 92 the models that define the class had changed. In the 123 transitions where the models changed, the price rose in 44, fell in 38 and stayed in 41. In the 92 where the models did not change, the price moved in 10 (Fig. 3b). A class price can therefore rise while every model's price is unchanged: a more expensive successor takes the class.

Fixed-capability evolution. The external series measure the price of a level of capability, and they fall steeply (Section 2). They do not conflict with a flat or rising class price, because the models in a class improve while the class price stays where it is. I did not compute a capability-adjusted series.

Answer. The cheapest listed price per token fell steeply in 2022–2024 through new models, not repricing; in 2025–2026 the price-defined classes did not get cheaper. This is a statement about published output-token list prices under this class definition. It does not say that inference got no cheaper per unit of capability.

4.3 AQ2: Where the record aligns, volume rose while class prices did not fall

Two comparisons align at provider and API scope. Alphabet's statements put direct API use of Google's own models at seven billion tokens per minute in October 2025, and each later statement continues the series ("up from 16 billion just a quarter ago" in July 2026, at about 22 billion). From the printed precision, the growth was about 4.2 to 5.5 times a year; the last value is "approximately", so the range is a floor on the uncertainty. Over that window the flagship output price on Google's Gemini API page was unchanged, while its middle- and economy-class prices rose 2.7 to 11.6 times a year (I01). OpenAI's DevDay page of 6 October 2025 gives "6B tokens per minute on the API platform" [OpenAI 2025], and its announcement of 31 March 2026 "more than 15 billion tokens per minute" [OpenAI 2026]. Both are OpenAI's own statements. Because the first value is rounded and the second is a lower bound, the growth was at least 5.6 times a year, with no upper bound from the sources. Its API class prices rose 2.3 to 9.8 times a year as successor models replaced older ones (I03).

Neither pair is aligned by token type or model mix. The statements say "tokens" without stating whether they count input, output or cached tokens, and the price series are output-token list prices of price-defined classes. The comparisons therefore support the direction of the two series for each provider's API, not a comparison of magnitudes.

Prices did fall over 2024–2025, but no aligned series covers that window. Google's tokens across all its surfaces rose from 9.7 trillion a month in May 2024 to over 480 trillion in May 2025, at least 40 times, as stated at its May 2025 keynote. Over that period, the economy class on its API price page fell 7 times. The volume includes free consumer surfaces with no per-token price, so the comparison remains partial (I02). Statements attributed to China's National Data Administration put tokens per day at about 100 billion in early 2024 and over 140 trillion in March 2026, at least 21 times a year (I05). Those statements are available only through state media and cover developers outside the price frame.

Answer. Where the record aligns at provider and API scope (2025–2026), API volume grew about 4.2 to 5.5 times a year (Google) and at least 5.6 times a year (OpenAI), while the price-defined classes did not get cheaper. These are descriptive statements. Volume growth does not show a price or efficiency response: it could come from new products, agents or integration, or from earlier price falls working through. The 2024–2025 price fall and the volume growth share no aligned window.

4.4 AQ3: No energy comparison is better than partial

One within-source energy comparison has a determinable boundary: Google's median Gemini Apps text prompt used 0.24 Wh in May 2025 on a facility boundary (C4), with energy per prompt reduced 33 times

over the year before [Elsworth et al. 2025]. During those twelve months, the economy class on Google's Gemini API price page fell 7 times while its flagship price was flat. That comparison is only partial: energy is per prompt, price is per token, and the consumer assistant is not the API (I09). The ML.ENERGY leaderboard reports 0.20 to 0.12 J per output token for Llama 3.1 8B at batch 64, but its boundary is undeterminable and Meta, the model's developer, is outside the price frame (I10). The widely repeated ChatGPT per-query figures of 2.9 Wh, 0.3 Wh and 0.34 Wh have undeterminable or different boundaries in the companion audit and cannot form a time series.

Answer. No energy-per-token slope on a stated boundary can yet be compared with a price slope for the same subject.

4.5 AQ4: Two comparisons align at provider scope; none reaches a branch

Under IRIF 1.0 the 12 comparisons give the outcomes in Table 6. None is descriptively comparable in the full sense, because no volume statement states the token type or model mix of the price series it sits beside. Two are aligned at provider and API scope (I01, I03), four are partly comparable and six are not. E is supported in none, K in none, and so Q in none. One comparison reaches E in part (I09). A reader who cites these series together as evidence for or against rebound is citing series that the public record cannot join. The figure shows, for every comparison, why it stops where it does (Fig. 4); the same matrix is released as a table.

Table 6

Outcomes of the 12 comparisons under IRIF 1.0: D descriptive comparability (aligned = provider-and-API-scope alignment), then E, K and Q, with what the record shows and the minimum missing evidence for the next step.

ID	Comparison	D	E / K / Q	What the Record Shows	Minimum Missing Evidence
I01	Google: direct API tokens vs Gemini API prices	Aligned	Not / Not / Not	Tokens of unstated type x4.2-5.5 a year (2025-10 to 2026-07); output-price classes: flagship flat, middle and economy x2.7-11.6	The token type and the models the volume counts, stated by Google [D: ME-D7, D8]; then energy per token for the same API on a stated boundary at two dates [E: ME-E1]
I02	Google: tokens across all surfaces vs Gemini API prices	Partly	Not / Not / Not	At least x40 (May 2024-May 2025); next year not bounded; economy price x0.14 then x2.7	A volume series restricted to priced API use, with token type [D: ME-D2, D5, D7, D8]
I03	OpenAI: API tokens vs API prices	Aligned	Not / Not / Not	Tokens of unstated type at least x5.6 a year (2025-10 to 2026-03), both statements first-party; class prices x2.3-9.8	The token type and the models the volume counts, stated by OpenAI [D: ME-D7, D8]; then energy per token for the API on a stated boundary at two dates [E: ME-E1]
I04	OpenAI: ChatGPT messages vs API prices	Not	Not / Not / Not	Messages, not tokens; no per-token price for the product	A token count for a priced scope [D: ME-D2]
I05	China (National Data Administration, via state media) vs Chinese developers in the frame	Partly	Not / Not / Not	Tokens at least x21 a year (early 2024 to March 2026)	Primary agency texts, a counting rule, token type, one price series [D: ME-D2, D7, D8]
I06	DeepSeek: one-day volume	Partly	Not / Not / Not	One 24-hour window	A second dated statement of the same quantity and scope [D: ME-D4, D7, D8]
I07	OpenRouter (platform)	Not	Not / Not / Not	Context: elasticities – 1.08 / -1.11 (short run) [Demirer et al. 2025]; 13.8x and 5.6x daily tokens for two discounted models vs 1.1x undiscounted in one window [OpenRouter 2026]	Same-period statements; a platform price series [D: ME-D1]

ID	Comparison	D	E / K / Q	What the Record Shows	Minimum Missing Evidence
I08	Baidu, Microsoft, Salesforce, Volcano Engine, Zhipu (outside the price frame)	Not	Not / Not / Not	Volume only	Their own per-token price series [D: ME-D1]
I09	Google: Gemini Apps energy per prompt vs Gemini API prices	Partly	Partly / Not / Not	Energy per prompt down 33x in a year (C4); economy price down 7x	Energy per token for the API on the same boundary, or a prompt-volume series for the app [D: ME-D2, D10]
I10	ML.ENERGY leaderboard: Llama 3.1 energy per token	Not	Not / Not / Not	0.20 to 0.12 J per output token (CU).	A price series; a stated boundary [D: ME-D1]
I11	IEA per-task, NVIDIA hardware-generation and AI Index hardware-efficiency claims	Not	Not / Not / Not	Stated rates, undated	A dated subject and a stated boundary [D: ME-D1]
I12	IEA, LBNL and EPRI data-centre totals vs volume	Not	Not / Not / Not	485 TWh in 2025 (+ 17%); AI-focused + 50%.	An inference share of a total and a same-scope volume series [D: ME-D1, D2]

What would advance each comparison. For the two aligned comparisons, the shortest step is documentation only: the provider states the token type and the models its volume counts (ME-D7, ME-D8), which would make the comparison descriptively comparable. After that, the next step is an energy-per-token series on a stated boundary for the same API at two dates (ME-E1). For K, none of the 12 has a stated change with a control. The one design with a control that I found is a discount window on one platform: two discounted models grew 13.8 and 5.6 times in daily tokens against 1.1 times for an undiscounted model in one window [OpenRouter 2026]. The post reports no energy and I did not audit whether other changes in the window were ruled out, so the ledger carries it as context only (I07). It is the kind of design K needs. The branches are independent: a design like that can identify a demand response without any energy data, and the same data, with an efficiency change and energy per unit, would be the route to Q.

Answer. Two comparisons align at provider and API scope, four partly and six not. One reaches energy accounting in part (I09); none reaches it in full, none identifies a causal demand response and none permits a quantified rebound.

4.6 AQ5: Totals bound the national picture, not inference

Data-centre electricity totals are available by year and scope: 176 TWh for the United States in 2023 [Shehabi et al. 2024], 415 TWh worldwide in 2024 and 485 TWh in 2025 [IEA 2026]. They are modelled on a facility boundary. None is attributed to inference, and no volume series shares their scope, so the ledger keeps them in a separate group and never joins them to per-token rows (I12).

Answer. The totals show growth but cannot be attributed to inference volume from the public record.

4.7 How far the ledger and the framework can be trusted

Agreement. For volume, Cohen's kappa was 1.00 on quantity, period, bound and confidence, 0.95 on scope and 0.93 on counting rule. For energy, it was 0.92 to 1.00 on component fields and 0.96 on normalisation; the basis of component codes (0.57), load (0.58) and uncertainty (0.43) had weaker agreement and were settled field by field. For price, the passes listed identical model sets in 89 of 127 captures (mean Jaccard 0.86). Of 1,337 models listed by both, input prices agreed exactly in 99.9% and output prices in 93.6%; listing status had kappa 0.75 and reasoning label 0.62. The largest adjudication rules were one row per model at the non-thinking price when a page prices modes separately (98 cases), excluding promotions as a model status (91), and excluding realtime audio models (40). This agreement shows consistency under the written rules. Checked independently of the coding passes: the price values against three open datasets (below), the 12 judgements against a rule-based checker that reads the ledger rows, and five archived price captures, re-fetched with matching hashes. Not checked: both coding passes were made by AI assistants working from written codebooks, and no independent person has re-coded or verified a sample of the price, volume or energy rows.

Cross-check and source independence. Of 595 price pairs also covered by the open datasets, 560 agreed. Agreement is not independence. The three datasets describe the same developers' price pages, and they differ in how they date a price (Table 7). The largest share of pairs (239) is against values of llm-prices.com that carry no start date, and for those 94.1% agree: the test is only that a price never changed. For the 11 pairs where that dataset states a date, 63.6% agree. The model-prices dataset dates a price by the commit date of a public model catalogue (292 pairs, 94.5% agree) or by a documented date (15 pairs, 93.3%); a commit date is when the catalogue changed, not when the page did, and 37 of its 811 entries cite an archived provider page, so a small part of its lineage is the same archive. The AI Signal dataset records current prices read from live pages and covers the last quarter end only (38 pairs, all agree). For 15 of the 35 differences, a dataset gave the model's first price without a start date and applied it to earlier dates. The captured value stands in every case, and the release records each difference as a separate row.

Table 7

Agreement of the ledger with the open datasets by how each dataset dates its values.

Dataset	How Its Values Are Dated	Pairs	Agree	Share
llm-prices.com	No start date for most rows	239	225	94.1%
llm-prices.com	A stated start or end date	11	7	63.6%
model-prices	Commit date of a public model catalogue	292	276	94.5%
model-prices	Documented date	15	14	93.3%
AI Signal	Verification date; last quarter end only	38	38	100%

Capture provenance. The release lists, for each of 127 capture-bearing grid cells, the archived page, its hashes, where the price text sits and how it was extracted. In 113 cells the prices are in the static text of the archived page; in 11 they sit in data blocks embedded in the same page; 2 are PDF text layers; one cell needed a separate archived route script and is not a priced cell. For that script, the archive serves its nearest capture and the script's own timestamp was not recorded, so same-date archiving of dependencies is not verified. No cell in the ledger depends on a separate script. Most of the 33 cells without a usable price are pages where scripts rendered the prices and the archive kept only the layout.

Reading sheet. The release includes a sheet of 40 archived price pages for an independent reader, with a key and a scoring script. The items oversample rows on which the coding passes disagreed, pages whose prices sit in embedded data or a PDF, and the Google and OpenAI cells of the two aligned comparisons. The sheet gives links only, asks for three readings of each page, carries no expected answer and requests no personal data. No independent reader has used it, so it is not part of the evidence of this paper.

The framework. The reference implementation reproduces all 12 judgements from the ledger rows and handles thirteen synthetic gold cases as expected: demand rising without a change (no rebound question); a price fall with an identified demand response of zero, which is K supported and Q not; a partial rebound of 50%, a backfire of 200% and a quantified rebound of zero under stated counterfactuals; a cheaper model entering while existing prices stay unchanged (composition, not repricing); energy figures on incompatible boundaries and units; a demand response identified with no energy series, and energy accounting with no identification design; a design whose other changes are not ruled out; aggregate tokens of unstated type beside an output-token price; volume in messages; and one dated observation. These tests show that the rules behave as written. They are not evidence about real-world rebound.

5. Discussion

The record is younger than the debate. The rebound argument is usually made about 2022–2024, when prices fell fastest. For those years the ledger has prices for few developers and no volume statement at

all. Volume statements began when volume became a selling point, in late 2024; by then, the price-defined classes had stopped getting cheaper. The windows in which price fell and the windows in which volume is visible barely overlap.

Composition is the mechanism of price change. A developer seldom cuts the price of a model it already sells (27 changes in 725 observations; 25 of 266 series). Prices fall when cheaper models are added and rise when successors are listed higher; 82 of the 92 class price moves coincide with a change of the models in the class. A curve of "the cheapest token" or of a price class therefore mixes product change with price change. An elasticity estimated against such a curve measures response to new products as much as response to price.

What the aligned series can and cannot carry. In I01 and I03, API volume grew while class prices did not fall. One reading is that demand grew for reasons other than price: new uses, agents, integration. Another is that earlier price falls were still working through, with a lag the window cannot see. The ledger cannot separate the two, and neither pair counts the same token type as the price. Neither comparison includes energy, so neither can show whether total energy rose because of efficiency gains. The pattern is consistent with the short-run marketplace estimates near unit elasticity [Demirer et al. 2025] and does not test the long run.

Why the branches are separate. Studies of how demand responds to a price change can be strong without energy data, and an energy account can be exact without a counterfactual. Requiring E before K, as the three-outcome procedure of record version 1.0.0 did, would have ruled out the first kind of study. Keeping them apart says what each can and cannot claim: a causal demand response says nothing about total energy, and an energy account says nothing about cause. A quantified rebound is the rare case where both hold. Energy per token and per task also differ, because longer reasoning outputs raise the energy of a task without changing the energy per token [Oviedo et al. 2026]; the unit has to be stated before the branches can be joined.

Energy is the weakest series. The best within-source energy comparison is on the right boundary but the wrong unit for price, per prompt rather than per token. Most energy figures do not state their boundary. Until a developer publishes energy per token on a stated boundary for a model it also prices, energy accounting for a priced scope is not possible from public sources.

What this does not show. It does not show that rebound or the Jevons paradox is absent or present. It shows that the public record supports provider-scope descriptive comparisons for about one year of Google's and OpenAI's API traffic, no full energy accounting, no identified demand response and no quantified rebound.

6. Recommendations

Efficiency claims and rebound warnings are already being used to support decisions on grid capacity, efficiency standards and AI investment. Yet no comparison of the 12 is descriptively comparable in the

full sense (2 are aligned at provider scope), and none supports energy accounting or a causal test. Of 122 energy figures, 69 (57%) cannot be placed on a boundary, and 40 of 49 volume statements (82%) do not say what they count. Applied to the coded statements, the minimum reporting schema of the release finds that 2 of 49 volume statements and none of 111 energy statements carry every required field. The organisations publishing these figures have the information needed to make the claims more useful. The practical first steps are to date prices, define volume counts and state energy boundaries. The nine recommendations below name who should act, which field or series to change, the first step, how to check it and the kind of effort it needs: documentation only, existing telemetry, additional instrumentation or independent research. The effort class is an assumption from the nature of the change, not a measured cost. The minimum reporting fields behind recommendations 1, 2, 4 and 8 are in the release as a schema with required, recommended and conditional fields, effort class and cost driver for each, and a validator.

Table 8
Recommendations, each traced to the finding it rests on.

#	Who	With What	First Step	How Success Is Checked	Finding (Trace)
1	API developers' pricing and documentation teams	The price page	Add an "effective from" date to every price and keep a public price-change log on the page.	An archived copy shows dated prices; a ledger cell can be filled from the page alone.	2022–2024 prices exist only in archive captures; open datasets start in 2025 (T05, T06).
2	Developers and platforms that publish volume	Each volume statement	State the token type (input, output, cached), the counting rule, the period, the scope and whether the number continues an earlier series.	The next statement validates against the schema's volume record.	40 of 43 token statements do not say which tokens they count; 29 of 49 are bounds or approximations (T13, T47, T67).
3	Developers' web teams and web-archive operators	The archived price page	Serve the price table in the HTML (or archive the page's price data endpoint).	A new capture passes the price-figure test without embedded data or scripts.	33 cells have captures without a usable price; 11 priced cells rely on embedded data blocks (T03, T04, T63).
4	Researchers and analysts comparing energy with price	The boundary class, the unit and the framework's branches	Code each energy figure with the companion vocabulary and report D, E, K and Q separately with their missing-evidence codes before comparing.	A comparison table that carries class, unit, the four outcomes and the codes.	69 of 122 energy figures undeterminable; no comparison reaches E, K or Q (T15, T16, T27, T66).
5	Maintainers of open price datasets	The dataset's date fields	Stop applying an undated first price to earlier dates; mark first-observed dates; backfill 2022–2024 from archive captures.	The cross-check against this ledger shows no back-dated differences, and pairs against dated values rise from 11.	15 of 35 cross-check differences are back-dated first prices; 239 of 595 pairs are against undated values (T32, T64).
6	Economists and policy analysts	The same-model series	Report same-model repricing, price-defined	A study that states all three numbers and	698 of 725 same-model observations unchanged across

#	Who	With What	First Step	How Success Is Checked	Finding (Trace)
	estimating price response		composition and a fixed-capability series as three separate comparisons beside any class curve.	the number of independent series.	266 series; 82 of 92 class price moves follow a change of models (T09, T10, T58, T62).
7	Authors, journalists and policy readers who cite volume	A source-type label and the printed bound	Label every press-only number, name the primary tried, and give growth as a range from the printed precision.	A bounded range and a sensitivity line beside any rate built on such numbers.	15 of 49 statements press-only, including every agency statement; a single growth factor overstates precision (T14, T47).
8	Energy agencies and regulators designing AI energy reporting	The reporting template	Adopt the minimum reporting schema: volume with scope, counting rule, token type and window beside energy per request or token with its boundary class.	The validator returns "complete" for a first provider report, and a first comparison reaches E under the framework.	No comparison reaches energy accounting; totals carry no inference share; 0 of 111 energy statements carry every required field (T25, T28, T53, T65).
9	Enterprise buyers budgeting API spend	Their own usage records and invoices	Track tokens by model and type, the invoiced price and the list price of the model in use and of its successor; do not budget from market price curves.	Budget variance explained by volume, model mix and price separately.	In the 2025–2026 window median class factors were 1.00-1.28 a year, the highest 4.0 (MiniMax economy class, \$0.15 to \$1.20 per million output tokens, 2025-03 to 2026-09), while same-model prices held (T08, T10).

Table 9

Effort class, cost driver and limitation of each recommendation (effort and cost are assumptions from the nature of the change, not measured costs).

#	Effort Class	Cost Driver	Limitation
1	Documentation only	Keeping the log	A date shows when a provider says a price took effect, not what buyers paid.
2	Documentation, from existing billing telemetry	Aggregation and review	A stated token type does not make statements from different providers additive.
3	Existing web tooling	A page-template change	An embedded table helps future captures; it does not recover pages already lost.
4	Documentation	Coding time per figure	The codes record what is missing; they do not judge whether a design is credible.
5	Documentation and data work	Archive capture and review	Backfilling covers only the pages the archive holds.
6	Documentation	An extra series from the same data	The fixed-capability series is external; none is computed here.
7	Documentation only	None beyond editing	A range from printed precision does not make statements comparable.
8	Additional instrumentation for some reporters	Metering at the needed level, allocation to inference	The schema is proposed; it has been applied to coded statements, not tested with a reporting provider.
9	Existing telemetry	Usage exports and invoices already held	List prices are not invoiced prices; the check needs the buyer's own records.

Monday checklist. Date your price page. Define your token count, its type, and whether it continues an earlier number. Put your price table in the HTML. Tag every energy figure with a boundary class. Report same-model prices beside class curves. Give growth as a range, and label press-only numbers.

7. Open problems and roadmap

Open problems.

1. Volume before 2024. No dated volume statement for 2022 or 2023 was found. Without one, the years of the steepest price falls cannot enter any aligned series.

2. Developers outside the frame. Meta failed only the archive condition, and five organisations with volume statements (Baidu, Microsoft, Salesforce, Volcano Engine, Zhipu) have no per-token price series of their own models in the frame; their comparisons stay not comparable.

3. Undeterminable energy rows. 57% of energy figures cannot be placed on a boundary. Which providers could publish energy per token for a priced model on a stated boundary, and on what schedule?
4. A causal design. No comparison has a counterfactual for demand. A discount window with an undiscounted control, as one platform has published, is a price design; an efficiency change with a control is still missing.
5. Prices beyond the list. Batch, cached-input, regional and long-context prices are excluded here; tokens per correct answer, GPU rental prices and long-context price tiers are candidates for a versioned update of the same archive.
6. Archive fidelity. How much of the web's price record survives only as layout without numbers, and which capture methods would keep it?
7. Independent validation. The coding has been checked for consistency under written rules and against independent sources where they exist; a blinded human re-coding of a sample of price rows, energy rows and the 12 judgements would test it further. The release carries the sheet, key and scoring script for such a check of the price rows.

Roadmap. The framework turns the open problems into steps that can be taken in order and estimated separately (Table 10).

Table 10
A roadmap from provider-scope alignment to separately estimated rebound effects.

Step	Question	Evidence Needed (Missing-evidence Codes)	Who Can Supply It
1	Can volume and price be compared in magnitude?	Token type and model mix of each volume statement; price effective dates (ME-D5, ME-D7, ME-D8)	Providers
2	How much does demand respond to a price change?	A dated price change, a control or exogenous design, the series around it, other changes ruled out (ME-K1 to ME-K4)	Providers, platforms, researchers
3	How much energy does a unit of service use, over time?	Energy per token or request on one stated boundary at two or more dates for a priced model (ME-E1 to ME-E6)	Providers, independent measurers
4	How large is the direct energy rebound?	Steps 2 and 3 for one service and period, with an efficiency gain stated (ME-Q1 to ME-Q4)	Researchers with provider data
5	What are the indirect and economy-wide effects?	Spending and activity data outside this record; not attempted here	Economists

8. Limitations

List prices are what pages say, not what buyers pay. Discounts, commitments and free tiers are outside the ledger. The classes are defined by price and compare successive models; the same-model series follow model names, and a few names span model generations (the robustness checks in Table 5 vary the naming rule). Volume statements were published at times chosen by the organisations; they are neither a census nor values that can be summed, many are bounds rather than values, and most do not state their token type. The price frame depends on three open datasets and an archive condition, which favours developers with English-language or international pages. The 725 same-model observations are repeated observations of 266 series, and one developer contributes 450 of them.

Archive captures are irregular, so a cell can lag a price change by up to a quarter. The energy figures come from the sources and retain their stated boundaries; I measured nothing. The framework records that an identification design exists and that other changes are ruled out; it does not judge whether a design is credible. It covers direct rebound for one service. Both coding passes were made by AI assistants working from written codebooks. The checks in Section 4.7 are the only checks of the coded rows, and no independent person has verified a sample of them. The synthetic tests check the rules, not the world.

What is robust: the counts of what the record holds, and the finding that no comparison reaches E, K or Q, which follows from missing fields and not from coding choices. What is sensitive to coding: the unchanged share of same-model prices moves between 93% and 96% with the naming rule (Table 5), and the provider-scope label rests on my reading that no Google or OpenAI statement names a token type. What the evidence cannot establish: whether volume responded to price or efficiency, which needs a control or an exogenous change, and whether the coding is right as judged by someone other than the passes that made it.

9. Conclusion

The public record supports provider-scope descriptive comparisons of price and volume for about one year of Google's and OpenAI's API traffic. During that window API token volume grew about 4.2 to 5.5 and at least 5.6 times a year, while published output-token prices under a price-defined class definition did not fall. The steep price falls of 2022–2024 have no aligned volume series, no comparison reaches energy accounting, none identifies a causal demand response and none permits a quantified rebound. The public evidence cannot yet show whether AI inference exhibits rebound or the Jevons paradox.

The most practical next step is for developers to date their price pages, define what their volume figures count, including the token type, and state the boundary of each energy figure. The released framework, ledger, reporting schema and checkers can then take each new statement and recompute the outcomes of every comparison.

Declarations

Conflicts of Interest

The author develops commercial data-compression software. No other competing interests.

Data availability

The ledger (price, volume and energy tables, capture manifest with URLs, timestamps and SHA-256 values, capture-level provenance and source-independence tables, codebooks, schema, the identifiability table per branch with its failure-reason matrix, adjudication records, the identifiability framework with its specification, schema, gold cases and tests, the minimum reporting schema with its validator, the blinded reading sheet with key and scoring script, the reference checker and the scripts) is released under CC BY 4.0 as the Zenodo data record "Jevons Paradox in AI Inference: A Longitudinal Evidence Framework for Token Prices, Demand, and Energy Consumption (2022–2026): data record and reference checker" [Spivak 2026]. Version 1.0.0, frozen on 2026-10-08, is doi:10.5281/zenodo.23247986. Version 1.1, which carries the material described here, is cited through the version-independent DOI doi:10.5281/zenodo.23247985, which always resolves to the latest version; version 1.1 corrects the descriptive label of two comparisons (I01 and I03, from "comparable" to "provider-and-scope aligned") after the token-type audit, and its change log lists every change. The 2022–2026 observations are frozen as version 1; later observations go into new versions without changing it. Provider page texts are not redistributed; every value carries its archive URL, capture timestamp and a quote.

References

1. AI Signal (2026) LLM API pricing dataset (models.json and price-changes.json). CC BY 4.0. GitHub repository alvaroluceroo/llm-api-pricing-dataset, retrieved 2026-10-06.
2. Aubakirova M, Atallah A, Clark C, Summerville J, Midha A (2026) State of AI: An Empirical 100 Trillion Token Study with OpenRouter. arXiv:2601.10088.
3. Chung J-W (2026) LLM Inference Energy: A Longitudinal Analysis. ML.ENERGY Blog, 2026-02-09 (Internet Archive capture 20260630213214)
4. Companion audit (in preparation), Spivak E Boundary taxonomy, disclosure schema and audit of published AI inference energy claims, 2024–2026. Data record: Inference Energy Claim Boundaries: Disclosure Schema, Classification Rules and Audit Table (2024–2026), version 1.0.0. Zenodo. 10.5281/zenodo.23245903
5. Cottier B, Snodin B, Owen D, Adamczewski T (2025) LLM inference prices have fallen rapidly but unequally across tasks. Epoch AI data insight, 2025 (first archived 2025-03-13), retrieved 2026-10-06.
6. Demirer M, Fradkin A, Tadelis N, Peng S (2025) The Emerging Market for Intelligence: Pricing, Supply, and Demand for LLMs. NBER Working Paper 34608. 10.3386/w34608

7. Demirer M, Fradkin A, Tadelis N (2026) The Emerging Market for Intelligence: How Firms Buy and Sell AI. *J Economic Perspect.* 10.1257/jep.20261506
8. Du M (2026) Tiered Super-Moore's Law: Price Evolution, Production Frontiers, and Market Competition in Large Language Model Inference Services. *arXiv:2603.28576*
9. Ekeocha DO (2026) The Price Elasticity of Compute: Causal Evidence on the AI Inference Rebound and the Energy Limits of Efficiency. *OSF Preregistration.* 10.17605/osf.io/acjw8
10. Elsworth C, Huang K, Patterson DA, Schneider I et al (2025) Measuring the environmental impact of delivering AI at Google Scale. *arXiv:2508.15734*
11. Fradkin A (2025) Demand for LLMs: Descriptive Evidence on Substitution, Market Expansion, and Multihoming. *arXiv:2504.15440*
12. Gujral H, Bhardwaj E, Perera D, Becker C, Easterbrook S (2026) Plateau That Never Comes: When Efficiency Claims in Datacenters and AI Become Greenwashing. *arXiv:2606.04214*
13. Gundlach H, Lynch J, Mertens M, Thompson N (2025) The Price of Progress: Price Performance and the Future of AI. *arXiv:2511.23455*
14. IEA (2026) Key Questions on Energy and AI. International Energy Agency, Paris
15. Jevons WS (1865) *The Coal Question.* Macmillan, London
16. llm-prices.com (2026) Historical prices, historical-v1.json. Retrieved 2026-10-06.
17. Luccioni AS, Strubell E, Crawford K (2025) From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate. *ACM FAccT 2025.* 10.1145/3715275.3732007
18. model-prices (2026) LLM API prices, current and historical, from models.dev with sourced corrections. MIT licence. GitHub repository wakamex/model-prices, data snapshot of 2026-10-05
19. Morand C, Ligozat A-L, Neveol A (2025) The Environmental Impacts of Language Model Training Keep Rising: Now is the Time to Catch Impacts on the Rebound. *arXiv:2510.09022*
20. OpenAI (2025) DevDay 2025. *openai.com/devday* (2025-10-06; Internet Archive capture 20251007021332)
21. OpenAI (2026) Accelerating the next phase of AI. *openai.com*, 2026-03-31 (Internet Archive capture 20260331205253)
22. OpenRouter (2026) GPT 5.6 Discounts & Jevons Paradox. OpenRouter blog, 2026-08-25 (updated 2026-08-27)
23. Oviedo F, Kazhamiaka F, Choukse E, Kim A, Luers A, Nakagawa M, Bianchini R, Ferres L, J. M (2026) Energy use of AI inference, efficiency pathways, and test-time scaling. *Joule* 102430. 10.1016/j.joule.2026.102430
24. Shehabi A et al (2024) Lawrence Berkeley Natl Lab. 10.71468/P1WC7Q. 2024 United States Data Center Energy Usage Report
25. Sorrell S (2009) Jevons' Paradox revisited: The evidence for backfire from improved energy efficiency. *Energy Policy* 37(4):1456–1469. 10.1016/j.enpol.2008.12.003

26. Sorrell S, Dimitropoulos J (2008) The rebound effect: Microeconomic definitions, limitations and extensions. *Ecol Econ* 65(3):636–649. 10.1016/j.ecolecon.2007.08.013
27. Sorrell S, Dimitropoulos J, Sommerville M (2009) Empirical estimates of the direct rebound effect: A review. *Energy Policy* 37(4):1356–1371. 10.1016/j.enpol.2008.11.026
28. Spivak E (2026) Jevons Paradox in AI Inference: A Longitudinal Evidence Framework for Token Prices, Demand, and Energy Consumption (2022–2026): data record and reference checker, version 1.0.0 [data set and software]. Zenodo. Version 1.0.0 10.5281/zenodo.23247986; version-independent doi:10.5281/zenodo.23247985
29. Tokens Per Day (2026) The AI inference demand index. tokensperday.com, data last verified 2026-07-16
30. Yu D, Xu B (2026) The Jevons Paradox in the AI era: Artificial intelligence adoption for enhancing environmental sustainability at the firm level. *Econ Anal Policy* 90:946–966. 10.1016/j.eap.2026.01.060
31. Zhang Y, Zhang T (2026) The Economics of Digital Intelligence Capital: Endogenous Depreciation and the Structural Jevons Paradox. arXiv:2601.12339

Figures

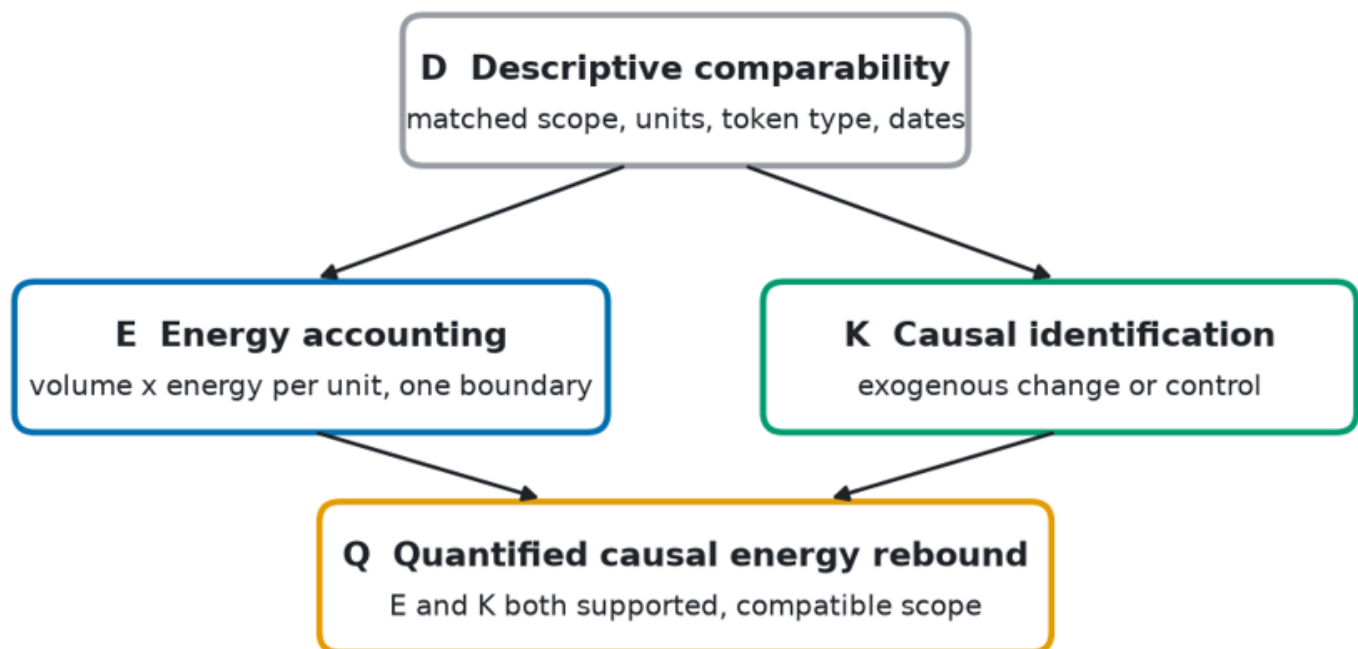


Figure 1

The structure of IRIF 1.0: descriptive comparability first, then two independent branches, energy accounting and causal identification; a quantified causal energy rebound needs both.

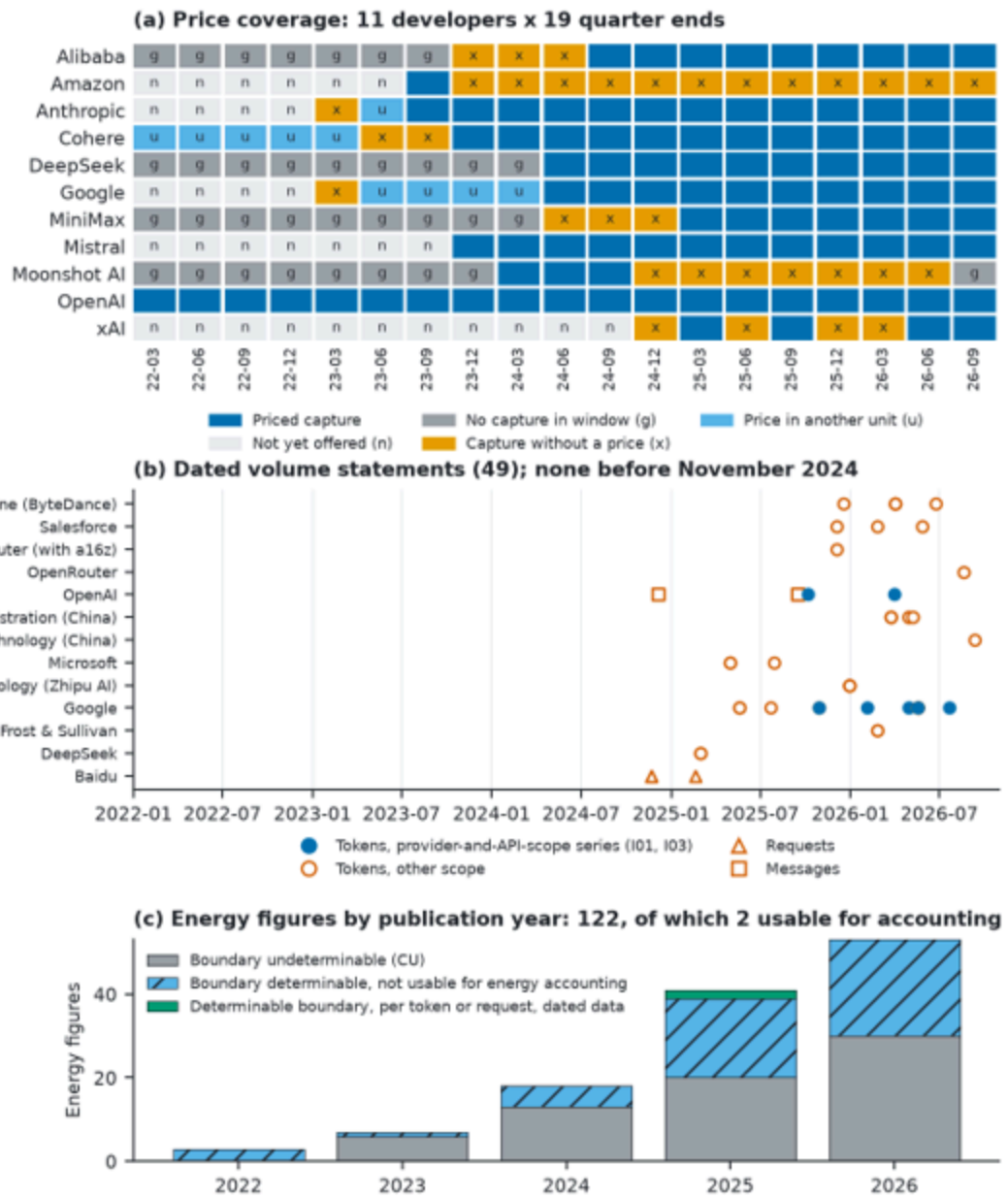


Figure 2

The missing-data window, 2022-2026. (a) Price coverage by developer and quarter end. (b) Dated volume statements; filled markers are the provider-and-API-scope series of Google and OpenAI. (c) Energy figures by publication year: undeterminable boundary, determinable boundary but not usable for accounting, and usable (determinable boundary, per token or request, dated data).

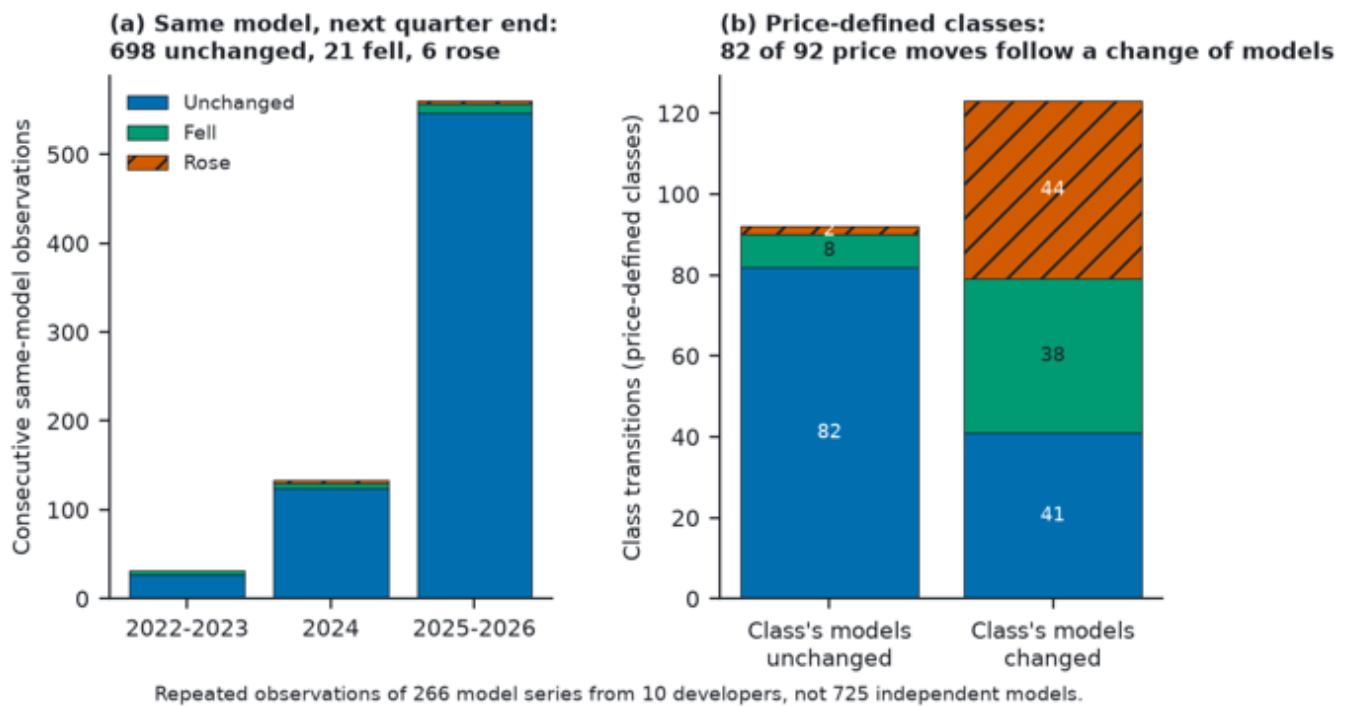


Figure 3

Same-model repricing and price-defined composition. (a) Output-price changes of the same model between consecutive quarter ends: repeated observations of 266 model series, not 725 independent models. (b) Class price moves by whether the models that define the class changed.

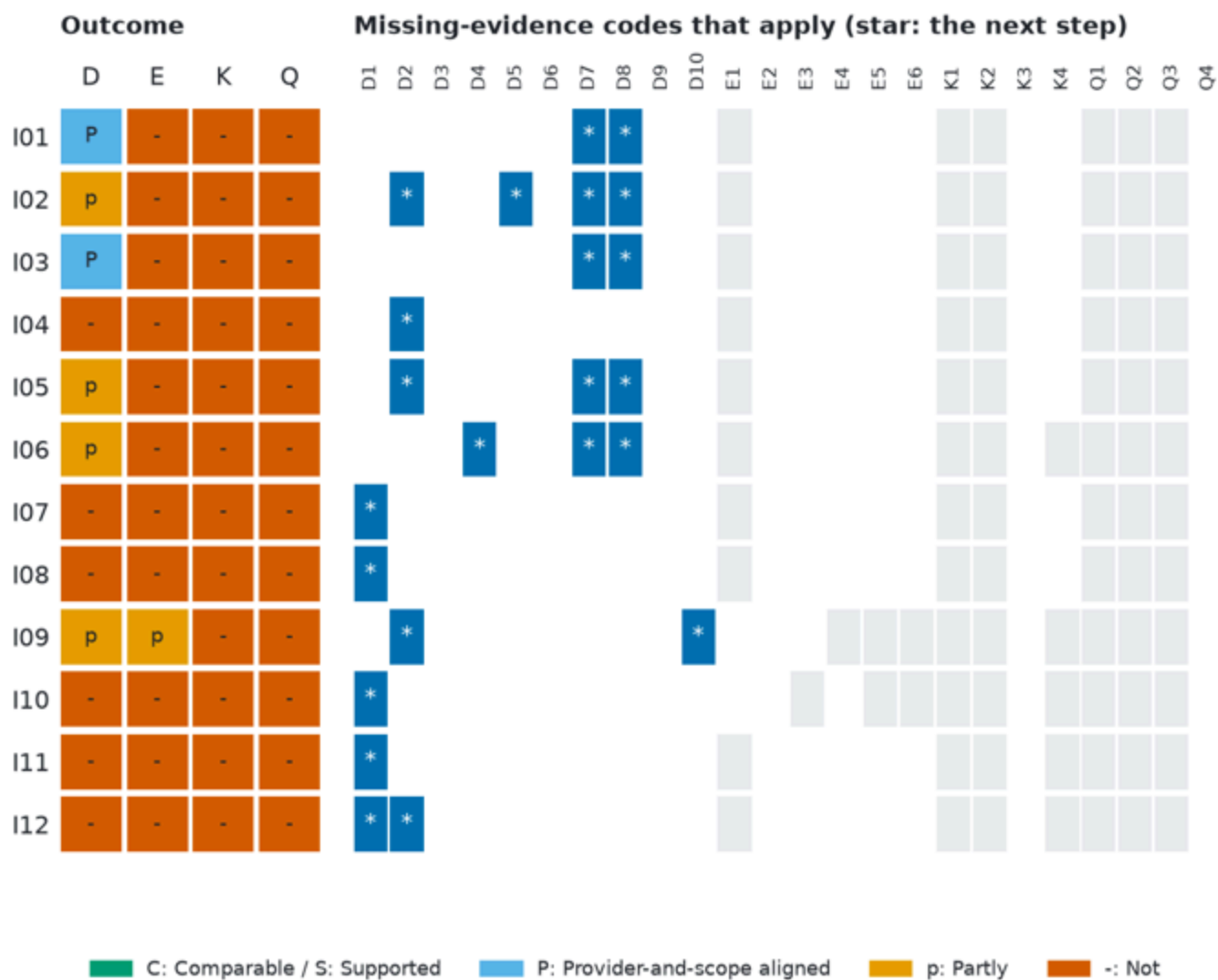


Figure 4

Why each of the 12 comparisons stops. Left: outcome per stage (C comparable or S supported, P provider-and-scope aligned, p partly, - not). Right: the missing-evidence codes that apply, with a star on the codes of the next step.