ROGER'S
RABBIT HOLE



A ROGER'S RABBIT HOLE WHITE PAPER

JEVONS PARADOX

Why more efficient AI can drive greater computational demand



BY ROGGIE D

NOVEMBER 2026

rogersrabbithole.com

JEVONS PARADOX

Why more efficient AI can drive greater computational demand

By Roggie D | November 2026

WHO THIS IS FOR

Local AI hobbyists, home lab builders, IT people and small-business technology folks who keep hearing that cheaper, faster AI will mean less computing, and want to know what history and the public numbers actually say.

HOW THE NUMBERS ARE LABELED

Historical figures come from Jevons's own book. Rebound figures come from published research reviews. AI price and usage figures are the named sources' own numbers. The dollar, token and unit tables in sections 4 and 5 are illustrative models, not measured data. No new CyberTron benchmarks were run for this paper. Calculations are labeled as calculations.

Related reading on Roger's Rabbit Hole: Bigger Isn't Always Smarter, Building AI on Aging Hardware, GPU VRAM: The Real Currency of Local AI, Local AI Model Racks and Mixture-of-Experts Models, all at rogersrabbithole.com.

THE SHORT VERSION

Jevons Paradox is an economic principle stating that improvements in the efficiency of using a resource can actually increase its total consumption rather than decrease it.

It sounds counterintuitive, but the underlying economics are straightforward: when something becomes more efficient, it becomes cheaper to use, encouraging greater adoption, increased usage and entirely new applications.

AI is a live test of that idea. Stanford's AI Index found that the price of getting GPT-3.5-level answers fell more than 280-fold in about 18 months [9]. Separately, Google says the tokens it processes each month went from 9.7 trillion in 2024 to more than 3.2 quadrillion in 2026 [12]. Cheaper per unit, far more units.

That doesn't prove cheaper AI caused the growth, and it doesn't make Jevons Paradox a law. Economists who study it say it happens in some cases and not in others. This paper

walks through where the idea came from, how it works, what the evidence shows, and why it matters for anyone building local AI.

> **Efficiency reduces the cost of an individual operation, but it does not guarantee a reduction in aggregate consumption.**

1. THE HISTORICAL ORIGIN

In 1865, English economist William Stanley Jevons published "The Coal Question" [1][3]. Jevons observed that improvements to the steam engine, including those made by James Watt, significantly reduced the amount of coal required to produce mechanical power.

He said so in plain terms. "Watt's two chief inventions of the condenser and the expansive mode of working are, simply, two modes of economising heat," he wrote, and he added that "the whole history of the steam-engine is one of economy" [1].

He backed it with a table of engine "duty," the pounds of water an engine could raise one foot by burning one bushel of coal. Table 1 shows a few of his rows.

YEAR	ENGINE	DUTY
1769	Average of old atmospheric engines	5,590,000
1772	Smeaton's atmospheric engine	9,450,000
1776	Watt's improved engine	21,600,000
1779-1788	Watt's engine working expansively	26,600,000
1859	Best engine on record	80,000,000

Table 1. Steam engine efficiency, from Jevons's table of engine duty (pounds of water raised one foot per bushel of coal)

By that measure, Watt's 1776 engine did about 3.9 times the work of the 1769 average on the same coal (calculated). Jevons summed it up: "In less than one hundred years, then, the efficiency of the engine has been increased at least ten-fold" [2].

Instead of declining, coal consumption increased as efficient steam engines became economically viable across more industries. The efficiency improvement expanded the market for the resource. Jevons's own figures show British coal use climbing from about 64.7 million tons in 1854 to about 86.3 million tons in 1863, an average increase he calculated at 3.26 percent a year [1].

Then came the line the paradox is named for. "It is wholly a confusion of ideas to suppose that the economical use of fuel is equivalent to a diminished consumption. The very contrary is the truth" [1]. A few sentences later he called it a "paradox" himself.

His example was the blast furnace. If a furnace needs less coal for each ton of iron, profits rise, new capital comes in, the price of iron falls and demand grows, "and eventually the greater number of furnaces will more than make up for the diminished consumption of each" [2].

Jevons wasn't right about everything. His larger worry was that Britain would run short of cheap coal. Encyclopaedia Britannica notes that this prediction was wrong because it didn't account for better mining technology [3]. The efficiency argument is the part that lasted.

2. HOW THE PARADOX WORKS

The chain Jevons described has six links:

1. Technological improvement
2. Increased efficiency
3. Lower cost per operation
4. Increased demand and adoption
5. New markets and applications
6. Higher total resource consumption

The mechanism behind this phenomenon is known as the rebound effect. When the rebound effect exceeds 100% of the expected resource savings, it becomes Jevons Paradox.

Economists measure rebound as a share of the savings an efficiency gain should have delivered. A 2007 review of the evidence for the UK Energy Research Centre, led by Steven Sorrell, puts it this way: "a rebound effect of 20% means that only 80% of the expected energy savings are achieved," and a rebound of 100 percent means the savings are "entirely offset" [4]. When efficiency leads to an overall increase in use, energy economists call it "backfire," the modern name for the outcome Jevons described.

REBOUND	WHAT HAPPENS TO TOTAL USE	NAME
0%	Falls by the full expected amount	No rebound
Between 0% and 100%	Falls, but by less than expected	Partial rebound
100%	Stays the same; savings fully offset	Break-even
Above 100%	Rises	Backfire (Jevons Paradox)

Table 2. Reading a rebound number

The modern debate started with two economists. Daniel Khazzoom brought the direct rebound effect to energy economists' attention in a 1980 paper on efficiency standards for household appliances [4][5]. Len Brookes, a strong advocate of the idea at the level of whole economies, made his case in a 1990 paper titled "The greenhouse effect: the fallacies in the energy efficiency solution" [4][6]. In 1992, Harry Saunders stated it formally as what's now called the Khazzoom-Brookes postulate: "with fixed real energy prices, energy efficiency gains will increase energy consumption above what it would be without these gains" [4][7].

The same UKERC review is careful about it. It found that the postulate "cannot be considered to have been verified," and that there is "no a priori reason to believe that 'backfire' is an inevitable outcome in all cases." It also found that backfire is more likely with pervasive, general-purpose technologies, and that there is some evidence steam engines and electric motors contributed to backfire in the past [4].

3. REAL-WORLD EXAMPLES

Table 3 lists places where the same pattern can show up. These are potential consequences, not measured results for each row.

TECHNOLOGY	EFFICIENCY IMPROVEMENT	POTENTIAL CONSEQUENCE
Steam engines	Less coal per unit of power	Increased coal consumption
Automobiles	Better fuel economy	More miles driven
LED lighting	Less electricity per light	More lighting installed
Cloud computing	Cheaper computing resources	More workloads deployed
Data storage	Lower cost per terabyte	More data retained
Network infrastructure	Cheaper bandwidth	Increased network traffic
Artificial intelligence	Lower cost per token	Increased AI consumption
Microprocessors	Lower energy per computation	More computational devices

Table 3. Where the pattern can appear

Not every efficiency improvement produces a full Jevons Paradox. It depends on how strongly demand responds to lower costs.

The research record shows how wide that range is:

- **Lighting:** Per person, lighting use in the United Kingdom grew by a factor of 6,566 between 1800 and 2000, driven in part by better lamps and cheaper fuel that pushed the price of light down, according to economic historians Roger Fouquet and Peter Pearson, as summarized by UKERC [4][8].
- **Driving:** The direct rebound for personal car travel in developed countries is likely between 10 and 30 percent, and may be closer to 10 percent, UKERC found [4]. People do drive more when cars get more efficient, but nowhere near enough to cancel the savings.
- **Electronics:** UKERC found direct rebound tends to be smaller where energy is a small part of total cost, and it named electricity used in electronic appliances as a good example [4].

So the steam engine and the lamp are the classic examples of demand outrunning efficiency. The fuel-efficient car is a rebound story, but not a Jevons story.

4. JEVONS PARADOX AND ARTIFICIAL INTELLIGENCE

This is particularly relevant to local AI, LLM optimization, quantization, Mixture-of-Experts architectures and the CyberTron project.

WHAT THE PUBLIC NUMBERS SHOW

AI is getting cheaper per unit, and fast.

- Stanford's 2025 AI Index found the cost of querying a model that scores at GPT-3.5's level on the MMLU benchmark fell from $20.00 per million tokens in November 2022 to $0.07 by October 2024, a more than 280-fold drop in about 18 months [9].
- Epoch AI, a research group that tracks AI trends, found in March 2025 that the price of reaching a fixed benchmark score fell anywhere from 9 to 900 times per year, depending on the task. It cautioned that the fastest drops were the least certain to last [10].
- In September 2026, Epoch AI estimated that the cost of a given level of AI performance has fallen about 47 percent per quarter since 2023, or about 13 times per year [11].

Meanwhile, use has exploded.

- Google says it processed 9.7 trillion tokens a month two years before its May 2026 developer conference, roughly 480 trillion a month a year later, and more than 3.2 quadrillion a month by May 2026 [12]. That's about 330 times in two years (calculated).
- In its October 2026 earnings release, Alphabet said its first-party model APIs were processing nearly 32 billion tokens per minute, more than four times the rate a year earlier [13].

Those two trends fit the Jevons pattern, but the sources don't show that one caused the other. Token counts also grow because of new products, free features and models that simply write more. Epoch AI noted that reasoning models tend to generate a much larger number of tokens than other models [10]. No source found for this paper measures how much of AI's growth is a response to lower prices.

AN ILLUSTRATIVE MODEL

Suppose AI inference becomes 90% cheaper through improvements in model architecture, GPU utilization and quantization. Instead of spending 90% less, businesses might deploy more AI agents, automate additional workflows, increase context windows and process significantly more tokens.

METRIC	BEFORE OPTIMIZATION	AFTER OPTIMIZATION
Cost per million tokens	$10.00	$1.00
Daily tokens processed	1 million	25 million
Daily operating cost	$10.00	$25.00
Annual operating cost	$3,650	$9,125

Table 4. Illustrative economic model; not measured benchmark data

Despite a 90% reduction in unit cost, overall expenditure increased by 150%, while computational output increased by 2,400%. This illustrates the Jevons mechanism applied to AI spending. Physical resource consumption would require separate measurement.

That last point matters. Dollars aren't kilowatt-hours. The International Energy Agency estimated that data centers used about 415 terawatt-hours of electricity in 2024, around 1.5 percent of the world's total, and projected that figure would more than double to about 945 terawatt-hours by 2030, with AI as the most important driver [14]. That's a physical number, and it's the kind of number a real Jevons test for AI would need.

5. JEVONS PARADOX CALCULATOR

This paper uses reusable formulas and a worked example in place of an interactive calculator.

PARAMETER	EXAMPLE
Original consumption	100 units
Resource savings per operation	90%
Workload demand multiplier	25x
New consumption	250 units
Consumption change	+150%
Break-even demand	10x

Table 5. Worked example (illustrative)

The formulas:

- New consumption = Original consumption x (1 - savings fraction) x demand multiplier
- Consumption change (%) = (New consumption / Original consumption - 1) x 100

- Break-even demand multiplier = 1 / (1 - savings fraction)

Checking the example: 100 x (1 - 0.90) x 25 = 250 units. (250 / 100 - 1) x 100 = +150 percent. 1 / (1 - 0.90) = 10x.

Above break-even: Jevons Paradox (if demand growth is induced by the efficiency gain). At break-even: savings are completely offset. Below break-even: net resource savings.

Table 6 holds the savings at 90 percent and changes only the demand multiplier, so you can see the line get crossed.

DEMAND MULTIPLIER	NEW CONSUMPTION	CHANGE	RESULT
5x	50 units	-50%	Net savings
10x	100 units	0%	Break-even
25x	250 units	+150%	Jevons Paradox

Table 6. Same 90% savings, different demand responses (calculated, illustrative)

The bigger the efficiency gain, the more demand it takes to cancel it out. Table 7 runs the break-even formula for four savings levels.

SAVINGS PER OPERATION	BREAK-EVEN DEMAND MULTIPLIER
50%	2x
75%	4x
90%	10x
99%	100x

Table 7. Break-even demand multiplier by savings level (calculated)

The worked example can also be stated as a rebound number, using the definition in section 2. The expected savings were 90 units. The actual change was an increase of 150 units, or savings of minus 150. Rebound = (90 - (-150)) / 90 = about 267 percent (calculated). Anything over 100 percent is backfire.

6. WHY IT MATTERS TO LOCAL AI

Jevons Paradox provides a compelling economic argument for the expansion of local AI infrastructure. As AI becomes more efficient through the following techniques, the barrier to entry decreases:

- Model quantization and reduced precision (FP16, INT8, Q4).

- Mixture-of-Experts architectures.
- Specialized model racks instead of monolithic LLMs.
- Improved inference engines.
- Hardware acceleration and memory optimization.
- Autonomous multi-agent systems.

Earlier papers in this series cover several of these. The case that bigger isn't automatically smarter is in the first paper [15]. Running real AI on old hardware is the second [16]. Why GPU memory is the budget everything else spends is the third [17]. Specialized model racks are the subject of another [18], and Mixture-of-Experts models, which do the math of a small model while holding a big one, are the most recent [19].

CyberTron, the budget local AI build behind this series, is a small example of that lower barrier. It runs on an old AMD FX-8300 PC with 32 GB of DDR3 and a 12 GB RTX 3060 [19].

A business that previously operated one AI assistant could eventually operate 50 specialized agents. A home user who occasionally used ChatGPT could operate a continuously running local AI ecosystem.

The same logic applies on a home rig. Each efficiency gain makes the next workload look affordable. The resources that matter there are GPU memory, electricity, heat and time, and a cheaper token doesn't put a ceiling on how many tokens get run.

7. THE ECONOMIC OPPORTUNITY

Jevons Paradox is not necessarily a negative phenomenon. From a business perspective, it can represent substantial productivity growth, market expansion and new revenue opportunities.

That was Jevons's own view of coal. He wrote that "the more we render it efficient and economical, the more will our industry thrive, and our works of civilization grow" [2]. His worry was the coal running out. For AI, the open questions are electricity, hardware and whether the extra work is worth doing.

> **Efficiency reduces the cost of an individual operation, but it does not guarantee a reduction in aggregate consumption.**

WHERE THIS PAPER FITS

This paper ties together the earlier white papers in this series on model size, aging hardware, GPU VRAM, local model racks and Mixture-of-Experts architectures [15][16][17] [18][19].

WHAT THIS PAPER DOESN'T PROVE

- It doesn't show that AI demand grows because prices fall. The price and usage trends in section 4 run side by side, but no source cited here isolates price as the cause.
- The dollar and token tables in sections 4 and 5 are illustrative models. They show how the math works, not what any real system did.
- Spending isn't the same as physical resource use. A Jevons claim about electricity or hardware needs electricity and hardware measurements.
- Backfire isn't automatic. The research reviewed for this paper found it in some cases and partial rebound in many others [4].

SOURCES

All links checked Nov. 26, 2026.

[1] William Stanley Jevons, "The Coal Question," first edition (London: Macmillan, 1865), Internet Archive scan https://archive.org/details/bub_gb_gAAKAAAAIAAJ

[2] William Stanley Jevons, "The Coal Question," second edition, revised (London: Macmillan, 1866), Online Library of Liberty https://oll.libertyfund.org/titles/jevons-the-coal-question

[3] Encyclopaedia Britannica, "William Stanley Jevons" https://www.britannica.com/biography/William-Stanley-Jevons

[4] Steven Sorrell, "The Rebound Effect: An Assessment of the Evidence for Economy-wide Energy Savings from Improved Energy Efficiency," UK Energy Research Centre, October 2007 https://ukerc.ac.uk/publications/the-rebound-effect-an-assessment-of-the-evidence-for-economy-wide-energy-savings-from-improved-energy-efficiency/

[5] J. Daniel Khazzoom, "Economic Implications of Mandated Efficiency in Standards for Household Appliances," The Energy Journal 1(4), 1980 https://doi.org/10.5547/ISSN0195-6574-EJ-Vol1-No4-2

[6] Len Brookes, "The greenhouse effect: the fallacies in the energy efficiency solution," Energy Policy 18(2), 1990 https://doi.org/10.1016/0301-4215(90)90145-T

[7] Harry D. Saunders, "The Khazzoom-Brookes Postulate and Neoclassical Growth," The Energy Journal 13(4), 1992 https://doi.org/10.5547/ISSN0195-6574-EJ-Vol13-No4-7

[8] Roger Fouquet and Peter J.G. Pearson, "Seven Centuries of Energy Services: The Price and Use of Light in the United Kingdom (1300-2000)," The Energy Journal 27(1), 2006 https://doi.org/10.5547/ISSN0195-6574-EJ-Vol27-No1-8

[9] Stanford Institute for Human-Centered AI, "Artificial Intelligence Index Report 2025" https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

[10] Epoch AI, "LLM inference prices have fallen rapidly but unequally across tasks," March 12, 2025 https://epoch.ai/data-insights/llm-inference-price-trends

[11] Epoch AI, "The plunging price of thought," Sept. 22, 2026 https://epoch.ai/publications/the-plunging-price-of-thought

[12] Google, "I/O 2026: Welcome to the agentic Gemini era," Sundar Pichai's keynote remarks, May 19, 2026 https://blog.google/innovation-and-ai/sundar-pichai-io-2026/

[13] Alphabet Inc., "Alphabet Announces Third Quarter 2026 Results," Oct. 28, 2026 (SEC filing) https://www.sec.gov/Archives/edgar/data/1652044/000165204426000080/googexhibit991q32026.htm

[14] International Energy Agency, "Energy and AI," executive summary, April 2025 https://www.iea.org/reports/energy-and-ai/executive-summary

[15] Roger's Rabbit Hole, "Bigger Isn't Always Smarter" https://www.rogersrabbithole.com/Rabbit-Hole/ArtMID/391/ArticleID/97/Bigger-Isnt-Always-Smarter

[16] Roger's Rabbit Hole, "Building AI on Aging Hardware" https://www.rogersrabbithole.com/Rabbit-Hole/ArtMID/391/ArticleID/99/Building-AI-on-Aging-Hardware

[17] Roger's Rabbit Hole, "GPU VRAM: The Real Currency of Local AI" https://www.rogersrabbithole.com/Rabbit-Hole/ArtMID/391/ArticleID/102/GPU-VRAM-The-Real-Currency-of-Local-AI

[18] Roger's Rabbit Hole, "Local AI Model Racks" https://www.rogersrabbithole.com/Rabbit-Hole/ArtMID/391/ArticleID/106/Local-AI-Model-Racks

[19] Roger's Rabbit Hole, "Mixture-of-Experts Models" https://www.rogersrabbithole.com/Rabbit-Hole/ArtMID/391/ArticleID/110/Mixture-of-Experts-Models

Test note: No CyberTron benchmarks were run for this paper. The CyberTron hardware described here is the configuration already published in the Mixture-of-Experts paper. The illustrative tables are models built to show the math and are not measurements of any system.

ABOUT THE AUTHOR

Roggie D has been messing with computers since the early 1980s and working in IT since the mid-1990s, from network engineering to the CEO's chair. These days he's retired, which mostly means he gets to pick which problems are worth solving. He writes Roger's Rabbit Hole, where CyberTron, his budget local AI build, keeps sending him down new ones.

More from the Rabbit Hole at rogersrabbithole.com.