The Cost Shift: Jevons' Paradox and the Restructuring of Enterprise AI Spend from Labor to Compute

Do Tuan Anh
Data & AI Engineering
Hanoi, Vietnam
tuananh071219@gmail.com

Abstract—When DeepSeek's efficiency breakthrough briefly convinced markets in January 2025 that AI compute demand would fall, Microsoft's CEO answered with a 160-year-old economic argument: "Jevons paradox strikes again." The paradox—that resource efficiency increases rather than decreases total consumption by stimulating demand—is now widely invoked in AI. This paper examines what it means for the organizations paying the bills. Our central claim is the cost shift thesis: AI efficiency gains do not reduce total enterprise AI spend; they restructure it, moving cost from human labor line items to compute line items while the rebound effect keeps aggregate consumption growing. We support the thesis with a taxonomy of five cases spanning quantization, API pricing, open-source releases, frontier training efficiency, and AI-assisted development, and with aggregate evidence: inference prices falling by roughly 40× per year for fixed capability while token volumes grew sevenfold year over year and hyperscaler capital expenditure guidance reached $725 billion for 2026. We derive consequences for enterprise AI budgeting and FinOps, cloud capacity planning, and AI governance policy.

Index Terms—Jevons Paradox, AI infrastructure, compute scaling, model efficiency, rebound effect, cost shift, enterprise AI

I. Introduction

In late January 2025, the Chinese lab DeepSeek released R1, a reasoning model competitive with OpenAI's o1, built on a base model whose final training run reportedly cost $5.6 million—a small fraction of the budgets assumed necessary for frontier capability [1], [2]. Markets read the efficiency breakthrough as a signal that AI compute demand would collapse, and NVIDIA lost roughly $600 billion in market value in a single day. Microsoft CEO Satya Nadella disagreed, in four words: "Jevons paradox strikes again!" [3]. Within eighteen months, the market had sided with Nadella: token volumes and infrastructure investment reached record levels even as per-unit costs collapsed.

The paradox Nadella invoked dates to 1865, when economist William Stanley Jevons observed that improvements in the efficiency of steam engines did not reduce coal consumption in Britain—they increased it [4]. More efficient engines made coal-powered machinery economically viable for a broader range of applications, dramatically expanding total demand. The phenomenon, also known as the rebound effect, has since been documented across energy, water, and transportation systems [5], [6].

Whether the paradox applies to AI is no longer the interesting question: industry leaders invoke it [3], and researchers have analyzed its role in AI's environmental debate [7], building on earlier work documenting the rising energy and carbon costs of AI training [8], [9]. The question this paper addresses is what the paradox means for the organizations paying the bills. Our central claim is the cost shift thesis: AI efficiency gains do not reduce total enterprise AI spend—they restructure it. Cost transitions from human labor (engineers, analysts, data workers) to compute (GPU hours, cloud inference, storage, networking), while the rebound effect ensures that aggregate compute consumption grows rather than shrinks. The total bill does not disappear; the line items change. Organizations that model AI return on investment purely as labor savings will systematically underestimate total cost of ownership.

The mechanism behind the thesis is the rebound effect itself. As inference becomes cheaper, organizations deploy more models. As models become smaller, they run on hardware that was previously insufficient—expanding the installed base of AI infrastructure. As API costs drop, applications that were previously uneconomical become viable, generating orders of magnitude more token throughput. Total compute consumption grows faster than individual efficiency gains would predict.

The remainder of this paper is organized as follows. Section II reviews the AI efficiency trajectory. Section III presents a taxonomy of five cases documenting the rebound effect. Section IV develops the cost shift thesis. Section V discusses implications for enterprise strategy, cloud planning, and AI policy.

II. The AI Efficiency Trajectory

The past four years have produced dramatic efficiency improvements across every dimension of AI model development and deployment. We summarize four major vectors.

A. Quantization

Quantization reduces the numerical precision of model weights and activations, shrinking memory footprint and

accelerating inference. Moving from 32-bit floating point (FP32) to 16-bit (FP16) halves memory requirements with negligible quality loss. INT8 quantization, popularized by methods such as LLM.int8() [10], reduces requirements by a further factor of two. INT4 formats, used in tools such as GGUF and llama.cpp, enable models with tens of billions of parameters to run on consumer hardware. A model that required an 8-GPU server in 2022 can run on a single consumer GPU in 2025.

B. Knowledge Distillation and Smaller Models

Knowledge distillation [11] enables a larger "teacher" model to transfer its learned representations to a smaller "student" model, yielding compact models with disproportionate capability. The release of GPT-4 in 2023 was followed rapidly by a generation of smaller, highly capable models: GPT-4o-mini, Llama 2 [12] and Llama 3 [13], Mistral 7B [14], and Microsoft's Phi series. These models achieve a substantial fraction of frontier model capability at a fraction of the cost. Chinchilla scaling laws [15] provided theoretical grounding for training smaller models on more data, shifting the efficiency frontier.

C. API Price Reduction

OpenAI's pricing for frontier-class models fell sharply between 2023 and 2024 [16]. GPT-4 input tokens launched at $30 per million in early 2023; GPT-4o launched at $5 per million in May 2024 and fell to $2.50 by August 2024—a 92% reduction in eighteen months—while capable smaller models such as GPT-4o-mini fell below $0.20 per million. These are not isolated data points. Epoch AI's analysis of inference prices at fixed capability levels finds that the price of GPT-4-level performance fell by roughly 40× per year, and that across tasks and capability thresholds prices declined between 9× and 900× per year, with a median of approximately 50× [17]. Competitive pressure from Anthropic, Google, DeepSeek, and open-source alternatives drove the trend industry-wide.

D. Compute Scaling Laws

Despite efficiency gains per model, total compute used to train frontier models has continued to grow. Epoch AI's tracking of training compute across AI systems shows a persistent doubling time of approximately six months [18], consistent with Kaplan et al.'s scaling law predictions [19]. Per-unit costs fall; units deployed increase.

Taken together, these four vectors establish that the AI efficiency trajectory is real and substantial. The question this paper addresses is whether that trajectory has reduced total compute consumption—or whether it has accelerated it.

III. Case Taxonomy: Five Instances of the Rebound Effect

We present five cases in which AI efficiency improvements led to increased, not decreased, total compute consumption. Each case follows the same structure: efficiency gain → mechanism → rebound outcome. Together they establish a pattern, not an exception.

A. Case 1: Model Quantization Enables Mass Deployment

Efficiency gain. INT8 quantization reduces GPU memory requirements by approximately 4× compared to FP32 [10]. A model previously requiring 80 GB of VRAM can run in 20 GB.

Mechanism. Organizations that previously ran a single large model due to hardware constraints can now run multiple specialized models on the same hardware budget: one for customer service, one for document processing, one for internal search, one for code generation.

Rebound outcome. Rather than freeing GPU capacity, quantization fills it with additional models. Total inference load per hardware unit increases. Organizations that hit their GPU ceiling begin provisioning more hardware to accommodate the expanded model fleet. The efficiency gain enables a scale of deployment that was not previously economically or logistically feasible, driving net compute consumption higher.

B. Case 2: Cheaper Inference APIs Drive Volume Explosion

Efficiency gain. OpenAI's frontier model input pricing fell 92% in eighteen months, from $30 per million tokens (GPT-4, early 2023) to $2.50 per million (GPT-4o, August 2024), with capable smaller models falling below $0.20 [16]. At fixed capability levels, inference prices fell by roughly 40× per year [17].

Mechanism. Applications that were previously uneconomical at $30 per million tokens became trivially affordable at $2.50 or less. Product teams that had rate-limited or capped AI features to control costs removed those restrictions. New applications—ambient AI features, background document processing, real-time content moderation, and increasingly autonomous agent workloads—became viable that did not exist before.

Rebound outcome. Volume growth overwhelmed the price decline. Google reported processing 480 trillion tokens per month across its products in May 2025 and 3.2 quadrillion tokens per month in May 2026—a sevenfold increase in a single year, during which per-token prices continued to fall [20]. This is the textbook Jevons mechanism: lower cost per unit drives higher total consumption.

C. Case 3: Open-Source Small Models Broaden Adoption

Efficiency gain. Llama 3 8B [13] and comparable open-source models run on a single consumer GPU (24 GB VRAM), enabling self-hosted AI inference without cloud API dependency.

Mechanism. Organizations that could not justify cloud LLM costs due to data privacy requirements, latency constraints, or budget limitations gained access to capable models through self-hosting. Healthcare providers, financial institutions, and government agencies that could

not send data to external APIs began deploying GPU infrastructure for the first time.

Rebound outcome. The total installed base of AI inference hardware expanded to include organizations and use cases that were previously excluded from the market. Open-source model releases did not redistribute existing compute demand—they created new demand from previously unserved segments. Total GPU infrastructure in production grew.

D. Case 4: Frontier Training Efficiency—The DeepSeek Shock

Efficiency gain. DeepSeek-V3 was trained for a reported $5.6 million in final-run compute [2], and DeepSeek-R1 matched OpenAI o1-class reasoning performance at a fraction of the assumed frontier training budget, with weights released openly [1].

Mechanism. The release was initially read as evidence that AI compute demand would fall: if frontier capability costs 10–20× less to train, the reasoning went, the installed base of GPUs was overbuilt. NVIDIA lost roughly $600 billion in market value in a single trading day in January 2025. Nadella's counterargument invoked this paper's subject directly: cheaper frontier capability would expand use, not shrink infrastructure [3].

Rebound outcome. The rebound prediction held. Open frontier-class weights spawned thousands of distilled variants and self-hosted deployments; reasoning-style inference—which consumes an order of magnitude more tokens per query than standard completion—became affordable enough to deploy broadly; and within a year the major cloud providers had raised, not lowered, their capital expenditure guidance (Section IV). The sharpest efficiency shock in the industry's history produced the largest subsequent investment in compute.

E. Case 5: AI Coding Assistants Reduce Dev Hours, Multiply Output

Efficiency gain. A controlled study by Peng et al. found that developers using GitHub Copilot completed tasks 55% faster than those not using it [21]. The magnitude is contested: a 2025 randomized controlled trial found experienced open-source developers were 19% slower with early-2025 AI tools on familiar codebases [22], indicating that productivity effects vary sharply with context.

Mechanism. The rebound argument does not depend on the precise productivity multiplier. Under either estimate, AI-assisted development changes what teams attempt and ship: more features, more prototypes, more services. Each additional feature requires backend services, data pipelines, storage, monitoring infrastructure, and ongoing compute to operate—and the assistant itself consumes inference on every keystroke. The AI assistant reduces (or redistributes) the human cost of building software; it does not reduce the infrastructure cost of running it.

Rebound outcome. Teams using AI coding assistants do not ship the same number of features faster—they ship more features, plus the inference load of the assistants themselves. Each feature adds to the cumulative infrastructure footprint. Total infrastructure spend grows in proportion to product output.

IV. The Cost Shift Thesis

The five cases in Section III share a common structure: efficiency gains reduce the cost of individual AI operations while stimulating demand sufficient to increase total compute consumption. We now develop a broader claim: AI efficiency gains do not reduce total organizational AI spend. They restructure it.

A. Pre-AI and Post-AI Cost Structures

Before widespread AI adoption, organizations bore heavy human labor costs for tasks that AI now automates: document review, data entry, customer support triage, code review, content moderation, report generation. These costs appeared as headcount, salaries, and contractor fees.

As AI systems absorb these tasks, headcount costs decline or grow more slowly. However, operating AI systems at scale requires compute infrastructure: GPU clusters, cloud inference endpoints, vector databases, embedding pipelines, monitoring systems. These costs are real, recurring, and scale with usage volume. Even before large-scale deployment, the compute cost of training frontier models is substantial—measured in megawatt-hours and millions of dollars [8], [9].

The total cost does not disappear. It transitions from a labor line item to a compute line item. Organizations that model AI ROI purely as labor savings will consistently underestimate total cost of ownership, because the compute cost of AI operations at production scale is not negligible.

B. The Efficiency-Adoption Feedback Loop

Efficiency improvements lower the barrier to AI adoption. Lower barriers drive broader adoption. Broader adoption increases total compute consumption. Increased compute consumption drives investment in infrastructure, which enables further efficiency improvements. This feedback loop is structurally identical to the mechanism Jevons identified in steam engine adoption.

C. Cost Visibility and Accounting

Labor costs are highly visible in organizational accounting: headcount is tracked, salaries are budgeted, contractor hours are invoiced. Compute costs are less visible by default: they appear as cloud bills with opaque line items, or as capital expenditure amortized over hardware lifecycles. This asymmetry means that the cost shift from labor to compute is often experienced as "AI savings" in the short term and "unexpected infrastructure costs" in the medium term.

Aggregate infrastructure investment is consistent with this view. The four largest hyperscalers—Amazon, Microsoft, Alphabet, and Meta—plan a combined $725 billion in capital expenditure for 2026, up 77% from approximately $410 billion in 2025, the large majority directed at AI infrastructure [23]. This investment surge occurred during the same period in which inference prices at fixed capability fell by 40× per year or more [17], and in the year following the industry's sharpest training-efficiency shock (Section III, Case 4). Record efficiency alongside record infrastructure investment is the aggregate signature of the rebound effect: volume growth more than offsets per-unit savings.

The cost-shift pattern has practical consequences for how organizations should plan and govern AI infrastructure, which we address next.

V. Implications

If the cost shift thesis holds, several common planning assumptions fail. We trace the consequences for enterprise strategy, infrastructure planning, and policy.

A. Enterprise AI Strategy

Organizations should budget for compute growth that is proportional to AI adoption, not inversely proportional. The common assumption—that AI efficiency improvements will reduce infrastructure spend over time—is not supported by historical patterns. AI efficiency improvements reduce per-unit costs while expanding the volume of AI operations, leaving net spend flat or growing.

AI FinOps practices must evolve accordingly. Tracking license costs and API subscriptions is insufficient; organizations need visibility into token consumption, GPU utilization, inference latency, and model deployment sprawl. Total cost of ownership analysis for AI initiatives should include multi-year compute projections at adoption-realistic scale, not benchmark scale.

B. Cloud and Infrastructure Planning

Cloud providers and data center operators should plan AI infrastructure capacity assuming continued demand growth, even as efficiency improves. The experience of energy utilities under energy efficiency programs is instructive: conservation mandates reduced per-unit consumption while total consumption continued rising due to population growth and new applications. AI efficiency mandates are unlikely to produce different results.

GPU supply chains, networking infrastructure, and data center cooling capacity should be planned for the demand levels that efficient, widely adopted AI systems will generate—not for the demand levels of today's less efficient, less adopted systems.

C. AI Governance and Policy

Regulatory frameworks that assume AI efficiency improvements will reduce AI resource consumption are likely to systematically underestimate demand. The International Energy Agency projects that data centre electricity consumption will more than double from 415 TWh in 2024 (about 1.5% of global electricity) to around 945 TWh by 2030, with AI the most important driver of that growth—a projection made in full awareness of ongoing efficiency gains [24]. Luccioni et al. identify the same dynamic as a central challenge in AI's environmental governance [7]. Compute governance proposals that rely on efficiency mandates as a primary lever for constraining AI resource use face the same structural challenge Jevons identified in 1865. The mandate achieves its narrow goal—lower per-unit use—while the broader goal of lower total use is undermined by the demand expansion that efficiency enables.

Effective AI resource governance likely requires demand-side interventions—caps, allocation systems, or pricing mechanisms—rather than relying on supply-side efficiency improvements to constrain total consumption.

Conclusion

The Jevons Paradox is not a historical curiosity. It is a structural property of resource systems in which efficiency improvements lower costs and stimulate demand, and AI infrastructure exhibits all the conditions for it to hold: per-unit costs falling by 40× per year, elastic demand, and a large population of potential adopters not yet served by current prices. For the organizations paying for AI, the operative consequence is the cost shift: efficiency does not shrink the AI bill, it moves the bill from labor line items to compute line items while total consumption grows. Practitioners who model AI value as labor savings alone will encounter compute budgets that grow despite each individual model becoming cheaper to run. The appropriate response is not to abandon efficiency work, but to plan infrastructure and budgets on the assumption that efficiency and adoption grow together—because for 160 years of resource economics and the entire deployment history of AI, they always have.

References

[1] DeepSeek-AI, "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," arXiv preprint arXiv:2501.12948, 2025.

[2] ——, "DeepSeek-V3 technical report," arXiv preprint arXiv:2412.19437, 2024.

[3] S. Nadella, "Jevons paradox strikes again! As AI gets more efficient and accessible, we will see its use skyrocket, turning it into a commodity we just can't get enough of," Post on X/LinkedIn, January 27, 2025, 2025, https://x.com/satyanadella/status/1883753899255046301.

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

[5] S. Sorrell, "Jevons' paradox revisited: The evidence for backfire from improved energy efficiency," Energy Policy, vol. 37, no. 4, pp. 1456–1469, 2009.

[6] B. Alcott, "Jevons' paradox," Ecological Economics, vol. 54, no. 1, pp. 9–21, 2005.

[7] A. S. Luccioni, E. Strubell, and K. Crawford, "From efficiency gains to rebound effects: The problem of Jevons' paradox in AI's polarized environmental debate," in Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, 2025.
[8] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in nlp," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3645–3650.
[9] D. Patterson, J. Gonzalez, Q. Le et al., "Carbon emissions and large neural network training," arXiv preprint arXiv:2104.10350, 2021.
[10] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Llm.int8(): 8-bit matrix multiplication for transformers at scale," arXiv preprint arXiv:2208.07339, 2022.
[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
[12] H. Touvron, L. Martin, K. Stone et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.
[13] A. Dubey, A. Jauhri, A. Pandey et al., "The llama 3 herd of models," arXiv preprint arXiv:2407.21783, 2024.
[14] A. Q. Jiang, A. Sablayrolles, A. Mensch et al., "Mistral 7b," arXiv preprint arXiv:2310.06825, 2023.
[15] J. Hoffmann, S. Borgeaud, A. Mensch et al., "Training compute-optimal large language models," arXiv preprint arXiv:2203.15556, 2022.
[16] OpenAI, "Openai api pricing," https://openai.com/api/pricing, 2024.
[17] Epoch AI, "LLM inference prices have fallen rapidly but unequally across tasks," https://epoch.ai/data-insights/llm-inference-price-trends, 2025.
[18] ——, "Compute trends across three eras of machine learning," https://epoch.ai/blog/compute-trends, 2023.
[19] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," arXiv preprint arXiv:2001.08361, 2020.
[20] S. Pichai, "Google I/O 2026 keynote," Google I/O, May 20, 2026, 2026, reported 3.2 quadrillion tokens processed monthly across Google products, a sevenfold year-over-year increase.
[21] S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer, "The impact of ai on developer productivity: Evidence from github copilot," arXiv preprint arXiv:2302.06590, 2023.
[22] J. Becker, N. Rush, E. Barnes, and D. Rein, "Measuring the impact of early-2025 AI on experienced open-source developer productivity," arXiv preprint arXiv:2507.09089, 2025.
[23] Tom's Hardware, "Google, microsoft, meta, and amazon capex spending to hit $725 billion in 2026, up 77% from last year," https://www.tomshardware.com/tech-industry/big-tech/big-techs-ai-spending-plans-reach-725-billion, 2026.
[24] International Energy Agency, "Energy and AI," IEA, Paris, Tech. Rep., 2025, https://www.iea.org/reports/energy-and-ai.