

The Covert Lever

How the artificial intelligence subsidy is being withdrawn, and who absorbs it

The question is not whether prices should rise. It is which instrument a seller reaches for when it decides to raise them, and what follows for a buyer who cannot observe the instrument being used.

TWO CORRECTIONS, OPPOSITE INCIDENCE

The correction everyone discusses would fall on capital. The correction actually occurring falls on adopters.

THE BUBBLE QUESTION

Whether the capital committed to model training and data center construction can be recovered. An asset price correction withdraws capital, lowers valuations, and lands on whoever holds the assets. A person who never invested is unaffected. It has not yet occurred and may not.

THE SUBSIDY WITHDRAWAL

What happens when the price a user pays for access approaches the cost of serving that user. It raises effective prices toward cost and lands on whoever built a dependency at the subsidized price. A person who never adopted is unaffected. It is already most of the way through its course.

No firm sustains an indefinite loss on its heaviest users, and whether prices should rise is not an interesting question. The interesting question is which instrument the seller reaches for — and what follows for a buyer who cannot observe the instrument being used.

METHOD, AND THE STARTING CONDITION STATED IN THE SELLER'S FAVOR

The subsidy is large, published, and not seriously contested. The paper's claim does not require the seller to be wrong about its costs.

The corpus

Every change between January 2025 and August 2026 that cut quantity at a given price, converted a subscription to metered billing, restricted signups, discontinued a tier, or altered the rate at which use consumes an allowance. Sellers named; the seller's own announcement is the source of record where it exists. Enterprise agreements, metered rate changes, and sellers with no consumer tier excluded. A lower bound, not a census — the change implemented without disclosure is the change least likely to be reported.

The magnitude

The flat plan loses money because of who uses it: the loss concentrates in a small share of subscribers whose consumption runs orders of magnitude above the median. For a developer consuming on the order of 140 million tokens a month against a \$200 plan, the implied subsidy runs in the low hundreds of dollars per user per month, an order of magnitude higher for the heaviest users. An interested counterparty put the Agent SDK subsidy at roughly fifteen to thirty times metered cost.

The clock

Aggregate capital expenditure across the major infrastructure providers has grown several times faster than operating cash flow, with the crossing point — combined free cash flow reaching zero — projected in the second half of 2026. Past that line the buildout runs on external capital, and the pricing question becomes a financing constraint. Sellers have said as much in their own voice.

None of this is hidden. It has been published, modeled, and argued in public for at least a year — a fact that matters later, because it determines the precise shape of the information asymmetry this paper alleges, which is narrower than the one usually asserted.

IV

The Lever, and Two Extensions

What the packaged goods literature knows, and where it stops

Consumers are four times as sensitive to a change in package price as to an equivalent change in package volume — and firms know it.

INVEST AND HARVEST, AND THE BUYER THE MODELS OMIT

The field models two buyers. One can negotiate. The other is, in the models, harvested — and has almost no literature, because there is nothing strategic to analyze about a party with no strategy available.

The pattern

Price below cost to acquire an installed base, above cost to harvest it. In the survey formulation: bargains followed by ripoffs. The consensus is that switching costs generally harm consumers by dampening price competition, deterring efficient entry, and narrowing variety.

Measured in the field

In Medicare Part D, premiums were consistent with insurers profiting from consumer inertia rather than cost, rose most in plans with the largest installed base, and cost the average inattentive consumer roughly \$1,050 over three years — under conditions considerably more favorable to the buyer than the ones considered here.

The stated boundary

The literature “focuses its analyses on passive individual buyers who behave as nonstrategic price takers.” That sentence is the hinge of this paper. A large buyer may be more locked in and still better positioned: it retains a procurement function, a legal department, alternative suppliers, and a credible threat. A solo practitioner has none of these.

Two findings cut against a naive version and belong in the body: toll free number portability showed larger contracts gained more from reduced switching costs, and with enough competing firms switching costs can be procompetitive. The response: this paper's claim concerns leverage and absorptive capacity, not the abstract degree of lock in.

THE ASYMMETRY, MEASURED

4×

as sensitive to a change in package price as to an equivalent change in package volume
— Chicago ice cream scanner data.

+8

percentage points of additional retail pass-through on downsized products. The covert instrument does not merely avoid resistance; it transmits cost more completely.

“Downsizing can be used as a hidden price increase to pass through rising input costs.”

The preference is durable rather than transitional: it does not attenuate over time, and in Kim's model even fully informed consumers would remain less sensitive to downsizing. The general form is obfuscation — sellers raising search costs deliberately to sustain margins, a lineage running back to Stigler — and where price separates into a visible headline and a hidden remainder, consumer protection requires constraining the hidden component directly. The European prohibition of drip pricing has rested on exactly that reasoning since 2011.

The seller's advantage does not survive discovery: consumers rate concealed downsizing as more unfair than the equivalent open increase, and abandon brands over it. The instrument trades a smaller present objection for a larger future one — conditional on the practice being found out. That conditionality is what the next two findings remove.

SOURCES Gourville & Koehler (2004); Çakır & Balagtas (2014); Kim (2023); Çakır (2021); Goedertier, Gorissen & Weijters (2026); Spiegler (2006, 2016); Stigler (1961); Petrikaitė (2018); Gu & Wenzel (2016); Inderst & Obradovits (2022).

QUANTITY REDUCTION AGAINST AN UNDISCLOSED BASELINE

A downsized grocery item still declares its quantity on the label. Several artificial intelligence subscriptions declare no quantity at all.

NOTION · LIVE 3 AUGUST 2026

The dashboard reports a percentage. The magnitude of which that percentage is a fraction appears nowhere in the documentation, which instead reserves the right to update “the size of the allowances, how they are calculated, and how they work.” A user can consume all of something without being told what the something was.

PERPLEXITY · NOV 2025 TO JUL 2026

Stated daily allowances converted into unstated weekly ones, without notice to subscribers who had prepaid annually against advertised terms; file uploads from unlimited to fifty per week; support staff reportedly unable to quantify the limits for users who asked. By July, the help documentation described the reductions as allocating more computing power per session “so that users receive deeper insights.”

Detection is impossible, not merely unlikely. The inattention the literature measures is inattention to a disclosed fact. No act of diligence is available when the fact is disclosed to nobody.

The corrections do not transfer. An index can be corrected only against a series of declared quantities. None exists, so the effective price increase is unmeasurable from outside the seller.

The backlash is foreclosed. The fairness cost is conditional on discovery. A practice that cannot be learned of does not carry it. The instrument's principal disadvantage has been engineered out.

SOURCES Imai & Watanabe (2014); Białek, Bobel & Oprych-Franków (2025); Notion (2026), *Manage your usage allowance for Notion AI*; The Covert Lever, Section VI.A (Perplexity sequence, reported rather than verified).

CONSUMPTION RATE REPRICING

The price holds. The allowance holds. What moves is the rate at which ordinary work consumes it — and the buyer sees only that the work runs out sooner.

approx. 1 Feb 2026	The Claude Code prompt cache default moves to the one hour cache (write price twice base input tokens, against 1.25× for five minutes).
approx. 6-7 Mar 2026	The default reverts to five minutes. No changelog entry, no deprecation notice, no API version bump. Any pause over five minutes in a long session now triggers full cache recreation — and these sessions are long and interrupted by construction.
11 Jan - 11 Apr 2026	A developer parses 119,866 of his own API calls across two machines and files issue 46829 on the seller's own repository: a 20 to 32 percent rise in cache creation costs, and quota exhaustion among subscribers who had never previously reached their limits. Trade coverage reports the company's position that the change "should not increase costs."
The remediation	A later release adds an environment variable restoring the one hour window — coupled, also without announcement, to telemetry: disabling telemetry removed the extended cache duration. In June 2026, Fable 5 produced the same effect by a different route, a more capable model consuming more tokens per unit of work against unchanged limits; the company adjusted its treatment the following month.

A posted price increase is legible to every buyer. A reduction against a stated allowance is legible to a diligent one. A reduction against an undisclosed baseline is legible to none. Consumption rate repricing is not merely illegible — there is no term anywhere in the buyer's contract or interface whose movement the buyer could look for.

THE COVERTNESS IS A CHOICE

“These quantities cannot be disclosed” fails on the record, because sellers in the same market and the same months disclosed anyway.

Z.ai

Approximate prompt limits per five hour window and per week, per tier — and a published schedule of peak and off peak quota multipliers, with a promotional reduction to a stated date. Consumption rate repricing, disclosed in advance and scheduled.

Alibaba Cloud

Publishes that each query consumes quota by the number of model calls it generates: simple tasks typically five to ten calls, complex tasks ten to thirty or more. A disclosure of the numerator's distribution rather than a promise about it — exactly what the technical objection says is impossible.

Windsurf

Published per plan message estimates, said plainly that consumption varies, held existing subscribers' prices while the list price rose, converted pre-purchased credits, and gave every paid subscriber a free week to test. Every element the opposite of the covert instrument, in the same month.

MiniMax

Publishes monthly token quotas by tier — a quantity term in the most literal sense available.

Uncertainty about consumption does not prevent disclosure. It requires the disclosure to be expressed as an estimate, a range, or a rate schedule rather than a guarantee — which is what these four sellers did. A percentage of an unstated magnitude, a stated allowance converted to an unstated one, a consumption rate moved without notice: each is a choice made against a demonstrated alternative.

SOURCES The Covert Lever, Section VI.C (Z.ai multiplier schedule and MiniMax quotas drawn from documentation aggregators; reported rather than verified); Alibaba Cloud (2026), *Coding plan*; Cognition (Mar. 18, 2026).

VII

The Record, 2025 and 2026

Same market, same cost pressure, same months, opposite choices

The chronology is a lower bound rather than a census.

THE PRECEDENT, AND THE WAVE

Nine sellers, eighteen months, one correction — administered through different instruments.

Cursor · Jun-Jul 2025	Pro converted to a \$20 usage pool at metered rates; overage charges; public apology on 4 July and refunds. The earliest reversal under user pressure, and the template the later cases are measured against.
Windsurf · Mar 2026	Credit pool replaced with daily and weekly quotas — with published message estimates, grandfathered prices, converted credits, and a free test week. Four days on, the docs still showed old pricing: a disclosure failure of a more ordinary kind.
Anthropic · Mar-Jun 2026	Quota windows exhausting early; postmortem, resets, and a credit campaign read as compensation; Claude Code removed from the \$20 tier; the silent cache reversion surfacing through issue 46829; May remediation doubling limits; the Agent SDK separate-credit change announced 13 May and paused on 15 June, its effective date.
GitHub · Apr-Jun 2026	New Copilot signups frozen, limits tightened, top model pulled from the paid individual tier; premium request units retired for usage-based AI Credits at one cent each, effective 1 June. Took effect as announced.
Perplexity, Notion, Alibaba Cloud, Z.ai, Moonshot	Unstated weekly limits; a percentage of an unpublished magnitude; a discontinued entry tier with daily restocked slots; weekly ceilings and a material price rise, with published multipliers and an apology credit; a capacity pause on new consumer subscriptions.

What the wave shows: open weight sellers rationed hosted access on the same timeline as closed ones, so the binding constraint is serving capacity rather than licensing philosophy. The instruments were not uniform. And the disclosure practice was not uniform either — Windsurf, Z.ai, Alibaba Cloud, and MiniMax published quantity terms; Notion, Perplexity, and Anthropic did not, or converted stated terms into unstated ones.

SOURCES Cognition (2026); Alibaba Cloud (2026); Claude Code issue 46829; The Register (2026); Anthropic (2026), *Use the Claude Agent SDK with your Claude plan*; The Covert Lever, Section VII (entries resting on trade coverage marked reported rather than verified).

PRICE CHANGE LEGIBILITY

Legibility is necessary for organized objection. It is not sufficient for reversal.

CHANGE	COULD BUYERS COMPUTE IT?	OUTCOME
Cursor, June 2025	Yes — announced as a change in how usage would be charged; subscribers computed what it meant.	Objection; apology; refunds.
Agent SDK, May 2026	Yes — surfaces, effective date, and dollar credit per plan all stated; vendors computed exposure in public.	Withdrawn on its effective date.
GitHub, April 2026	Yes — units retired, credits at a stated one cent rate, stated date.	Proceeded on schedule.
Perplexity; Notion	No — stated allowances converted to unstated ones; a percentage of an unpublished magnitude.	Proceeded; neither reversed.

Observational: five events, five sellers. It cannot separate legibility from seller size, community organization, or competitive position — and the developer population that objected in May 2026 is unusually capable of quantifying its own exposure, which is part of the mechanism rather than a confound to be regretted. Prediction Two is written to test the regularity.

WHY THIS BUYER CANNOT SELF PROTECT

Cannot negotiate, cannot exit, cannot measure, cannot forecast.

No leverage

Standard advice — use renewed negotiating position to redesign terms around shared risk — presupposes a procurement function, a contract large enough to interest the seller, and a credible alternative. A sole practitioner clicking a consumer checkout page has none of the three.

No substitute

A finding of the chronology, not an assumption: open weight sellers rationed too, so the constraint is serving capacity. Free weights do not make the hardware free, and the only escape — running weights on owned hardware — reintroduces capital cost, skill, and capability limits this population is least equipped to absorb.

No quantity term

No disclosed magnitude to check — and in the consumption rate case, no term anywhere in the contract or interface whose movement the buyer could look for.

No forecastability

A property of the technology: one instruction can trigger many model calls behind the interface, and sellers differ in whether they meter messages or requests. This holds even under perfect disclosure of the allowance, because it concerns the numerator rather than the denominator.

That is not a market participant making a poor decision. It is the condition under which no decision procedure available to the party would have produced a different outcome.

THE ASYMMETRY, STATED PRECISELY

The one figure necessary to price one's own exposure is the one figure withheld, while the general warning sits in the open and does the work of appearing to have informed everybody.

THE RECRUITMENT

On 13 May 2026, Claude for Small Business launched: small businesses as 44 percent of GDP, a ten city training tour with free workshops, a month of the highest consumer tier for attendees, partnerships with three community development financial institutions. Nothing in that is false, and the training is a real transfer of value.

The point is structural and narrow: the launch came six weeks after the same company removed its coding agent from its entry tier, one month after it shortened the prompt cache, three weeks after a major competitor froze individual signups, and on the same day as the Agent SDK announcement. Entrants were brought in at the moment of a documented correction, by a seller who could compute the correction, and placed where they could not compute it.

THE MECHANISM REACHES LITIGATION

Kahn v. Anthropic, PBC, filed 14 June 2026 in the Northern District of California, breach of contract. The allegations are unproven, and nothing here is a finding. What matters is the shape: actual usage alleged to fall far below the advertised amount, real caps “nearly impossible to pin down,” changed without clear notice — and arithmetic computed from the seller's own July 2025 emails, the one occasion in this record on which something close to a quantity term was published, suggesting Max 5x delivered roughly 3.5 times Pro's hours rather than five, and Max 20x roughly six rather than twenty.

That complicates the remedy: disclosure is necessary for a buyer to price an exposure, but a published figure inconsistent with the marketing it accompanies relocates the problem rather than solving it.

LIMITATIONS, AND FIVE PREDICTIONS WRITTEN TO FAIL

The record understates the covert instruments by construction, and the paper sizes nothing it cannot see.

LIMITATIONS

Both consumption rate instances reached the record because sophisticated users parsed their own logs — not a method available to this population, so the record is biased toward the visible instrument. The magnitude is unmeasurable from outside under current disclosure; a reader who wants a number should be suspicious of anyone who supplies one from public sources. What could be built: a longitudinal audit of published terms at fixed captures, coded on whether a quantity term exists, in what unit, and whether it moved. Two corrected errors are recorded in the paper itself: the discontinued tier is Alibaba Cloud's, not Z.ai's, and the cache reversion dates to early March, fixed by the issue tracker record.

PREDICTIONS

One. Posted prices at prevailing tiers stay roughly flat through 2027 while effective quantity per dollar declines.

Two. Legibility necessary but not sufficient: no reversal follows a change buyers could not compute; some computable changes proceed anyway.

Three. Consumption rate repricing becomes more common than allowance reduction — cheaper, unannounced, detectable only by self-instrumenting users.

Four. At least one further open weight seller rations, pauses, or discontinues hosted access during a surge before the end of 2027.

Five. Disclosure regulation arrives before price regulation, first in the European Union, on the drip pricing precedent.

THREE REMEDIES, NONE TOUCHING PRICE

A seller that must raise the effective price should be able to. What is at issue is whether the buyer can see it happen.

Quantity term disclosure

The magnitude, in whatever unit the seller meters, and notice when it changes. Feasible, because sellers in this market already do it as estimates, ranges, and rate schedules. With a consistency condition: the disclosed quantity and the advertised comparison must be commensurable, or the disclosure relocates the problem.

Notice for consumption rate changes

The one instrument no disclosure regime currently reaches, because nothing the buyer was told has changed. A cache lifetime, a default model, and a metering rule are terms of the bargain in substance; a change to any of them is a change to the price, and should carry the same notice as an allowance change.

Portability

Export of conversation history, configuration, and stored context as a default capability. Constrains no seller's pricing; lowers the switching cost on which the entire invest and harvest dynamic depends — the intervention the literature has recommended since 1995.

A buyer told neither the size of what they purchase nor when a consumption rule moves is left to infer a transfer from the fact that the work stopped getting done. Some sellers chose disclosure under the same cost pressure. The rest chose the second condition — and the first group is the reason we know that was a choice.

References

Travis Gilly, *The Covert Lever: How the Artificial Intelligence Subsidy Is Being Withdrawn, and Who Absorbs It*, working paper, July 2026 (v0.5). Complete list of 38 entries, in the order the paper gives it, across 3 slides.

Anthropic. (2026). *Use the Claude Agent SDK with your Claude plan* [Help Center article, updated June 16, 2026]. <https://support.claude.com/en/articles/15036540-use-the-claude-agent-sdk-with-your-claude-plan>

Arize AI. (2026, July 1). *Model subsidies are ending. What do you do now?* <https://arize.com/blog/ai-model-subsidies-ending-llm-inference-costs/>

Ayinaddis, S. G. (2025). Artificial intelligence adoption dynamics and knowledge in SMEs and large firms: A systematic review and bibliometric analysis. *Journal of Innovation & Knowledge*.

Badghish, S., et al. (2024). Artificial intelligence adoption by SMEs to achieve sustainable business performance: Application of technology-organization-environment framework. *Sustainability*.

Białek, J., Bobel, A., & Oprych-Franków, D. (2025). Automatic downsizing and upsizing detection using scanner data and their impact on price indices. *Statistical Journal of the IAOS*, 41(3), 858-872. <https://doi.org/10.1177/18747655251342654>

Broda, E. (2026, May 14). *The brutal reality of token economics*. AgenticMesh.

Çakır, M. (2021). Retail pass-through of package downsizing. *Agribusiness*, 38(2), 259-278. <https://doi.org/10.1002/agr.21724>

Çakır, M., & Balagtas, J. V. (2014). Consumer response to package downsizing: Evidence from the Chicago ice cream market. *Journal of Retailing*. <https://doi.org/10.1016/j.jretai.2013.06.002>

Dinh, T., et al. (2025). Artificial intelligence in SMEs: Enhancing business functions through technologies and applications. *Information*.

Farrell, J., & Klemperer, P. (2007). Coordination and lock-in: Competition with switching costs and network effects. In M. Armstrong & R. Porter (Eds.), *Handbook of Industrial Organization* (Vol. 3, pp. 1967-2072). Elsevier.

Goedertier, F., Gorissen, K., & Weijters, B. (2026). How downsizing a typical product at a stable unit price recalibrates price perceptions. *Psychology & Marketing*, 43(7), 1654-1674. <https://doi.org/10.1002/mar.70134>

Gourville, J. T., & Koehler, J. J. (2004). *Downsizing price increases: A greater sensitivity to price than quantity in consumer markets*. <https://doi.org/10.2139/ssrn.559482>

References, continued

- Gu, Y., & Wenzel, T. (2016). Consumer confusion, obfuscation and price regulation. *Scottish Journal of Political Economy*, 64(2), 169-190. <https://doi.org/10.1111/sjpe.12121>
- Ho, K., Hogan, J., & Scott Morton, F. (2017). The impact of consumer inattention on insurer pricing in the Medicare Part D program. *The RAND Journal of Economics*, 48(4), 877-905. <https://doi.org/10.1111/1756-2171.12207>
- Imai, S., & Watanabe, T. (2014). Product downsizing and hidden price increases: Evidence from Japan's deflationary period. *Asian Economic Policy Review*, 9(1), 69-89. <https://doi.org/10.1111/aep.12047>
- Inderst, R., & Obradovits, M. (2022). Excessive competition on headline prices. *International Economic Review*, 64(2), 783-808. <https://doi.org/10.1111/iere.12616>
- Kahn v. Anthropic, PBC, No. 3:26-cv-05763 (N.D. Cal. filed June 14, 2026). Docket retrieved from CourtListener, <https://www.courtlistener.com/docket/73482575/kahn-v-anthropic-pbc/>
- Claude Code issue 46829. (2026). Cache TTL silently regressed from 1h to 5m around early March 2026, causing quota and cost inflation. <https://github.com/anthropics/claude-code/issues/46829>
- Kim, I. K. (2023). Consumers' preference for downsizing over package price increases. *Journal of Economics & Management Strategy*, 33(1), 25-52. <https://doi.org/10.1111/jems.12548>
- Alibaba Cloud. (2026). Coding plan [Model Studio documentation]. <https://www.alibabacloud.com/help/en/model-studio/coding-plan>
- Cognition. (2026, March 18). Introducing our new Windsurf pricing plans. <https://devin.ai/blog/windsurf-pricing-plans/>
- The Register. (2026, April 13). Anthropic: Claude quota drain not caused by cache tweaks. https://www.theregister.com/2026/04/13/claude_code_cache_confusion/
- Klemperer, P. (1987). Markets with consumer switching costs. *The Quarterly Journal of Economics*, 102(2), 375-394.
- Klemperer, P. (1995). Competition when consumers have switching costs. *The Review of Economic Studies*, 62(4), 515-539.
- Lewis, T. R., & Yildirim, H. (2005). Managing switching costs in multiperiod procurements with strategic buyers. *International Economic Review*, 46(4), 1233-1269. <https://doi.org/10.1111/j.1468-2354.2005.00366.x>

References, concluded

- Nield, D. (2026). *Prepare to expect less from your cheap AI subscription*. PCWorld.
- Notion. (2026). *Manage your usage allowance for Notion AI* [Help Center article].
<https://www.notion.com/help/manage-your-usage-allowance-for-notion-ai>
- Pearcy, J. (2015). Bargains followed by bargains: When switching costs make markets more competitive. *Journal of Economics & Management Strategy*, 25(4), 826-851.
<https://doi.org/10.1111/jems.12158>
- Petríkaitė, V. (2018). Consumer obfuscation by a multiproduct firm. *The RAND Journal of Economics*, 49(1), 206-223. <https://doi.org/10.1111/1756-2171.12225>
- Rajaram, K., et al. (2024). *Generative artificial intelligence in small and medium enterprises: Navigating its promises and challenges*. arXiv.
- Sánchez, E., et al. (2025). Artificial intelligence adoption in SMEs: Survey based on TOE-DOI framework, primary methodology and challenges. *Applied Sciences*.
- Schmidt, R. C. (2010). On the value of a large customer base in markets with switching costs. *The Journal of Industrial Economics*, 58(3), 627-641.
<https://doi.org/10.1111/j.1467-6451.2010.00427.x>
- Soomro, R., et al. (2025). A SEM-ANN analysis to examine impact of artificial intelligence technologies on sustainable performance of SMEs. *Scientific Reports*.
- Spiegler, R. (2006). Competition over agents with boundedly rational expectations. *Theoretical Economics*, 1(2), 207-231.
- Spiegler, R. (2016). Choice complexity and market competition. *Annual Review of Economics*, 8, 1-25.
- Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3), 213-225.
- Viard, V. B. (2008). Do switching costs make markets more or less competitive? The case of 800-number portability. *The RAND Journal of Economics*, 38(1), 146-163.
<https://doi.org/10.1111/j.1756-2171.2007.tb00049.x>
- Zavodna, L. S., et al. (2024). Barriers to the implementation of artificial intelligence in small and medium-sized enterprises: Pilot study. *Journal of Economics and Management*.