

The Covert Lever:

How the Artificial Intelligence Subsidy Is Being Withdrawn, and Who Absorbs It

Travis Gilly

ORCID: 0009-0007-2954-6313

Real Safety AI Foundation

Illinois, United States

♣ Working paper, July 2026 (v0.5)

ABSTRACT

Switching costs, consumer inattention, and shrinkflation together predict how a seller will pass a cost increase to a locked in buyer: by reducing the quantity supplied rather than raising the posted price, because consumers are measurably less sensitive to the first. This paper applies that prediction to the withdrawal of the flat rate subsidy in artificial intelligence services across 2025 and 2026, and reports three findings the packaged goods literature does not contain. First, a limiting case. A downsized grocery item still declares its quantity on the label, which is what makes downsizing measurable and what allows a price index to be corrected for it. Notion's usage allowance, live 3 August 2026, presents consumption as a percentage of a magnitude the company does not publish and reserves the right to change; Perplexity converted stated daily search allowances into unstated weekly ones. Quantity reduction against an undisclosed baseline is not merely unnoticed, as shrinkflation is; it is undetectable in principle and no series exists to correct. Second, a third instrument with no packaged goods analogue, in which the posted price and the stated allowance both hold and the rate at which ordinary work consumes the allowance moves. Anthropic reverted the Claude Code prompt cache default from one hour to five minutes in early March 2026 with no changelog entry, a change reconstructed by a user parsing 119,866 of his own API calls and associated with a 20 to 32 percent rise in cache creation cost, which the company then publicly said should not increase costs. Third, and decisively, the covertness is a choice rather than a property of the good. In the same market and the same months, Windsurf published per plan message estimates and grandfathered existing subscribers' prices through a structural change, Z.ai publishes approximate prompt limits per window and a schedule of peak and off peak quota multipliers, and Alibaba Cloud publishes the range of model calls a query consumes. The technical objection that these quantities cannot be disclosed is answered by sellers who disclose them. The paper assembles a chronology across eight sellers, tests whether the legibility of a change predicts its reversal and finds legibility necessary for organized objection but not sufficient for reversal with GitHub's fully disclosed June 2026 credit transition as the counterexample, and closes on Kahn v. Anthropic, PBC, a proposed class action filed 14 June 2026 alleging that Claude Max usage limits are hard to determine and were changed without clear notice. Limitations, five falsifiable predictions, and three disclosure remedies are set out.

Keywords: industrial organization; consumer protection; switching costs; invest and harvest pricing; shrinkflation; consumer inattention; quantity reduction against an undisclosed baseline; price change legibility; consumption rate repricing; small and medium enterprises; solo practitioners; quantity disclosure

JEL codes: D18; D82; D83; L11; L15; L86; M31

I. INTRODUCTION

Two arguments about the economics of artificial intelligence are running at once and are being confused for one another. The first concerns a possible asset price correction, generally called the bubble question, and it asks whether the capital committed to model training and data center construction can be recovered. The second concerns the withdrawal of a consumer subsidy, and it asks what happens when the price a user pays for access begins to approach the cost of serving that user.

These are not the same event and their incidence runs in opposite directions. An asset price correction withdraws capital, lowers valuations, and lands on whoever holds the assets. A person who never invested is unaffected. A subsidy withdrawal raises effective prices toward cost and lands on whoever built a dependency at the subsidized price. A person who never adopted is unaffected. Public attention is directed almost entirely at the first, which has not yet occurred and may not. The second is already most of the way through its course, and this paper is about it.

The withdrawal itself is not in dispute and this paper does not argue against it. The arithmetic of the subsidy has been published and is not seriously contested. No firm sustains an indefinite loss on its heaviest users, and the question of whether prices should rise is not an interesting one. The interesting question is which instrument a seller reaches for when it decides to raise them, and what follows from that choice for a buyer who cannot observe the instrument being used.

The argument proceeds in four moves. First, the choice of instrument is not neutral and is well studied outside this market: consumers respond less to quantity reductions than to equivalent price increases, and firms know it. Second, artificial intelligence subscriptions extend that practice past the point where the existing literature can follow, because several sellers publish no quantity term at all. Third, this market has a third instrument the packaged goods literature has no case for, in which price and stated allowance both hold and the rate of consumption against the allowance moves. Fourth, the population absorbing all of this sits in the gap between two literatures, one of which states plainly that it does not model such buyers and the other of which stops its analysis at the moment the exposure begins.

The paper sets out its limitations, offers five predictions written so that they can fail, and proposes three remedies aimed at disclosure rather than at price.

II. METHOD AND CORPUS

The theoretical material in Sections IV through VI is drawn from the published economics literature on switching costs, package downsizing, and obfuscation, selected by following citation chains outward from the surveys in each area rather than by systematic protocol. Where a finding reaches this paper through a citing paper rather than the original, the text says so.

The chronology in Section VII is a compiled record. An event was included where a seller of an artificial intelligence service offered on a recurring subscription made a change between January 2025 and August 2026 that reduced the service quantity available at a given price, converted a subscription surface to metered billing, restricted new subscriptions, discontinued a tier, or altered the rate at which ordinary use consumes an

allowance. Sellers are named. Where a seller's own announcement, help center page, or docket entry could be retrieved, it is the source of record and the citation points to it; where only trade coverage exists, the text says so and the entry is marked as reported rather than verified.

Three exclusions apply. Enterprise agreements are excluded, because their terms are negotiated and unpublished and no outside party can code them. Changes to published per unit metered rates are excluded, because such a change is by construction visible to the buyer and is not the instrument this paper is about. Sellers with no consumer or small business subscription tier are excluded.

Two limits follow and are carried into Section XI. The record is drawn from English language sources and is weighted toward sellers with large developer communities, because those are the events that generated coverage and, in one instance, litigation. Absence from this chronology is therefore not evidence that a seller made no change; it is evidence that no change was reported. That asymmetry runs in the direction of the paper's own thesis, since a change implemented without disclosure is precisely the change least likely to be reported. The chronology is a lower bound rather than a census.

III. THE SUBSIDY AND ITS MAGNITUDE

The starting condition should be stated in the seller's favor, because the paper's claim does not require the seller to be wrong about its costs.

Subscription and metered access to the same models are different businesses with different economics, and only the first has been running below cost. Metered access is priced against consumption and does not carry the exposure. A unit of inference is not itself expensive to serve. The flat plan loses money because of who uses it and how hard they use it, so the loss concentrates in a small share of subscribers whose consumption runs orders of magnitude above the median (Arize, 2026).

Estimates of the per user subsidy on heavy coding plans have circulated widely enough to constitute a shared premise in the trade literature. Broda (2026) works through the arithmetic for a developer consuming on the order of one hundred and forty million tokens per month against a two hundred dollar plan and arrives at an implied provider subsidy in the low hundreds of dollars per user per month, rising by an order of magnitude for the heaviest users in the sample. A second estimate arrives from an interested party on the other side of the transaction: writing after Anthropic's May 2026 Agent SDK announcement, the editor tooling vendor Zed put the subsidy on Agent SDK usage at roughly fifteen to thirty times the equivalent metered cost.

The timing pressure is structural rather than discretionary. Aggregate capital expenditure across the major infrastructure providers has been growing several times faster than operating cash flow, and projections place the crossing point, where combined free cash flow reaches zero, in the second half of 2026 (Arize, 2026). Past that line the buildout runs on external capital, which converts the pricing question from a strategic preference into a financing constraint.

Sellers have said as much in their own voice. Announcing the retirement of premium request units in April 2026, GitHub described agentic usage as becoming the default and driving higher compute and

inference demand. Anthropic's head of Claude Code, explaining an earlier restriction on third party harnesses, said that the company's subscriptions were not built for the usage patterns of those tools.

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

IV. SWITCHING COSTS AND THE BUYER THE MODELS OMIT

The pattern of pricing below cost to acquire an installed base and above cost to harvest it is among the oldest results in the economics of switching costs. Klemperer (1987, 1995) established the two period structure and Farrell and Klemperer (2007) consolidated the field. The seller invests, by pricing low to build market share, and then harvests, by raising price on customers for whom switching now carries a cost. In the survey formulation the sequence is bargains followed by ripoffs, and the consensus finding is that switching costs generally, though not universally, harm consumers by dampening price competition, deterring efficient entry, and narrowing product variety (Lewis and Yildirim, 2005, summarizing Farrell and Klemperer).

The pattern has been confirmed outside of theory. Ho, Hogan, and Scott Morton (2017) study the Medicare Part D prescription drug market, where consumers switch plans infrequently and search imperfectly. Estimating a model of plan choice with inattentive consumers, they find that observed premiums are consistent with insurers profiting from consumer inertia rather than from underlying cost. Premiums rose across the study period, and rose most in the plans with the largest installed base and therefore the strongest incentive to harvest. Their counterfactual places the cost to an average consumer of that inattention at roughly one thousand and fifty dollars over three years. This is the invest and harvest dynamic measured in a regulated market with a public enrollment infrastructure, which is to say under conditions considerably more favorable to the buyer than the ones considered here.

The literature also contains, in its own words, the boundary of what it has studied. Lewis and Yildirim (2005) open their analysis of strategic procurement by observing that a buyer capable of affecting market outcomes will act to avoid or manage switching costs, and that this possibility is not addressed by the extant switching cost literature, which focuses its analyses on passive individual buyers who behave as nonstrategic price takers. That sentence is the hinge of the present paper. The field contains two buyer types. One is large enough to negotiate, to credibly threaten replacement of an incumbent supplier, and to extract concessions from the threat. The other is modeled as a price taker and is, in the models, harvested. The first has an extensive literature. The second has almost none, because from a modeling standpoint there is nothing strategic to analyze about a party with no strategy available to it.

Two findings cut against a naive version of the argument and belong in the body rather than a footnote. Viard (2008), studying toll free number portability, found that prices on larger contracts fell more when switching costs were reduced, which is consistent with larger buyers having been more locked in, not less. Percy (2015) shows that with a sufficiently large number of competing firms, switching costs can be procompetitive, and Schmidt (2010) shows that a firm may not benefit from an installed base beyond half the market. The response is that this paper's claim concerns leverage and absorptive capacity rather than

the degree of lock in. A large buyer may be more locked in and still be better positioned, because it retains a procurement function, a legal department, alternative suppliers under contract, and the ability to make a threat the seller has reason to believe. A solo practitioner has none of these and is not made whole by having a lower switching cost in the abstract.

V. THE COVERT LEVER

A firm passing through a cost increase has two classical instruments. It can raise the posted price for an unchanged quantity, or it can reduce the quantity supplied at an unchanged price. The two are arithmetically equivalent in unit terms and are not equivalent in consumer response, and the asymmetry has been measured.

Gourville and Koehler (2004) show, across marketplace and experimental data, that consumers are more sensitive to changes in price than to equivalent changes in quantity, and that a seller therefore encounters less resistance holding the price tag and shrinking the package than holding the package and raising the price. Working with scanner data from the Chicago ice cream market, Çakır and Balagtas (2014) estimate the two elasticities directly and find consumers on average four times as sensitive to a change in package price as to an equivalent change in package volume. Their own conclusion states the managerial implication without euphemism: downsizing can be used as a hidden price increase to pass through rising input costs.

The preference is durable rather than transitional. Kim (2023), exploiting divergent responses by South Korean milk manufacturers to a 2018 cost shock, finds that consumers strongly prefer downsizing to package price increases and that the preference does not attenuate over time, and develops a model in which even fully informed consumers would remain less sensitive to downsizing. Çakır (2021) finds that downsized products carry a retail pass through rate roughly eight percentage points higher than nondownsized products, so the covert instrument does not merely avoid resistance, it transmits cost more completely.

The seller's advantage does not survive discovery. Evangelidis, as reported and synthesized by Goedertier, Gorissen, and Weijters (2026), finds that consumers rate downsizing without a price adjustment as more unfair than an equivalent price increase would have been, and the broader packaging literature documents brand abandonment following discovery. The instrument therefore trades a smaller present objection for a larger future one, conditional on the practice being found out. That conditionality is what makes Sections VI and VII load bearing.

The general form of this behavior is obfuscation, and it has its own literature. Sellers raise consumer search costs deliberately in order to sustain margins (Spiegler, 2006, 2016), a lineage running back to the observation that price search is costly in the first place (Stigler, 1961). The mechanisms studied include offering incommensurable package configurations across retailers, requiring consumers to work through option trees before a price is visible, and disguising basket level increases behind many small movements (Petrikaitė, 2018; Gu and Wenzel, 2016). Where price components separate into a visible headline and a less visible remainder, competition on the headline can be intense while the remainder is where the margin lives, and consumer protection then requires constraining the hidden component directly rather than trusting competition to do it (Inderst and Obradovits, 2022). The prohibition of drip pricing in European consumer law dates to 2011 and rests on exactly that reasoning.

VI. TWO EXTENSIONS THE PACKAGED GOODS LITERATURE DOES NOT CONTAIN

A. *The Undisclosed Baseline*

Every empirical result in the preceding section depends on a condition that has gone unremarked because in packaged goods it is always satisfied. The quantity is printed on the package.

That declaration is what makes downsizing a measurable object. Imai and Watanabe (2014), working with scanner data from roughly two hundred Japanese supermarkets over a decade, could establish that about one third of product replacements involved a reduction in size or weight, and could estimate that a one percentage point larger size reduction was associated with a price decline of roughly half that magnitude, producing an effective price increase. They could then argue that the consumer price index was biased downward as a result. Statistical agencies have since built automated procedures to detect downsizing and upsizing in scanner data precisely so the index can be corrected (Białek, Bobel, and Oprych Franków, 2025). All of this rests on grams being comparable to grams.

Several artificial intelligence subscriptions declare no quantity at all.

Notion's usage allowance, which took effect on 3 August 2026 for its Business and Enterprise plans, is the clearest instance. The company's help documentation describes two windows, a rolling six hour window and a monthly billing cycle window, and directs the user to a dashboard to observe consumption. The dashboard reports a percentage. The magnitude of which that percentage is a fraction appears nowhere in the documentation, which states instead that Notion may update how these allowances work, including the features impacted, the size of the allowances, how they are calculated, and how they work alongside Notion credits. A user can therefore observe that they have consumed all of something without being told what the something was, and without assurance that the unstated quantity will be the same next month.

Perplexity supplies the other form, which is conversion of a stated quantity into an unstated one. Between November 2025 and February 2026 the company reduced Pro plan limits, changed the underlying models, and redesigned the interface without notifying subscribers who had prepaid annually against previously advertised terms. File uploads moved from unlimited to fifty per week. Pro Search moved from stated daily allowances to weekly limits which, on contemporaneous reporting, the company's own support staff said they could not quantify for users who asked. A further reduction was reported in mid May 2026. By July 2026 Perplexity's help documentation described the adjustment of usage limits as a means of allocating more computing power per session so that users receive deeper insights.

The consequence is a difference in kind rather than degree, in three parts.

First, detection is impossible rather than merely unlikely. The consumer inattention the downsizing literature measures is inattention to a disclosed fact. A shopper who reads the label sees the change. No equivalent act of diligence is available to a subscriber, because the fact is disclosed to nobody.

Second, the correction methods do not transfer. Imai and Watanabe could correct an index because a series of declared quantities existed to correct against. No such series exists here, so the magnitude of the effective price increase across this market is not merely unmeasured, it is unmeasurable from outside the

seller under current disclosure practice.

Third, the fairness backlash the packaged goods literature identifies as the seller's constraint is conditional on discovery, and discovery is what has been foreclosed. If consumers rate concealed downsizing as more unfair than an open price increase once they learn of it, then a practice that cannot be learned of does not carry that cost. The instrument's principal disadvantage has been engineered out.

I use the term quantity reduction against an undisclosed baseline for this practice, in preference to extending the word shrinkflation, because shrinkflation names a reduction in a stated quantity and the stated quantity is exactly what is absent.

B. Consumption Rate Repricing

There is a third instrument, and it has no packaged goods analogue, because it requires a good whose consumption is measured by the seller rather than counted by the buyer.

The posted price holds. The stated allowance, where one exists, holds. What moves is the rate at which ordinary work consumes the allowance. The buyer sees no announcement, because nothing the buyer was told has changed. The buyer sees only that the work runs out sooner.

The clearest instance is documented in the seller's own public issue tracker. Anthropic's prompt caching offers two durations, five minutes at a write price of one and a quarter times base input tokens and one hour at twice base. Around 1 February 2026 the Claude Code default moved to the one hour cache. Around 6 or 7 March 2026 it reverted to five minutes. There was no changelog entry, no deprecation notice, and no API version bump.

The mechanism is arithmetic. Cached context that survives is cheap to reuse. Cached context that has expired must be written again, so under a five minute default any pause longer than five minutes in a long session triggers a full cache recreation. Claude Code sessions are long and interrupted by construction.

The change was reconstructed, not announced. A developer parsed 119,866 Claude Code API calls across two machines from 11 January to 11 April 2026 and filed the result as issue 46829 on Anthropic's own repository, reporting a 20 to 32 percent increase in cache creation costs and a measurable rise in quota consumption among subscription users who had never previously reached their limits, including himself. Trade coverage in April reported the company's position that the change should not increase costs, against users reporting faster depleting quotas. A subsequent Claude Code release introduced an environment variable restoring the one hour window for users who set it. Reporting also identified a coupling that had not been announced either: disabling telemetry removed the extended cache duration at the same time.

Every element of the pattern this paper describes is present in that one episode. The posted price did not move. The stated usage limits did not move. A parameter neither published in the plan nor visible in the interface moved, the effective quantity fell by a fifth to a third for affected workloads, the buyer could learn of it only by parsing months of their own logs, and the seller's own account of the effect differed from the buyers'. That last point is the paper's epistemic problem in miniature: where a term is undisclosed, a dispute about whether it moved cannot be settled by the party who paid for it.

In June 2026 the arrival of Fable 5 produced a similar effect by a different route, a more capable model consuming more tokens for a given unit of work against usage limits that had not changed, so that the same subscription bought less. Anthropic adjusted the treatment of Fable 5 against weekly limits the following month.

I use the term consumption rate repricing for this instrument. It stands at the far end of a spectrum. A posted price increase is legible to every buyer. A quantity reduction against a stated allowance is legible to a diligent one. A quantity 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.

C. The Covertness Is a Choice

The obvious defense of every practice described above is that these quantities cannot be disclosed. Consumption depends on task complexity, on context length, on which model handled the request, and on how many calls a single instruction generates behind the interface. A seller cannot promise a number it does not control.

The defense fails on the record, because sellers in the same market and the same months disclosed anyway.

Z.ai publishes, for its coding subscription, approximate prompt limits per five hour window and per week for each tier, together with separate monthly quotas for connector usage. It also publishes the thing this paper argues is hardest to see: a schedule of quota multipliers, under which its flagship models consume at a stated multiple during a stated daily peak window and a lower multiple outside it, with a promotional reduction to parity running to a stated date. That is consumption rate repricing, disclosed in advance and scheduled, in the same market and the same quarter in which another seller changed a cache lifetime with no changelog entry at all.

Alibaba Cloud publishes, for its coding plan, that each query consumes quota according to the number of model calls it generates, that simple tasks typically use five to ten calls and complex tasks ten to thirty or more, and that actual consumption varies by task complexity, context, and tool usage. That is a disclosure of the numerator's distribution rather than a promise about it, which is exactly what the technical objection says is impossible.

Windsurf, announcing on 18 March 2026 the replacement of its credit pool with daily and weekly quotas effective the following day, published estimates of the number of messages a subscriber could expect under each plan, stated plainly that consumption would vary with task complexity and model choice, held existing subscribers' prices unchanged while the list price for new subscribers rose, converted pre-purchased credits into usage under the new system, and gave every existing paid subscriber an additional free week to test the new arrangement before committing. Every element of that is the opposite of the instrument this paper describes, executed by a seller under the same cost pressure in the same month.

MiniMax publishes monthly token quotas by tier, which is a quantity term in the most literal sense available.

The conclusion is narrow and it is the paper’s central point. 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 as a guarantee, which is what these four sellers did. Where a seller instead publishes a percentage of an unstated magnitude, converts a stated allowance into an unstated one, or moves a consumption rate without notice, that is a choice made against a demonstrated alternative.

VII. THE RECORD, 2025 AND 2026

What follows is the compiled record described in Section II.

A. The Precedent: Cursor, June and July 2025

On 16 June 2025 Anysphere changed the Cursor Pro plan from an allotment of five hundred fast requests plus slower unlimited use to a twenty dollar monthly usage pool consumed at rates equivalent to metered pricing. Subscribers who had budgeted against the earlier structure encountered overage charges. On 4 July 2025 the chief executive published a public apology acknowledging that the rollout had been communicated poorly, and the company refunded unexpected charges incurred between 16 June and 4 July. This is the earliest instance in the record of a seller reversing course under user pressure, and it is the template the later cases are measured against.

B. The Wave, March 2026 to August 2026

Windsurf, March 2026. Announced 18 March and effective 19 March, the monthly credit pool was replaced with daily and weekly quotas across the Free, Pro, and Teams plans and a new Max tier was added. The company published per plan message estimates, stated that consumption would vary with task complexity and model choice, left existing subscribers’ prices unchanged while the list price for new Pro subscribers moved from fifteen to twenty dollars, converted pre-purchased credits into usage under the new system, and gave existing paid subscribers a free additional week. Verified against the company’s own announcement. Four days after the change, the product documentation still displayed the old pricing, which is a disclosure failure of a different and more ordinary kind.

Anthropic, March and April 2026. Subscribers to the Max tiers reported five hour quota windows exhausting in a small fraction of that time. The company published a postmortem, reset subscriber quotas, and ran a credit campaign of up to two hundred dollars per subscriber, framed by the company as a discount celebration and read in the user community as compensation for a defect that had consumed credits in error. A public apology from the company’s product lead reached over a million views and was widely characterized as shifting responsibility onto users. In early April the company removed Claude Code from its twenty dollar Pro tier. The Claude Code prompt cache default had reverted from one hour to five minutes around 6 or 7 March, silently; the effect surfaced in April through issue 46829 on the company’s own repository and in trade coverage on 13 and 14 April. On 6 May the company announced remediation including a doubling of Claude Code limits, removal of peak hour throttling, and additional compute capacity.

GitHub, April and June 2026. On 20 April GitHub froze new individual signups for Copilot, tightened usage limits, and removed the most capable model from the paid individual tier. On 27 April it announced that premium request units would retire and every Copilot plan would move to usage based GitHub AI Credits, at one credit to one cent, effective 1 June 2026. The transition took effect as announced.

Perplexity, November 2025 to July 2026. As set out in Section VI.A.

Anthropic, May and June 2026. On 13 May the company announced, and on 14 May formalized in its help center, that Claude Agent SDK usage, the `claude -p` noninteractive command, the Claude Code GitHub Actions integration, and third party applications authenticating through the Agent SDK would cease drawing on subscription usage limits from 15 June 2026 and would instead draw on a separate monthly credit, sized at twenty dollars for Pro, one hundred for Max 5x, two hundred for Max 20x, and correspondingly for Team and Enterprise seats, billed at standard metered rates once exhausted. On 15 June the company updated the same page: it was pausing the changes, nothing had changed, the previously announced monthly credit was not available, and it would share an update before anything took effect.

Notion, July and August 2026. As set out in Section VI.A.

Alibaba Cloud, March and April 2026. The Model Studio coding plan's Lite tier stopped accepting new subscriptions on 20 March 2026 and renewals and upgrades were discontinued on 13 April 2026, with existing subscribers permitted to run to expiry. Remaining slots on the surviving tier are restocked daily in limited quantity on a first come basis. Verified against the company's own documentation, which also publishes the range of model calls a query consumes.

Z.ai, February to June 2026. Weekly ceilings were added to the coding plan in February 2026 at a stated multiple of the five hour limit. The Pro tier's price rose materially across the period. The company publishes approximate prompt limits per window and per week for each tier and a schedule of peak and off peak quota multipliers. It also credited subscribers with the equivalent of fifteen days of their subscription fee as compensation for a period of service problems, which is the third instance of an apology credit in this record.

Moonshot, July 2026. On 19 July Moonshot paused new Kimi K3 consumer subscriptions, citing capacity constraints after a demand surge, and announced a split of its membership into separate general and coding tiers; wire coverage followed the next day.

C. What the Wave Shows

Three features matter more than the length of the list.

The correction is not confined to firms with a particular business model. Sellers distributing open model weights rationed hosted access on the same timeline as sellers distributing none, which indicates the binding constraint is serving capacity rather than licensing philosophy.

The instruments were not uniform, which permits the comparison in the next section.

And the disclosure practice was not uniform either. Windsurf, Z.ai, Alibaba Cloud, and MiniMax published quantity terms in some form. Notion, Perplexity, and Anthropic did not, or converted stated terms into unstated ones. Same market, same cost pressure, same months, opposite choices. That is the observation

Section VI.C rests on, and it is what removes the technical impossibility defense.

VIII. PRICE CHANGE LEGIBILITY

I use the term price change legibility for the extent to which a change is expressed in units the buyer can convert into an expected cost before it takes effect. The chronology permits a first test of whether legibility predicts anything, and the honest result is more qualified than I expected when I began.

Two legible changes met organized objection. Cursor's June 2025 conversion was announced as a change in how usage would be charged; subscribers computed what it meant, objected, and received an apology and refunds. Anthropic's May 2026 Agent SDK announcement stated the surfaces affected, the effective date, and the dollar credit per plan, so an affected developer could compute their exposure directly, and third party vendors did so in public. It was withdrawn on the day it was to take effect.

One legible change proceeded unchanged. GitHub's April 2026 announcement was at least as computable: premium request units retired, credits substituted at a stated rate of one cent each, effective 1 June. It took effect on schedule.

The two illegible changes proceeded. Perplexity reduced twice, converting stated daily allowances into unstated weekly ones, without notice to prepaid annual subscribers, and neither reduction was reversed. Notion's allowance took effect on schedule against an unpublished magnitude.

The correct statement of the finding is therefore narrower than a symmetry. Legibility appears to be necessary for organized objection, because a buyer who cannot compute an exposure cannot mobilize against it, and every instance of organized objection in this record followed a computable announcement. Legibility is not sufficient for reversal. GitHub's transition was computable and proceeded, which is consistent with legibility determining whether objection is possible rather than whether it succeeds.

This is observational and involves five events across five sellers. It cannot separate legibility from seller size, from the organization and technical sophistication of the affected community, or from competitive position at the moment of announcement. 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. What the comparison supports is a regularity worth testing, and Prediction Two in Section XII is written to test it.

IX. WHY THIS BUYER CANNOT SELF PROTECT

Four conditions hold simultaneously for the population at issue, and each independently forecloses a standard response.

There is no leverage, and the literature concedes the point about itself. Section IV set out the Lewis and Yildirim observation. Standard advice to buyers facing a repricing, which is to use renewed negotiating position to redesign commercial terms around shared risk, presupposes a procurement function, a contract of sufficient size to interest the seller, and a credible alternative. A sole practitioner clicking a consumer checkout page has none of the three.

There is no substitute, and this is a finding of the chronology rather than an assumption. If the constraint on hosted inference were a business model choice, a buyer could route around it by selecting a seller with a different philosophy. Moonshot's July 2026 capacity pause establishes that the constraint is serving capacity. Open weights do not help a buyer who is renting the inference, because the weights being free does not make the hardware free. The only position that escapes the constraint is running the weights on hardware the buyer already owns, which reintroduces capital cost, technical skill, and capability limits the population at issue is least equipped to absorb.

There is no disclosed quantity term, and in the consumption rate case there is no term at all. Section VI.

There is no forecastability, and this is a property of the technology rather than of anyone's conduct. Agentic workloads consume unpredictably by construction. A single user instruction in a coding assistant can trigger many model calls behind the interface as the tool searches, plans, generates, and validates, and sellers differ in whether they meter user messages or underlying requests. A buyer cannot construct a budget line for a quantity they can neither observe in advance nor infer from their own behavior, and this remains true even under perfect disclosure of the allowance, because it concerns the numerator rather than the denominator.

The four together describe a party that cannot negotiate, cannot exit, cannot measure, and cannot forecast. 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.

X. WHAT THE SELLERS KNEW

This section makes an information asymmetry claim, which is an economic one. Whether any of it amounts to a legal duty is a separate inquiry not undertaken here.

The asymmetry must be stated precisely, because the loose version is false. It is not the case that the industry concealed its economics. The subsidy arithmetic has been published, the capital expenditure trajectory has been modeled publicly, and sellers themselves said in April 2026 that agentic usage was becoming the default and driving compute demand. Any sufficiently attentive buyer could have concluded that a correction was coming.

The asymmetry is narrower and it is sufficient. What was not published is the per seller quantity term. Aggregate industry economics tells a buyer that prices will rise somewhere. It does not tell them what they currently receive, and therefore does not permit them to compute what a reduction would cost them, when they would encounter it, or whether the workflow they are about to build will survive it. 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.

Against that background, the acquisition activity is the relevant fact. On 13 May 2026 Anthropic launched Claude for Small Business. The announcement materials stated that small businesses account for forty four percent of United States gross domestic product and employ nearly half the private sector workforce, that their adoption of artificial intelligence had lagged that of larger enterprises, that tools and training are rarely tailored to how small businesses operate, and that as a result their use often stops at the chat window. The

company's president stated that small businesses have never had the resources of bigger companies and that artificial intelligence is the first technology that can close that gap. The company's head of the small business segment described the target as businesses largely left out of the artificial intelligence boom. The launch was accompanied by a ten city training tour offering free half day workshops to a hundred local business leaders per stop, with attendees receiving a one month subscription to the company's highest consumer tier, and by partnerships with three community development financial institutions.

Nothing in that is false and the paper does not suggest otherwise. The characterization of the population is accurate, the stated intention is plausible, and the training is a real transfer of value. The point is structural and narrow. This recruitment occurred six weeks after the same company removed its coding agent from its entry tier, one month after it shortened the Claude Code prompt cache, three weeks after a major competitor froze new individual subscriptions and pulled its most capable model from its paid individual plan, and on the same day the company announced the Agent SDK change. Entrants were brought into a market at the moment of a documented correction by a seller who could compute the correction, and were placed where they could not compute it, because the quantity term against which they would have had to compute it was not published.

A. The Mechanism Reaches Litigation

On 14 June 2026 a proposed class action was filed against Anthropic in the United States District Court for the Northern District of California, *Kahn v. Anthropic, PBC*, No. 3:26-cv-05763, pleaded in diversity on a breach of contract theory. The allegations are unproven and the company has not been shown to have responded on the merits, so nothing here should be read as a finding.

What matters for this paper is the shape of the complaint, because it is the mechanism described above stated as a grievance. On contemporaneous reporting, the plaintiff alleges that the actual usage provided by the Max 5x and Max 20x plans falls far below the advertised amount, that the real caps are nearly impossible for a user to pin down, and that they changed without clear notice. The complaint is reported to reason from Anthropic's own July 2025 emails, which gave per tier weekly hour ranges by model, and to compute from those published figures that the Max 5x plan delivered roughly three and a half times the Pro plan's hours rather than five, and that Max 20x delivered roughly six times rather than twenty.

That last point deserves emphasis, because it complicates the remedy proposed in Section XIII. The July 2025 emails are the one occasion in this record on which a seller published something close to a quantity term. The complaint's arithmetic, if it holds, is that the published quantity did not match the advertised multiple. Disclosure is a necessary condition for a buyer to price an exposure. It is not by itself a sufficient one, and a disclosure regime that produces figures inconsistent with the marketing has moved the problem rather than solved it.

XI. LIMITATIONS

The chronology is a lower bound and not a census, for the reasons given in Section II. It is drawn from reported events, and the events least likely to be reported are precisely the ones this paper argues are hardest

to detect. That biases the record toward the visible instrument and understates the prevalence of the covert ones. The consumption rate instrument in Section VI.B is the clearest case: both instances reached the record because a technically sophisticated user community reconstructed them from logs, which is not a detection method available to the population this paper is about.

The comparison in Section VIII is observational and involves five events across five sellers. It cannot separate legibility from seller size, community organization, or competitive position, and the GitHub counterexample is the reason the finding is stated as necessity rather than sufficiency.

The magnitude of the effect is not estimated anywhere in this paper, and by the argument of Section VI it cannot be estimated from outside the seller under current disclosure practice. This paper establishes that a transfer is occurring and identifies three instruments. It does not size it. A reader who wants a number will not find one here and should be suspicious of anyone who supplies one from public sources. What could be measured, and would be worth building, is a longitudinal audit of published subscription terms and usage documentation across sellers at fixed capture intervals, coded on whether a quantity term is published at all, in what unit, and whether it moved between captures. That recovers no unpublished magnitude. It does establish how many sellers publish a quantity term and whether that share is rising or falling, which would convert a claim about several documented instances into a rate.

The subsidy figures in Section III are estimates produced by trade analysis and by an interested counterparty rather than counts from seller accounts, and are used only to establish that the subsidy is large, which no party contests.

An earlier draft of this paper attributed the discontinued entry tier to Z.ai. It belongs to Alibaba Cloud, and Z.ai's entry tier remains available. The error is recorded here because it is the kind that trade coverage propagates: two sellers of similar products in the same segment, one of which discontinued a tier, reported together. An earlier draft also dated the cache reversion to April 2026 rather than early March, having taken it from a secondary account; the issue tracker record fixes it at 6 or 7 March.

Entries resting on trade coverage rather than a seller's own materials, and therefore weaker than the rest: Perplexity's November 2025 to February 2026 sequence and the July 2026 help center wording; Anthropic's March and April quota reports, the subscriber credit campaign, and the May remediation; MiniMax's token quotas; and Z.ai's multiplier schedule, which is drawn from documentation aggregators rather than from the company's own page. The cache reversion, by contrast, is documented in the seller's own public repository. The Kahn allegations are unproven and are used here only to characterize a filed grievance, not to establish any fact about the seller's conduct.

Finally, the argument is confined to markets where the quantity term is undisclosed or where consumption rates move without notice. Where a seller publishes its allowance and holds its consumption rates fixed, the ordinary shrinkflation analysis applies and this paper's distinctive claims do not.

XII. PREDICTIONS

One. Posted subscription prices at the prevailing consumer tiers will remain approximately flat through 2027 while the effective quantity delivered per dollar declines. Falsified by a general round of posted price increases at the twenty dollar and two hundred dollar tiers unaccompanied by quantity restoration.

Two. Legibility will be necessary but not sufficient for reversal. No reversal will follow a change that buyers could not compute in advance; some computable changes will proceed anyway. Falsified by a quantity reduction against an undisclosed baseline being reversed under user pressure without the underlying magnitudes first being disclosed.

Three. Consumption rate repricing will become more common than allowance reduction, because it is cheaper to implement, requires no announcement, and is detectable only by users who instrument their own consumption. Falsified by sellers converging on published allowances with stable or published consumption rates, on the pattern Z.ai already follows.

Four. Because the binding constraint is serving capacity rather than licensing model, at least one further seller distributing open model weights will ration, pause, or discontinue hosted subscription access during a demand surge before the end of 2027. Falsified by the open weight segment absorbing demand surges without access restriction while closed sellers continue to restrict.

Five. Disclosure regulation will arrive before price regulation, and will arrive first in the European Union, on the drip pricing precedent discussed in Section V. Falsified by a jurisdiction reaching subscription price levels before reaching quantity term disclosure.

XIII. REMEDIES

None of these touches price. A seller that must raise the effective price to cover its costs should be able to do so. What is at issue is whether the buyer can see it happen.

Quantity term disclosure, with a consistency condition. The buyer is told the magnitude of what they are purchasing, in whatever unit the seller meters, and is told when that magnitude changes. Section VI.C establishes that this is feasible, because sellers in this market already do it, expressed as estimates, ranges, and rate schedules rather than guarantees. Section X.A supplies the necessary qualification: a published figure that contradicts the marketing claim it accompanies relocates the problem rather than fixing it, so the disclosed quantity and the advertised comparison have to be commensurable. The nearest existing analogue is the European prohibition on drip pricing, in force since 2011, resting on the principle that a price the buyer cannot see is not a price they agreed to.

Notice for consumption rate changes. The instrument in Section VI.B is the one no disclosure regime currently reaches, because nothing the buyer was told has changed. A cache lifetime, a default model, and a metering rule are all terms of the bargain in substance, and a change to any of them is a change to the price. Notice for these should be on the same footing as notice for an allowance change, and Z.ai's published multiplier schedule shows what the disclosed form of this instrument looks like when a seller chooses to publish it.

Portability. Export of conversation history, configuration, and stored context as a default capability. This constrains no seller's pricing. It lowers the switching cost on which the entire invest and harvest dynamic depends, which is the intervention the switching cost literature has recommended since Klemperer (1995).

XIV. CONCLUSION

The subsidy withdrawal now under way in artificial intelligence services is necessary at the level of the firm and this paper does not contest it. What the paper contests is the frame in which the withdrawal is discussed. The public conversation is organized around an asset price correction whose incidence would fall on capital. The correction actually occurring falls on adopters, and it is being administered through instruments adopters cannot see.

The packaged goods literature has known for two decades that consumers respond less to quantity reductions than to price increases and that firms therefore prefer the former. This market removes the last condition under which that preference remained checkable, and then adds an instrument the older literature has no case for, in which nothing the buyer was told changes and the work simply runs out sooner. The population absorbing the result is one the switching cost literature says it does not model and the adoption literature stops following at precisely the point the exposure begins.

What follows is a disclosure question rather than a pricing one. A buyer told the size of what they purchase, and told when a rule governing consumption moves, can observe a reduction, price it, object to it, and decide whether to stay. A buyer told neither is left to infer a transfer from the fact that the work stopped getting done. Some sellers in this record chose the first condition and published estimates, ranges, and rate schedules under the same cost pressure. The rest chose the second, and the first group is the reason we know that was a choice.

• •

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic’s Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author’s own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

REFERENCES

- 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-c>
- 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>
- 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/aepr.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>
- 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>
- Petrikaitė, 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*.