Who Captures AI Deflation? Quality-Adjusted Prices, Markups, and Pass-Through in AI Services.

Zijie Wang*

This version: March 2026

Abstract

AI inference costs are falling rapidly, yet how much of this deflation reaches downstream firms depends on market structure. I construct the first constant-quality price index for AI services and show that, once quality improvement is accounted for, effective AI prices fell by approximately 78 percent from early 2024 through early 2026, roughly twice the nominal price decline. The AI services market is oligopolistic, and markups meaningfully moderate cost transmission: GPU cost pass-through is positive and significant, but higher-markup providers absorb a disproportionately larger share of upstream cost declines. Embedding the quality-adjusted price index in an input-output framework, I find a roughly 120 basis points cumulative deflationary contribution to the PPI over the same period, with indirect supply-chain linkages accounting for the majority of the aggregate effect. A welfare decomposition reveals that the entry of new, higher-capability models, not price reductions on existing ones, accounts for the substantial majority of consumer surplus gains. This pattern points to a Schumpeterian tradeoff: the same markup rents that attenuate cost pass-through also sustain the investment in frontier models that generates value for downstream firms.

JEL classification: D43, E31, L13, L86, O33.

Keywords: Artificial Intelligence; Markups; Cost Pass-Through; Hedonic Prices; Oligopoly Pricing.

*zijie.wang.1@warwick.ac.uk, Warwick Business School

I am grateful to Philippe Mueller and Ganesh Viswanath-Natraj for their time, patience and guidance. The author has no financial relationships or other potential conflicts of interest to disclose.

1 Introduction

Artificial intelligence (AI) services are rapidly emerging as a significant intermediate input for firms across a growing number of sectors, and their costs are falling rapidly. Between early 2024 and early 2026, the nominal price of AI inference declined by approximately 39 percent. Yet the true cost reduction is far larger: once quality is held constant, the effective price of a unit of AI capability fell by roughly 78 percent over the same period. This dramatic deflation in the quality-adjusted price of AI services raises a central question: how much of it actually reaches downstream firms, and who captures the rest?

The answer depends critically on market structure. AI services are supplied by a handful of providers, most notably Anthropic, Google, OpenAI, and xAI, each operating a portfolio of differentiated models. These providers face steep scale economies, invest heavily in proprietary model development, and compete in a market where quality and capabilities are rapidly evolving. In such a setting, the pass-through of input cost reductions is not determined by cost shares alone, as in competitive input markets, but is moderated by endogenous markups. A provider with sufficient market power can absorb a decline in graphics processing unit (GPU) rental costs into its margin rather than reducing its API prices, attenuating the deflationary effect before it reaches downstream sectors. This paper measures that attenuation and traces its economy-wide consequences.

I establish three main results. First, quality-adjusted AI prices have fallen by dramatically more than nominal prices suggest. Using hedonic regressions that control for model capabilities across four domains (coding, general reasoning, math and science, and reliability), I construct a constant-quality AI services price index that falls by approximately 78 percent from early 2024 through early 2026, against a nominal decline of roughly 39 percent over the same period. The gap quantifies the extent to which naive price comparisons understate true AI deflation. The hedonic framework also yields time-varying implicit prices for each capability dimension. Coding and reliability are associated with persistently positive price premia throughout the sample, while general reasoning and mathematical ability enter with negative coefficients once coding and reliability are controlled for. This hierarchy suggests that as the capability frontier advances, market premia concentrate on

dimensions tied to specific high-value tasks rather than on general cognitive ability per se.

Second, the AI services market is oligopolistic, and markups meaningfully moderate cost pass-through. Estimating a demand system using model-level token usage shares on OpenRouter and instrumenting for price with GPU rental cost shifters interacted with model-size categories, I recover model-level and provider-level markups, as well as a market-level aggregate markup index spanning all mainstream AI models. The aggregate index fluctuates substantially over the sample, ranging from approximately 1.4 in late 2024 to a trough in mid-2025 before recovering. Using provider-level markups as a moderating variable, I show that GPU cost pass-through is positive and significant ($\hat{\theta}_0 = 0.2285$), but its interaction with provider-level markups is negative and significant ($\hat{\theta}_1 = -0.1520$), indicating that higher-markup providers absorb a larger share of upstream cost shocks. The linear estimates imply a zero-pass-through threshold at a gross markup of approximately $\mu^* = 1.50$: providers pricing above this level fully absorb GPU cost declines into their margins, transmitting nothing downstream. Several providers, most notably Anthropic following its frontier model releases in late 2025, sustain markups well above this threshold for extended periods.

Third, AI deflation propagates through the production network and delivers a meaningful but attenuated deflationary contribution to economy-wide producer prices. Embedding the quality-adjusted price index in a Leontief input–output framework, I estimate that the first-order deflationary effect from AI cost reductions amounts to roughly 120 basis points of cumulative PPI decline over 2024–2026, with indirect supply-chain linkages accounting for the majority of the aggregate effect. Under competitive pricing, the cumulative deflationary contribution would have been larger in magnitude; the difference represents deflation absorbed by oligopoly margins rather than passed through to downstream sectors.

Identifying demand for AI services and recovering markups requires addressing the endogeneity of prices: providers observe model-specific demand conditions that the econometrician does not, and set prices accordingly, biasing naive demand estimates. I instrument for model-level prompt prices using daily rental prices for NVIDIA H100 GPUs, obtained from Silicon Data via Bloomberg, interacted with model size categories. The

GPU rental rate is the most direct market-based cost measure for AI inference because it is an all-in market price for usable compute capacity that bundles several cost layers that would otherwise be difficult to observe at high frequency: the amortized acquisition cost of the GPUs, the electricity consumed by energy-intensive inference workloads, and the capital and operating costs of the data center. Because larger models require proportionally more GPU-hours per token than smaller ones, a common movement in GPU rental prices produces differential cost shocks across model size categories, generating the cross-sectional variation in cost needed for identification. The exclusion restriction requires that these size-specific cost shifters affect token usage shares only through their effect on prices, conditional on month fixed effects that absorb the common time-series component of GPU costs and on the quality and operational characteristics that model size influences. This restriction is plausible because model size determines computational cost per token but does not independently shift user preferences once capability and operational characteristics are controlled for.

My analysis makes three contributions. The first is to provide, to the best of my knowledge, the first systematic measurement of quality-adjusted AI services prices. AI inference is not a homogeneous input with an observable unit price: it is a differentiated cognitive service whose quality is rapidly improving and whose effective units are not well-defined. A dollar spent on AI inference in late 2025 purchases dramatically more capability than in early 2024. Ignoring this renders nominal price comparisons uninformative about the true cost of AI to downstream firms. I construct a constant-quality AI services price index via hedonic regressions, using benchmark scores from Artificial Analysis across fifteen established capability evaluations, and show that true AI deflation is roughly twice as large as nominal prices suggest.

The second contribution is the first empirical investigation of market power and markup dynamics in the AI services market. The structure of this market, its concentration, the extent of providers' pricing power, and how that power responds to competitive entry, are largely taken as given or ignored in the growing literature that treats AI as a production input. I address this gap by estimating a demand system using model-level

token usage shares on OpenRouter and recovering markups via multi-product Bertrand inversion. The resulting estimates reveal a market with consequential and time-varying competitive dynamics: the entry of DeepSeek in early 2025 compressed rivals' markups, while Anthropic's successive frontier-model releases in late 2025 pushed its markup to roughly three times that of its nearest competitors.

The third contribution is the first empirical estimate of how AI cost reductions propagate through the production network to downstream sectors. Combining sector-level AI exposure measures derived from BEA use tables with the quality-adjusted price index, I show that the indirect supply-chain channel accounts for the majority of the aggregate deflationary effect, and that oligopolistic pricing absorbs a non-trivial share of the deflationary effect that would otherwise reach downstream firms.

A welfare analysis complements the pass-through results. While the static welfare cost of oligopoly pricing is non-trivial, decomposing cumulative surplus gains reveals that new model entry accounts for the substantial majority of consumer welfare gains over the sample period, pointing to a Schumpeterian trade-off in which the rents that sustain premium pricing also finance the frontier-model development that benefits downstream users.

Roadmap The remainder of the paper is organized as follows. Section 2 reviews the related literature. Section 3 describes the OpenRouter marketplace and the construction of the price, usage, and cost data. Section 4 develops the quality measurement framework and hedonic pricing results. Section 5 presents the demand estimation and markup recovery. Section 6 reports the micro pass-through estimates, welfare analysis, and macro propagation results. Section 7 concludes.

2 Literature Review

This paper connects four literatures: quality-adjusted price measurement for technology inputs, the measurement and implications of firm markups, incomplete pass-through of cost shocks, and the propagation of sectoral shocks through production networks.

My measurement framework builds on the hedonic price index tradition. The hedonic framework for quality adjustment, developed by Rosen (1974) and applied to technology goods by Pakes (2003), has been used extensively to construct constant-quality deflators for personal computers (Berndt and Rappaport, 2001), semiconductors (Flamm, 1993), and telecommunications equipment (Doms, 2004). These exercises consistently find that quality-adjusted deflation exceeds nominal price declines by large margins, a measurement gap with direct consequences for real output measurement, real interest rate estimation, and inflation assessment (Advisory Commission to Study the Consumer Price Index, 1996; Byrne et al., 2016; Syverson, 2017). I extend this tradition to AI services, a differentiated cognitive input whose quality is captured by capability benchmarks rather than hardware specifications, and for which no constant-quality price index previously existed.

A second literature studies markups and market power. Because the quality-adjusted index still embeds markups, understanding market power is essential for interpreting the price decline. A growing body of work in financial economics examines the rise of firm markups and its implications for profitability, factor shares, and asset prices. De Loecker et al. (2020) document the secular increase in aggregate markups; Baqaee and Farhi (2020) show that markup dispersion generates misallocation with first-order productivity effects; Corhay et al. (2020) demonstrate that markup dynamics are priced in equity returns; and Edmond et al. (2023) quantify the aggregate welfare cost of markups. Despite this interest, the market structure of AI provision, arguably the most consequential new intermediate input market in decades, has only recently attracted empirical attention. The closest work to mine is Demirer et al. (2025), who use OpenRouter and Microsoft Azure data to document six stylized facts about the LLM market, including rapid entry, persistent price heterogeneity across intelligence tiers, and reduced-form demand elasticities. My paper complements theirs by moving from descriptive facts to structural estimation: I estimate

a full demand system, recover model-level and provider-level markups via multi-product Bertrand inversion (Berry et al., 1995; Nevo, 2001), and trace the consequences of those markups for cost pass-through and aggregate price dynamics.

A third strand examines incomplete pass-through of cost shocks. The exchange rate pass-through literature provides a close analogue to my micro pass-through exercise. Gopinath et al. (2010) link pass-through to invoicing currency; Amiti et al. (2019) show that high-market-share importers absorb more of exchange rate movements into markups; and Gopinath et al. (2020) establish that dominant currency pricing amplifies the role of market structure in determining pass-through rates. De Loecker et al. (2016) find similar markup-absorbing behavior following trade liberalization, and Nakamura and Zerom (2010) and Ganapati et al. (2020) document analogous patterns in commodity markets. I bring this framework to the AI supply chain, where the upstream cost shock (GPU rental prices) is observed daily and the moderating variable (provider-level markups) is estimated structurally rather than proxied by market share.

A fourth literature studies production networks and the aggregate propagation of sectoral shocks. Acemoglu et al. (2012) and Baqaee and Farhi (2019) establish the theoretical foundations for how sectoral shocks propagate through input-output linkages. In finance, Herskovic (2018) shows that production network centrality is priced in equity returns, Barrot and Sauvagnat (2016) provide causal evidence that supply chain disruptions propagate to downstream firm valuations, and Ozdagli and Weber (2017) document that monetary policy transmits through customer-supplier relationships. The literature connecting AI to macroeconomic outcomes, including the task-based frameworks of Acemoglu and Restrepo (2018, 2019) and the growth-theoretic analysis of Aghion et al. (2018), typically treats AI as a general productivity shifter without modeling the market structure of AI provision. I contribute to these literatures by embedding an imperfectly competitive AI sector into the Leontief framework, showing that oligopolistic pricing absorbs a non-trivial share of the deflationary effect.

3 Institutional Detail and Data

3.1 OpenRouter and the Market for AI APIs

OpenRouter as a multi-model gateway. OpenRouter, founded in 2023, is a unified API gateway that provides access to multiple AI model providers through a single interface. The platform provides developers and downstream applications with access to over 600 models from more than 60 providers, including OpenAI, Anthropic, Google, Meta, Mistral, xAI, and DeepSeek. OpenRouter provides a single, standardized API endpoint which allow users to send inference requests to OpenRouter's routing layer, which dispatches them to the selected model provider.

OpenRouter charges users on a pay-per-token basis, with per-million-token prices posted separately for prompt (input) and completion (output) tokens. These prices are passed through from the upstream model providers without markup, e.g the per-token rates on OpenRouter are identical to those on each provider's own API[1]. OpenRouter's revenue instead derives from a small fee (5.5% as of February 2026) levied at the point of credit purchase. Billing is transparent and model-specific: users observe the exact per-token price for each model before making a request, and costs are metered and deducted from a prepaid credit balance in real time.

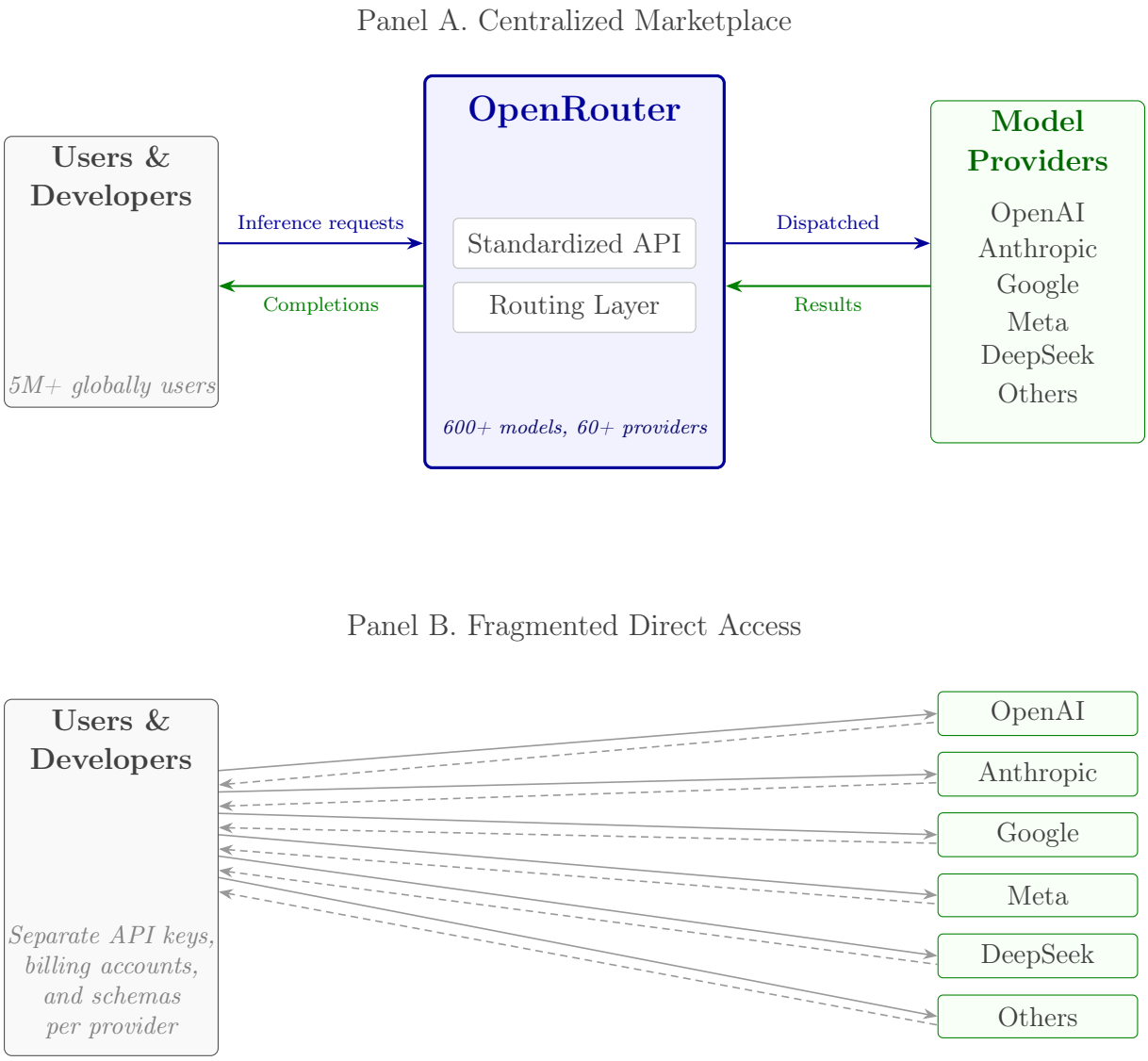
Two features of OpenRouter make it a particularly useful empirical setting for studying the market for AI services. First, *product comparability*: all models are accessible through an identical API interface, and the standardized schema minimizes the integration frictions that typically segment demand across model providers. Second, *revealed demand*: because users can costlessly switch between models at any point, token usage on OpenRouter reflects active choice among substitutes. The platform thus functions as a global marketplace in which heterogeneous AI models compete for user demand on price, quality, speed, and capabilities[2]. Figure 1 illustrates this structure: Panel A shows OpenRouter in which users

[1]Some models on OpenRouter are offered for free with rate limits; I exclude free-tier usage from my empirical analysis.

[2]As of mid-2025, OpenRouter serves over five million developers and end-users globally, processing more than 100 trillion tokens annually and routing roughly one trillion tokens per day as of late 2025. (Aubakirova et al., 2026). Over 50% of usage originates outside the United States.

access multiple providers through a single standardized API, while Panel B shows the fragmented alternative direct access, in which each provider requires a separate integration, billing relationship, and schema.

Figure 1: OpenRouter and the Market for AI Inference



Notes. This figure provides a comparison of OpenRouter with direct access to upstream model providers. Panel A shows a centralized marketplace in which users and downstream applications access multiple AI providers through a single standardized API, while OpenRouter performs routing and billing. OpenRouter passes through provider inference prices and earns revenue from a fee on prepaid credits rather than from per-token markups. Panel B shows the fragmented alternative, in which users must maintain separate API integrations, billing relationships, and schemas with each provider. The comparison illustrates how OpenRouter reduces integration frictions.

3.2 Data

Pricing and usage data on OpenRouter. OpenRouter provies two key variables for my analysis. The first is the *posted prompt price*, p_{mt}^{prompt}, defined as the listed price (in USD per one million prompt tokens) for model m at time t. Because OpenRouter directly passes through provider pricing, this measure captures the price charged by the upstream AI company and is comparable across AI models. The second variable is *token usage*, $tokens_{mt}^{OR}$, defined as the total number of prompt and completion tokens consumed by model m at time t on OpenRouter.

OpenRouter does not provide a downloadable historical dataset, so I constructed the price and usage panels from archived snapshots of the platform. For prices, I queried the Wayback Machine for all snapshots of OpenRouter's model catalog page from November 2023 through February 2026. From each snapshot I parsed the per-token prompt and completion prices for every listed model. For usage, I applied the same archival approach to each provider's page on OpenRouter, which displays daily model-level token usage.

AI model quality measures. Artificial Analysis is an independent benchmarking platform that evaluates AI models under standardized conditions across a comprehensive suite of capability tests, covering reasoning, mathematics, coding, and reliability.[3] I collected model-level benchmark scores and operational characteristics from Artificial Analysis leaderboard page, which yields a cross-section of 399 models. For each model, I obtain scores on 15 individual benchmarks as well as operational characteristics such as context window length, output speed (tokens per second), latency (time to first token), whether the model is classified as a reasoning model, and model size category (as classified by Artificial Analysis). Table 2 details the individual benchmarks and their domain classifications. I merged these quality measures with the OpenRouter price and usage panels by matching on model identifiers, producing a matched sample of 208 models that covers the majority of token volume observed on the platform.

[3]See https://artificialanalysis.ai/methodology/intelligence-benchmarking. All evaluations are conducted independently by Artificial Analysis using consistent evaluation criteria across models.

Table 1: Summary Statistics of Model Characteristics, Prices, and Token Usage

	Panel A: Model Characteristics					
	Speed	Latency	Open	US Company	Context Window	Reasoning
Mean	71.489	3.201	0.529	0.611	0.378	0.351
Std. Dev.	90.686	12.410	0.500	0.489	0.787	0.478

	Panel B: Model Price and Monthly Sum of Token Usage						
	Mean	Median	Std. Dev.	P1	P25	P75	P99
Prompt Price	3.139	0.400	12.632	0	0.100	2.000	75.000
Tokens usage	48.700	0.601	249.206	0	0	10.789	892.899
Token Value	0.018	0	0.091	0	0	0.003	0.414

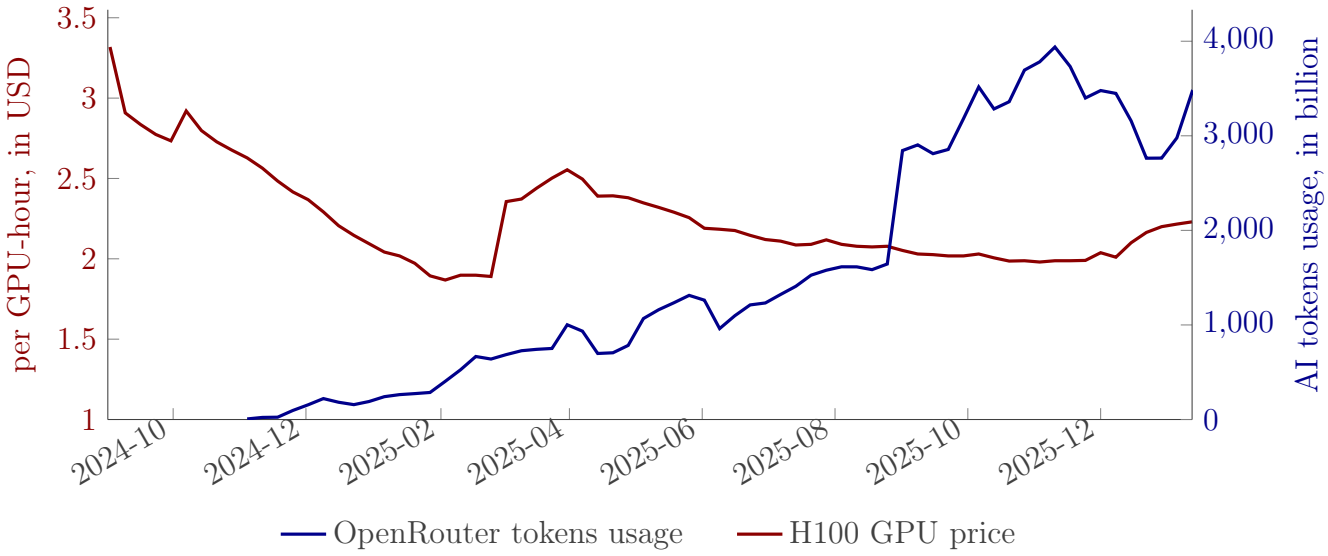
Notes. This table reports summary statistics for the OpenRouter model panel from January 2024 to February 2026. Panel A reports model-level characteristics using the latest available observation for each model ($N = 208$). "Open," "US Company," and "Reasoning Model" are indicator variables equal to one when the condition is met; "Context Window" is the maximum context length, measured in millions of tokens. Panel B reports the distribution across model-month observations. Prompt price is the monthly average price. Token usage and token value are monthly totals. Total usage is reported in billions of tokens per month. Token value is reported in millions of U.S. dollars per month. Columns P1, P25, P75, and P99 denote the 1st, 25th, 75th, and 99th percentiles.

Cost of producing AI services. I obtained daily rental prices for NVIDIA H100 GPUs from Silicon Data via Bloomberg. Silicon Data aggregates quoted prices from major cloud providers, yielding a market-wide spot rental price series for GPU compute capacity (USD per GPU-hour).

GPU rental prices serve as my primary measure of the marginal cost of producing AI services. Importantly, a GPU rental price is not merely a hardware lease: it is an all-in price for usable GPU capacity that bundles several cost components that would otherwise be difficult to observe at high frequency. First, it embeds the amortized acquisition cost of the GPU itself, along with supporting hardware. Second, it incorporates electricity costs, which constitute a substantial and growing share of AI compute expenditure; AI training and inference workloads are energy-intensive, and cloud providers price power consumption into their per-hour rates. Third, the rental price reflects the capital and operating costs of the data centres, including land, construction, and maintenance. Because these cost layers are jointly priced in the GPU rental market, the GPU rental rate provides a market-based

measure of the upstream cost of supplying AI services. In addition, the H100 functions as a unit of account for accelerator capacity: the cost of a more powerful GPU, such as the H200 by Silicon Data, is mapped into equivalent H100 GPU-hours, so that shifts in the frontier hardware mix are reflected in the index rather than lost to compositional change.

Figure 2: GPU Rental Price and AI Model Usage



Notes. The figure plots the GPU rental price and AI usage on OpenRouter. The left axis (red) shows the posted rental price of an NVIDIA H100 GPU (USD per GPU-hour). The right axis (blue) shows OpenRouter's total weekly token usage (billions of tokens).

BEA input–output tables. I use the Bureau of Economic Analysis (BEA) 2024 Supply-Use tables at the summary industry level, which cover 71 industries corresponding to three-digit NAICS codes. For each industry, the Use table records intermediate purchases Z_{js}, the dollar value of commodity j used as an input by industry s in a given year, alongside gross output, compensation, and value added. Because the BEA revises these tables with a one-to-two year lag, the 2024 edition is the most recent available and reflects intermediate-input relationships prior to the rapid diffusion of AI services. The BEA input-output tables predate the rapid diffusion of AI services, so the sectoral exposure measures used in the macro propagation exercise reflect an earlier stage of AI adoption. The implied PPI effects are therefore likely to understate the magnitude of the deflationary channel going forward.

4 Measuring AI Prices

Nominal prices for AI inference have declined substantially since late 2023, falling by approximately 39 percent through early 2026. Over the same period, however, the AI models have become dramatically more capable. A downstream firm purchasing AI inference cares not about the nominal token price but about the price of a *constant-quality* unit of AI capability. Without quality adjustment, declining nominal prices may overstate or understate true cost reductions, depending on whether the decline reflects genuine cost savings or the entry of lower-quality models.

The approach proceeds in two steps. Subsection 4.1 constructs model-level capability indices using principal component analysis (PCA) across four domains that map to distinct downstream use cases. Subsection 4.2 embeds these indices in a hedonic price regression that serves two purposes: it nets out quality variation to recover a constant-quality AI services price index, and it reveals which dimensions of model capability command a market premium and how those implicit prices evolve over time.

4.1 Quality Measurement via Principal Components

AI models differ along multiple capability dimensions that are relevant to downstream users: a software firm values coding proficiency, a financial analyst values reasoning accuracy, a scientific laboratory values mathematical and scientific problem-solving, and any production deployment requires reliability and instruction compliance. To capture this heterogeneity parsimoniously, I group 15 established benchmarks from the AI research literature into four domains: *General Reasoning*, *Math & Science*, *Coding*, and *Reliability & Compliance*. Table 2 lists the benchmarks assigned to each domain.

Within each domain, I first standardize every benchmark score to have zero mean and unit variance (z-score). I then extract the first principal component (PC1) of the standardized scores within the domain, re-orienting the loading vector so that higher values consistently indicate stronger performance. The resulting PC1 score is re-standardized to mean zero and standard deviation one. This procedure yields four model-level quality

indices Q_{mt}^{code}, Q_{mt}^{gen}, $Q_{mt}^{math/sci}$, and Q_{mt}^{relia}, each of which summarizes the model's capability in its respective domain. Across the four domains, the first principal component explains between 76 and 81 percent of the within-domain variance, indicating that a single factor captures the substantial majority of cross-model capability variation in each domain[4].

Figure 3 plots each model at its first observed date in the sample against these four indices. Two patterns are evident. First, the capability frontier has expanded rapidly in all four domains: models released in late 2025 routinely score two to four standard deviations above the sample mean, whereas the earliest models cluster near zero. Second, the rate of improvement differs across domains, with Coding and Math & Science exhibiting particularly steep advances. This heterogeneous quality expansion is precisely why naive price comparisons are misleading. A model priced at $1 per million tokens in late 2025 is a fundamentally different product from one priced at $1 per million tokens in early 2025.

4.2 Hedonic Pricing and Quality-adjusted AI price index

The quality indices constructed in Subsection 4.1 document rapid and heterogeneous capability growth, which lead to the question that matters for downstream cost measurement: how much does a constant-quality unit of AI inference cost over time? To answer this, I adopt a hedonic pricing approach.

The hedonic framework (Rosen, 1974; Pakes, 2003), widely used for product quality adjustment, decomposes the observed price of a differentiated product into a component attributable to quality characteristics and a common time effect that captures the movement in prices conditional on quality. The latter, recovered from the month fixed effects γ_t, yields the quality-adjusted price index.

I estimate a hedonic price function to net out observed quality:

$$\log p_{mt}^{prompt} = \alpha + \gamma_t + \sum_k \beta_k Q_{mt}^k + \Phi' X_{mt} + \varepsilon_{mt}, \tag{1}$$

[4]PCA loadings are estimated on complete cases, while observations with partial benchmark coverage are retained by setting missing standardized benchmark inputs to zero when projecting onto PC1 (equivalent to imputing the within-domain mean after standardization).

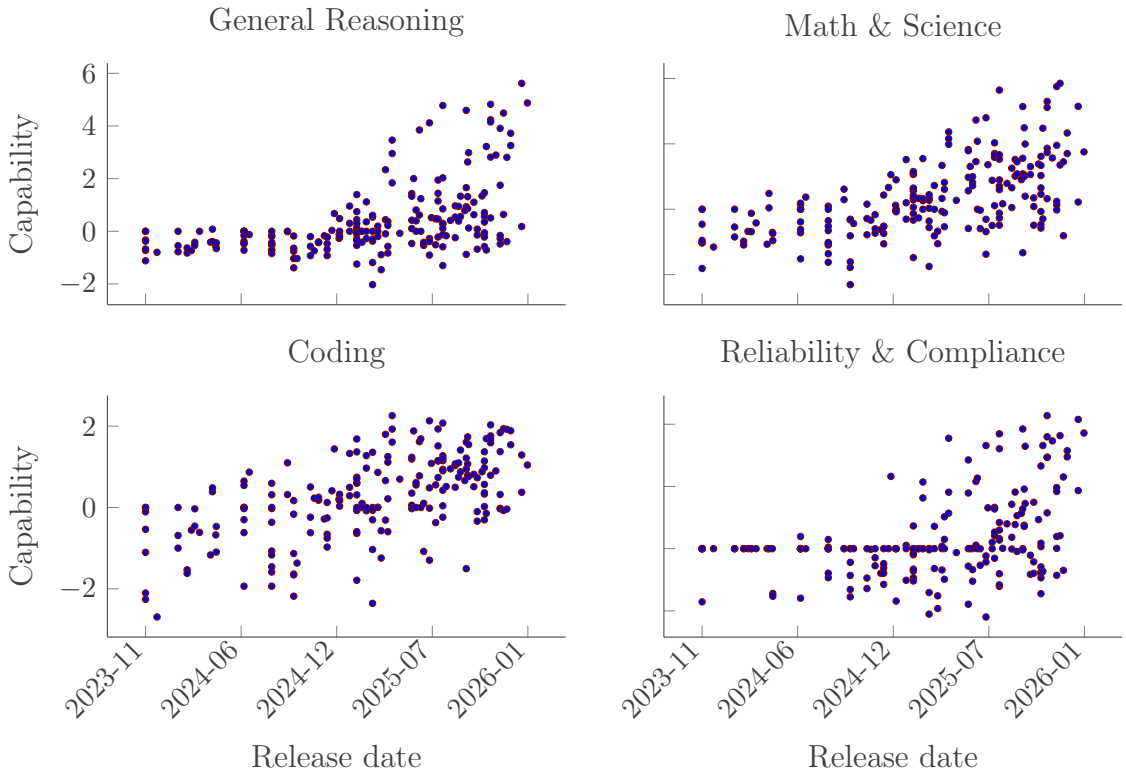
Table 2: AI Capability Benchmarks

Benchmark type	Benchmark name	What it measures	Original paper
General Reasoning	`mmlu_pro`	Hard multi-domain knowledge + reasoning via challenging multiple-choice questions.	Link
	`hle`	Frontier-level closed-ended academic exam (broad subjects, multimodal).	Link
	`mmmu_pro`	Robust multi-discipline multimodal understanding/reasoning.	Link
Math & Science	`aime25`	Competition-style math problems; requires multi-step reasoning to exact integer answers.	Link
	`math_500`	A 500-problem subset of the MATH competition-math dataset used to evaluate mathematical reasoning.	Link
	`gpqa`	Graduate-level, "Google-proof" science Q&A (biology/chemistry/physics) written by domain experts.	Link
	`critpt`	Frontier physics research problems designed to probe research-level reasoning beyond standard exams.	Link
Coding	`humaneval`	Python function synthesis from docstrings; functional correctness measured by unit tests.	Link
	`livecodebench`	Continuously-updated, contamination-aware coding benchmark (code generation, repair, execution, test prediction).	Link
	`scicode`	Scientist-curated research coding tasks across natural sciences; tests realistic scientific code generation.	Link
Reliability & Compliance	`ifbench`	Verifiable instruction-following under diverse constraints; focuses on out-of-domain generalization.	Link
	`lcr`	Long-context reasoning over very long documents; tests synthesis/inference across many sources.	Link
	`tau2`	Interactive agent evaluation in a dual-control environment (agent + user both act via tools in shared state).	Link
	`terminalbench_hard`	Hard, realistic command-line tasks for autonomous agents; tests long-horizon tool use and robustness.	Link
	`omniscience`	Knowledge reliability + calibration (penalizes hallucinations, rewards abstention when uncertain).	Link

Notes. This table reports the AI capability benchmarks used to construct Q_{mt}^{code} (coding), Q_{mt}^{gen} (general reasoning), $Q_{mt}^{math/sci}$ (math and science), and Q_{mt}^{relia} (reliability/robustness). Model-level benchmark scores are obtained from Artificial Analysis.

where $\log p_{mt}^{prompt}$ is the log posted prompt price (USD per 1M prompt tokens) for model m in month t. Q_{mt}^{k} collects model capability measures including Q_{mt}^{code} (coding), Q_{mt}^{gen} (general reasoning), $Q_{mt}^{math/sci}$ (math and science), and Q_{mt}^{relia} (reliability/robustness). X_{mt} includes operational characteristics such as *speed* (output tokens per second), *latency* (response delay), an indicator for *reasoning models*, and *context length* (maximum context window, in tokens). Table 2 lists each benchmark and its domain classification. The four domain-level quality measures (Q_{mt}^{code}, Q_{mt}^{gen}, $Q_{mt}^{math/sci}$, and Q_{mt}^{relia}) are constructed as

Figure 3: AI Model Capabilities



Notes. The figure plots one observation per model at its release data. The y-axis reports four model-level capability indices: General Reasoning, Math & Science, Coding, and Reliability & Compliance, respectively. For each domain, I construct the index as the first principal component of benchmark scores after standardizing each benchmark (z-score). PCA loadings are sign-normalized so that higher values indicate stronger capability, and the resulting component is re-standardized to have mean zero and unit variance.

described in Subsection 4.1.

The time fixed effect γ_t captures the component of $\log p_{mt}^{prompt}$ that is common to all models in month t after controlling for observed quality and other characteristics. I define the constant-quality AI services input price index as:

$$P_t^{AI} \equiv \exp(\gamma_t - \gamma_{t_0}), \tag{2}$$

so that $P_{t_0}^{AI} = 1$ (equivalently, $\log P_{t_0}^{AI} = 0$).

Hedonic estimation results. Table 3 reports the hedonic regression estimates. The preferred specification for constructing the quality-adjusted index is column (5), which

includes all four capability indices and operational characteristics (speed, latency, context window length, and a reasoning-model indicator) but excludes non-operational characteristics that do not directly contribute to the quality of AI inference output. Column (6) adds such non-operational characteristics including an open-model indicator, a US-company indicator, and model size categories (large and medium) for comparison.

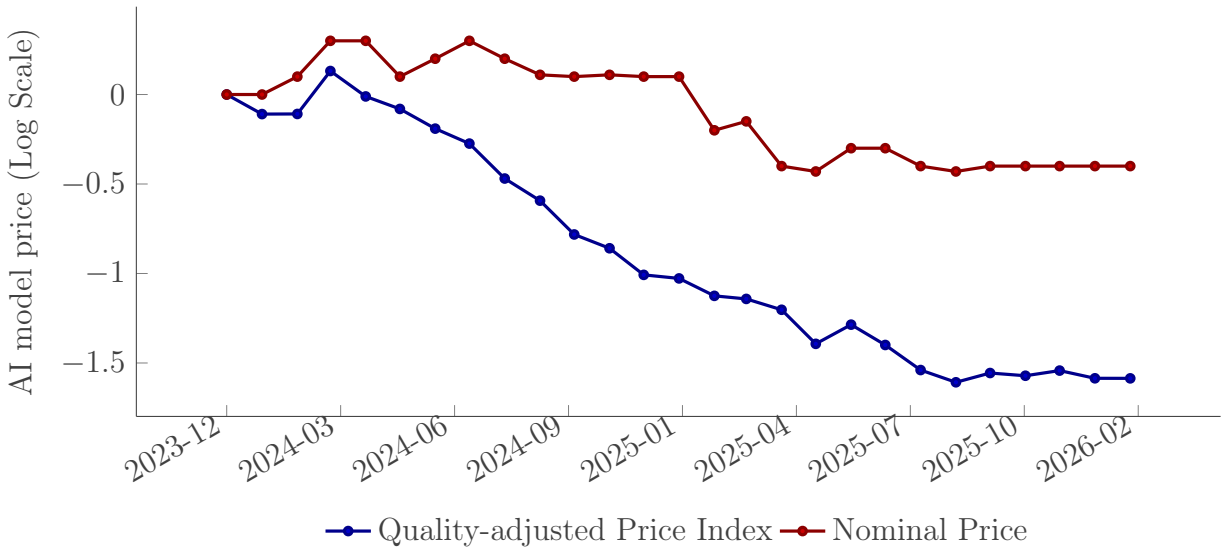
Coding and reliability enter with large positive coefficients, indicating that these dimensions carry a premium. A one-standard-deviation improvement in coding capability is associated with approximately 87 log points (139%) higher prices; the corresponding figure for reliability is 103 log points (179%). General reasoning and math/science, by contrast, carry negative coefficients (-0.654 and -0.329), suggesting that conditional on coding ability and reliability, additional reasoning or quantitative capability does not independently command a price premium. One possible interpretation is that downstream users treat reasoning and mathematical proficiency as auxiliary: these capabilities matter for model selection only insofar as they complement coding or reliability, and users are unwilling to pay more for incremental gains along these dimensions once the task-critical attributes are held fixed.

Among operational characteristics, the signs reflect conditional correlations after quality is controlled for. Speed enters negatively and latency positively, a pattern consistent with faster models being smaller and cheaper to serve, conditional on capability. Context window length also enters negatively, consistent with longer context becoming a standard feature of lightweight models rather than a premium differentiator. Reasoning-model classification is associated with higher prices, reflecting the additional inference cost of chain-of-thought. Column (6) further adds non-operational characteristics including open model, US company, and model size indicators. These variables are excluded from the quality-adjusted index specification because they do not directly determine the quality of inference output as experienced by downstream users. Rather, they proxy for the competitive position of the provider. Because these characteristics capture cross-provider variation in pricing power rather than in output quality, they belong not in the hedonic quality adjustment but in the structural demand and markup analysis of Section 5, where

I model market power directly.

The quality-adjusted price index. Figure 4 compares the quality-adjusted index $\log P_t^{AI}$ with the log nominal prompt price over the sample period. Both series decline, but the quality-adjusted index falls far more steeply by roughly 1.5 log points (approximately 78%) from early 2024 through early 2026, while the nominal price declines by approximately 0.5 log points (approximately 39%) over the same period. The wedge between the two series measures the cumulative quality improvement: holding capabilities fixed, the effective price of AI inference has fallen by roughly twice as much as the nominal price suggests. This gap quantifies the extent to which naive nominal price comparisons understate the true deflation in AI services and motivates the use of the quality-adjusted index in the downstream pass-through analysis of Section 6.

Figure 4: Quality-adjusted Index vs. Nominal Prompt Price



Notes. The figure compares a quality-adjusted AI services price index $\log P_t^{AI}$ in equation (2) (blue) to the median nominal prompt price (red), both on a log scale. Vertical bars indicate confidence intervals for the quality-adjusted index.

Dynamic implicit prices of capabilities. The pooled hedonic regression in equation (1) reveals that the market does not value all capability dimensions equally, but it imposes time-invariant implicit prices β_k over the full sample. In a market where

Table 3: AI Prices and Quality: Hedonic Pricing

Q^{coding}	0.723***				0.870***	0.170***
	(0.015)				(0.023)	(0.018)
Q^{gen}		0.512***			-0.654***	-0.346***
		(0.013)			(0.030)	(0.019)
$Q^{math/sci}$			0.649***		-0.329***	-0.249***
			(0.015)		(0.038)	(0.028)
Q^{relia}				0.831***	1.027***	0.329***
				(0.013)	(0.019)	(0.016)
speed					-0.007***	-0.003***
					(0.000)	(0.000)
latency					0.018***	0.008***
					(0.002)	(0.001)
context window length					-0.095***	-0.154***
					(0.012)	(0.013)
reasoning model					0.119**	0.480***
					(0.050)	(0.037)
open model						-0.829***
						(0.024)
US						-0.215***
						(0.019)
size large						2.714***
						(0.033)
size medium						1.359***
						(0.026)
N companies	21	21	21	21	21	21
N models	200	200	200	200	200	200
N	63828	63828	63828	63828	63828	63828
R-squared Adj.	0.154	0.096	0.129	0.211	0.361	0.668

Notes. This table reports estimates from the hedonic pricing regression in equation 1. The dependent variable is the log posted prompt price (USD per 1M prompt tokens), $\log p_{mt}^{prompt}$. Regressors include model capability measures including Q_{mt}^{code} (coding), Q_{mt}^{gen} (general reasoning), $Q_{mt}^{math/sci}$ (math and science), and Q_{mt}^{relia} (reliability/robustness) as well as operational characteristics: *speed* (output tokens per second), *latency* (response delay), an indicator for *reasoning models*, and *context length* (maximum context window, in tokens). *size large* and *size medium* indicate the size of the AI model based on the number of parameters. White heteroskedasticity-robust standard errors are reported in parentheses. Significance levels: *** 1%, ** 5%, * 10%.

both the supply of capabilities and the willingness to pay for them are evolving rapidly, this restriction may obscure economically important dynamics: both model providers and downstream users care about how the market's valuation of each capability domain changes over time. To allow for such variation, I estimate a rolling-window version of the hedonic regression.

Let τ define a 3-month window $\mathcal{W}(\tau) \equiv \{t : t \in (\tau - 3, \tau]\}$. For each τ, estimate on

the subsample $t \in \mathcal{W}(\tau)$:

$$\log p_{mt}^{prompt} = \alpha^{(\tau)} + \gamma_t^{(\tau)} + \sum_k \beta_k^{(\tau)} Q_{mt}^k + \delta^{(\tau)\prime} X_{mt} + \varepsilon_{mt}^{(\tau)}. \tag{3}$$

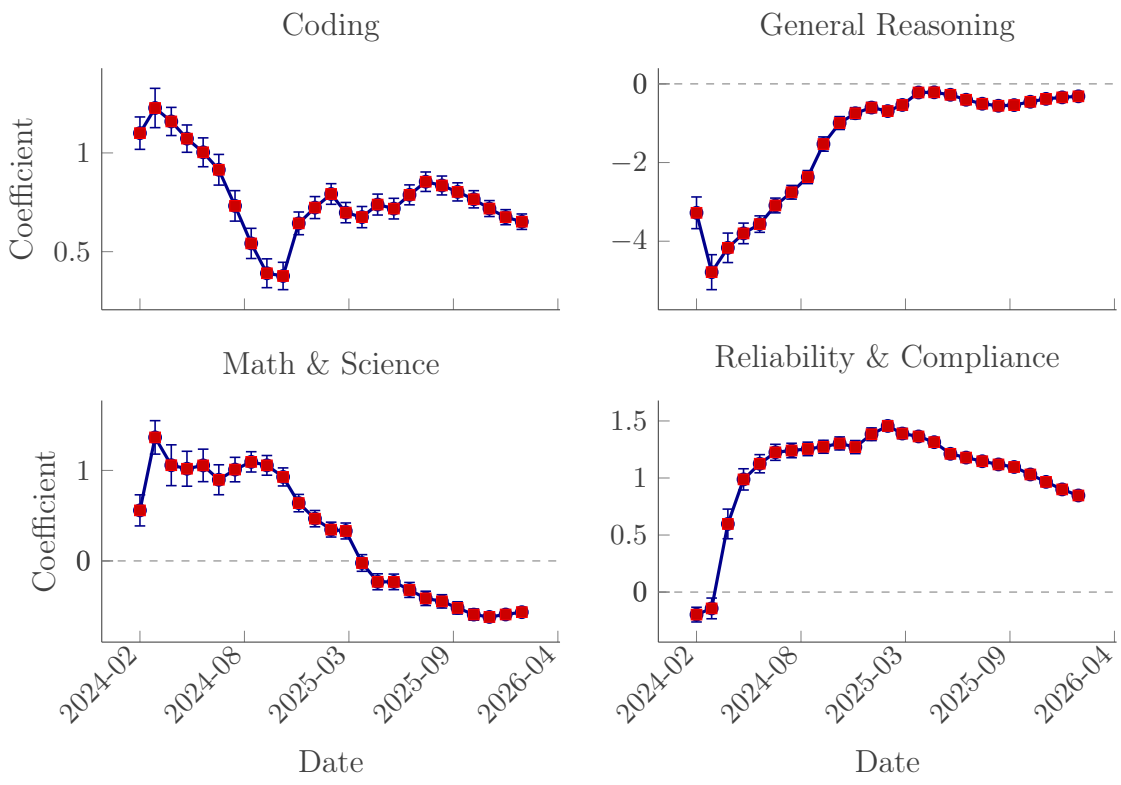
This yields a time series of rolling-window coefficients $\{\widehat{\beta}_k^{(\tau)}\}_\tau$.

Figure 5 plots the results. The time-varying estimates reinforce the cross-sectional finding. Coding capability commands a positive premium throughout the sample period, consistent with persistent downstream demand for code-generation services. Reliability and compliance maintains a similarly stable and positive premium, consistent with its role as a prerequisite for production deployment. General reasoning and math & science, by contrast, enter with negative coefficients once other quality attributes are controlled for.

Taken together, the pooled and rolling-window results suggest a clear hierarchy in how the market prices AI capabilities. Dimensions tied to specific, high-value downstream tasks (coding) or to production readiness (reliability) are persistently priced, whereas more general cognitive dimensions (reasoning, mathematical ability) are competed away as the capability frontier advances and supply broadens. These time-varying implicit prices also underscore the importance of the quality-adjustment exercise: not only are capabilities improving, but the market's valuation of each dimension is shifting, so that a fixed-weight quality index would mismeasure the effective quality of the marginal dollar spent on AI services.

What the quality-adjusted index does and does not capture. The index P_t^{AI} measures the price of a constant-quality unit of AI inference, netting out capability improvements via the hedonic regression. This is the effective price faced by downstream firms and the relevant object for measuring sectoral cost exposure in Section 6. This price, however, reflects both the marginal cost of serving inference and the markup charged by oligopolistic providers. A decline in the quality-adjusted price could therefore reflect falling compute costs, falling markups, or both, and the welfare implications differ sharply across these channels. Disentangling marginal cost from markup, and quantifying how market structure moderates the pass-through of upstream cost shocks, requires the demand

Figure 5: Rolling Implicit Prices of Capability



Notes. The figure reports rolling-window estimates of the implicit price of capability (hedonic coefficients) in four domains: Coding, General Reasoning, Math & Science, and Reliability & Compliance. Each line plots the point estimate by month. Vertical error bars show the 95% confidence intervals.

estimation and markup recovery developed in Section 5.

5 Market Structure and Markups

The quality-adjusted price index P_t^{AI} constructed in Section 4 measures the price of a constant-quality unit of AI inference, but it still embeds the markups charged by oligopolistic AI providers. A decline in P_t^{AI} could reflect falling marginal costs, falling markups, or both, and the welfare implications differ sharply across these channels. This section develops the demand estimation and markup recovery framework required to disentangle market power from cost.

The approach proceeds in two steps. Subsection 5.1 maps OpenRouter token usage to market-level shares, estimates a logit demand system using GPU-cost shifters interacted with model size as instruments for price, and derives the demand Jacobian required for markup recovery.[5] Subsection 5.2 inverts the multi-product Bertrand first-order conditions to recover model-level markups and implied marginal costs, and aggregates them to provider and market levels.

5.1 Demand Estimation

From OpenRouter usage to market shares. Let tokens_{mt}^{OR} denote OpenRouter token volume for model m in month t. Define total OpenRouter volume and within-OpenRouter shares as

$$\text{volume}_t^{OR} \equiv \sum_{j=1}^{J_t} \text{tokens}_{jt}^{OR}, \qquad s_{mt}^{OR} \equiv \frac{\text{tokens}_{mt}^{OR} \cdot p_{mt}^{prompt}}{\sum_{j=1}^{J_t} \text{tokens}_{jt}^{OR} \cdot p_{jt}^{prompt}}. \tag{4}$$

To map OpenRouter usage to the full AI API market, I assume: (i) OpenRouter captures a constant market penetration $\lambda \in (0, 1)$, and (ii) its within-month model composition is representative of the broader market. I then scale OpenRouter shares by λ to obtain market shares:

$$\tilde{s}_{mt} \equiv s_{mt}^{MKT} \approx \lambda \, s_{mt}^{OR}, \qquad s_{0t} \equiv 1 - \sum_{m=1}^{J_t} \tilde{s}_{mt} = 1 - \lambda, \tag{5}$$

[5] Appendix A provides a robustness check using a random coefficient logit model (BLP) that relaxes the independence of irrelevant alternatives (IIA) property; the resulting markups and pass-through estimates are qualitatively and quantitatively similar.

where s_{0t} is the outside share (non-OpenRouter demand). In the baseline, I set $\lambda = 0.10$ for all months.[6]

Figure 6 plots accumulated token usage and token value by company on OpenRouter over the sample period. The upper panel reveals a highly concentrated and rapidly shifting usage distribution: Google and xAI account for the largest share of total token volume, with several providers experiencing rapid growth. The lower panel, which weights usage by per-token prices, paints a somewhat different picture of revenue concentration. Because premium models (e.g., flagship models from OpenAI and Anthropic) carry higher per-token prices, these providers account for a larger share of token value than of raw token volume. This distinction between volume shares and value shares will matter for the markup aggregation in Subsection 5.2, where I weight by token value to construct economically relevant market-level markup indices.

Logit demand specification. I model demand with a multinomial logit system. The baseline share equation is

$$y_{mt}^{MKT} \equiv \log \tilde{s}_{mt} - \log s_{0t} = -\alpha \log p_{mt}^{prompt} + \beta' Q_{mt} + \Phi' X_{mt} + \tau_t + \xi_{mt}. \tag{6}$$

Under equation (5), $\tilde{s}_{mt} = \lambda s_{mt}^{OR}$ and $s_{0t} = 1 - \lambda$, so

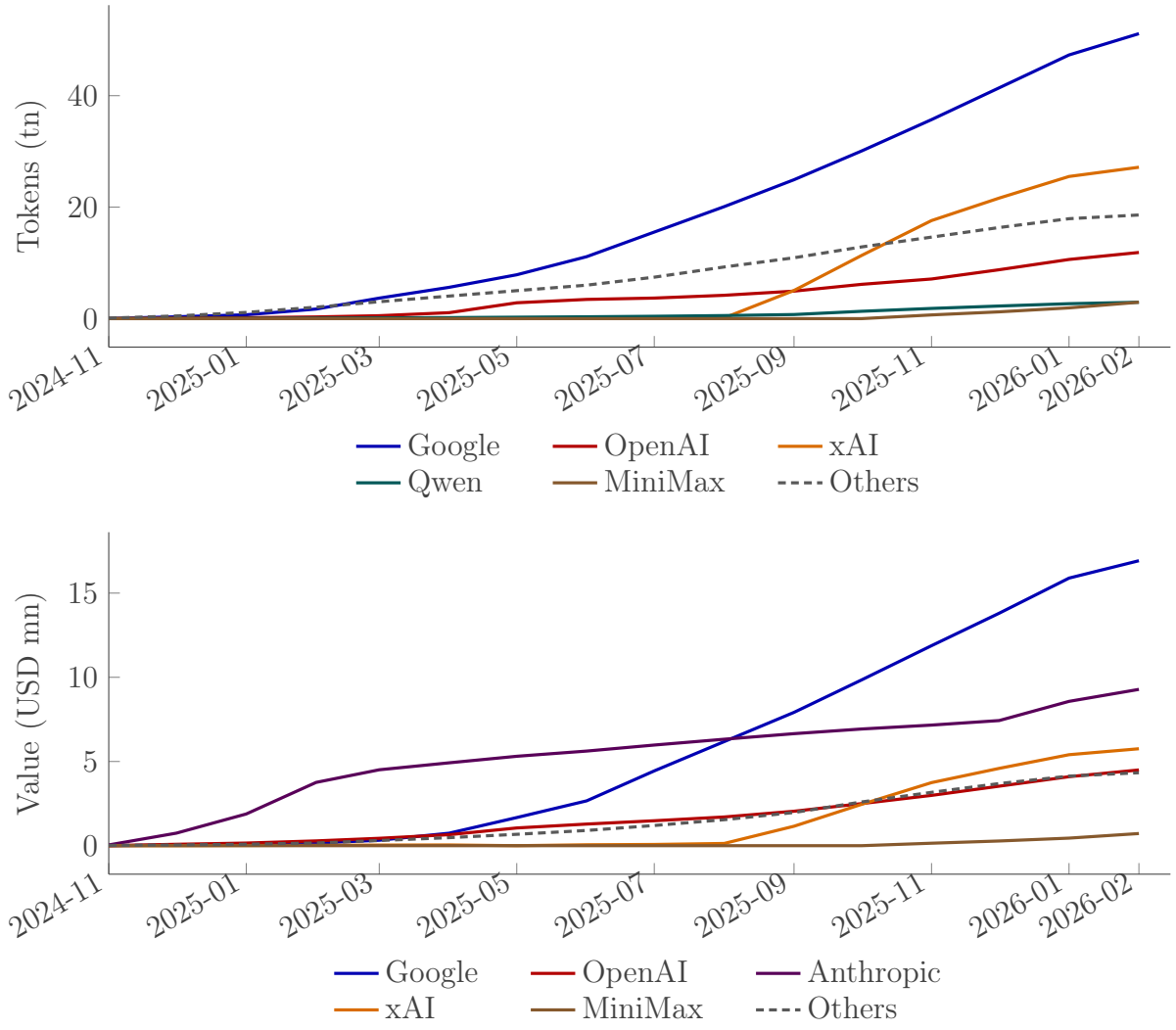
$$\log \tilde{s}_{mt} - \log s_{0t} = \log s_{mt}^{OR} + \big[\log \lambda - \log(1 - \lambda)\big]. \tag{7}$$

Thus, with month fixed effects τ_t, estimating equation (6) using $\log s_{mt}^{OR}$ on the left-hand side yields the same slope coefficients (α, β, Φ). The penetration parameter λ shifts only levels (market shares and the outside share), not demand elasticities.

Price endogeneity and instruments. The log prompt price $\log p_{mt}^{prompt}$ in equation (6) is potentially endogenous: providers have private information about ξ_{mt} and set prices

[6]The true market penetration of OpenRouter is unobservable because the broader AI API market is fragmented across providers' proprietary platforms, none of which publicly disclose usage volume. The choice of λ, however, has negligible impact on the estimated demand elasticities. The demand slope α is identified without λ, as equation (7) shows that λ enters only as an additive constant absorbed by month fixed effects.

Figure 6: OpenRouter Token Usage



Notes. The upper panel plots accumulated total token usage by company on OpenRouter (in trillion); the lower panel plots accumulated token value by company (in million USD).

accordingly, biasing OLS estimates of α. To address this, I instrument for price using cost shifters derived from the GPU rental market.

The instrument exploits the fact that GPU rental costs constitute a primary component of the marginal cost of serving AI inference. Variation in the H100 GPU rental price shifts providers' marginal costs but, conditional on time fixed effects and quality controls, should not directly affect the relative demand for any given model. Because models of different sizes have different cost sensitivities to GPU price movements, a large model requires proportionally more GPU-hours per token than a small model, I interact the

monthly GPU rental price index with model size category indicators (large, medium, and small). This produces a set of instruments $\{gpu_t \times \mathbb{1}[\text{size}_m = k]\}_k$ that vary across both models and months. The exclusion restriction requires that, conditional on month fixed effects (which absorb the common time-series component of GPU costs) and observed quality attributes, the size-specific GPU cost shifters affect token shares only through their effect on prices. This restriction is plausible because, while model size is correlated with capability in the raw data, the demand equation directly controls for model quality through the four capability indices as well as operational characteristics including speed, latency, and context length. Conditional on these controls, model size determines computational cost per token but does not independently shift user preferences: a user choosing between two models with identical capability scores, speed, and latency has no reason to prefer one over the other on the basis of parameter count alone.

Table 4 reports the demand estimates. Columns (1)–(3) present OLS results; columns (4)–(6) present the corresponding 2SLS specifications. The OLS price coefficient is negative and significant throughout, ranging from -0.325 without quality controls to -0.665 with controls and company fixed effects. The 2SLS estimate in column (5), which includes quality controls and month fixed effects but excludes company fixed effects, yields a price coefficient of -0.366.

Column (5) serves as the baseline for markup recovery and all subsequent Jacobian computations, for the following reason. Provider identity is closely tied to pricing power: company fixed effects absorb time-invariant cross-provider markup differences, which is precisely the variation that the markup inversion in Subsection 5.2 aims to recover. Including them would partial out this heterogeneity before it enters the demand estimates, attenuating the markup dispersion that is central to the pass-through analysis. Column (6) is consistent with this reasoning: once company fixed effects are added, the price coefficient attenuates from -0.366 to -0.291 and loses statistical significance at conventional levels, though the point estimate remains negative and economically non-trivial. This attenuation is expected if provider dummies absorb the cross-provider price variation that reflects market power.

Table 4: Demand System

	OLS			2SLS		
$\log p^{prompt}$	-0.325***	-0.602***	-0.665***	0.040	-0.366***	-0.291
	(0.021)	(0.026)	(0.119)	(0.127)	(0.134)	(0.222)
Q^{code}		1.054***	0.854**		0.862**	0.590
		(0.070)	(0.370)		(0.385)	(0.397)
Q^{gen}		0.372***	0.456*		0.481	0.502**
		(0.105)	(0.238)		(0.361)	(0.253)
$Q^{math/sci}$		-0.550***	-0.326		-0.503	-0.288
		(0.097)	(0.363)		(0.383)	(0.384)
Q^{relia}		0.544***	0.324		0.323	0.155
		(0.096)	(0.279)		(0.268)	(0.216)
controls	No	Yes	Yes	No	Yes	Yes
month FE	Yes	Yes	Yes	Yes	Yes	Yes
company FE	No	No	Yes	No	No	Yes
N	1231	1231	1231	1123	1123	1123
R^2	0.117	0.437	0.593	0.037	0.417	0.551

Notes. This table reports estimates of a logit demand equation using OpenRouter usage data. The dependent variable is $\log s_{mt}^{OR}$, where s_{mt}^{OR} is model m's token share in month t. $\log p_{mt}^{prompt}$ is the log posted prompt price (USD per 1M prompt tokens). Q_{mt}^{code}, Q_{mt}^{gen}, $Q_{mt}^{math/sci}$, and Q_{mt}^{relia} measure model performance in coding, general reasoning, math/science, and reliability/robustness, respectively. "Controls" include speed, latency, an indicator for reasoning models, and maximum context-window length (# of tokens). "Company FE" indicates whether company fixed effects are included. Columns report OLS and 2SLS estimates; the 2SLS specifications instrument $\log p_{mt}^{prompt}$ with GPU-cost shifters (size-interacted GPU price indices). Standard errors are reported in parentheses and clustered by month; in the company-FE specification, standard errors are clustered by month and company. Significance levels: *** 1%, ** 5%, * 10%.

Demand Jacobian. For markup recovery I require the Jacobian of *level market shares* with respect to *level prices*:

$$\Delta_{ij,t} \equiv \frac{\partial \tilde{s}_{it}}{\partial p_{jt}^{prompt}}.$$

Since $\log p_{jt}^{prompt}$ enters the logit share equation linearly with coefficient $-\alpha$ (equation (6)), the derivative of *log share* with respect to *log price* is

$$\frac{\partial \log \tilde{s}_{it}}{\partial \log p_{jt}^{prompt}} = -\alpha\Big(\mathbb{1}\{i=j\} - \tilde{s}_{jt}\Big). \tag{8}$$

Converting log-elasticities into level-price derivatives,

$$\frac{\partial \tilde{s}_{it}}{\partial p_{jt}^{prompt}} = \tilde{s}_{it}\,\frac{\partial \log \tilde{s}_{it}}{\partial p_{jt}^{prompt}}, \qquad \frac{\partial \log \tilde{s}_{it}}{\partial p_{jt}^{prompt}} = \frac{\partial \log \tilde{s}_{it}}{\partial \log p_{jt}^{prompt}} \cdot \frac{1}{p_{jt}^{prompt}},$$

I obtain

$$\frac{\partial \tilde{s}_{it}}{\partial p_{jt}^{prompt}} = -\frac{\alpha}{p_{jt}^{prompt}} \tilde{s}_{it}\Big(\mathbb{1}\{i=j\} - \tilde{s}_{jt}\Big). \tag{9}$$

Stacking equation (9) across all (i, j) gives the estimated demand Jacobian:

$$\Delta_t = \Big(\operatorname{diag}(\tilde{\mathbf{s}}_t) - \tilde{\mathbf{s}}_t\tilde{\mathbf{s}}_t'\Big) \cdot \operatorname{diag}\left(-\frac{\alpha}{\mathbf{p}_t^{prompt}}\right). \tag{10}$$

Note that λ enters *only* through $\tilde{\mathbf{s}}_t = \lambda \mathbf{s}_t^{OR}$ in equation (10); the slope parameter α is identified in equation (6) without λ.

Evaluated at $\tilde{\mathbf{s}}_t$ and $\mathbf{p}_t$, the own- and cross-price derivatives are:

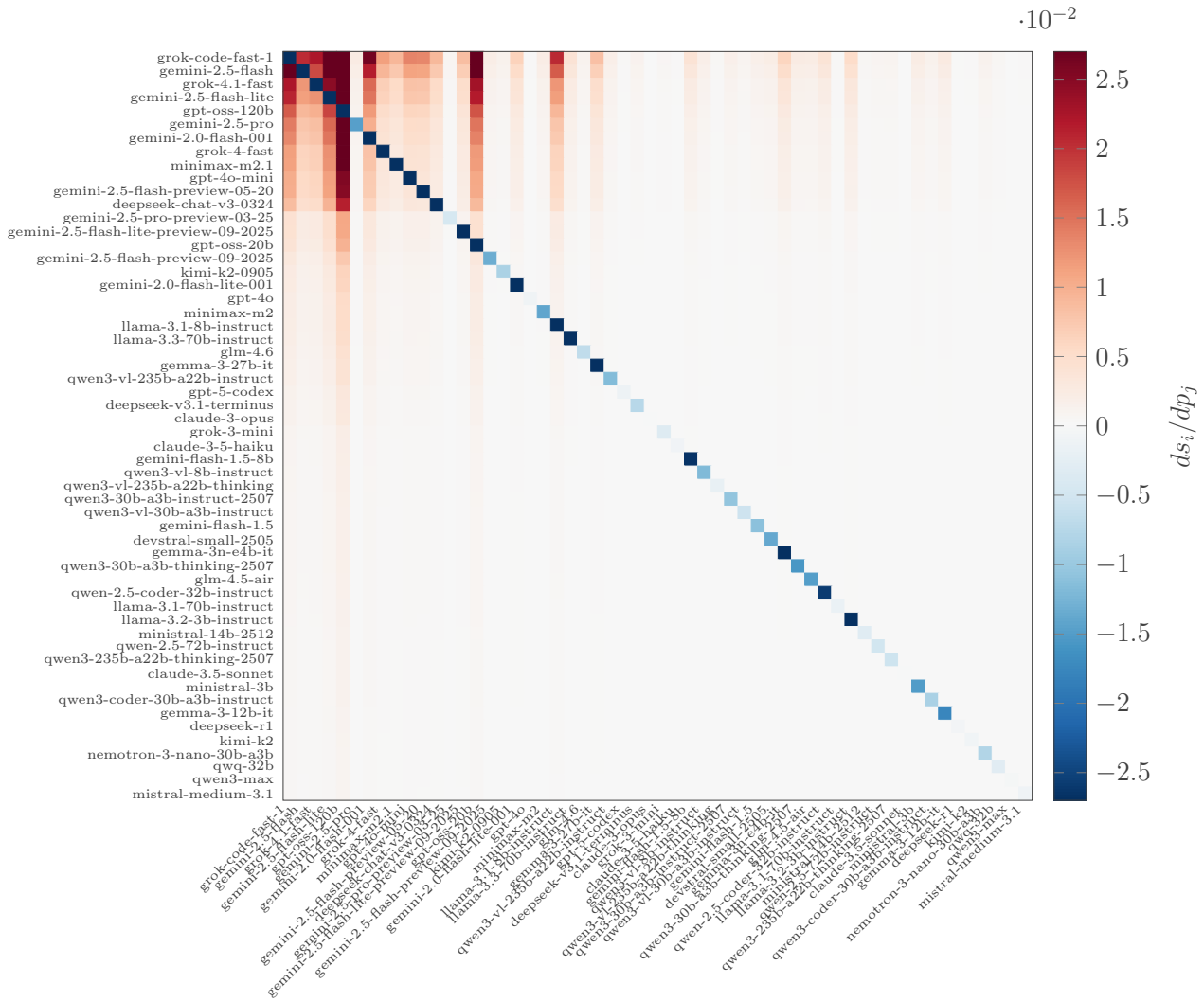
$$\frac{\partial \tilde{s}_{mt}}{\partial p_{mt}^{prompt}} = -\frac{\widehat{\alpha}}{p_{mt}^{prompt}} \tilde{s}_{mt}\big(1 - \tilde{s}_{mt}\big), \tag{11}$$

$$\frac{\partial \tilde{s}_{jt}}{\partial p_{mt}^{prompt}} = \frac{\widehat{\alpha}}{p_{mt}^{prompt}} \tilde{s}_{jt}\tilde{s}_{mt} \qquad (j \neq m). \tag{12}$$

The own-price derivative in equation (11) is negative: a price increase for model m reduces its market share. The cross-price derivative in equation (12) is positive: a price increase for model m diverts demand to every rival, with the magnitude proportional to the rival's own share. This proportional substitution is a well-known limitation of the logit model. Appendix A estimates a random coefficient logit (BLP) model that relaxes IIA by allowing consumer-level heterogeneity in price and quality valuations; the resulting markups and pass-through estimates are qualitatively and quantitatively similar to those reported here.

Figures 7 and 8 visualize the estimated Jacobian $\widehat{\Delta}_t$ for the latest month in the sample. In Figure 7, models are ordered by descending OpenRouter token share. Three patterns are evident. First, own-price effects (the diagonal) are large and negative across most models, indicating that price changes meaningfully affect a model's own demand throughout the usage distribution. Second, among the highest-share models in the upper-left block, cross-price effects are visibly positive and non-trivial, reflecting strong competitive interdependence at the top of the market: a price change by one high-volume model meaningfully redirects demand toward its rivals. Third, cross-price effects decay sharply as one moves away from the upper-left corner, indicating that lower-volume models

Figure 7: Demand Jacobian Heatmap: Models Ordered by Token Share

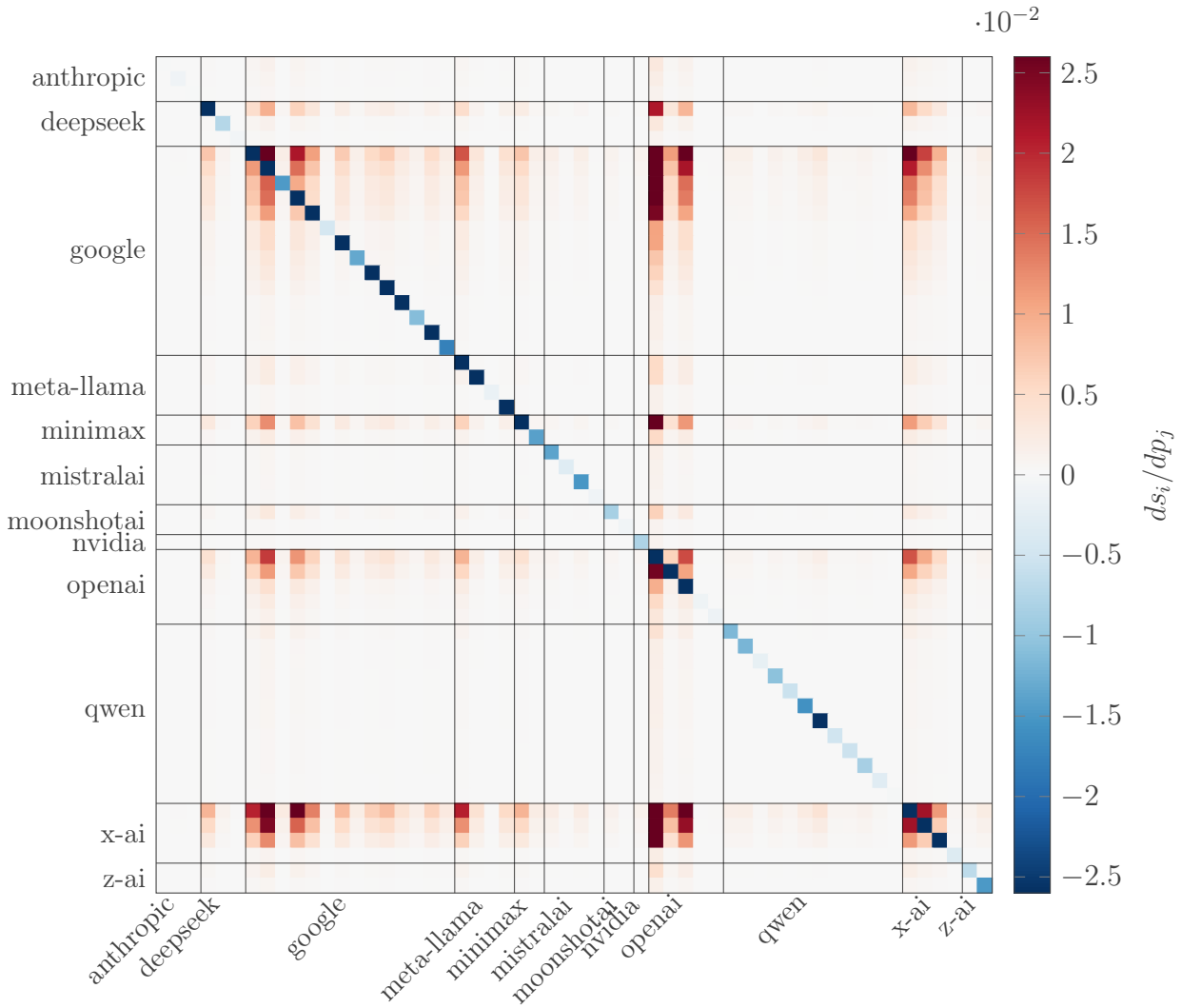


Notes. Each cell (i, j) shows $\partial s_i / \partial p_j$ for model i and j . Models are ordered by their OpenRouter token share (largest first). Rows index share i; columns index price j. For visual clarity, the figure displays the 56 models with the largest token shares, which together account for 98% of total OpenRouter token volume.

exert negligible competitive pressure on one another. This heterogeneity suggests that meaningful price competition is concentrated among the handful of models that account for the bulk of token volume, while the long tail of lower-volume models operates in relative competitive isolation.

Figure 8 reorders the same Jacobian by provider, revealing both the multi-product structure of competition and the heterogeneity in cross-provider substitution. Within each provider block (delineated by black lines), own-model effects dominate; cross-model

Figure 8: Demand Jacobian Heatmap: Models Ordered by Provider



Notes. Same Jacobian as Figure 7, with models reordered by provider (alphabetical) and then by token share within each provider. Black lines delineate provider blocks.

effects within the same provider are positive but modest, reflecting the cannibalization channel through which multi-product pricing attenuates competitive pressure: when a provider changes the price of one model, some diverted demand flows to its own other models rather than to rivals.

The cross-provider blocks reveal additional structure. Anthropic's row exhibits near-zero cross-price effects with respect to all other providers, suggesting that its models occupy a sufficiently differentiated product space that demand is largely insulated from rivals' pricing decisions, consistent with strong market power in the high-capability segment. By

contrast, the cross-provider blocks linking Google, OpenAI, and xAI display visibly larger positive entries, indicating meaningful demand substitution among these three providers. This pattern suggests that users treat the models of Google, OpenAI, and xAI as closer substitutes, so that price changes among them generate non-trivial demand reallocation across providers.

5.2 Markups and Marginal Costs

I recover model-level markups and implied marginal costs from pricing first-order conditions (FOCs). The approach follows the standard conduct-based markup recovery in differentiated-product oligopoly (e.g., Berry et al., 1995; Nevo, 2001), adapted here to the AI inference market. Given observed prices, market shares, and the estimated demand Jacobian, I infer the markup vector that rationalizes observed prices under multi-product Bertrand-Nash pricing.

Ownership structure and pricing. Let mc_{mt} denote marginal cost for model m in month t. Each AI provider (e.g., OpenAI, Google, Anthropic) operates a portfolio of models and sets prices to maximize the joint profit across its model lineup. Define the ownership matrix Ω_t, where $\Omega_{ij,t} = 1$ if models i and j are owned by the same provider and $\Omega_{ij,t} = 0$ otherwise. Under multi-product Bertrand-Nash pricing, each provider internalizes the demand externalities among its own models when setting prices. The system of first-order conditions can be written compactly as:

$$\tilde{\mathbf{s}}_t + \big(\Omega_t \odot \Delta_t\big)\big(\mathbf{p}_t - \mathbf{mc}_t\big) = \mathbf{0}, \tag{13}$$

where $\odot$ denotes the element-wise product. The matrix $\Omega_t \odot \Delta_t$ retains only those demand derivatives that are internalized by the same provider: for a single-product firm, the only nonzero entry in its row is the own-price derivative; for a multi-product firm, the relevant entries include both own- and cross-model derivatives within its portfolio.

Markup inversion. Define the level-markup vector $\mathbf{m}_t \equiv \mathbf{p}_t - \mathbf{mc}_t$. Rearranging equation (13), I recover $\mathbf{m}_t$ by solving the linear system month by month.

$$\left(\Omega_t \odot \widehat{\Delta}_t\right)\widehat{\mathbf{m}}_t = -\tilde{\mathbf{s}}_t. \tag{14}$$

For each model-month (m, t), implied marginal costs and Lerner indices are then

$$\widehat{\mathbf{mc}}_t = \mathbf{p}_t - \widehat{\mathbf{m}}_t, \qquad \widehat{\boldsymbol{\ell}}_t \equiv \frac{\widehat{\mathbf{m}}_t}{\mathbf{p}_t} = \mathbf{1} - \frac{\widehat{\mathbf{mc}}_t}{\mathbf{p}_t}, \tag{15}$$

and the gross markup (the ratio of price to marginal cost) is

$$\widehat{\mu}_{mt} \equiv \frac{p_{mt}^{prompt}}{\widehat{mc}_{mt}} = \frac{1}{1 - \widehat{\ell}_{mt}}. \tag{16}$$

Aggregation to provider-level and market-level markups. Model-level markups are heterogeneous across products and providers. To construct economically meaningful provider- and market-level summaries, I aggregate using token value as weights. Define token value for model m in month t as $v_{mt} \equiv \text{tokens}_{mt}^{OR} \times p_{mt}^{prompt}$, which captures the economic size of each model in the marketplace. The provider-level token-value-weighted markup for provider j is

$$\widehat{\mu}_{jt}^{prov} \equiv \frac{\sum_{m \in \mathcal{M}_{jt}} v_{mt}\, \widehat{\mu}_{mt}}{\sum_{m \in \mathcal{M}_{jt}} v_{mt}}, \tag{17}$$

where $\mathcal{M}_{jt}$ is the set of models owned by provider j in month t. The market-level markup index is constructed analogously by weighting across all models:

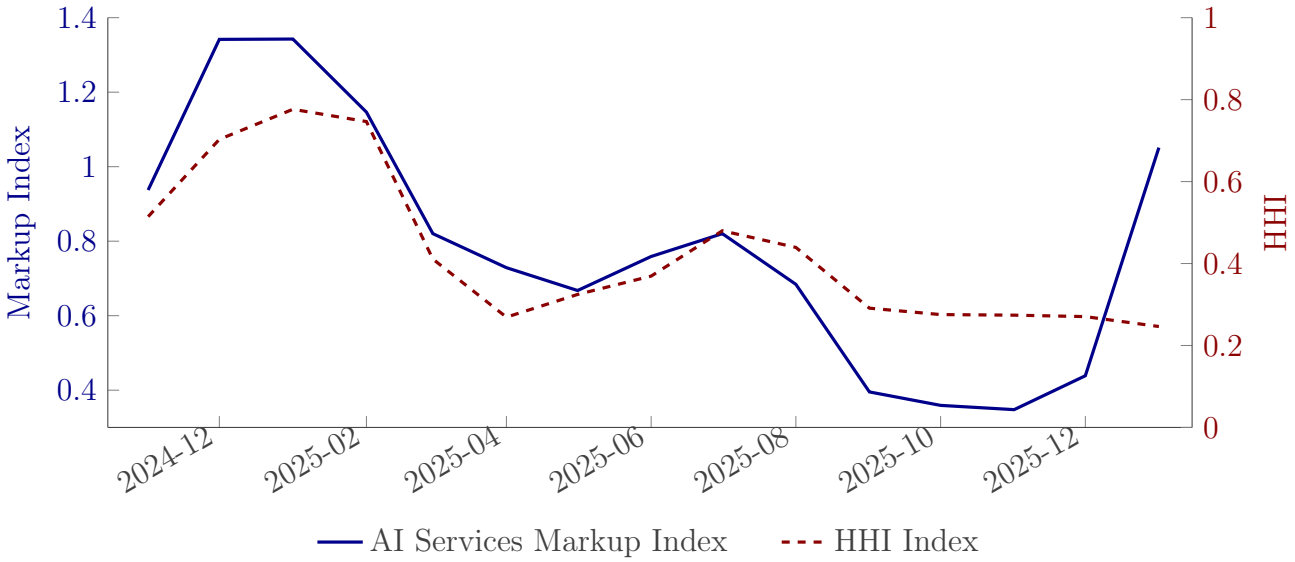
$$\widehat{\mu}_{t}^{mkt} \equiv \frac{\sum_{m} v_{mt}\, \widehat{\mu}_{mt}}{\sum_{m} v_{mt}}. \tag{18}$$

Token-value weighting ensures that markups on high-usage, high-price models receive proportionally greater weight. I additionally compute the Herfindahl-Hirschman Index (HHI) based on provider-level token-value shares to measure concentration in the AI services market.

Markup results. Figure 9 plots the market-level markup index alongside provider concentration. Two features stand out. First, the markup index and HHI comove closely throughout the sample period: both decline as new entrants and open models gain token share through mid-2025, and both rise subsequently as the market reconcentrates. This comovement lends external validity to the markup estimates. If the recovered markups were spurious, one would not expect them to track an independently constructed concentration measure so closely. Second, the markup index contains information beyond what HHI captures: while HHI measures the distribution of market shares, the markup index reflects providers' pricing power, accounting for demand elasticities and multi-product internalization. This distinction is salient in the sharp markup increase from December 2025 through February 2026, which occurs without a commensurate rise in HHI, suggesting that the episode reflects intensified pricing power among specific model providers rather than a broad increase in concentration.

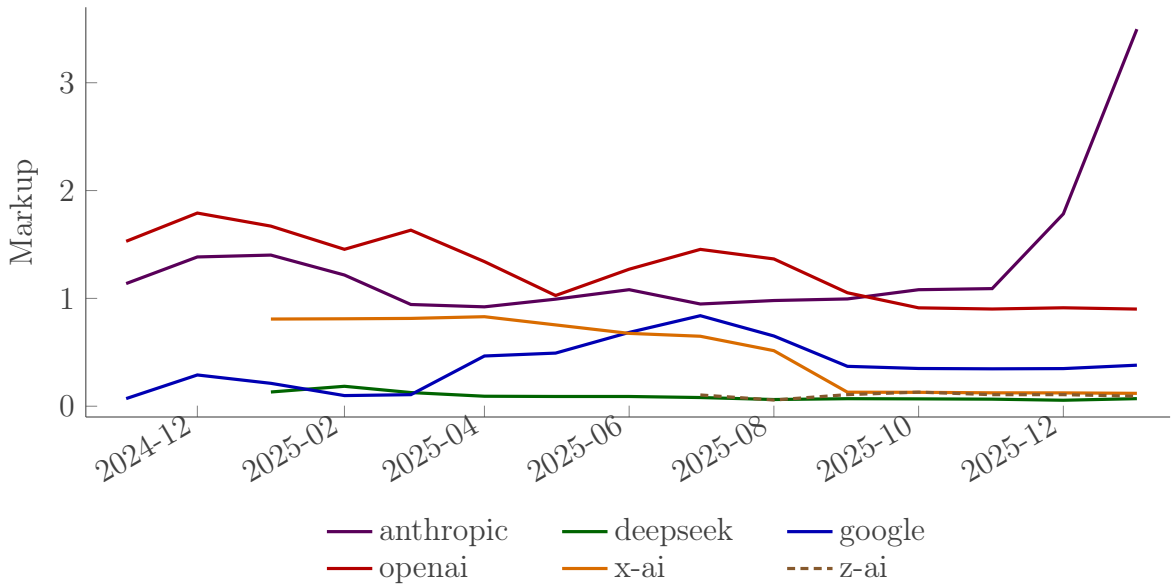
Figure 10 disaggregates markups by provider and identifies Anthropic as the primary driver of the markup surge from December 2025 through February 2026: its token-value-weighted markup rises sharply, pulling up the market-level index while other providers' markups remain comparatively stable. This surge coincides with Anthropic's successive releases of frontier models, Claude Opus 4.5 in late November 2025 followed by Claude Opus 4.6 and Sonnet 4.6 in February 2026, which reinforced its market position and allowed it to sustain premium pricing in an otherwise competitive market. More broadly, several patterns emerge across the full sample. Markups differ substantially across providers at any given point in time, with providers operating premium frontier models maintaining higher markups, consistent with differentiated product power in the high-capability segment. Within providers, markups are time-varying, responding to competitive entry, model releases, and shifts in the cost environment. In particular, the entry of aggressively priced competitors, notably the DeepSeek R1 release in early 2025, is associated with compression in rivals' markups, consistent with the competitive pressure implied by the demand estimates. These provider-level markup series serve as the key moderating variable in the micro pass-through analysis of Section 6.

Figure 9: Market Markup Index and Provider Concentration



Notes. The blue series (left axis) shows the aggregate AI services markup index. The red dashed series with square markers (right axis) shows provider concentration measured by token-value HHI.

Figure 10: Token-Value-Weighted Markups by Provider



Notes. The figure plots provider-level token-value-weighted markup series over time.

6 From Micro AI Pricing to Macro Inflation Pass-through

The previous sections established that AI inference is a rapidly improving, differentiated intermediate input whose constant-quality price has fallen steeply (Section 4), and that the market for AI services is oligopolistic, with meaningful and time-varying markups across providers (Section 5). This section investigates how these features jointly determine the transmission of upstream cost shocks to downstream prices, and who captures the value created by falling AI costs.

The analysis proceeds in three steps, progressively widening the lens from the micro mechanism to economy-wide implications. Subsection 6.1 estimates the micro pass-through of GPU cost shocks into quality-adjusted AI prices and documents that higher-markup providers absorb more of the cost decline. Subsection 6.2 quantifies value creation and distribution within the AI services market, finding that while oligopolistic pricing imposes a measurable welfare cost through markup rents and deadweight loss, the entry of new, higher-capability models accounts for the substantial majority of cumulative consumer surplus gains. Finally, Subsection 6.3 embeds the AI services sector in the U.S. production network using input–output tables, combines sector-level AI exposure measures with the quality-adjusted price index to derive implied PPI effects, and decomposes these into direct, indirect, and markup-absorption channels.

6.1 Micro Pass-Through

I estimate the pass-through of GPU costs into quality-adjusted prompt prices using the following specification:

$$\Delta \log \tilde{p}_{mt} = \theta_0 \Delta \log gpu_t + \theta_1 \Big(\Delta \log gpu_t \times \widehat{\mu}^{\text{prov}}_{j(m),t-1} \Big) + \theta_2 \widehat{\mu}^{\text{prov}}_{j(m),t-1} + \alpha_{j(m)} + \varepsilon_{mt}, \quad (19)$$

where $\log \tilde{p}_{mt}$ is the quality adjust price for model m at time t, $\widehat{\mu}^{\text{prov}}_{j(m),t-1}$ is the lagged company-level token-value-weighted markup (for company j that owns model m), and $\alpha_{j(m)}$ denotes company fixed effects.

Table 5 reports the results. A negative θ_1 implies that higher-markup providers absorb

Table 5: AI Services Cost Pass-through

$\Delta \log gpu_t$	0.1566*	0.2285**
	(0.0839)	(0.0938)
$\Delta \log gpu_t \times \widehat{\mu}^{\text{prov}}_{j(m),t-1}$		-0.1520**
		(0.0661)
$\widehat{\mu}^{\text{prov}}_{j(m),t-1}$		0.0041
		(0.0255)
company FE	Yes	Yes
R-squared Adj.	0.0037	0.0026
N	1003	1003

Notes. This table reports estimates of GPU cost pass-through into quality-adjusted AI prompt prices. The dependent variable is $\Delta \log \tilde{p}_{mt}$, the monthly change in the log quality-adjusted prompt price (USD per 1M tokens) for model m. $\Delta \log gpu_t$ is the monthly change in the log H100 GPU rental price (USD per GPU-hour). $\widehat{\mu}^{\text{prov}}_{j(m),t-1}$ is the lagged provider-level gross markup, constructed as a token-value-weighted average across all models owned by company j. Company fixed effects are included in all columns. Standard errors are clustered at the provider level and reported in parentheses. Significance levels: *** 1%, ** 5%, * 10%.

more of a common GPU cost shock, i.e., they exhibit lower pass-through. The direct pass-through coefficient is $\hat{\theta}_0 = 0.229$, while the interaction term $\hat{\theta}_1 = -0.152$ is negative and significant, confirming that higher-markup providers absorb a larger share of GPU rental cost movements. The level effect of markups is economically small and statistically insignificant, indicating that markup levels do not independently trend with AI price changes after controlling for their interaction with cost shocks.

The estimates imply that the effective pass-through rate varies substantially across the markup distribution. For a hypothetical competitive provider pricing at marginal cost ($\mu = 1$), the implied pass-through rate is $\hat{\theta}_0 + \hat{\theta}_1 \times 1 = 0.077$: roughly 8 cents of each dollar of GPU cost decline would reach API prices. As markups rise, pass-through declines. Setting $\hat{\theta}_0 + \hat{\theta}_1 \mu^* = 0$ yields a zero-pass-through threshold of $\mu^* = \hat{\theta}_0 / |\hat{\theta}_1| \approx 1.50$: providers with gross markups above this threshold fully absorb GPU cost declines into their margins, transmitting none of the upstream cost reduction to downstream users. Figure 10 shows that several providers, notably Anthropic after its frontier-model releases in late 2025, sustain markups well above this threshold for extended periods, consistent with complete markup absorption of cost shocks in the premium segment of the market.

One feature of the estimation warrants comment. The adjusted R^2 values are low, reflecting the fact that GPU cost variation explains only a small share of the total month-to-month variation in quality-adjusted AI prices. This is expected: model-level price changes are also driven by competitive entry and provider-specific strategic decisions that are orthogonal to the common GPU cost shock. The pass-through coefficients should therefore be interpreted as capturing the *GPU-cost-driven* component of price adjustment, not the totality of price dynamics.

6.2 Welfare in the AI Services Market

The micro pass-through results establish that market power attenuates the transmission of GPU cost declines to AI API prices. But how large is the welfare cost of this attenuation, and how does it compare to the welfare gains from the entry of new, higher-capability models? This subsection quantifies value creation and distribution *within* the AI services market.

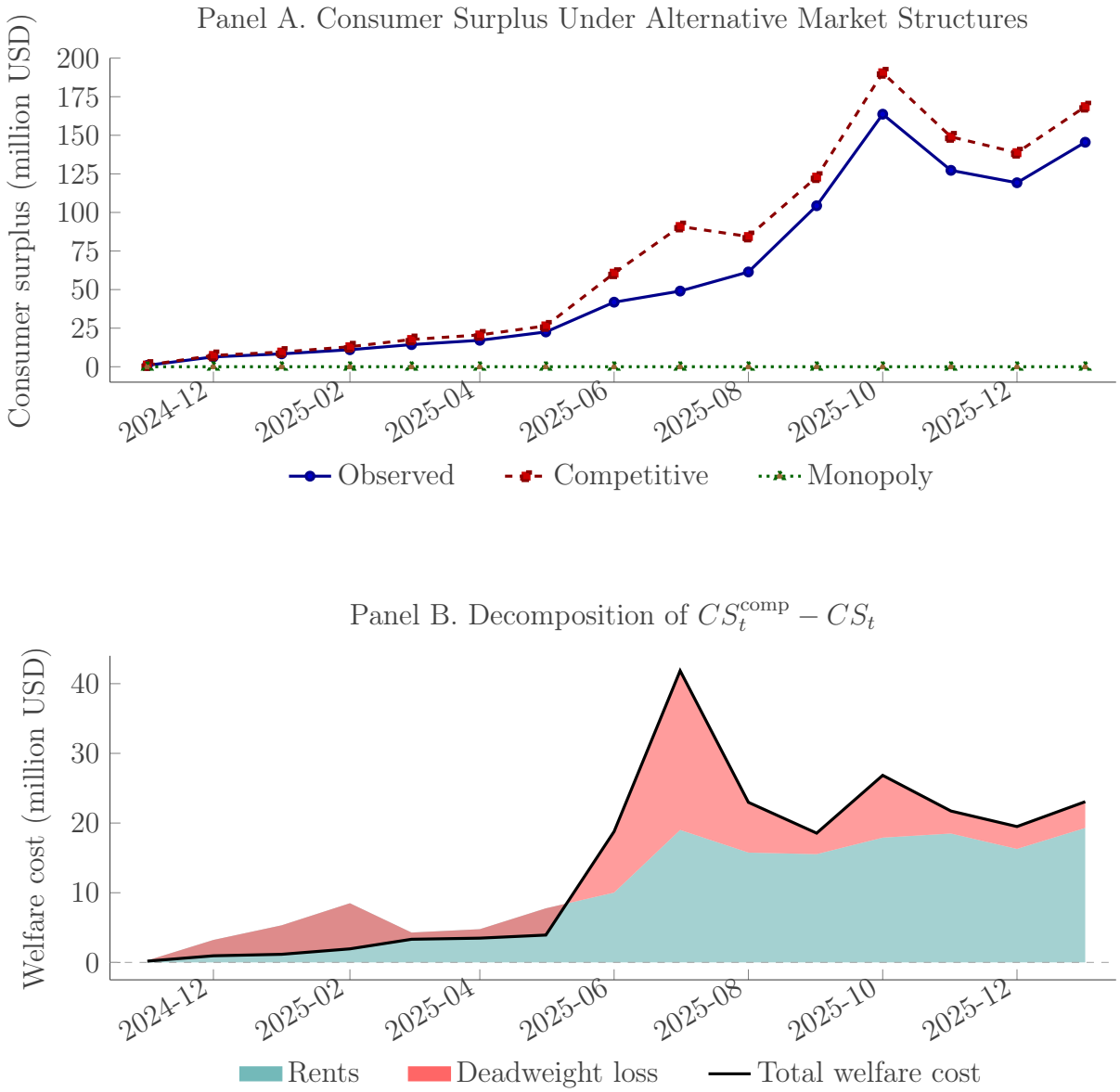
The logit demand system estimated in Section 5 yields a closed-form expression for aggregate consumer surplus. Following Small and Rosen (1981), the market-level surplus in month t is

$$CS_t = \frac{1}{\hat{\alpha}} \ln\left(1 + \sum_{m=1}^{J_t} \exp\left(\hat{\delta}_{mt}\right)\right) \times \frac{\text{tokens}_t^{OR}}{\lambda}, \tag{20}$$

where $\hat{\alpha}$ is the estimated price coefficient from equation (6); $\hat{\delta}_{mt} = -\hat{\alpha} \log p_{mt}^{\text{prompt}} + \hat{\beta}'\mathbf{Q}_m + \hat{\delta}'\mathbf{X}_{mt} + \hat{\tau}_t$ is the estimated mean utility of model m in month t; and tokens_t^{OR} is aggregate OpenRouter token volume in month t, divided by λ to scale up to total market volume. Consumer surplus is thus expressed in dollar terms and reflects the value that developers and firms purchasing AI inference derive from access to the available set of models at prevailing prices.

Counterfactual surplus under competitive pricing. To quantify the welfare cost of market power, I construct a competitive counterfactual in which all AI providers price at marginal cost. Using the implied marginal costs $\widehat{mc}_{mt}$ recovered from the demand

Figure 11: Consumer Surplus Under Alternative Market Structures



Notes. Panel A plots consumer surplus CS_t under three regimes: observed prices (blue), the competitive counterfactual CS_t^{comp} in which all providers price at marginal cost (red dashed), and the monopoly counterfactual CS_t^{mono} in which a single provider owns all models (green dotted). Surplus is computed from the logit demand estimates in Section 5 using equation (20). The gap between the competitive and observed series measures the total welfare cost of market power; the gap between observed and monopoly measures the welfare gain from existing competition. Panel B decomposes $CS_t^{\text{comp}} - CS_t$ into transfers (markup rents) and deadweight loss as defined in equations (23)–(24); the black line reports total welfare cost.

estimates, I replace observed prices with marginal costs and recompute mean utilities:

$$\hat{\delta}_{mt}^{\text{comp}} = -\hat{\alpha} \log \widehat{mc}_{mt} + \hat{\beta}' \mathbf{Q}_m + \hat{\delta}' \mathbf{X}_{mt} + \hat{\tau}_t. \tag{21}$$

The competitive consumer surplus is then

$$CS_t^{\text{comp}} \;=\; \frac{1}{\hat{\alpha}}\;\ln\!\left(\,1+\sum_{m=1}^{J_t}\exp\!\big(\hat{\delta}_{mt}^{\text{comp}}\big)\right)\;\times\;\frac{\text{tokens}_t^{OR}}{\lambda},. \tag{22}$$

Similarly, I compute a monopoly counterfactual CS_t^{mono} using the prices that would prevail if all models were owned by a single provider, obtained by setting the ownership matrix $\Omega_{ij,t} = 1$ for all (i, j) and re-solving for the monopoly price vector p_t^{mono}.

Welfare decomposition: surplus, rents, and deadweight loss. The difference between competitive and observed surplus decomposes the welfare effects of market power into two components. Define the *transfer* from AI users to providers as the aggregate markup revenue:

$$T_t \;=\; \sum_{m=1}^{J_t}(p_{mt}^{\text{prompt}} - \widehat{mc}_{mt})\;\times\;\frac{\text{tokens}_t^{OR}}{\lambda}, \tag{23}$$

where $p_{mt}^{\text{prompt}} - \widehat{mc}_{mt}$ is the level markup and tokens_{mt}^{OR} is token volume. The *deadweight loss* is the residual welfare destruction that accrues to neither users nor providers:

$$DWL_t \;=\; \big(CS_t^{\text{comp}} - CS_t\big)\;-\;T_t. \tag{24}$$

The identity $CS_t^{\text{comp}} - CS_t = T_t + DWL_t$ provides a complete accounting of the welfare cost of oligopoly pricing: part of the surplus loss is a transfer to AI providers (rents), and the remainder is destroyed.

Price versus variety/quality channels. The welfare gains that AI users have realized over the sample period reflect two distinct sources: falling prices on existing models (the price channel) and the entry of new, higher-capability models that expand the choice set (the variety/quality channel). In the AI services market, quality improvement does not operate through existing models becoming more capable as each model's capability indices $\mathbf{Q}_m$ are fixed at release, but rather through releases of new models that supplant or supplement older ones. The counterfactual therefore freezes the *product set*, not individual model qualities.

Let $\mathcal{J}_{t_0}$ denote the set of models available in the initial period. To isolate the price channel, I construct a *product-set-frozen* counterfactual that restricts the choice set to the models available at t_0, with their time-invariant capabilities, while allowing prices and operational characteristics to evolve as observed for models that remain listed:

$$\hat{\delta}_{mt}^{\text{set-frozen}} \;=\; -\hat{\alpha}\,\log p_{mt}^{\text{prompt}} \;+\; \hat{\beta}'\mathbf{Q}_m \;+\; \hat{\delta}'\mathbf{X}_{mt} \;+\; \hat{\tau}_t, \qquad m \in \mathcal{J}_{t_0} \cap \mathcal{J}_t. \tag{25}$$

The summation runs over $\mathcal{J}_{t_0} \cap \mathcal{J}_t$—the original models that remain active in month t—so that delisted models exit the counterfactual naturally rather than requiring phantom price assumptions.[7] The corresponding product-set-frozen consumer surplus is

$$CS_t^{\text{set-frozen}} \;=\; \frac{1}{\hat{\alpha}}\,\ln\!\left(1 + \sum_{m \in \mathcal{J}_{t_0} \cap \mathcal{J}_t} \exp\!\left(\hat{\delta}_{mt}^{\text{set-frozen}}\right)\right). \tag{26}$$

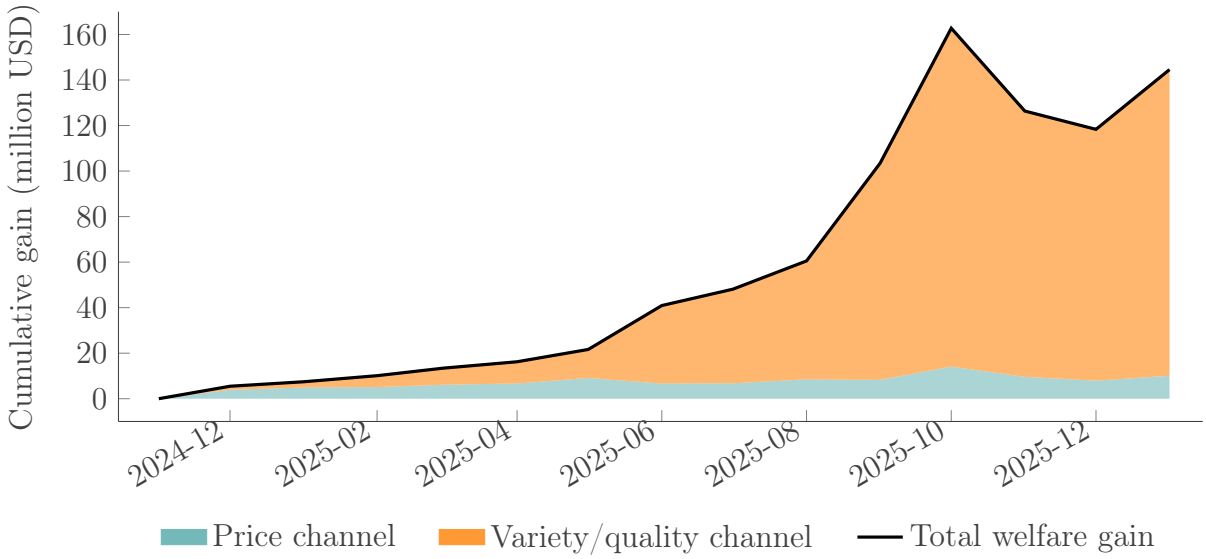
The decomposition of cumulative welfare gains is then:

$$\underbrace{CS_t - CS_{t_0}}_{\text{Total welfare gain}} \;=\; \underbrace{CS_t^{\text{set-frozen}} - CS_{t_0}}_{\text{Price channel}} \;+\; \underbrace{CS_t - CS_t^{\text{set-frozen}}}_{\text{Variety/quality channel}}. \tag{27}$$

The price channel captures the surplus gain from price declines on the original product set. The variety/quality channel captures the surplus gain from access to new, higher-capability models that entered after t_0, reflecting the cumulative product innovation by AI providers over the sample period. This channel benefits AI users regardless of the prevailing level of markups, conditional on entry occurring. Figure 12 plots the cumulative decomposition and reveals that the variety/quality channel accounts for the substantial majority of total welfare gains throughout the period: by the end of the sample, new model entry explains most of the cumulative surplus gain, while price declines on the original product set contribute a comparatively modest share.

[7] This conservative treatment means that the product-set-frozen surplus $CS_t^{\text{set-frozen}}$ may *decline* toward the end of the sample if a large share of initial-period models are delisted, attributing part of the delisting effect to the variety/quality channel. In practice, most high-usage models from $\mathcal{J}_{t_0}$ remain listed throughout the sample.

Figure 12: Welfare Decomposition: Price Channel vs. Variety/Quality Channel



Notes. The figure decomposes cumulative consumer surplus gains into the price channel (restricting the choice set to models available in the initial period) and the variety/quality channel (the residual gain from the entry of new, higher-capability models), as defined in equation (27). The shaded area between the two series represents the variety/quality channel; the area below the product-set-frozen series represents the price channel.

Scope of the counterfactual and the Schumpeterian trade-off. The competitive and monopoly counterfactuals hold the set of available models and marginal costs fixed, isolating the static pricing effect of market structure. This approach is standard in the differentiated-products literature (Berry et al., 1995; Nevo, 2001) and provides a well-defined welfare benchmark, but it abstracts from dynamic responses, notably changes in entry incentives and quality investment. This abstraction is salient here because the markup rents captured by T_t also finance the R&D expenditure that produces the new, higher-capability models driving the variety/quality channel. The static welfare cost $T_t + DWL_t$ therefore overstates the dynamic welfare cost to the extent that lower rents would slow new model development—and the dominance of the variety/quality channel in Figure 12 suggests this dynamic offset may be substantial. Fully resolving this trade-off would require a dynamic model of endogenous innovation that is beyond the scope of this paper.

6.3 Macro Impact: AI Deflation, Production Networks, and PPI Contributions

The final step is to ask how a change in the price of AI services ripples through the rest of the economy. Because AI is an intermediate input, a cost reduction does not stay within the AI sector: it lowers costs for any firm that purchases AI services directly, and, through supply-chain linkages, for any firm whose suppliers do. This cost-push propagation through production networks is well-documented in the literature (Acemoglu et al., 2012; Baqaee and Farhi, 2019); tracing it here requires embedding the quality-adjusted AI price index in a Leontief input–output framework.

I start with the BEA 2024 Use table at the summary level, which records the dollar value of intermediate purchases Z_{js} used by industry s alongside gross output x_s for each of the 71 industries. Normalizing by gross output gives the direct-requirements coefficient

$$A_{js} \;=\; \frac{Z_{js}}{x_s}, \tag{28}$$

which is the dollar value of commodity j required per dollar of industry s's output. A single round of input purchases, however, captures only the immediate cost push: sector s buys AI services, but s's own suppliers do too, as do their suppliers. To aggregate these successive rounds of indirect exposure into a single summary statistic, I invert the direct-requirements matrix. The price Leontief inverse,

$$\mathbf{L}^p \;\equiv\; \left(\mathbf{I} - \mathbf{A}^\top\right)^{-1}, \tag{29}$$

captures how a cost shock to one commodity flows into the unit costs of every downstream industry: entry $[L^p]_{sc}$ gives the total increase in industry s's unit cost per dollar increase in the price of commodity c, summing the direct purchase and all indirect supply-chain

routes.[8]

Mapping AI services and constructing sector exposure. AI inference services do not have a dedicated industry code in the BEA classification. The closest match is NAICS 514 ("Data processing, internet publishing, and other information services"), the industry group within which cloud AI API providers are classified.[9] Because NAICS 514 also covers non-AI activities such as traditional data hosting and web publishing, I introduce a calibration parameter $\kappa \in (0, 1]$ representing the share of NAICS 514 expenditure attributable to AI services, set to $\kappa = 0.50$ in the baseline. Letting c index the NAICS 514 commodity, I define two exposure measures for each downstream industry s. The *direct* AI input share,

$$b_s^{AI} \;=\; \kappa \;\cdot\; A_{cs}, \tag{30}$$

measures the share of sector s's gross output spent on AI services in a single round of production. The *total* AI supply-chain exposure,

$$E_s^{AI} \;=\; \kappa \;\cdot\; \left[L^p\right]_{sc}, \tag{31}$$

adds to b_s^{AI} the AI content embedded in every intermediate input that sector s purchases; the difference $E_s^{AI} - b_s^{AI}$ is thus a measure of indirect exposure.

Figure 13 plots E_s^{AI} for the top 20 sectors. NAICS 514 itself tops the ranking, followed by other information-intensive sectors such as publishing, computer systems design, and professional and technical services. This cross-sectoral heterogeneity in E_s^{AI} is the key source of variation exploited below: combining the exposure measures with the quality-adjusted AI price index P_t^{AI} from Section 4 yields a sector-specific implied PPI effect that

[8]The price and quantity Leontief inverses are transposes of each other: $\mathbf{L}^p = (\mathbf{I} - \mathbf{A}^\top)^{-1} = [(\mathbf{I} - \mathbf{A})^{-1}]^\top = \mathbf{L}^\top$, so $[L^p]_{sc} = L_{cs}$. This means the cost exposure of sector s to a price shock in commodity c equals the total input requirement of commodity c per dollar of final demand for sector s—the same object as an entry of the quantity Leontief inverse, with indices transposed. In practice, this equivalence means only one matrix inverse needs to be computed: the standard quantity Leontief inverse $\mathbf{L} = (\mathbf{I} - \mathbf{A})^{-1}$, with E_s^{AI} read off as the (c, s) entry rather than the (s, c) entry. I verify numerically that the maximum element-wise difference $|\mathbf{L}^p - \mathbf{L}^\top|$ is below 10^{-10}, confirming this equivalence holds in the data. I also verify that the spectral radius $\rho(\mathbf{A}) < 1$, ensuring that both inverses are well-defined.

[9]NAICS 514 includes 5142 (Data processing, hosting, and related services) and 5141 (Information services). Major AI API providers such as OpenAI, Anthropic, and Google's cloud AI division fall within this classification.

can be aggregated into an economy-wide deflationary contribution.

Implied PPI effects. The Leontief framework delivers a first-order approximation of the implied PPI effect for each downstream sector s:

$$\Delta \log PPI_{s,t} \;\approx\; E_s^{AI} \;\times\; \Delta \log P_t^{AI}, \tag{32}$$

where $E_s^{AI} = \kappa \cdot L_{cs}$ is the total AI supply-chain exposure from equation (31), and $\Delta \log P_t^{AI}$ is the change in the quality-adjusted AI services price index from equation (2).

Aggregating to an economy-wide PPI contribution by weighting sectors by their gross-output shares $\omega_s \equiv x_s / \sum_{s'} x_{s'}$ (row T018 of the BEA Use table) gives

$$\Delta \log PPI_t^{agg} \;\approx\; \sum_s \omega_s \;\cdot\; E_s^{AI} \;\cdot\; \Delta \log P_t^{AI}. \tag{33}$$

Given the steep decline in the quality-adjusted index, roughly -1.5 log points over 2024–2026 (Figure 4), this channel delivers a meaningful deflationary contribution.

Three-way decomposition. I decompose the *observed* implied PPI contribution into two channels based on the observed quality-adjusted AI price change $\Delta \log P_t^{AI}$, and separately construct GPU-based counterfactual benchmarks that isolate the incremental pass-through that would arise under alternative market structures.

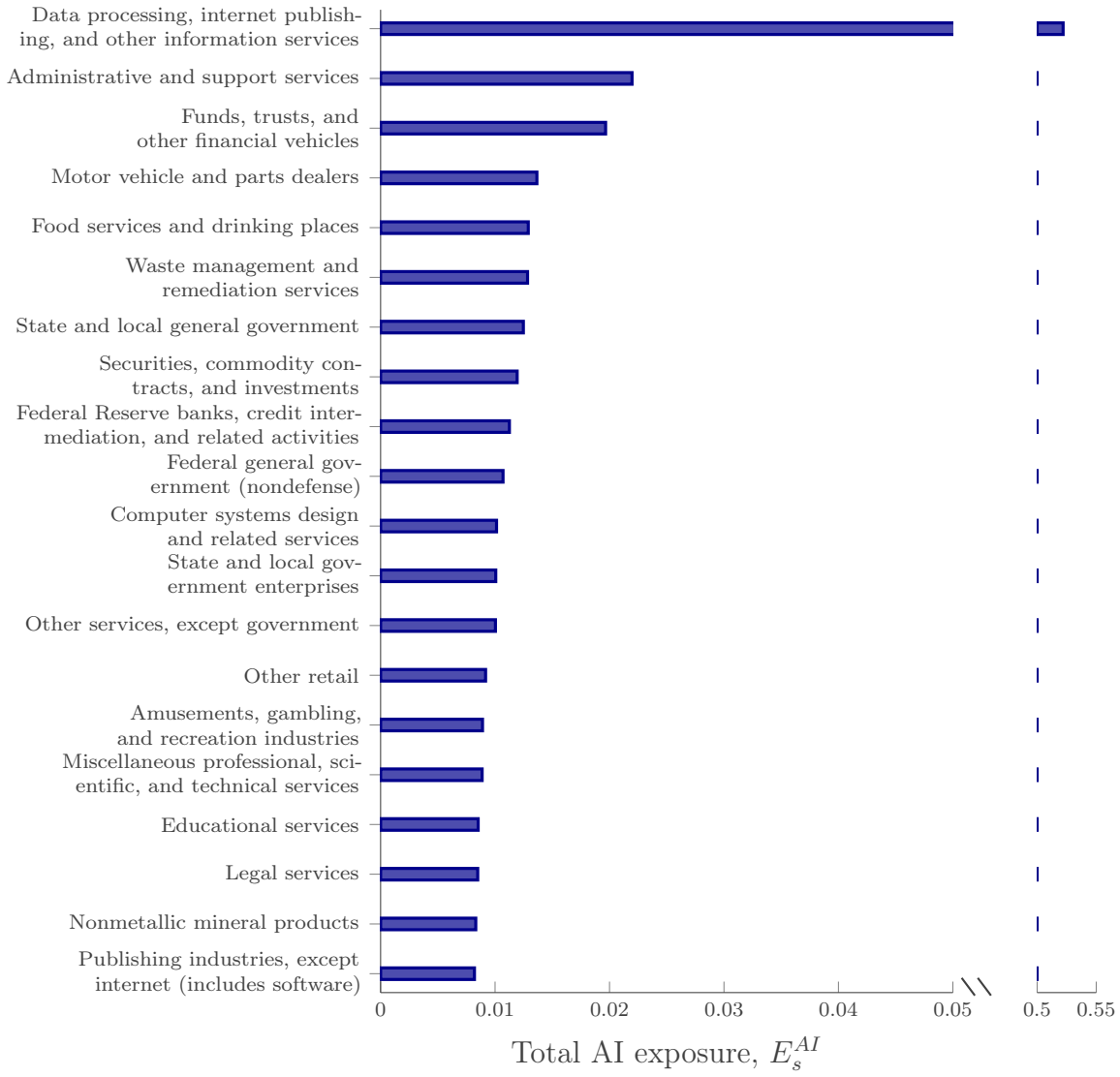
(i) Direct channel. The direct effect captures the cost push through sector s's own purchases of AI services:

$$\text{Direct}_{s,t} \;=\; b_s^{AI} \;\times\; \Delta \log P_t^{AI}. \tag{34}$$

(ii) Indirect (supply-chain) channel. The indirect effect captures the additional exposure arising because sector s's suppliers themselves purchase AI services:

$$\text{Indirect}_{s,t} \;=\; \left(E_s^{AI} - b_s^{AI}\right) \;\times\; \Delta \log P_t^{AI}. \tag{35}$$

Figure 13: Cross-Sector Distribution of Total AI Exposure



Notes. The figure plots total AI supply-chain exposure $E_s^{AI} = \kappa \cdot L_{cs}$ for the top 20 sectors, ranked in descending order. The underlying input–output coefficients are from the 2024 BEA Use table at the summary level. NAICS 514 ("Data processing, internet publishing, and other information services") is used as the AI-producing sector. The calibration parameter is $\kappa = 0.50$. The x-axis is broken between 0.10 and 0.50 so the lower-exposure sectors remain legible.

By construction, the observed PPI effect $E_s^{AI} \times \Delta \log P_t^{AI}$ is exactly the sum of these two channels.

(iii) Market-power wedge. This channel quantifies how much of the upstream GPU cost decline is absorbed by AI providers' market power and therefore does *not* reach downstream sectors. Using the pass-through rate from equation (19), $\partial \Delta \log \tilde{p} / \partial \Delta \log gpu = \hat{\theta}_0 + \hat{\theta}_1 \mu$, I construct GPU-driven AI price changes under prevailing oligopolistic markups, perfect

competition ($\mu = 1$), and a monopoly counterfactual μ^{mono}:[10]

$$\Delta \log P_t^{AI,\,gpu,\,k} \;\equiv\; \left(\hat{\theta}_0 + \hat{\theta}_1\, \mu^k\right) \cdot \Delta \log gpu_t, \qquad \mu^k \;\in\; \left\{\bar{\mu}_{t-1},\; 1,\; \mu^{\text{mono}}\right\}. \tag{36}$$

The *market-power wedge* measures the PPI effect absorbed by oligopoly pricing relative to the observed pass-through:

$$\text{Wedge}_{s,t}^{k} \;\equiv\; E_s^{AI} \;\times\; \left[\Delta \log P_t^{AI,\,gpu,\,k} \;-\; \Delta \log P_t^{AI,\,gpu,\,\bar{\mu}}\right], \qquad k \;\in\; \{\text{comp},\, \text{mono}\}. \tag{37}$$

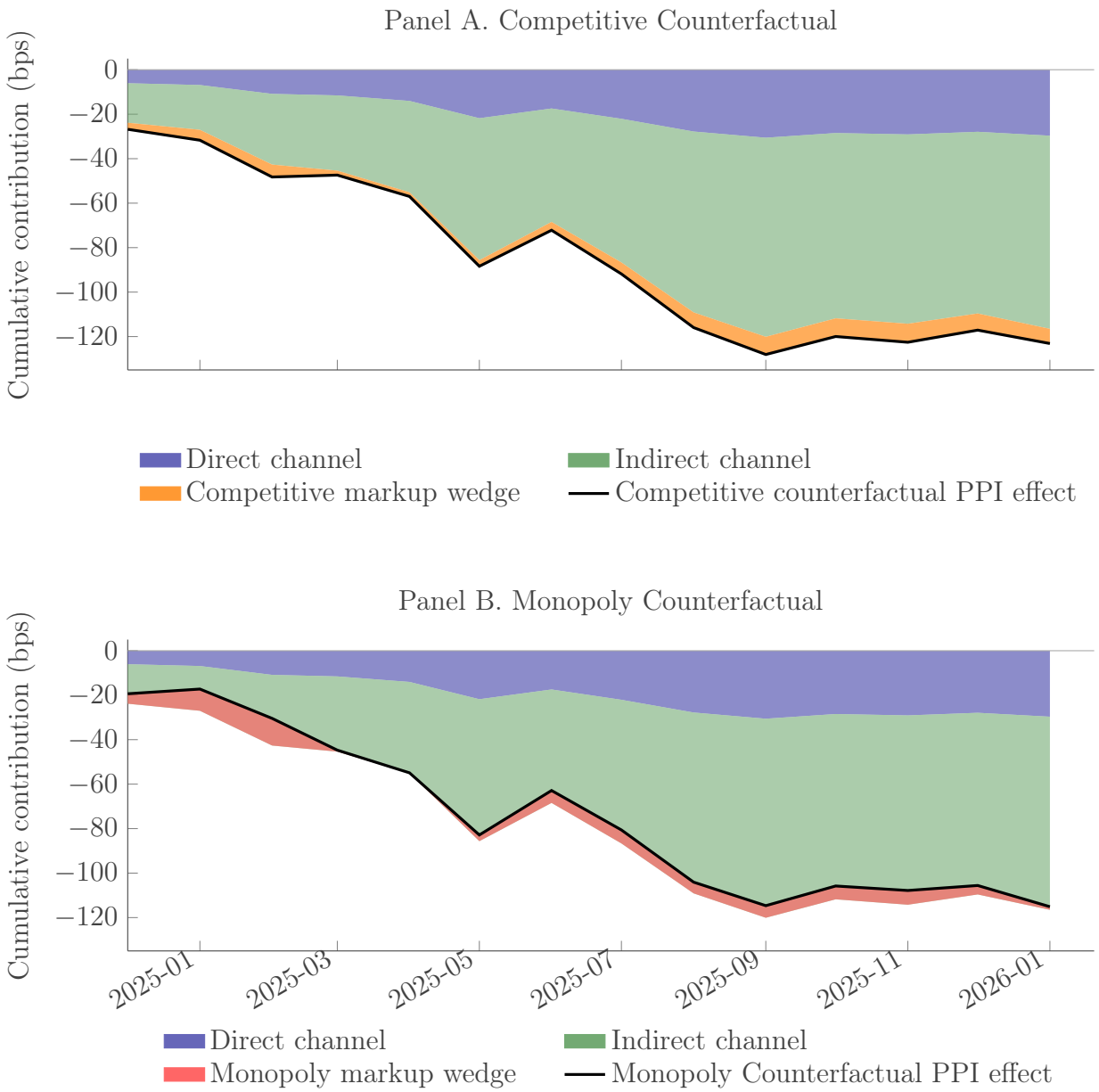
Since $\hat{\theta}_1 < 0$ (Table 5) and actual markups exceed unity, $\text{Wedge}_{s,t}^{\text{comp}} < 0$ when GPU costs fall indicates that competitive pricing would have transmitted more deflation downstream, while $\text{Wedge}_{s,t}^{\text{mono}} > 0$ indicates monopoly pricing would have transmitted less. The counterfactual total PPI effect under each scenario is obtained by adding the respective wedge to the observed effect. Aggregating each object with gross-output weights ω_s as in equation (33) yields economy-wide time series for the direct, indirect, and market-power wedge channels.

Figure 14 translates the micro pass-through result into economy-wide terms. Three features stand out. First, the indirect supply-chain channel accounts for the majority of the cumulative deflationary contribution in both panels, reflecting the fact that AI cost reductions propagate beyond the sectors that purchase AI services directly. Second, and central to the paper's argument, the competitive wedge in Panel A shows that oligopolistic pricing has meaningfully attenuated this deflationary impulse: under marginal-cost pricing, the cumulative PPI contribution would have been appreciably more negative, with the gap between the competitive counterfactual and the observed effect representing deflation that was absorbed into provider margins rather than passed through to downstream firms. Third, the monopoly wedge in Panel B provides a lower bound on pass-through:

[10] $\bar{\mu}_{t-1}$ is the usage-weighted average markup from Section 5. The monopoly counterfactual is set to $\mu^{\text{mono}} = 3.498$, the highest provider markup observed in the sample. The theoretical monopoly markup, recovered by re-solving the multi-product Bertrand system under a single owner for all models in Section 6.2, is far outside the support of the pass-through regression in equation (19), which is estimated over the observed markup range of approximately 1 to 3.5; extrapolating the linear approximation $\hat{\theta}_0 + \hat{\theta}_1\, \mu$ to that value yields implausibly large wedges. The maximum observed markup therefore serves as a conservative upper bound on the market-power wedge that is credibly identified by the data.

even the most concentrated market structure consistent with the data would still deliver a substantial deflationary contribution through the supply chain, because the indirect channel operates on the quality-adjusted price index P_t^{AI} which has fallen steeply regardless of market structure, rather than solely on the GPU-cost-driven component moderated by markups. Taken together, the two panels bound the economy-wide deflationary effect of AI: market power does not eliminate it, but it does attenuate it, and the magnitude of that attenuation is economically non-trivial relative to the direct and indirect channels.

Figure 14: Cumulative PPI Contribution of AI: Direct, Indirect, and Market-Power Wedge



Notes. The figure plots cumulative economy-wide implied PPI contributions of AI, aggregated using sector gross-output weights. Panel A shows the competitive counterfactual (what pass-through would have been under marginal-cost pricing) and Panel B shows the monopoly counterfactual. In both panels, the blue area is the cumulative direct channel, the green area is the additional cumulative indirect supply-chain channel, and the third shaded area is the cumulative market-power wedge—the deflation absorbed by oligopoly pricing rather than passed through to downstream sectors, as defined in equation (37). The black line is the cumulative counterfactual total PPI effect, equal to the sum of the three components. Units are basis points.

7 Conclusion

This paper constructs the first constant-quality price index for AI inference services using hedonic regressions across four capability domains, and shows that quality-adjusted AI prices fell by approximately 78 percent from early 2024 through early 2026, roughly twice the decline suggested by nominal prices alone. Within this rapidly improving market, coding proficiency and reliability command persistent price premia, while general reasoning and mathematical ability enter with negative coefficients once other quality attributes are controlled for, suggesting that the market's willingness to pay concentrates on dimensions tied to specific high-value downstream tasks rather than on general-purpose reasoning capability. How much of this deflation actually reaches downstream firms, however, depends on market structure.

The AI services market is oligopolistic, and this market structure has first-order consequences for cost transmission. Estimating a demand system on OpenRouter token usage data and recovering markups via multi-product Bertrand inversion, I document substantial and time-varying markup heterogeneity across providers. GPU cost pass-through is positive and significant, but higher-markup providers absorb a disproportionately larger share of upstream cost declines: the estimates imply a zero-pass-through threshold at a gross markup of approximately $\mu^* \approx 1.50$, above which providers fully absorb GPU cost reductions into their margins. Several frontier-model providers sustain markups well above this threshold for extended periods, most notably following major model releases. The mechanism, endogenous markup absorption of cost shocks by firms with market power, echoes a central finding of the exchange rate pass-through literature. This paper documents it operating in a new and increasingly consequential intermediate input market.

Embedding the quality-adjusted price index in a Leontief input–output framework, I find that AI deflation delivers a cumulative implied PPI contribution of roughly 120 basis points over 2024–2026, with the indirect supply-chain channel accounting for the majority of the aggregate effect. Under competitive pricing, this deflationary contribution would have been appreciably larger; the difference represents deflation absorbed by oligopoly margins rather than transmitted to downstream sectors. Turning to welfare, a

decomposition reveals that the entry of new, higher-capability models, not price declines on existing ones, accounts for the substantial majority of cumulative consumer surplus gains. This is consistent with a Schumpeterian trade-off in which the markup rents that attenuate cost pass-through also finance the frontier-model development that generates the bulk of downstream value.

These findings carry implications for price measurement and competition policy. On measurement, the roughly twofold gap between nominal and quality-adjusted AI price declines echoes a longstanding concern in the price index literature: statistical agencies that rely on nominal token prices will understate AI deflation in both the PPI and, to the extent that AI-intensive goods enter final consumption, the CPI. As AI input shares grow, this mismeasurement will bias estimates of real output growth, sectoral productivity, and the real interest rate. The hedonic approach developed here, which relies on publicly available benchmark scores and posted API prices, offers a readily implementable framework for producing quality-adjusted AI deflators suitable for incorporation into official price statistics.

On competition policy, the finding that providers with gross markups above approximately 1.50 fully absorb GPU cost declines suggests that upstream compute savings may be slow to reach downstream firms absent competitive pressure. The DeepSeek entry episode, which compressed rivals' markups and accelerated pass-through, illustrates how entry can transmit deflation that market concentration otherwise attenuates. The dominance of the variety and quality channel in the welfare decomposition, however, counsels caution: policies that compress margins may also weaken the incentive to develop frontier models, the primary source of welfare gains over the sample period.

Taken together, the results provide evidence that AI is a deflationary intermediate input whose pass-through to downstream prices is moderated by endogenous market power, and that the welfare gains from this rapidly evolving market arise predominantly from the entry of new, higher-capability models rather than from the repricing of existing ones.

References

Acemoglu, Daron and Pascual Restrepo, "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, 2018, *108* (6), 1488–1542.

_ **and** _ , "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives*, 2019, *33* (2), 3–30.

_ **, Vasco M. Carvalho, Asuman Ozdaglar, and Ali Tahbaz-Salehi**, "The Network Origins of Aggregate Fluctuations," *Econometrica*, 2012, *80* (5), 1977–2016.

Advisory Commission to Study the Consumer Price Index, "Toward a More Accurate Measure of the Cost of Living," Technical Report, Senate Finance Committee 1996. (Boskin Commission Report).

Aghion, Philippe, Benjamin F. Jones, and Charles I. Jones, "Artificial Intelligence and Economic Growth," in "The Economics of Artificial Intelligence: An Agenda," University of Chicago Press, 2018, pp. 237–282.

Amiti, Mary, Oleg Itskhoki, and Jozef Konings, "International Shocks, Variable Markups, and Domestic Prices," *Review of Economic Studies*, 2019, *86* (6), 2356–2402.

Aubakirova, Malika, Alex Atallah, Chris Clark, Justin Summerville, and Anjney Midha, "State of AI: An Empirical 100 Trillion Token Study with OpenRouter," *arXiv preprint arXiv:2601.10088*, 2026.

Baqaee, David Rezza and Emmanuel Farhi, "The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem," *Econometrica*, 2019, *87* (4), 1155–1203.

_ **and** _ , "Productivity and Misallocation in General Equilibrium," *Quarterly Journal of Economics*, 2020, *135* (1), 105–163.

Barrot, Jean-Noël and Julien Sauvagnat, "Input Specificity and the Propagation of Idiosyncratic Shocks in Production Networks," *Quarterly Journal of Economics*, 2016, *131* (3), 1543–1592.

Berndt, Ernst R. and Neal J. Rappaport, "Price and Quality of Desktop and Mobile Personal Computers: A Quarter-Century Historical Overview," *American Economic Review*, 2001, *91* (2), 268–273.

Berry, Steven, James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, July 1995, *63* (4), 841–890.

Byrne, David M., John G. Fernald, and Marshall B. Reinsdorf, "Does the United States Have a Productivity Slowdown or a Measurement Problem?," *Brookings Papers on Economic Activity*, 2016, *2016* (1), 109–182.

Corhay, Alexandre, Howard Kung, and Lukas Schmid, "Competition, Markups, and Predictable Returns," *The Review of Financial Studies*, 05 2020, *33* (12), 5906–5939.

Demirer, Mert, Andrey Fradkin, Nadav Tadelis, and Sida Peng, "The Emerging Market for Intelligence: Pricing, Supply, and Demand for LLMs," Working Paper 34608, National Bureau of Economic Research 2025.

Doms, Mark E., "Communications Equipment: What Has Happened to Prices?," *FRBSF Working Paper*, 2004, (2003-15).

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How Costly Are Markups?," *Journal of Political Economy*, 2023, *131* (7), 1619–1675.

Flamm, Kenneth, "Measurement of DRAM Prices: Technology and Market Structure," 1993, pp. 157–206.

Ganapati, Sharat, Joseph S. Shapiro, and Reed Walker, "Energy Cost Pass-Through in US Manufacturing: Estimates and Implications for Carbon Taxes," *American Economic Journal: Applied Economics*, 2020, *12* (2), 303–342.

Gopinath, Gita, Emine Boz, Camila Casas, Federico J. Díez, Pierre-Olivier Gourinchas, and Mikkel Plagborg-Møller, "Dominant Currency Paradigm," *American Economic Review*, 2020, *110* (3), 677–719.

_ , Oleg Itskhoki, and Roberto Rigobon, "Currency Choice and Exchange Rate Pass-Through," *American Economic Review*, 2010, *100* (1), 304–336.

Herskovic, Bernard, "Networks in Production: Asset Pricing Implications," *Journal of Finance*, 2018, *73* (4), 1785–1818.

Loecker, Jan De, Jan Eeckhout, and Gabriel Unger, "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, 2020, *135* (2), 561–644.

_ , Pinelopi K. Goldberg, Amit K. Khandelwal, and Nina Pavcnik, "Prices, Markups, and Trade Reform," *Econometrica*, 2016, *84* (2), 445–510.

Nakamura, Emi and Dawit Zerom, "Accounting for Incomplete Pass-Through," *Review of Economic Studies*, 2010, *77* (3), 1192–1230.

Nevo, Aviv, "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 2001, *69* (2), 307–342.

Ozdagli, Ali K. and Michael Weber, "Monetary Policy Through Production Networks: Evidence from the Stock Market," *NBER Working Paper*, 2017, (23424).

Pakes, Ariel, "A Reconsideration of Hedonic Price Indexes with an Application to PC's," *American Economic Review*, 2003, *93* (5), 1578–1596.

Rosen, Sherwin, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 1974, *82* (1), 34–55.

Small, Kenneth A. and Harvey S. Rosen, "Applied Welfare Economics with Discrete Choice Models," *Econometrica*, 1981, *49* (1), 105–130.

Syverson, Chad, "Challenges to Mismeasurement Explanations for the US Productivity Slowdown," *Journal of Economic Perspectives*, 2017, *31* (2), 165–186.

A Random Coefficient Logit (BLP) Demand Specification

The baseline demand system in Section 5 employs a multinomial logit model, which imposes proportional substitution via the Independence of Irrelevant Alternatives (IIA) property. This appendix generalizes the demand system by estimating a random coefficient logit model in the spirit of Berry et al. (1995), which allows consumer-level heterogeneity in price sensitivity and quality valuations, generating flexible substitution patterns. I show that the main empirical results, positive GPU cost pass-through moderated by markups, are robust to this richer demand specification.

A.1 Model and Estimation

Consumer utility. The indirect utility of consumer i for model m in month t is

$$u_{imt} = \delta_{mt} + \mu_{imt} + \varepsilon_{imt}, \tag{38}$$

where δ_{mt} is mean utility common to all consumers,

$$\delta_{mt} = -\alpha \log p_{mt}^{\text{prompt}} + \sum_k \beta_k Q_{mt}^k + \delta' X_{mt} + \tau_t + \xi_{mt}, \tag{39}$$

and μ_{imt} collects individual-specific deviations from mean utility:

$$\mu_{imt} = -\sigma_p \nu_{ip} \log p_{mt}^{\text{prompt}} + \sum_k \sigma_k \nu_{ik} Q_{mt}^k. \tag{40}$$

The terms ν_{ip} and $\{\nu_{ik}\}_k$ are independent standard-normal draws representing unobserved consumer-level taste heterogeneity, and ε_{imt} is an i.i.d. shock. The parameters $\{\sigma_p, \sigma_k\}$ are the standard deviations of the random coefficients. I place random coefficients on $\log p_{mt}^{\text{prompt}}$ and all four quality indices ($Q^{\text{code}}, Q^{\text{gen}}, Q^{\text{math/sci}}, Q^{\text{relia}}$), so each consumer has an individual price sensitivity ($\alpha + \sigma_p \nu_{ip}$) and individual quality valuations ($\beta_k + \sigma_k \nu_{ik}$).

The mean utility δ_{mt} in equation (39) contains the same regressors as the baseline logit in equation (6): log prompt price, the four quality indices, and operational characteristics X_{mt} (speed, latency, context-window length, and a reasoning-model indicator). When all σ parameters are zero, equation (38) collapses to the standard logit.

Market shares and outside good. Following Section 5.1, I set the market penetration $\lambda = 0.10$, so the outside good share is $s_{0t} = 1 - \lambda = 0.90$. Conditional on consumer draws ν_i, the probability that consumer i chooses model m is logit:

$$s_{mt}(\delta_t, \sigma \mid \nu_i) = \frac{\exp\big(\delta_{mt} + \mu_{imt}\big)}{1 + \sum_{j=1}^{J_t} \exp\big(\delta_{jt} + \mu_{ijt}\big)}. \tag{41}$$

The predicted market share integrates over consumer heterogeneity via simulation:

$$\tilde{s}_{mt}(\delta_t, \sigma) \approx \frac{1}{R} \sum_{r=1}^{R} s_{mt}(\delta_t, \sigma \mid \nu^r), \tag{42}$$

using $R = 500$ scrambled Halton draws.

Contraction mapping. For a candidate σ, I recover $\delta_t(\sigma)$ via the BLP contraction:

$$\delta_{mt}^{h+1} = \delta_{mt}^{h} + \log \tilde{s}_{mt}^{\text{obs}} - \log \tilde{s}_{mt}\big(\delta_t^h, \sigma\big), \tag{43}$$

which iteratively adjusts each model's mean utility upward when its predicted share falls short of the observed share, and downward when it exceeds it.[11] Observed shares $\tilde{s}_{mt}^{\text{obs}}$ are constructed from OpenRouter token usage, scaled by $\lambda = 0.10$.

Two-step GMM. Given $\delta_t(\sigma)$, I recover linear parameters (α, β, δ) and the structural error $\xi_{mt}(\sigma)$ from equation (39) via IV regression, then minimize the GMM objective

$$\hat{\sigma} = \arg\min_{\sigma} \; \xi(\sigma)' Z \, W \, Z' \xi(\sigma). \tag{44}$$

The first step uses $W = (Z'Z)^{-1}$; the second step updates W using the first-step residuals.

Instruments. I use the same cost-side instruments as Section 5.1: GPU rental price indices interacted with model-size dummies, $\{gpu_t \times \mathbf{1}[\text{size}_m = k]\}_k$. To aid identification of the random coefficients, I additionally construct BLP-style instruments based on rival- and own-firm product characteristics:

$$z_{m,x,t}^{\text{rival}} = \sum_{j \neq m} x_{jt}, \qquad z_{m,x,t}^{\text{own}} = \sum_{\substack{j \neq m \\ j \in \mathcal{F}_m}} x_{jt}, \tag{45}$$

where x_{jt} runs over the four quality indices and operational characteristics, and $\mathcal{F}_m$ is the set of models owned by the same provider as m. I also include the total number of competitor models and the number of own-firm models as instruments. Instruments are pruned to a linearly independent set before estimation.

Demand Jacobian and markup recovery. The BLP own- and cross-price derivatives are computed via simulation:

$$\frac{\partial \tilde{s}_{mt}}{\partial p_{mt}^{\text{prompt}}} = -\frac{1}{R} \sum_{r=1}^{R} \frac{\alpha + \sigma_p \nu_p^r}{p_{mt}^{\text{prompt}}} \, s_{mt}(\cdot \mid \nu^r)\big[1 - s_{mt}(\cdot \mid \nu^r)\big], \tag{46}$$

$$\frac{\partial \tilde{s}_{jt}}{\partial p_{mt}^{\text{prompt}}} = \frac{1}{R} \sum_{r=1}^{R} \frac{\alpha + \sigma_p \nu_p^r}{p_{mt}^{\text{prompt}}} \, s_{jt}(\cdot \mid \nu^r) \, s_{mt}(\cdot \mid \nu^r), \quad j \neq m. \tag{47}$$

Unlike the logit, each simulated consumer contributes a different price sensitivity $(\alpha + \sigma_p \nu_p^r)$, so substitution patterns depend on product similarity in characteristic space rather than on market shares alone. Given the BLP Jacobian $\hat{\Delta}_t^{\text{BLP}}$, markups are recovered exactly as in Section 5.2 by solving the multi-product Bertrand system:

$$\big(\Omega_t \odot \hat{\Delta}_t^{\text{BLP}}\big) \, \hat{m}_t^{\text{BLP}} = -\tilde{s}_t. \tag{48}$$

[11] I set the convergence tolerance to 10^{-8} with a maximum of 2,000 iterations.

A.2 Results

Demand estimates. Table A.1 reports the BLP demand estimates across three specifications mirroring those in Table 4. The mean price coefficient $\hat{\alpha}$ across specifications is comparable in magnitude to the baseline, confirming that average price sensitivity is robust to demand specifications.

Table A.1: Demand System: BLP

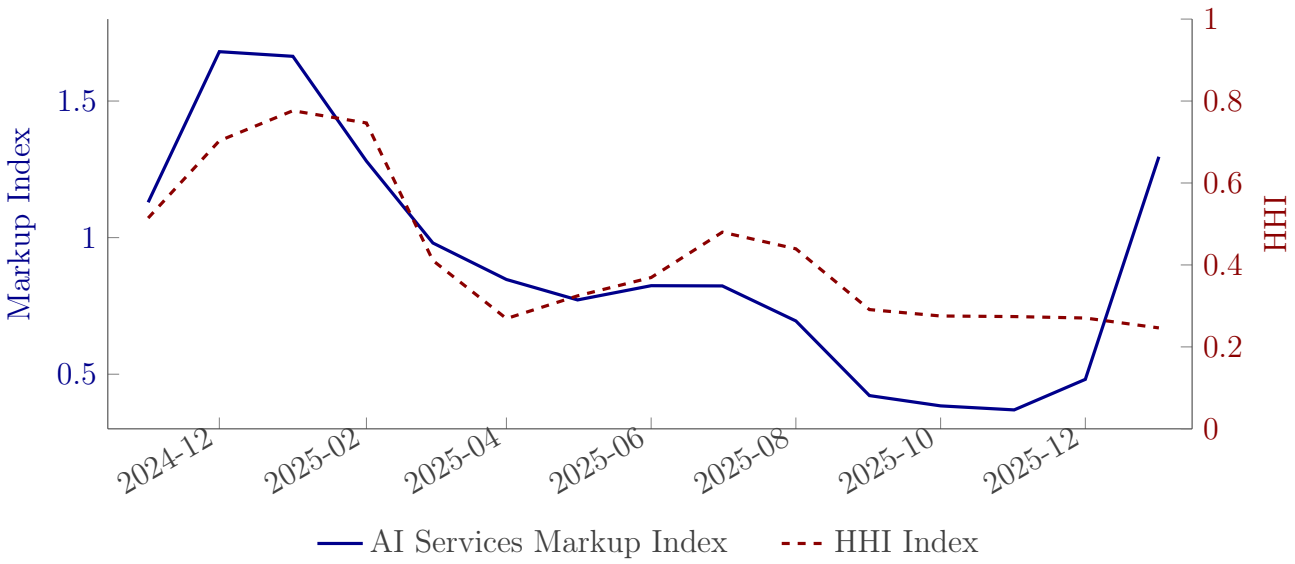
	BLP		
$\log p^{prompt}$	-0.033	-0.242***	-0.315***
	(0.041)	(0.048)	(0.063)
Q^{code}		0.924***	0.608***
		(0.134)	(0.127)
Q^{gen}		0.459***	0.500***
		(0.137)	(0.114)
$Q^{math/sci}$		-0.415***	-0.292*
		(0.144)	(0.151)
Q^{relia}		0.346***	0.166
		(0.105)	(0.111)
controls	No	Yes	Yes
Month FE	Yes	Yes	Yes
Provider FE	No	No	Yes
N	1123	1123	1123
R2	0.084	0.413	0.556

Notes. This table reports estimates of a random coefficient logit (BLP) demand equation using OpenRouter usage data. The dependent variable is $\log s_{mt}^{OR}$, where s_{mt}^{OR} is model m's token share in month t. $\log p_{mt}^{\text{prompt}}$ is the log posted prompt price (USD per 1M prompt tokens). Q_{mt}^{code}, Q_{mt}^{gen}, $Q_{mt}^{\text{math/sci}}$, and Q_{mt}^{relia} are the four quality indices from Section 4. "Controls" include speed, latency, an indicator for reasoning models, and context-window length. "Provider FE" indicates whether provider fixed effects are included. The reported mean coefficients are from the linear IV step at the estimated nonlinear parameters; $\hat{\sigma}$ rows report estimated random-coefficient standard deviations. Instruments include size-interacted GPU rental price indices (cost-side) and BLP-style rival/own-firm characteristic sums. Standard errors are clustered by month; in the provider-FE specification, by month and provider. Significance levels: *** 1%, ** 5%, * 10%.

Markups. Figure A.1 plots the BLP-implied aggregate markup index alongside the HHI, directly comparable to Figure 9 in Section 5. The two series are qualitatively identical: the market-level markup declines over the sample period, and markup variation co-moves positively with concentration. Quantitatively, the BLP markups are also similar to those implied by the baseline logit.

Demand Jacobian. Figure A.2 visualizes the BLP demand Jacobian for the latest sample month, directly comparable to Figure 8 in Section 5. The qualitative patterns are the same: large negative own-price effects along the diagonal, small positive cross-price effects, and visible within-provider blocks. The BLP Jacobian does exhibit richer off-diagonal heterogeneity relative to the logit, consistent with consumer heterogeneity generating

Figure A.1: Market Markup Index and Provider Concentration: BLP Demand



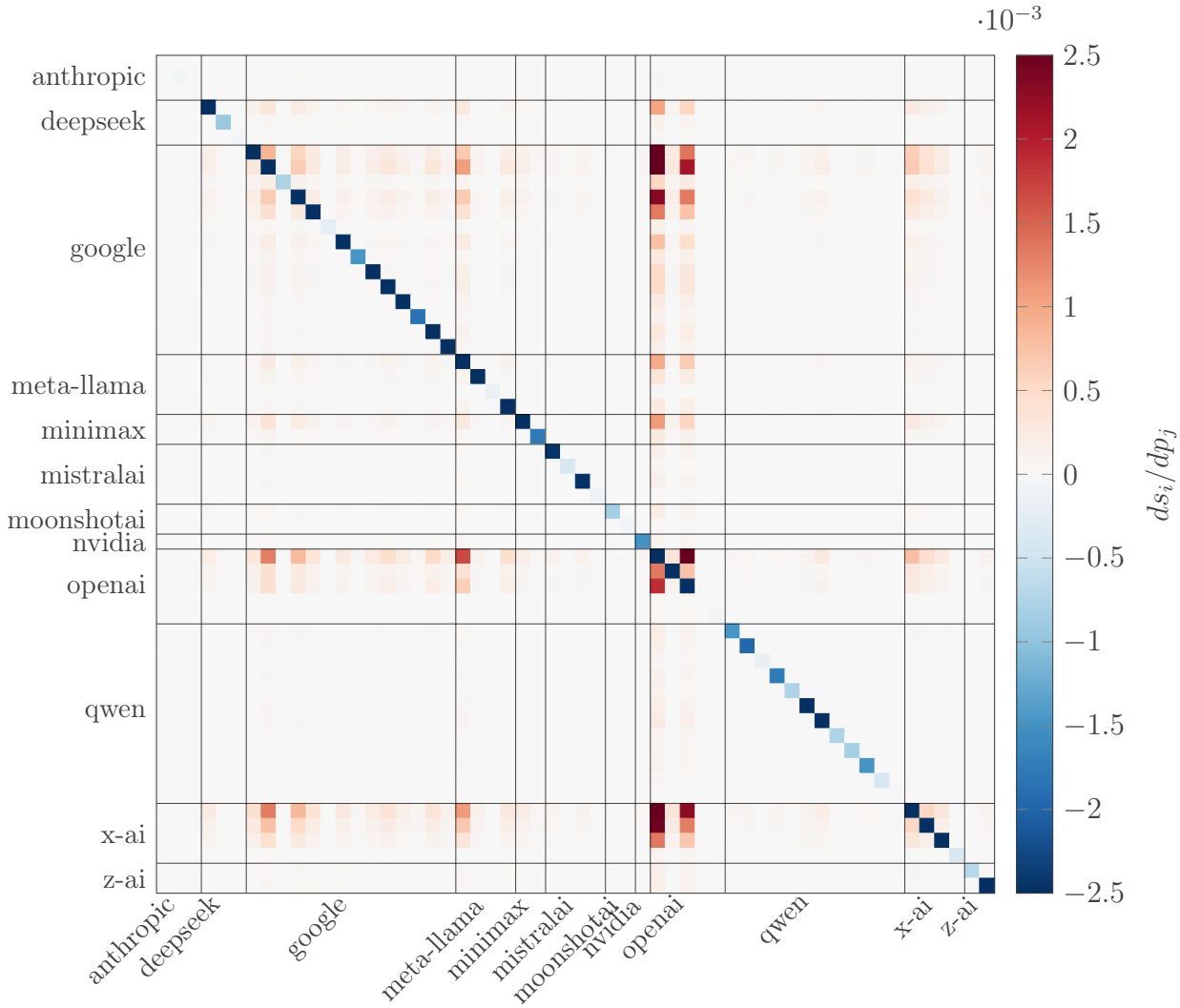
Notes. The blue series (left axis) shows the aggregate AI services markup index constructed using BLP demand estimates. The red dashed series with square markers (right axis) shows provider concentration measured by token-value HHI.

characteristic-space-based substitution that departs from pure share-proportionality.[12]

Cost Pass-through robustness. Table A.2 re-estimates the cost pass-through specification in equation (19) using BLP-implied provider-level markups $\hat{\mu}^{\text{BLP, prov}}_{j(m),t-1}$ in place of the baseline logit markups. The results are directly comparable to Table 5 in Section 6. The interaction coefficient $\hat{\theta}_1$ remains negative, confirming that the central finding, higher-markup providers absorb more of the GPU cost shock, is robust to the demand specification. The magnitude and statistical significance of $\hat{\theta}_1$ are quantitatively similar across the two demand specifications.

[12]The BLP Jacobian entries are roughly an order of magnitude smaller than the logit's because consumer heterogeneity compresses the average of the conditional term $s^r_m(1-s^r_m)$: when random coefficients are large, each simulated consumer's conditional share s^r_m is pushed toward zero or one, where $s(1-s)\to 0$, yielding a smaller average than the logit's $\tilde{s}_m(1-\tilde{s}_m)$ (by Jensen's inequality). The lower estimated mean price coefficient $\hat{\alpha}$ under BLP moves in the opposite direction, so the implied demand elasticities, and hence the recovered Lerner indices and gross markups remain quantitatively similar across the two specifications.

Figure A.2: Demand Jacobian Heatmap by Provider: BLP Demand



Notes. BLP demand Jacobian $\partial s_i/\partial p_j$ for the latest sample month, with models ordered by provider (then by token share within provider). Black lines delineate provider blocks. Compare to Figure 8, which uses the baseline logit Jacobian.

Table A.2: AI Services Cost Pass-through: BLP

$\Delta \log gpu_t$	0.1566*	0.2279**
	(0.0839)	(0.0960)
$\Delta \log gpu_t \times \widehat{\mu}^{\text{prov}}_{j(m),t-1}$		-0.1301**
		(0.0565)
$\widehat{\mu}^{\text{prov}}_{j(m),t-1}$		0.0099
		(0.0174)
company FE	Yes	Yes
R-squared Adj.	0.0037	0.0027
N	1003	1003

Notes. This table reports regression results corresponding to equation (19), using BLP-implied provider-level markups $\hat{\mu}^{\text{BLP,prov}}_{j(m),t-1}$. The baseline specification includes $\Delta \log gpu_t$ only; the interaction specification additionally includes $\widehat{\mu}^{\text{prov}}_{j(m),t-1}$ and $\Delta \log gpu_t \times \widehat{\mu}^{\text{prov}}_{j(m),t-1}$. Provider fixed effects are included in all columns. Standard errors are clustered at the provider level and reported in parentheses. Compare to Table 5, which uses baseline logit markups. Significance levels: *** 1%, ** 5%, * 10%.