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

Zijie Wang*

This version: June 2026

Abstract

Between falling compute costs and the firms that use AI stands an oligopoly. I build a novel dataset linking prices, usage, and capabilities to GPU rental costs for over 200 models competing on OpenRouter, a centralized AI marketplace. Structural demand estimates recover provider markups: pass-through of cost declines falls with market power, and the premium segment passes through nothing, retaining the declines as margin. Quality-adjusted prices still fell 78 percent over 2024–2026, twice the nominal price decline, cutting U.S. PPI by a cumulative 160 basis points through the production network. Entry of more capable models, not price cuts, drives consumer surplus gains.

JEL classification: D43, E01, E31, E32, O33.

Keywords: Artificial Intelligence; Inflation Measurement; Markups; Cost Pass-Through; Production Networks; Market Power.

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

I thank Vatsala Shreeti, Patrick Coen, Jon Frost, Inaki Aldasoro, and seminar participants at the WBS brownbag for helpful comments and suggestions. 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

Between 2024 and 2026, artificial intelligence (AI) set off one of the largest capital deployments in economic history. Annual investment in AI infrastructure, led by Alphabet, Amazon, Meta, and OpenAI, is projected to reach $765 billion in 2026 (Goldman Sachs Global Institute, 2026). NVIDIA, the dominant supplier of AI accelerators, booked $197 billion of data center revenue in its fiscal year ending January 2026 (NVIDIA Corporation, 2026). The capital stock itself produces a single intermediate input: AI inference, sold per token to firms across a growing range of sectors. For the economy's other great intermediate input, oil, the transmission of upstream cost shocks into producer prices has occupied macroeconomics for four decades (Hamilton, 1983). For compute, no one has measured that transmission. Nor can it be measured the way oil pass-through is. Crude oil reaches the economy as a homogeneous commodity at a world price. Inference reaches the economy through a handful of providers, each pricing a portfolio of differentiated models. Between the collapsing cost of computation and the firms that buy its output stands an oligopoly. How much of the cost decline passes through is therefore not a cost-share calculation; it is an equilibrium pricing outcome.

I provide the first end-to-end accounting of AI deflation: how large it is, who absorbs it, and where it lands. I construct a novel granular dataset that links, for more than 200 AI models, token usage, inference prices, measured capabilities, and the graphics processing unit (GPU) rental cost of producing inference. These data support three estimates: a constant-quality price index for AI inference, structural estimates of demand and provider markups in the market for AI, and a measurement of how AI deflation propagates into U.S. producer prices. The deflation is large: holding quality constant, the effective price of AI inference fell by approximately 78 percent from early 2024 through early 2026, roughly twice the 39 percent nominal price decline. The accounting returns a three-way split. Providers with pricing power keep a substantial share: pass-through of GPU cost declines falls with markups and stops entirely above a gross markup of 1.26, a level the premium segment sustains for extended periods. Downstream producers receive a large but attenuated deflation, roughly 160 basis points of cumulative producer price decline

over 2024–2026, most of it arriving through supply-chain linkages. Users gain mainly from the entry of new, more capable models rather than from price cuts on existing ones; most of the cumulative consumer surplus traces to that entry.

AI is a general purpose technology (David, 1990; Bresnahan and Trajtenberg, 1995): an input that spreads through the economy while improving continuously. Spreading as an intermediate input means the price of inference enters other firms' costs and provider pricing has economy-wide consequence; the analysis therefore cannot stop at the AI sector and must follow costs through the production network. Continuous improvement means a token of 2026 inference is not a token of 2024 inference, so measurement must hold quality constant. Because improvement arrives embodied in newly released models rather than upgrades to existing ones, welfare gains flow through entry. What separates AI from earlier general purpose technologies is pace and market structure. Constant-quality AI prices are falling at roughly twice the annualized rate the hedonic literature documents for personal computers (Berndt and Rappaport, 2001). The supply side is concentrated and unregulated, so oligopoly pricing, not competition, decides whether the deflation reaches the economy. Earlier technology revolutions posed such questions only in retrospect; here, every link in the chain is observable as the technology diffuses: the daily cost of compute, posted output prices, public capability benchmarks, and marketplace quantities. I exploit that visibility to answer the question in real time: who captures the deflation.

The accounting begins with measurement. I compress 15 independent AI capability benchmarks from the computer science literature into four domain indices (coding, general reasoning, math and science, and reliability) and embed them in hedonic price regressions whose time fixed effects trace out the constant-quality price index for AI services. The wedge between the nominal and constant-quality declines is the cumulative quality improvement the market delivered in two years. The hedonic coefficients also show which capabilities the market prices: coding and reliability command persistent premia, while premia on general reasoning and mathematical ability erode as the frontier broadens. The index, however, records what a constant-quality token costs, not why its price fell. That price still embeds the markups of oligopolistic providers, and its decline could be due to

falling marginal costs, falling markups, or both; the welfare implications differ sharply across these channels. Prices alone cannot separate them.

Markups are not observed; they must be recovered from estimates of demand. The data come from OpenRouter, a centralized marketplace for AI inference. The platform relays each provider's posted per-token prices, presents every model behind the same standardized interface, and lets users switch among them costlessly. Observed token flows are revealed choices among close substitutes rather than artifacts of integration frictions. OpenRouter publishes no historical dataset, so I reconstruct the platform's price and usage panels from archived snapshots on the Wayback Machine and merge them with independent capability benchmarks for more than 200 models. On these data I estimate demand on monthly token-value shares, allowing users to differ in price sensitivity and in how they value capabilities. Markups follow from inverting each provider's multi-product pricing problem.

The identification problem is the classic one: providers observe demand for their models that I do not, and they set prices accordingly, biasing naive estimates of price sensitivity. I instrument prices with a cost shifter, the NVIDIA H100 GPU rental price interacted with each model's parameter count (disclosed for open-weight models, estimated for closed-source ones). The logic is that larger models require more GPU-hours to run, so a common shock to the rental rate raises marginal cost more for a 600-billion-parameter model than for an 8-billion-parameter one, and that cost variation traces out the demand curve. The exclusion restriction is that, conditional on time fixed effects and the measured capability and operational characteristics, parameter count moves usage only through price. A user choosing between two models with identical benchmark scores, speed, and latency has no independent reason to care how many parameters sit behind the interface, nor is this information available to her for closed models.

The markup estimates reveal an oligopoly in motion. DeepSeek's entry in early 2025 compressed rivals' margins; Anthropic's frontier releases in late 2025 pushed its markup to roughly three times that of its nearest competitors. The markups then govern transmission. Regressing changes in quality-adjusted model prices on GPU cost changes

and their interaction with lagged provider markups yields a direct pass-through coefficient of 0.315 and an interaction of −0.250, both statistically significant. Together they imply that pass-through declines steadily with market power and vanishes at a gross markup of 1.26. Above that line, each dollar of GPU cost decline becomes a dollar of margin. In the premium segment of the market, the deflation stops at the provider.

The final step traces the deflation that does pass through to the producer prices of U.S. industries. Mapping AI services into the BEA input–output tables and applying the Leontief price inverse converts the quality-adjusted decline into sector-level cost reductions. These aggregate to the 160 basis point cumulative PPI contribution over 2024–2026. The majority of the effect arrives indirectly, through firms whose suppliers purchase AI services rather than through direct purchases: the signature of a pervasive input. Under marginal-cost pricing the contribution would have been larger; the gap is deflation retained as oligopoly margin. And because the 2024 tables predate the rapid diffusion of AI, these magnitudes are lower bounds on a channel that is mechanically growing.

The welfare accounting weighs what market power cost users against what the market delivered them. Freezing the choice set at its initial composition and letting prices evolve isolates the price channel of consumer surplus growth. The remainder, the gain from access to newly entered, more capable models, accounts for the bulk of cumulative surplus over the sample. The static cost of market power, markup rents plus deadweight loss, is real and widens as the market grows. But it is small next to the value created by frontier entry, and that entry comes disproportionately from the highest-markup providers. Policy that compresses margins to speed pass-through therefore risks taxing the activity that generates most of the gains. For measurement, the implication is immediate: statistical agencies relying on nominal token prices will increasingly understate AI deflation as input shares grow. The hedonic framework developed here is directly implementable as a quality-adjusted deflator.

This paper contributes an end-to-end accounting of AI deflation in three steps. First, I measure the relevant price object. I construct a constant-quality price index for AI inference, extending the hedonic deflator tradition from computers and semiconductors to

an input whose quality is multidimensional, benchmark-based, and embodied in frontier-model entry rather than hardware specifications. The index measures the price faced by downstream firms for a constant-quality unit of AI capability, not the nominal price of a token. Second, I estimate the market-power mechanism that governs the incidence of this deflation. The closest work, Demirer et al. (2025), opens the empirical study of AI inference markets by documenting entry, price dispersion, and differentiation in observed prices and choices. I move from those facts to equilibrium objects by estimating demand and inverting providers' multi-product pricing problems. This delivers model- and provider-level markups, implied marginal costs, pass-through schedules that vary with market power, counterfactual prices under alternative market structures, and a welfare decomposition between price declines and frontier-model entry. Third, I trace the markup-mediated price decline through the production network. This brings the variable-markup pass-through framework from international finance (Amiti et al., 2019) to a setting in which the upstream cost shock is observed daily and market power is recovered structurally rather than proxied by shares or concentration. Together, these pieces provide the chain that prior work has not assembled: from GPU spot prices, to oligopolistic markups in AI inference, to downstream producer prices. For oil, pass-through can often be approximated from cost shares because crude is traded as a homogeneous commodity. For compute, pass-through must be recovered from the pricing problem of differentiated, multi-product oligopolists.

Roadmap Section 2 reviews the related literature, and Section 3 describes the OpenRouter marketplace and the construction of the price, usage, and cost data. Section 4 develops the quality measurement, Section 5 the demand estimation and markup recovery, and Section 6 the pass-through, welfare, and propagation results; Section 7 concludes.

2 Related Literature

This paper speaks to three strands of literature. The first literature links market power to pass-through and the distribution of rents. In finance, product-market power matters because it insulates cash flows from aggregate shocks, affects expected returns, and shapes factor shares and welfare. Firms in concentrated industries earn lower average returns (Hou and Robinson, 2006), competition erodes the markups that shield profits from aggregate shocks (Bustamante and Donangelo, 2017; Dou et al., 2021), and markup dynamics forecast returns (Corhay et al., 2020). A complementary macro-finance literature documents the secular rise of markups and concentration and traces its consequences for misallocation, welfare, and factor shares (Grullon et al., 2019; Autor et al., 2020; Baqaee and Farhi, 2020; Barkai, 2020; De Loecker et al., 2020; Edmond et al., 2023). International finance supplies the pass-through mechanism. High-share firms adjust markups to absorb exchange-rate and cost shocks, so pass-through is incomplete and declines with market power (Atkeson and Burstein, 2008; Gopinath et al., 2010; Amiti et al., 2014, 2019; Gopinath et al., 2020), a logic that Weyl and Fabinger (2013) generalize as incidence under imperfect competition. The same markup absorption recurs after trade liberalization and in commodity and energy markets (Nakamura and Zerom, 2010; Berman et al., 2012; De Loecker et al., 2016; Auer and Schoenle, 2016; Ganapati et al., 2020). These literatures typically infer market power from market shares, concentration, or accounting residuals. I instead estimate demand for a single, consequential new input and recover model- and provider-level markups by inverting multi-product Bertrand pricing (Berry et al., 1995; Nevo, 2001). This lets me measure directly how an observed upstream cost shock, the GPU rental rate, is split between AI suppliers and the downstream firms that buy inference.

The second literature studies AI inference as a market. The closest empirical work, Demirer et al. (2025), documents rapid entry, falling prices, persistent price dispersion, and differentiation across LLM API markets; Aubakirova et al. (2026) reports similar marketplace patterns. Related work maps the layered AI supply chain and the geography of the firms that populate it (Gambacorta and Shreeti, 2025; Rishabh and Shreeti, 2026). Closest to the market-power mechanism, Korinek and Vipra (2025) argue that economies

of scale and scope in compute, data, and talent tilt foundation-model markets toward concentration. These papers establish that AI inference is not a homogeneous compute commodity but a differentiated product market with substantial dispersion in prices, capabilities, and provider positions. I ask the next incidence question: conditional on this market structure, who captures the deflation from falling compute costs? Answering that question requires equilibrium objects that prices and shares alone do not reveal: marginal costs, markups, pass-through schedules, and welfare counterfactuals. A broader literature treats AI as a force acting on firms and the macroeconomy, studying its productivity effects, growth implications, and asset-pricing relevance (Aghion et al., 2018; Eisfeldt et al., 2023; Acemoglu, 2025; Babina et al., 2024). That work motivates the importance of AI exposure. My contribution is to endogenize the price of the AI input itself.

The third literature concerns the measurement and propagation of prices for rapidly improving intermediate inputs. Hedonic methods (Rosen, 1974; Pakes, 2003) have produced constant-quality deflators for computers, semiconductors, and communications equipment (Flamm, 1993; Berndt and Rappaport, 2001; Doms, 2005), which typically fall faster than nominal prices. When improvement instead arrives through new varieties, the relevant gains are measured as consumer surplus from new goods (Bresnahan and Gordon, eds, 1997; Hausman, 1999; Brynjolfsson et al., 2019). These measurement issues matter for macro-finance because real output, real rates, and productivity are obtained by deflating nominal magnitudes, so mispricing a fast-improving input biases all three (Advisory Commission to Study the Consumer Price Index, 1996; Byrne et al., 2016; Syverson, 2017). AI combines this measurement problem with a propagation problem. It is a new general purpose technology whose quality is read from capability benchmarks rather than hardware specifications, whose improvements arrive largely through frontier-model entry, and whose output is sold as an intermediate input to downstream firms. Foundational work shows that sectoral shocks aggregate through input-output linkages rather than washing out (Gabaix, 2011; Acemoglu et al., 2012; Baqaee and Farhi, 2019), while a finance literature prices network position directly, showing that centrality and supply-chain exposure carry risk premia and transmit shocks to firm value (Barrot and Sauvagnat, 2016; Ozdagli and

Weber, 2025; Herskovic, 2018; Richmond, 2019; Gofman et al., 2020). Work linking AI to the macroeconomy, including the task-based frameworks of Acemoglu and Restrepo (2018, 2019), typically enters AI as a uniform productivity shifter and stops short of the market through which AI services are sold. I combine the measurement and network problems: the relevant object is not a nominal token price but a quality-adjusted price of AI inference, and the production network transmits not the raw decline in GPU costs, but the quality-adjusted AI price decline that remains after oligopolistic providers absorb part of the shock through markups.

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 gives developers and downstream applications access to over 600 models from more than 60 providers, including OpenAI, Anthropic, Google, Meta, Mistral, xAI, and DeepSeek. Users send inference requests to a single, standardized API endpoint, and OpenRouter's routing layer 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: 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.

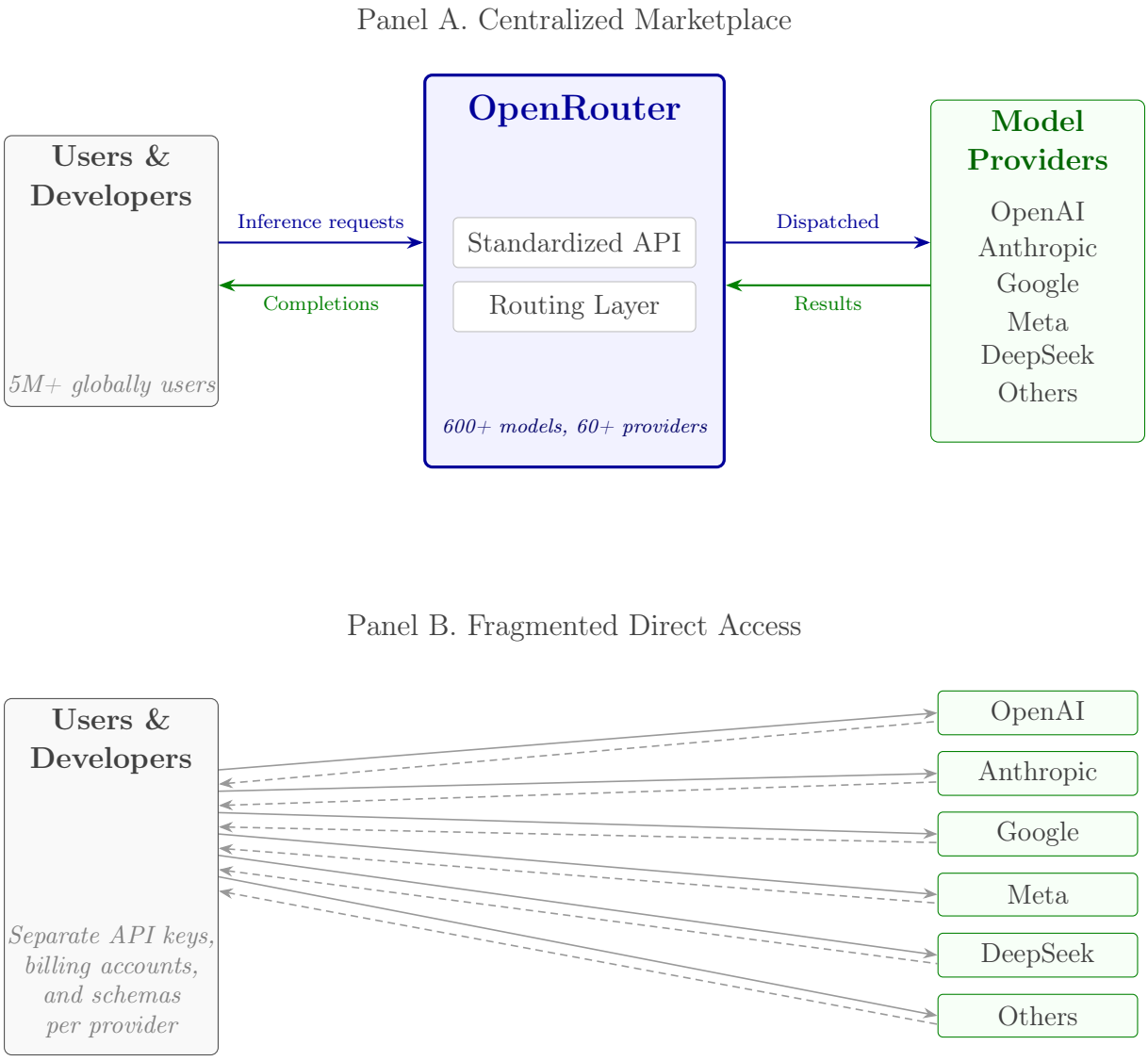
Two features of OpenRouter make it a particularly useful empirical setting for studying the market for AI services. First, *product comparability*: all models are accessible through an identical API interface, and the standardized schema minimizes the integration frictions that typically segment demand 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 is thus 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 access multiple providers through a single standardized API, while Panel B shows

[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 and processes more than 100 trillion tokens annually; by late 2025 it was routing roughly one trillion tokens per day (Aubakirova et al., 2026). Over 50% of usage originates outside the United States.

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 compares 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 provides 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 suite of capability tests covering reasoning, mathematics, coding, and reliability.[3] I collected model-level benchmark scores and operational characteristics from the 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	Params.	Open	US Co.	Context	Reason.
Mean	96.783	7.657	275.957	0.595	0.636	0.368	0.375
Std. Dev.	41.762	10.902	631.067	0.492	0.481	0.844	0.485

	Panel B: Model Price and Monthly Sum of Token Usage						
	Mean	Median	Std. Dev.	P1	P25	P75	P99
Prompt Price	1.993	0.250	7.713	0.000	0.050	1.000	30.000
Token Usage	15.384	0.000	146.765	0.000	0.000	0.000	368.742
Token Value	0.004	0.000	0.045	0.000	0.000	0.000	0.072

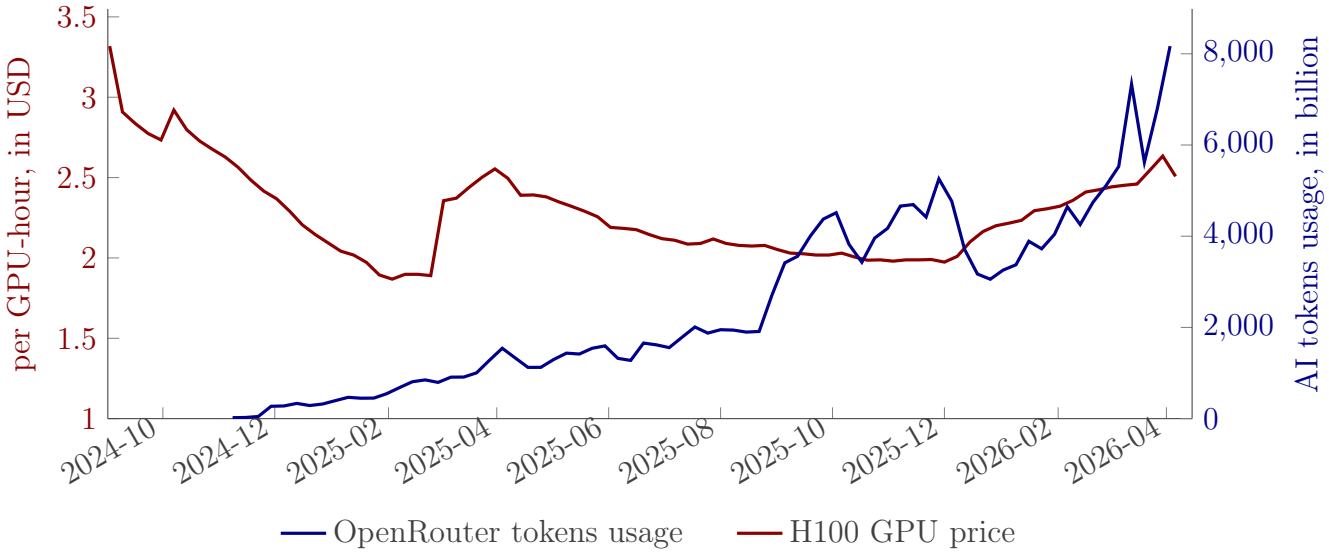
Notes. This table reports summary statistics for the OpenRouter model panel from January 2024 to May 2026. Panel A reports model-level characteristics for the matched sample of models with Artificial Analysis characteristics. Speed is output speed, measured as the number of output tokens generated per second after the first token is received. Latency is time to first token, measured in seconds between sending an API request and receiving the first response token. Params. is the total number of model parameters, measured in billions. "Open," "US Co.," and "Reason." are indicator variables equal to one for open-weight models, models developed by U.S.-headquartered companies, and reasoning models, respectively. Context is the maximum combined input-output context window, measured in millions of tokens. Panel B reports the distribution across model-month observations. Prompt prices are monthly average posted prices in USD per one million prompt and completion tokens. Token usage is total monthly prompt plus completion tokens, reported in billions. Token value is monthly token expenditure, reported in millions of U.S. dollars. 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 are my primary measure of the marginal cost of producing AI services. 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 centers, 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 is a unit of account for accelerator capacity: the cost of a more powerful GPU, such as the NVIDIA H200, 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 predates the rapid diffusion of AI services. The sectoral exposure measures used in the macro

propagation exercise therefore reflect an earlier stage of AI adoption, and the implied PPI effects are 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, AI models have become far 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 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 shows 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 bulk 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 the Quality-Adjusted AI Price Index

The quality indices constructed in Subsection 4.1 document rapid and heterogeneous capability growth, which leads 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).

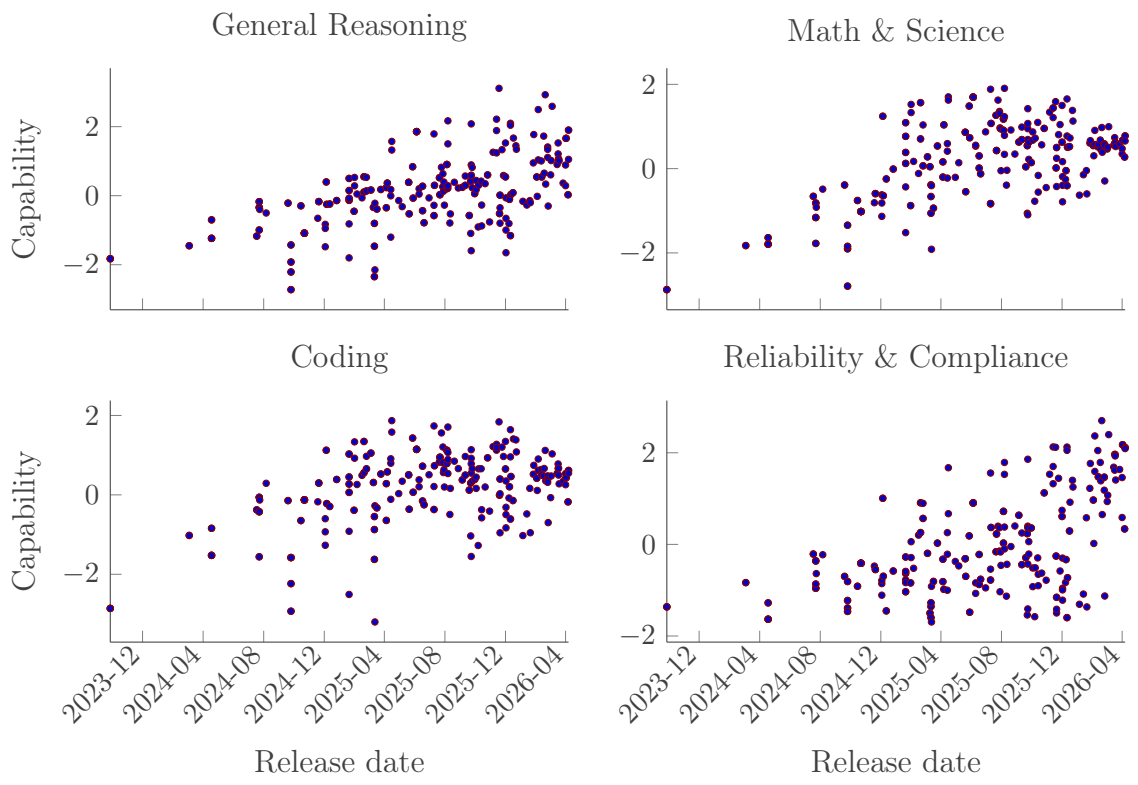
Table 2: AI Capability Benchmarks

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 date. 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.

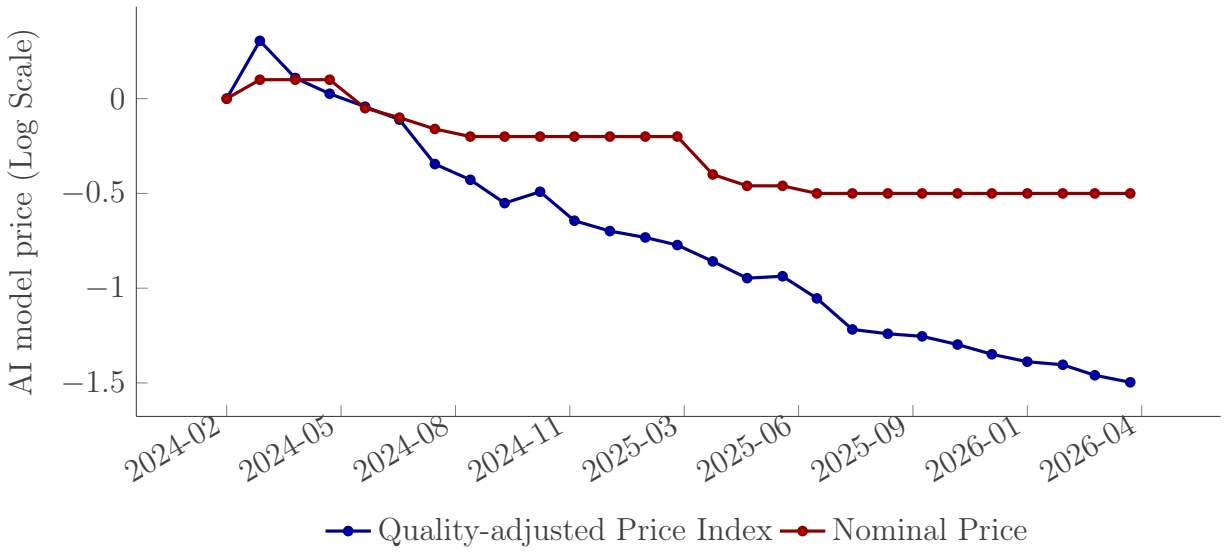
Coding and reliability enter with large positive coefficients, indicating that these dimensions carry a premium. A one-standard-deviation improvement in coding capability is associated with approximately 87 log points (139%) higher prices; the corresponding figure for reliability is 103 log points (179%). 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 about 0.5 log points (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) shows 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}	0.473***				0.596***	0.135***
	(0.012)				(0.021)	(0.014)
Q^{gen}		0.730***			0.204***	-0.047*
		(0.013)			(0.031)	(0.026)
$Q^{math/sci}$			0.375***		-0.638***	-0.686***
			(0.012)		(0.028)	(0.020)
Q^{relia}				1.200***	1.224***	-0.076***
				(0.015)	(0.021)	(0.020)
speed					-0.005***	-0.004***
					(0.000)	(0.000)
latency					0.007***	0.001
					(0.002)	(0.001)
context window length					-0.127***	-0.187***
					(0.009)	(0.012)
reasoning model					-0.526***	0.801***
					(0.035)	(0.032)
open model						-0.443***
						(0.028)
US						-0.065***
						(0.022)
log(parameters)						0.865***
						(0.009)
N companies	32	32	32	32	31	30
N models	264	264	264	264	263	241
N	89593	89593	89593	89593	89581	79179
R-squared Adj.	0.122	0.143	0.083	0.260	0.347	0.682

Notes. This table reports estimates from the hedonic pricing regression in equation 1. The dependent variable is the log posted prompt price, $\log p_{mt}^{prompt}$, measured in USD per 1M prompt tokens. Regressors include model capability measures: Q_{mt}^{code} for coding, Q_{mt}^{gen} for general reasoning, $Q_{mt}^{math/sci}$ for math and science, and Q_{mt}^{relia} for reliability/robustness. They also include operational characteristics: *speed* (output tokens per second), *latency* (response delay), an indicator for *reasoning models*, and *context length* (maximum context window, in tokens). *log(parameters)* denotes the log number of model 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 changing quickly, 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 $t \in \mathcal{W}(\tau)$:

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

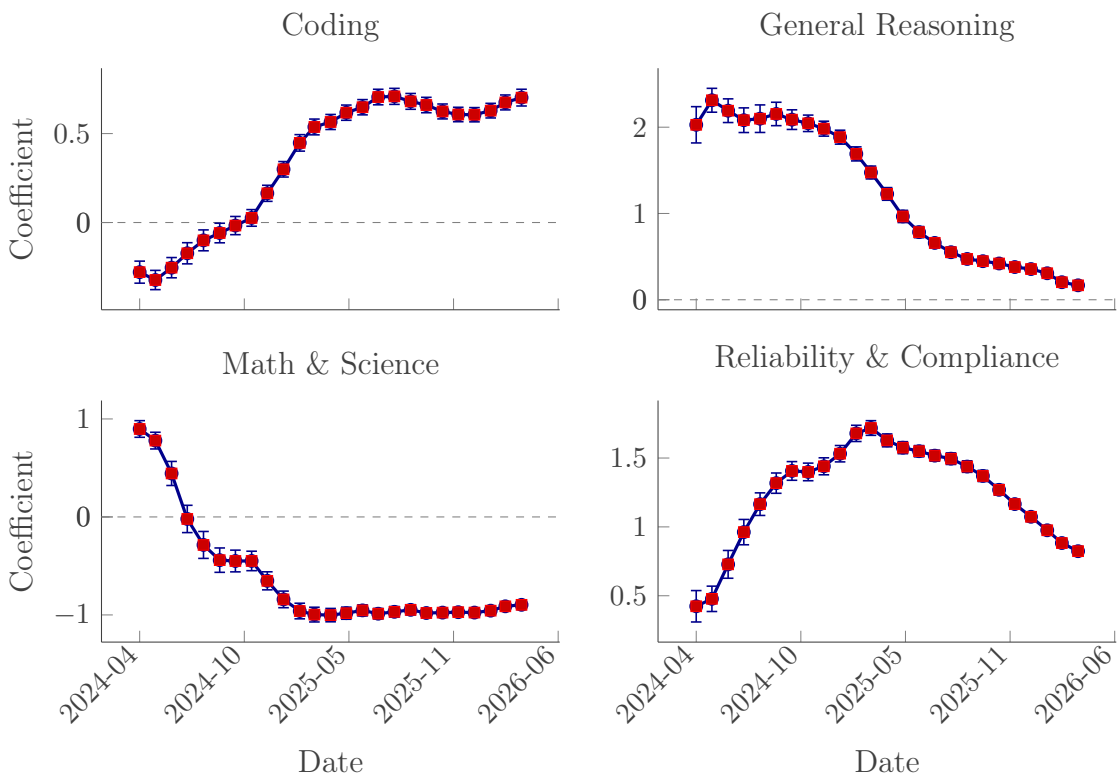
This yields a time series of rolling-window coefficients $\{\widehat{\beta}_k^{(\tau)}\}_\tau$.

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 reinforce the case for 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 random coefficient logit (BLP) demand system using GPU-cost shifters interacted with model size and BLP-style rival/own-firm characteristic instruments, and derives the demand Jacobian required for markup recovery. 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:

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

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

for all months.[5]

Figure 6 plots accumulated token usage and token value by company on OpenRouter over the sample period. The upper panel shows a highly concentrated and rapidly shifting usage distribution: Google and xAI account for the largest share of total token volume, with several providers growing quickly. The lower panel, which weights usage by per-token prices, gives 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.

Random coefficient logit (BLP) demand specification. I estimate a random coefficient logit demand system in the spirit of Berry et al. (1995), which allows consumer-level heterogeneity in price sensitivity and quality valuations and thereby relaxes the proportional substitution (IIA) property of multinomial logit. The indirect utility of consumer i for model m in month t is

$$u_{imt} = \delta_{mt} + \mu_{imt} + \varepsilon_{imt}, \tag{6}$$

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

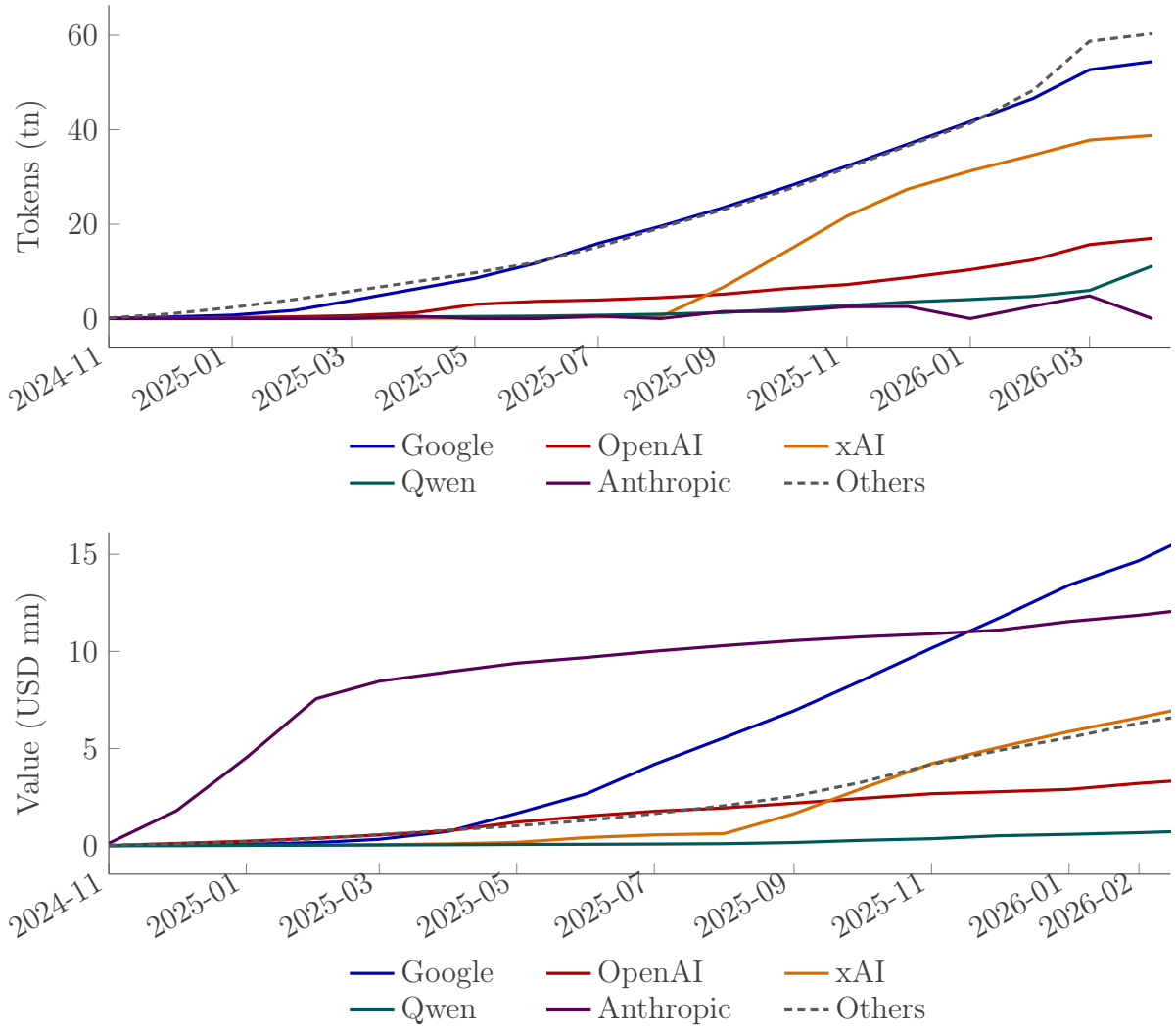
$$\delta_{mt} = \beta_0 + \sum_k \beta_k Q^k_{mt} + \Phi' X_{mt} + \tau_t + \xi_{mt}, \tag{7}$$

and μ_{imt} collects individual-specific deviations:

$$\mu_{imt} = -\alpha_i \log p^{\text{prompt}}_{mt} + \sum_k \sigma_k \nu_{ik} Q^k_{mt}, \qquad \alpha_i = \exp\big(\mu_\alpha + \sigma_\alpha \nu_{ip}\big). \tag{8}$$

[5]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; λ enters only as an additive constant absorbed by month fixed effects, and the demand slope parameters are identified separately from λ.

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).

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. extreme-value shock. Each consumer's price coefficient α_i is *lognormally distributed*, $\ln \alpha_i \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$, with median $\exp(\mu_\alpha)$ and mean $\exp(\mu_\alpha + \frac{1}{2}\sigma_\alpha^2)$; the lognormal specification guarantees $\alpha_i > 0$ for every simulated consumer, ruling out the upward-sloping demand draws that can occur under a normal random coefficient on price.[6] The four quality indices $(Q^{\text{code}}, Q^{\text{gen}}, Q^{\text{math/sci}}, Q^{\text{relia}})$ retain normal random coefficients with means β_k and dispersions σ_k. When all σ paramet-

[6]Under a normal random coefficient $\alpha_i = \alpha + \sigma_p \nu_{ip}$, a positive mass of the consumer distribution can have $\alpha_i < 0$ whenever $|\alpha|/\sigma_p$ is not large, yielding economically implausible substitution patterns and inflated welfare estimates. The lognormal parameterization is standard in recent applied IO work.

ers collapse to zero and α_i degenerates to a constant, equation (6) reduces to the standard logit.

Market shares and outside good. With market penetration $\lambda = 0.10$, the outside-good share is $s_{0t} = 1 - \lambda = 0.90$. Conditional on consumer draws ν_i, the choice probability is

$$s_{mt}(\delta_t, \sigma \mid \nu_i) = \frac{\exp\big(\delta_{mt} + \mu_{imt}\big)}{1 + \sum_{j=1}^{J_t} \exp\big(\delta_{jt} + \mu_{ijt}\big)}, \tag{9}$$

and the predicted market share integrates over consumer heterogeneity by simulation:

$$\tilde{s}_{mt}(\delta_t, \sigma) \approx \frac{1}{R} \sum_{r=1}^{R} s_{mt}(\delta_t, \sigma \mid \nu^r), \tag{10}$$

using $R = 500$ scrambled Halton draws. Observed shares $\tilde{s}^{\text{obs}}_{mt}$ are constructed from OpenRouter token usage, scaled by λ.

Estimation: contraction and two-step GMM. Collect the nonlinear parameters as $\theta_2 \equiv (\mu_\alpha, \sigma_\alpha, \{\sigma_k\}_k)$. For a candidate θ_2, I recover the mean utilities $\delta_t(\theta_2)$ via the BLP contraction:

$$\delta^{h+1}_{mt} = \delta^{h}_{mt} + \log \tilde{s}^{\text{obs}}_{mt} - \log \tilde{s}_{mt}\big(\delta^{h}_{t}, \theta_2\big), \tag{11}$$

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.[7] Because the price coefficient is fully nonlinear under the lognormal specification, the IV step in equation (7) recovers only the linear parameters (β, Φ) and the structural error $\xi_{mt}(\theta_2)$; the price parameters $(\mu_\alpha, \sigma_\alpha)$ enter θ_2 and are estimated jointly with the quality dispersions. The nonlinear vector $\hat{\theta}_2$ minimizes the GMM objective

$$\hat{\theta}_2 = \arg\min_{\theta_2} \; \xi(\theta_2)' Z \, W \, Z' \, \xi(\theta_2). \tag{12}$$

[7]The contraction tolerance is set to 10^{-8} with a maximum of 2,000 iterations. The outer GMM optimization uses L-BFGS-B with function- and gradient-tolerances $(10^{-9}, 10^{-8})$ and integration over $R = 500$ scrambled Halton draws (seeded for reproducibility).

The first step uses $W = (Z'Z)^{-1}$; the second step updates W to the heteroskedasticity-robust optimal weight matrix evaluated at the first-step residuals.

Price endogeneity and instruments. The log prompt price $\log p_{mt}^{prompt}$ entering μ_{imt} in equation (8) is potentially endogenous: providers have private information about ξ_{mt} and set prices accordingly, biasing estimates of the price-coefficient parameters $(\mu_\alpha, \sigma_\alpha)$. 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 larger models require proportionally more GPU-hours per token, their marginal cost moves more strongly with the GPU rental price. I therefore interact the monthly GPU rental price index with each model's parameter count to obtain a single continuous cost shifter that varies across both models and months:

$$z_{m,t}^{\text{cost}} \;=\; gpu_t \;\times\; \text{params}_m, \tag{13}$$

where params_m is the model's parameter count, measured in billions. For open-weight models, params_m is disclosed by the provider; for closed-source models it is estimated using the inference-cost knowledge prior of Li (2026). Figure 7 plots the resulting parameter counts against posted prompt prices for all AI models in the sample: the open-weight and closed-source groups overlap substantially in parameter–price space, with a clear positive relationship between size and posted price on a log–log scale, consistent with parameter count serving as a coarse proxy for inference cost per token. The exclusion restriction requires that, conditional on month fixed effects (which absorb the common time-series component of GPU costs) and observed quality attributes, $z_{m,t}^{\text{cost}}$ affects token shares only through its effect on prices. This restriction is plausible because, while parameter count 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, parameter count 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.

To aid identification of the random coefficients, I additionally construct BLP-style instruments based on rival- and own-firm product characteristics:

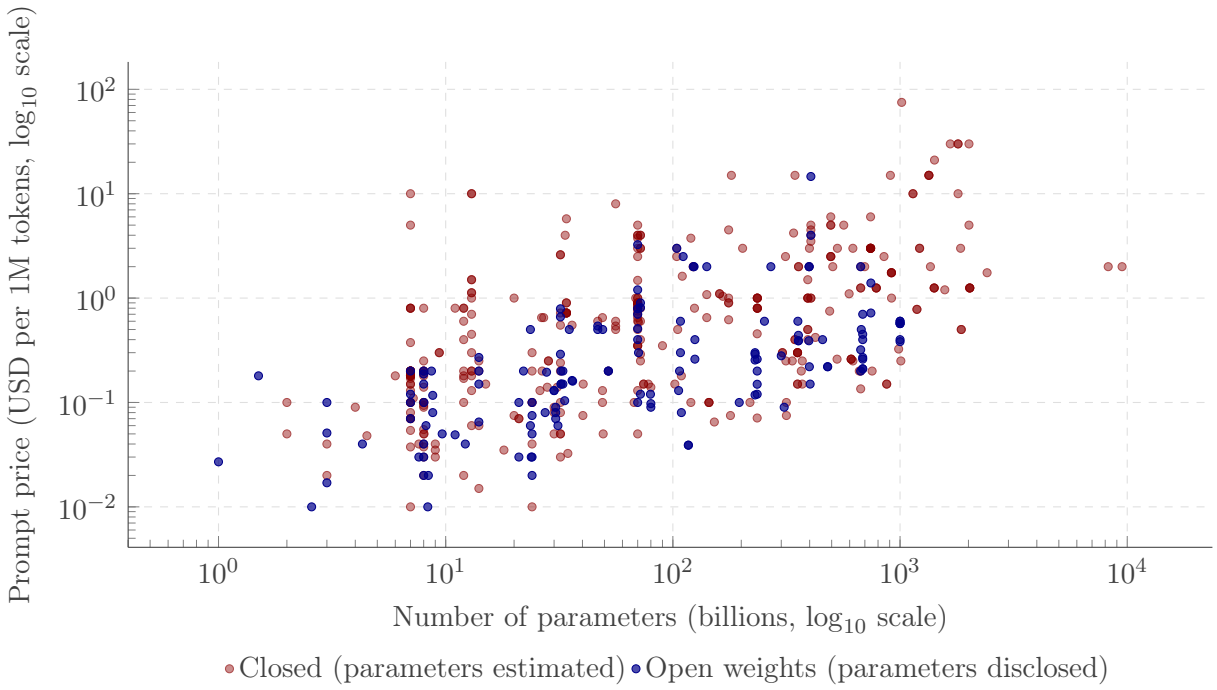
$$z^{\text{rival}}_{m,x,t} = \sum_{j \neq m} x_{jt}, \qquad z^{\text{own}}_{m,x,t} = \sum_{\substack{j \neq m \\ j \in \mathcal{F}_m}} x_{jt}, \tag{14}$$

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. Instruments are pruned to a linearly independent set before estimation. The GPU-cost shifters identify the location of the price-coefficient distribution $\hat{\mu}_\alpha$; the BLP characteristic instruments identify its dispersion $\hat{\sigma}_\alpha$ together with the quality random-coefficient dispersions $\{\hat{\sigma}_k\}$, which jointly govern how proximity in characteristic space shapes substitution patterns.

Table 4 reports the BLP mean coefficients across three specifications. The estimated price-coefficient parameters $(\hat{\mu}_\alpha, \hat{\sigma}_\alpha)$ imply a strictly negative price coefficient $\alpha_i = \exp(\hat{\mu}_\alpha + \hat{\sigma}_\alpha \nu_{ip})$ for every simulated consumer; the implied median own-price elasticity is comparable in magnitude to the multinomial logit estimates.

The specification without provider fixed effects is the baseline for markup recovery and all subsequent Jacobian computations, for the following reason. Provider identity is closely tied to pricing power: provider 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. Adding provider fixed effects attenuates the price coefficient, consistent with the interpretation that provider dummies absorb the cross-provider price variation that

Figure 7: Parameter Counts and Prompt Prices Across AI Models



Notes. Each marker represents one AI model in the sample. The vertical axis is the posted prompt price (USD per 1M tokens); the horizontal axis is the number of model parameters (in billions). Both axes are on a $\log_{10}$ scale. Blue markers denote open-weight models whose parameter counts are disclosed by the provider; red markers denote closed-source models whose parameter counts are estimated following Li (2026).

reflects market power.

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

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

Under random coefficients there is no closed-form expression analogous to the multinomial logit's stacked formula; derivatives are computed by simulation across the same $R = 500$ Halton draws used to compute predicted shares. Let $\alpha_i^r \equiv \exp(\mu_\alpha + \sigma_\alpha \nu_p^r)$ denote the (strictly positive) price sensitivity of simulated consumer r. The own-price derivative is

$$\frac{\partial \tilde{s}_{mt}}{\partial p_{mt}^{\text{prompt}}} = -\frac{1}{R}\sum_{r=1}^{R} \frac{\alpha_i^r}{p_{mt}^{\text{prompt}}}\, s_{mt}(\cdot \mid \nu^r)\big[1 - s_{mt}(\cdot \mid \nu^r)\big], \tag{15}$$

Table 4: Demand System: BLP

$\log p^{prompt}$	-0.000***	-0.903***	-1.297***
	(0.000)	(0.331)	(0.360)
Q^{code}		3.910**	0.449
		(1.865)	(1.532)
Q^{gen}		1.619***	2.858***
		(0.628)	(0.668)
$Q^{math/sci}$		-1.250	-0.119
		(1.020)	(0.629)
Q^{relia}		0.165	0.068
		(0.409)	(0.723)
Speed		0.002	-0.004
		(0.002)	(0.004)
Latency		-0.022	-0.026
		(0.028)	(0.046)
Reasoning model		-3.531***	-2.616***
		(0.499)	(0.653)
Context window		1.408***	0.559
		(0.346)	(0.462)
Extra controls	No	Yes	Yes
Month FE	Yes	Yes	Yes
Provider FE	No	No	Yes
N	1150	1150	1150
R^2	-0.000	0.735	0.737

Notes. This table reports mean-coefficient estimates from 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. Instruments include the H100 GPU rental price interacted with each model's parameter count (cost-side; see equation (13)) 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%.

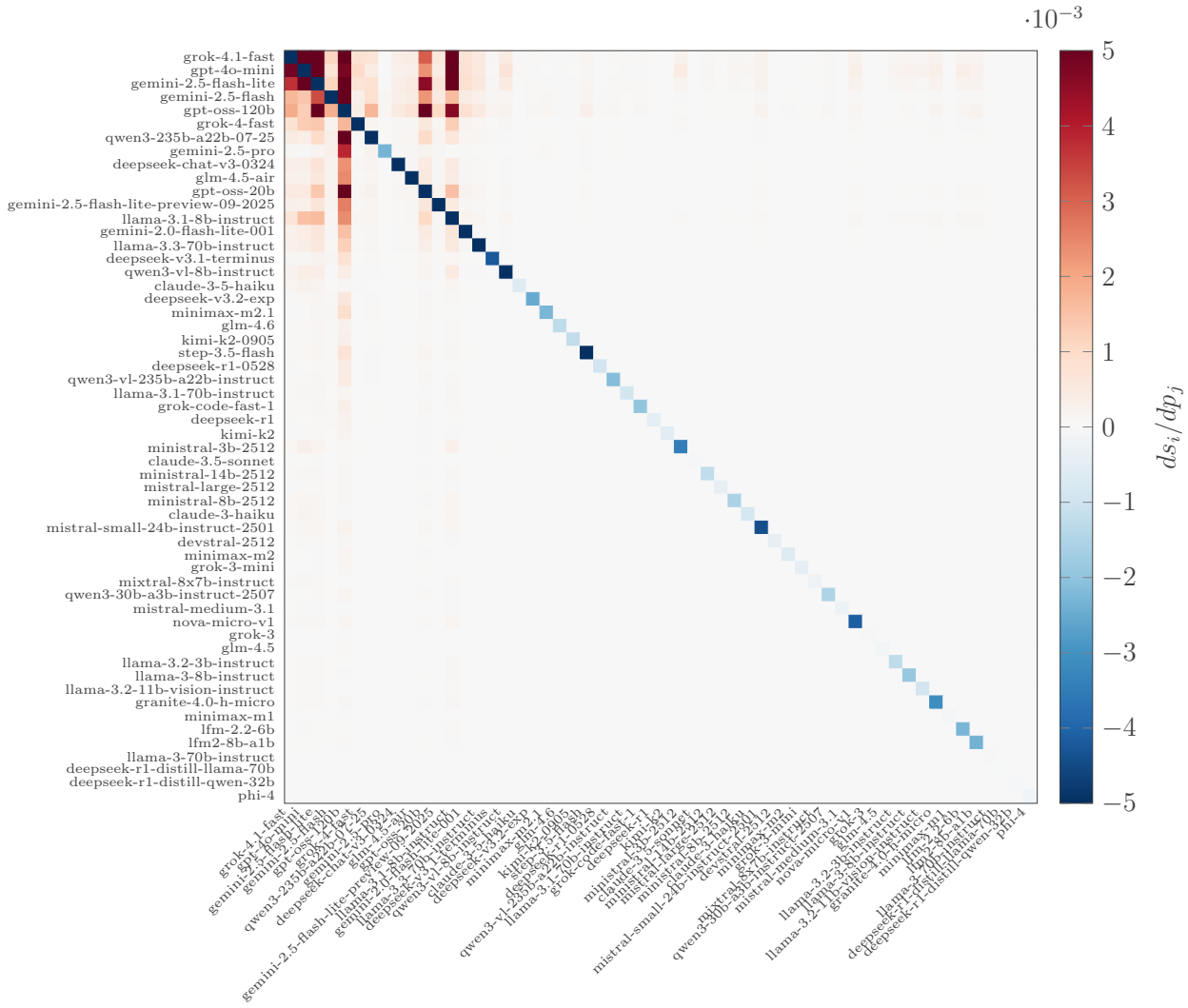
and the cross-price derivative is

$$\frac{\partial \tilde{s}_{jt}}{\partial p_{mt}^{\text{prompt}}} = \frac{1}{R}\sum_{r=1}^{R} \frac{\alpha_i^r}{p_{mt}^{\text{prompt}}}\, s_{jt}(\cdot \mid \nu^r)\, s_{mt}(\cdot \mid \nu^r), \quad j \neq m. \tag{16}$$

Two features distinguish this Jacobian from the logit's. First, each simulated consumer contributes a different price sensitivity $\alpha_i^r = \exp(\mu_\alpha + \sigma_\alpha \nu_p^r)$, so consumers with high ν_p^r amplify own-price effects in the segments they primarily occupy; because $\alpha_i^r > 0$ for every draw, all simulated own-price derivatives are strictly negative and all cross-price derivatives strictly positive. Second, the IIA restriction is relaxed: substitution patterns

now depend on *product similarity in characteristic space* rather than on observed market shares alone, so two models that are close in $(\mathbf{Q}, X)$-space are close substitutes for the consumers who value those characteristics, regardless of their overall market shares.

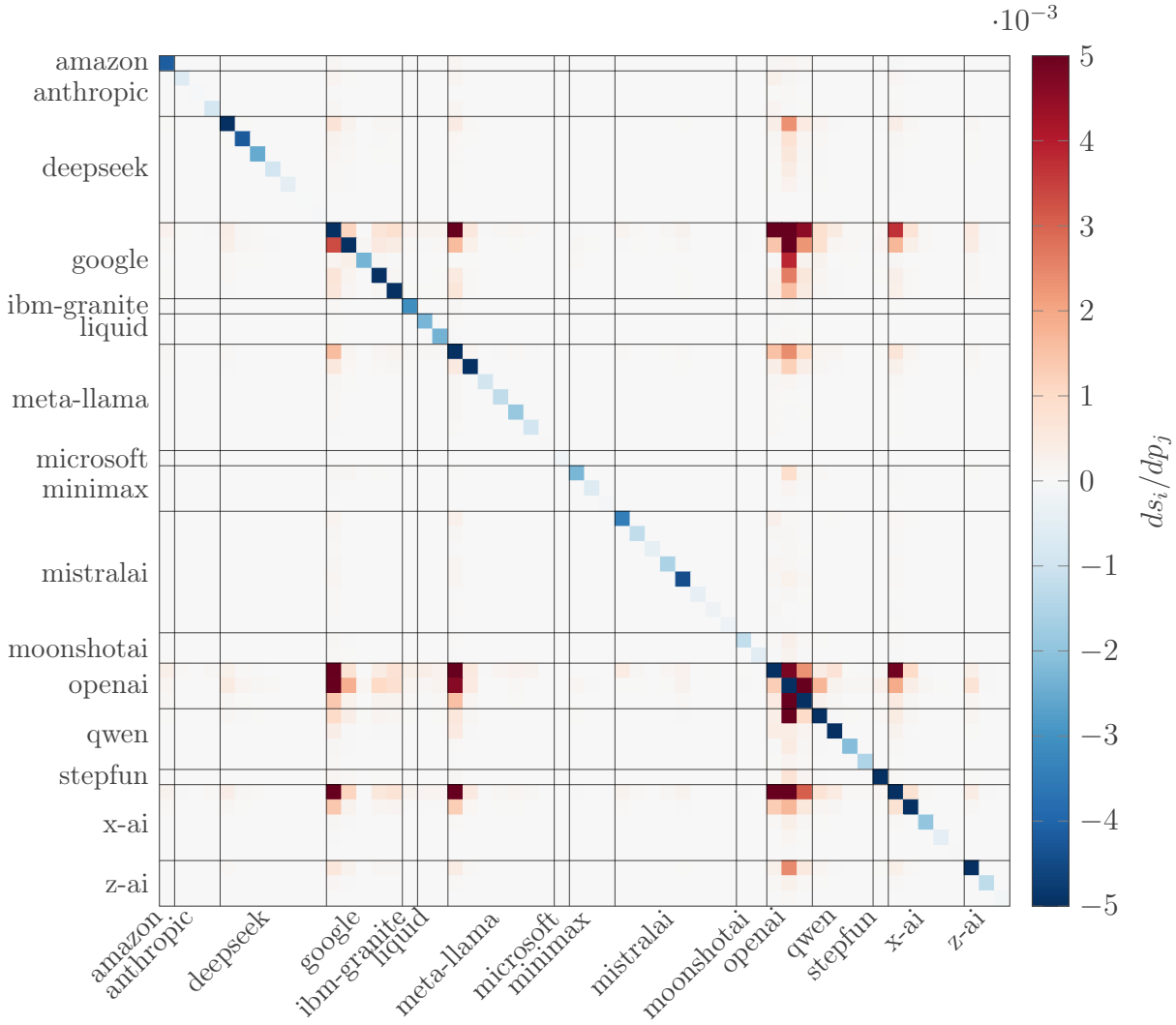
Figure 8: Demand Jacobian Heatmap: Models Ordered by Token Share



Notes. Each cell (i, j) shows the BLP-implied $\partial s_i / \partial p_j$ 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.

Figures 8 and 9 visualize the estimated BLP Jacobian $\widehat{\Delta}_t$ for the latest month in the sample. In Figure 8, 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 substantially affect a model's own

Figure 9: Demand Jacobian Heatmap: Models Ordered by Provider



Notes. Same BLP Jacobian as Figure 8, with models reordered by provider (alphabetical) and then by token share within each provider. Black lines delineate provider blocks.

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 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 exert negligible competitive pressure on one another. This heterogeneity suggests that 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 9 reorders the same Jacobian by provider, showing 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 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 show additional structure. The Google–OpenAI–xAI sub-block displays the largest off-diagonal entries among cross-provider pairs, indicating strong demand substitution among these three providers. Anthropic's row, by contrast, exhibits comparatively small cross-price effects, consistent with a sufficiently differentiated product position that insulates Anthropic's demand from rivals' pricing decisions and with strong market power in the high-capability segment. Cross-provider effects are smaller in absolute terms than under the multinomial logit, reflecting that random coefficients compress conditional shares toward zero or one for each simulated consumer; the relative ordering across provider pairs, which carries the substantive market-power interpretation, is preserved.

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 simulation-based BLP 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:

$$\tilde{\mathbf{s}}_t + \big(\Omega_t \odot \Delta_t\big)\big(\mathbf{p}_t - \mathbf{mc}_t\big) = \mathbf{0}, \tag{17}$$

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 (17), I recover $\mathbf{m}_t$ by solving the linear system month by month.

$$\big(\Omega_t \odot \widehat{\Delta}_t\big)\widehat{\mathbf{m}}_t = -\tilde{\mathbf{s}}_t. \tag{18}$$

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

$$\widehat{\mathbf{mc}}_t = \mathbf{p}_t - \widehat{\mathbf{m}}_t, \qquad \widehat{\boldsymbol{\ell}}_t \equiv \frac{\widehat{\mathbf{m}}_t}{\mathbf{p}_t} = \mathbf{1} - \frac{\widehat{\mathbf{mc}}_t}{\mathbf{p}_t}, \tag{19}$$

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

$$\widehat{\mu}_{mt} \equiv \frac{p_{mt}^{prompt}}{\widehat{mc}_{mt}} = \frac{1}{1 - \widehat{\ell}_{mt}}. \tag{20}$$

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

$$\widehat{\mu}_{jt}^{prov} \equiv \frac{\sum_{m \in \mathcal{M}_{jt}} v_{mt} \, \widehat{\mu}_{mt}}{\sum_{m \in \mathcal{M}_{jt}} v_{mt}}, \tag{21}$$

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:

$$\widehat{\mu}_{t}^{mkt} \equiv \frac{\sum_{m} v_{mt} \, \widehat{\mu}_{mt}}{\sum_{m} v_{mt}}. \tag{22}$$

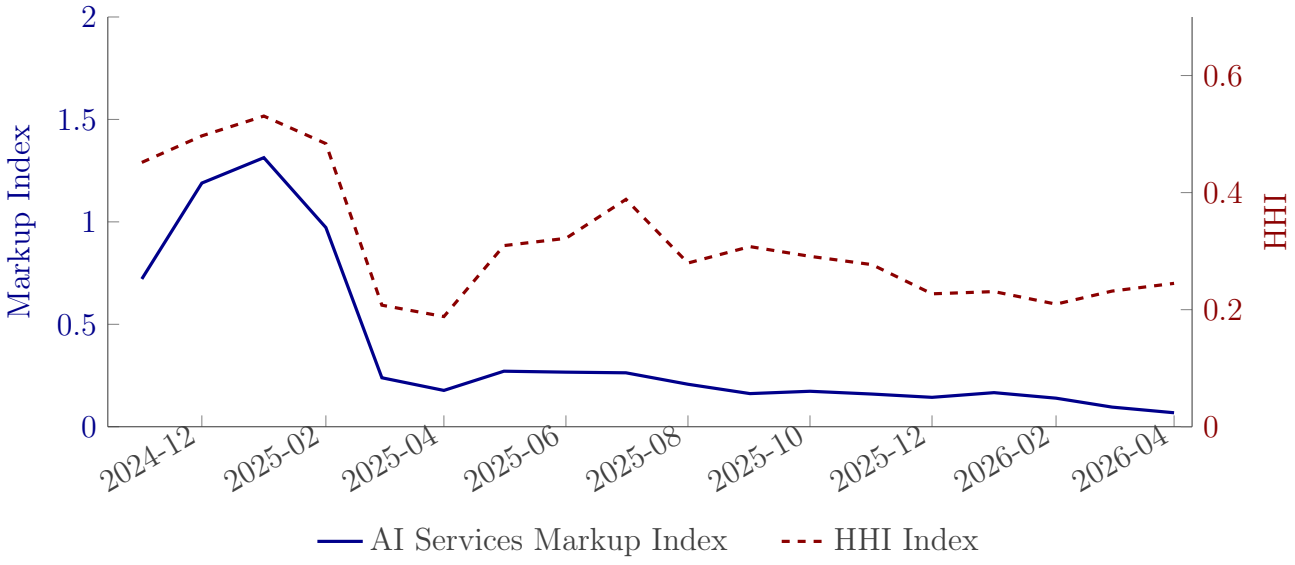
Token-value weighting ensures that markups on high-usage, high-price models receive proportionally greater weight. I also compute the Herfindahl-Hirschman Index (HHI) based on provider-level token-value shares to measure concentration in the AI services market.

Markup results. Figure 10 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 apparent 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 11 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

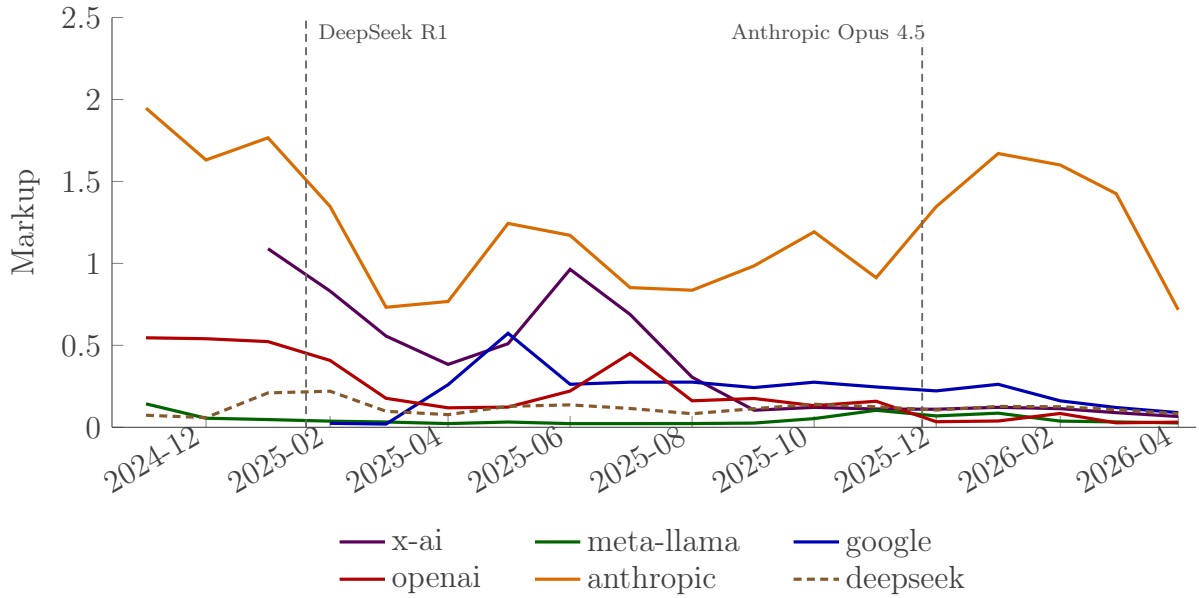
releases of frontier models, Claude Opus 4.5 in late November 2025 followed by Claude Opus 4.6 and Sonnet 4.6 in February 2026, which reinforced its market position and allowed it to sustain premium pricing in an otherwise competitive market. More broadly, several patterns emerge across the full sample. 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 are the key moderating variable in the micro pass-through analysis of Section 6.

Figure 10: Market Markup Index and Provider Concentration



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.

Figure 11: Token-Value-Weighted Markups by Provider



Notes. The figure plots provider-level token-value-weighted markup series over time, constructed using BLP demand estimates.

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 sizable, 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, moving 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 most of the 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:

$$\Delta \log \tilde{p}_{mt} = \theta_0 \Delta \log gpu_t + \theta_1 \Big(\Delta \log gpu_t \times \widehat{\mu}^{\text{prov}}_{j(m),t-1} \Big) + \theta_2 \widehat{\mu}^{\text{prov}}_{j(m),t-1} + \alpha_{j(m)} + \varepsilon_{mt}, \quad (23)$$

where $\log \tilde{p}_{mt}$ is the quality-adjusted 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

	BLP	
$\Delta \log gpu_t$	0.2156**	0.3150***
	(0.1004)	(0.1167)
$\Delta \log gpu_t \times \widehat{\mu}^{\text{prov}}_{j(m),t-1}$		-0.2502**
		(0.1002)
$\widehat{\mu}^{\text{prov}}_{j(m),t-1}$		-0.0435
		(0.0328)
company FE	Yes	Yes
R-squared Adj.	0.0451	0.0453
N	1011	1011

Notes. This table reports estimates of GPU cost pass-through into quality-adjusted AI prompt prices, using BLP-implied provider-level markups $\hat{\mu}^{\text{prov}}_{j(m),t-1}$. 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 $\hat{\theta}_0 = 0.315^{***}$, while the interaction term $\hat{\theta}_1 = -0.250^{**}$ 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 $\hat{\theta}_0 + \hat{\theta}_1 \times 1$: roughly this fraction 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 $\mu^* = \hat{\theta}_0 / |\hat{\theta}_1| \approx 1.26$: 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 11 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 because 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

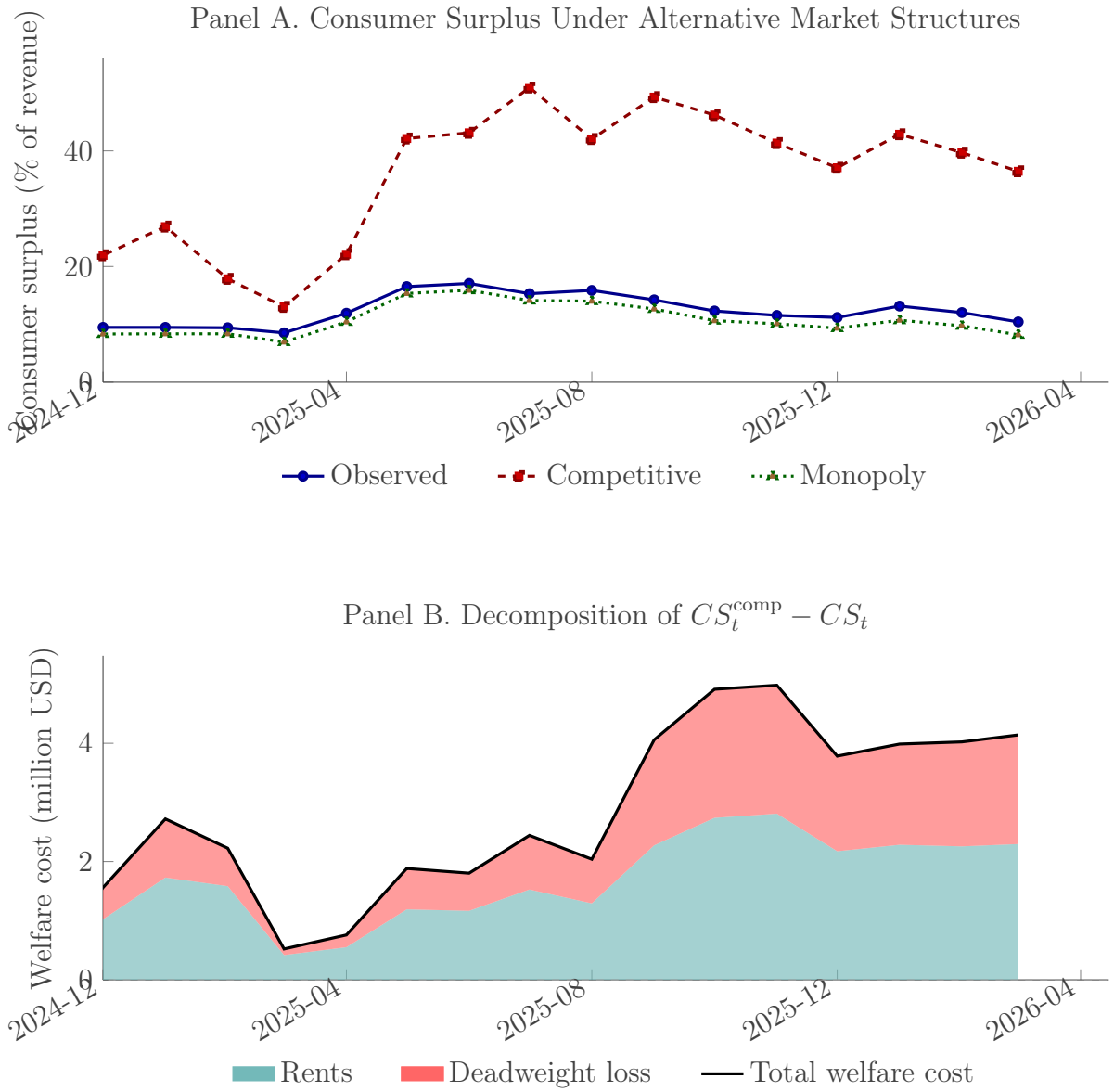
The micro pass-through results establish that market power attenuates the transmission of GPU cost declines to AI API prices. But how large is the welfare cost of this attenuation, and how does it compare to the welfare gains from the entry of new, higher-capability models? This subsection quantifies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

The BLP demand system estimated in Section 5 does not admit a closed-form expression for aggregate consumer surplus; under random coefficients, surplus must be computed by simulation across the same $R = 500$ Halton draws used in demand estimation. Following Small and Rosen (1981), the per-consumer log-sum-exp is integrated over consumer-level draws to yield market-level surplus:

$$CS_t \;=\; \frac{1}{R}\sum_{r=1}^{R} \frac{1}{\hat{\alpha} + \hat{\sigma}_p \nu_p^r} \ln\!\left(1 + \sum_{m=1}^{J_t} \exp\!\left(\hat{\delta}_{mt} + \hat{\mu}_{mt}^r\right)\right) \;\times\; \frac{\text{tokens}_t^{OR}}{\lambda}, \tag{24}$$

where $\hat{\delta}_{mt}$ is the estimated mean utility from equation (7); $\hat{\mu}_{mt}^r$ is the simulated individual deviation from equation (8) for draw r; $\hat{\alpha} + \hat{\sigma}_p \nu_p^r$ is consumer r's individual price coefficient, which converts log-utility units to dollars; 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.

Figure 12: 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 by simulation from the BLP demand estimates in Section 5 using equation (24). 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 (27)–(28); the black line reports total welfare cost.

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

and the random-coefficient deviations:

$$\hat{\delta}_{mt}^{\text{comp}} \;=\; -\hat{\alpha}\,\log \widehat{mc}_{mt} + \hat{\beta}'\mathbf{Q}_m + \hat{\Phi}'\mathbf{X}_{mt} + \hat{\tau}_t, \quad \hat{\mu}_{mt}^{r,\text{comp}} \;=\; -\hat{\sigma}_p \nu_p^r \log \widehat{mc}_{mt} + \sum_k \hat{\sigma}_k \nu_{ik}^r Q_{mt}^k. \tag{25}$$

The competitive consumer surplus is then

$$CS_t^{\text{comp}} \;=\; \frac{1}{R}\sum_{r=1}^{R} \frac{1}{\hat{\alpha} + \hat{\sigma}_p \nu_p^r}\; \ln\!\left(1 + \sum_{m=1}^{J_t} \exp\!\left(\hat{\delta}_{mt}^{\text{comp}} + \hat{\mu}_{mt}^{r,\text{comp}}\right)\right) \;\times\; \frac{\text{tokens}_t^{OR}}{\lambda}. \tag{26}$$

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}.

The gap between observed and competitive surplus is initially small but widens markedly from mid-2025 onward. Two forces account for this pattern. First, provider-level markups were lower and more compressed in the early part of the sample (Figure 11), so observed prices were closer to marginal cost and the static welfare cost of market power was correspondingly modest. The entry of DeepSeek in early 2025 further compressed markups, keeping the two series close through the first half of 2025. Second, as the market expanded rapidly in aggregate token volume (Figure 2), the same markup percentage applied to a much larger base generates a proportionally larger absolute surplus gap.

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:

$$T_t \;=\; \sum_{m=1}^{J_t} (p_{mt}^{\text{prompt}} - \widehat{mc}_{mt}) \;\times\; \frac{\text{tokens}_t^{OR}}{\lambda}, \tag{27}$$

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:

$$DWL_t \;=\; \left(CS_t^{\text{comp}} - CS_t\right) \;-\; T_t. \tag{28}$$

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:

$$\hat{\delta}_{mt}^{\text{set-frozen}} \;=\; -\hat{\alpha}\,\log p_{mt}^{\text{prompt}} \;+\; \hat{\beta}'\mathbf{Q}_m \;+\; \hat{\Phi}'\mathbf{X}_{mt} \;+\; \hat{\tau}_t, \qquad m \in \mathcal{J}_{t_0} \cap \mathcal{J}_t, \tag{29}$$

with the corresponding random-coefficient deviation $\hat{\mu}_{mt}^{r,\text{set-frozen}} = -\hat{\sigma}_p \nu_p^r \log p_{mt}^{\text{prompt}} + \sum_k \hat{\sigma}_k \nu_{ik}^r Q_{mt}^k$. 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.[8] The corresponding product-set-frozen consumer surplus is

$$CS_t^{\text{set-frozen}} \;=\; \frac{1}{R}\sum_{r=1}^{R} \frac{1}{\hat{\alpha} + \hat{\sigma}_p \nu_p^r} \ln\!\left(1 + \sum_{m \in \mathcal{J}_{t_0} \cap \mathcal{J}_t} \exp\!\left(\hat{\delta}_{mt}^{\text{set-frozen}} + \hat{\mu}_{mt}^{r,\text{set-frozen}}\right)\right) \times \frac{\text{tokens}_t^{OR}}{\lambda}. \tag{30}$$

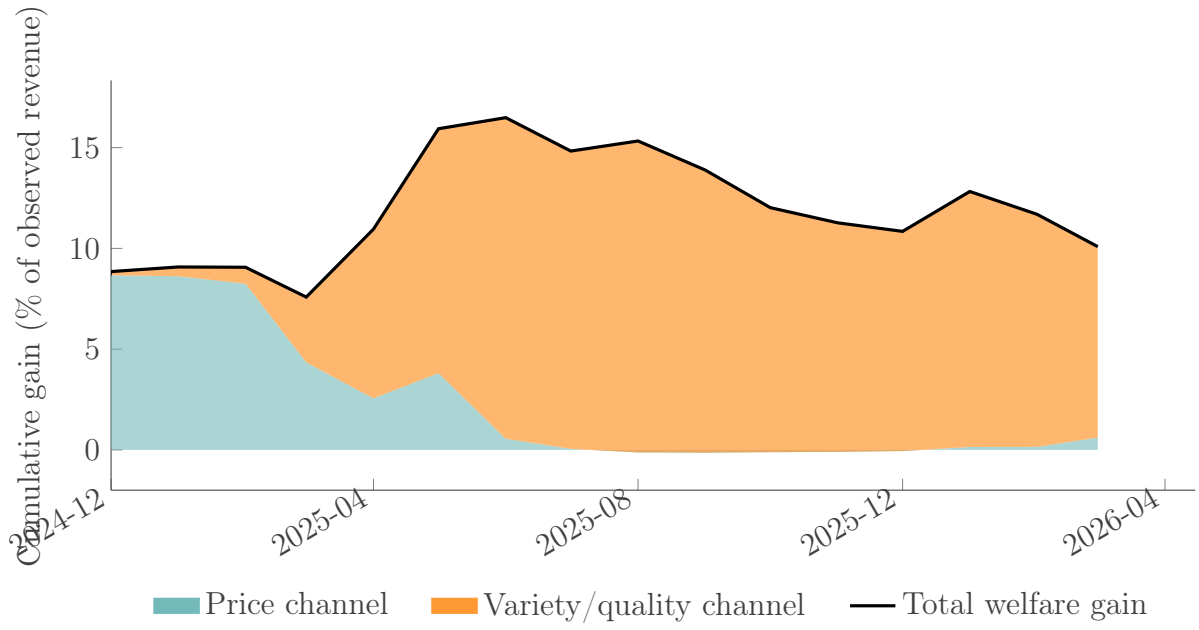
[8] This conservative treatment means that the product-set-frozen surplus $CS_t^{\text{set-frozen}}$ may *decline* toward the end of the sample if a large share of initial-period models are delisted, attributing part of the delisting effect to the variety/quality channel. In practice, most high-usage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

The decomposition of cumulative welfare gains is then:

$$\underbrace{CS_t - CS_{t_0}}_{\text{Total welfare gain}} = \underbrace{CS_t^{\text{set-frozen}} - CS_{t_0}}_{\text{Price channel}} + \underbrace{CS_t - CS_t^{\text{set-frozen}}}_{\text{Variety/quality channel}}. \tag{31}$$

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 13 plots the cumulative decomposition. The variety/quality channel dominates 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.

Figure 13: 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 (31). The shaded area between the two series represents the variety/quality channel; the area below the product-set-frozen series represents the price channel.

The welfare analysis above is subject to an important limitation. 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 matters here because Figure 13 shows that the entry of new, higher-capability models, not price reductions on existing ones, drives most of the cumulative welfare gains. The static welfare cost $T_t + DWL_t$ therefore likely overstates the true welfare cost of market power if lower margins would slow the pace of frontier-model development. The same providers that sustain the highest markups also drive the product innovation that generates most of the consumer surplus gains documented above. Whether policies that compress provider margins would accelerate cost pass-through or slow the product innovation that accounts for most welfare gains remains an open question; resolving it would require a dynamic model of endogenous entry and investment, which 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

$$A_{js} = \frac{Z_{js}}{x_s}, \tag{32}$$

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,

$$\mathbf{L}^p \equiv \left(\mathbf{I} - \mathbf{A}^\top\right)^{-1}, \tag{33}$$

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.[9]

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.[10] 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,

$$b_s^{AI} = \kappa \cdot A_{cs}, \tag{34}$$

[9]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.

[10]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.

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,

$$E_s^{AI} \;=\; \kappa \;\cdot\; \left[L^p\right]_{sc}, \tag{35}$$

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 14 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 can be aggregated into an economy-wide deflationary contribution.

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

$$\Delta \log PPI_{s,t} \;\approx\; E_s^{AI} \;\times\; \Delta \log P_t^{AI}, \tag{36}$$

where $E_s^{AI} = \kappa \cdot L_{cs}$ is the total AI supply-chain exposure from equation (35), and $\Delta \log P_t^{AI}$ is the change in the quality-adjusted AI services price index from equation (2).

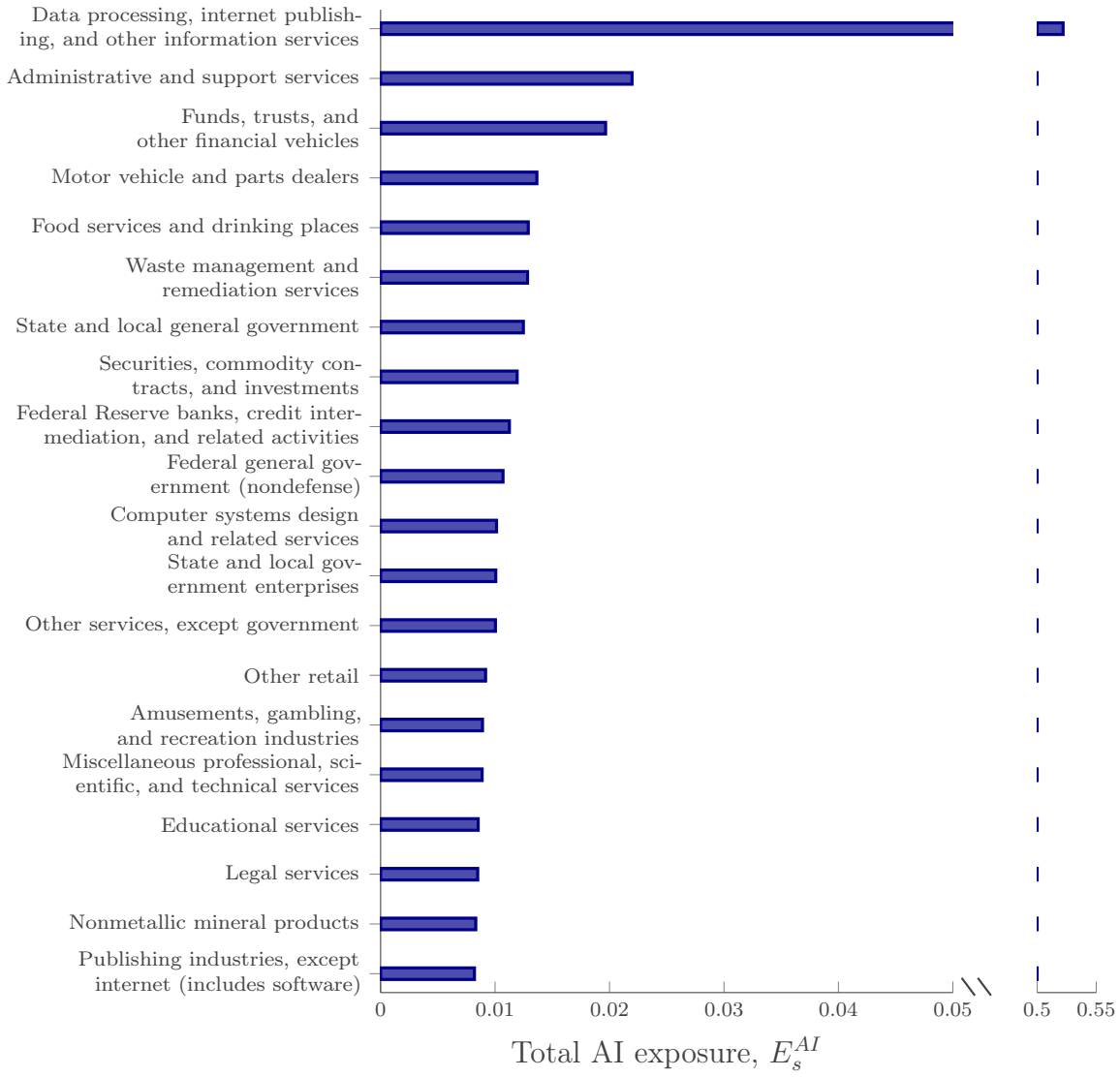
Aggregating to an economy-wide PPI contribution by weighting sectors by their gross-output shares $\omega_s \equiv x_s / \sum_{s'} x_{s'}$ (row T018 of the BEA Use table) gives

$$\Delta \log PPI_t^{agg} \;\approx\; \sum_s \omega_s \;\cdot\; E_s^{AI} \;\cdot\; \Delta \log P_t^{AI}. \tag{37}$$

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

Figure 14: 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.

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:

$$\text{Direct}_{s,t} \;=\; b_s^{AI} \;\times\; \Delta \log P_t^{AI}. \tag{38}$$

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

$$\text{Indirect}_{s,t} \;=\; \left(E_s^{AI} - b_s^{AI}\right) \;\times\; \Delta \log P_t^{AI}. \tag{39}$$

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 (23), $\partial \, \Delta \log \tilde{p} / \partial \, \Delta \log gpu = \hat{\theta}_0 + \hat{\theta}_1 \, \mu$, I construct GPU-driven AI price changes under prevailing oligopolistic markups, perfect competition ($\mu = 1$), and a monopoly counterfactual μ^{mono}:[11]

$$\Delta \log P_t^{AI,\, gpu,\, k} \;\equiv\; \left(\hat{\theta}_0 + \hat{\theta}_1 \, \mu^k\right) \cdot \Delta \log gpu_t, \qquad \mu^k \;\in\; \left\{\bar{\mu}_{t-1},\; 1,\; \mu^{\text{mono}}\right\}. \tag{40}$$

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

$$\text{Wedge}_{s,t}^{k} \;\equiv\; E_s^{AI} \;\times\; \left[\Delta \log P_t^{AI,\, gpu,\, k} \;-\; \Delta \log P_t^{AI,\, gpu,\, \bar{\mu}}\right], \qquad k \;\in\; \{\text{comp},\, \text{mono}\}. \tag{41}$$

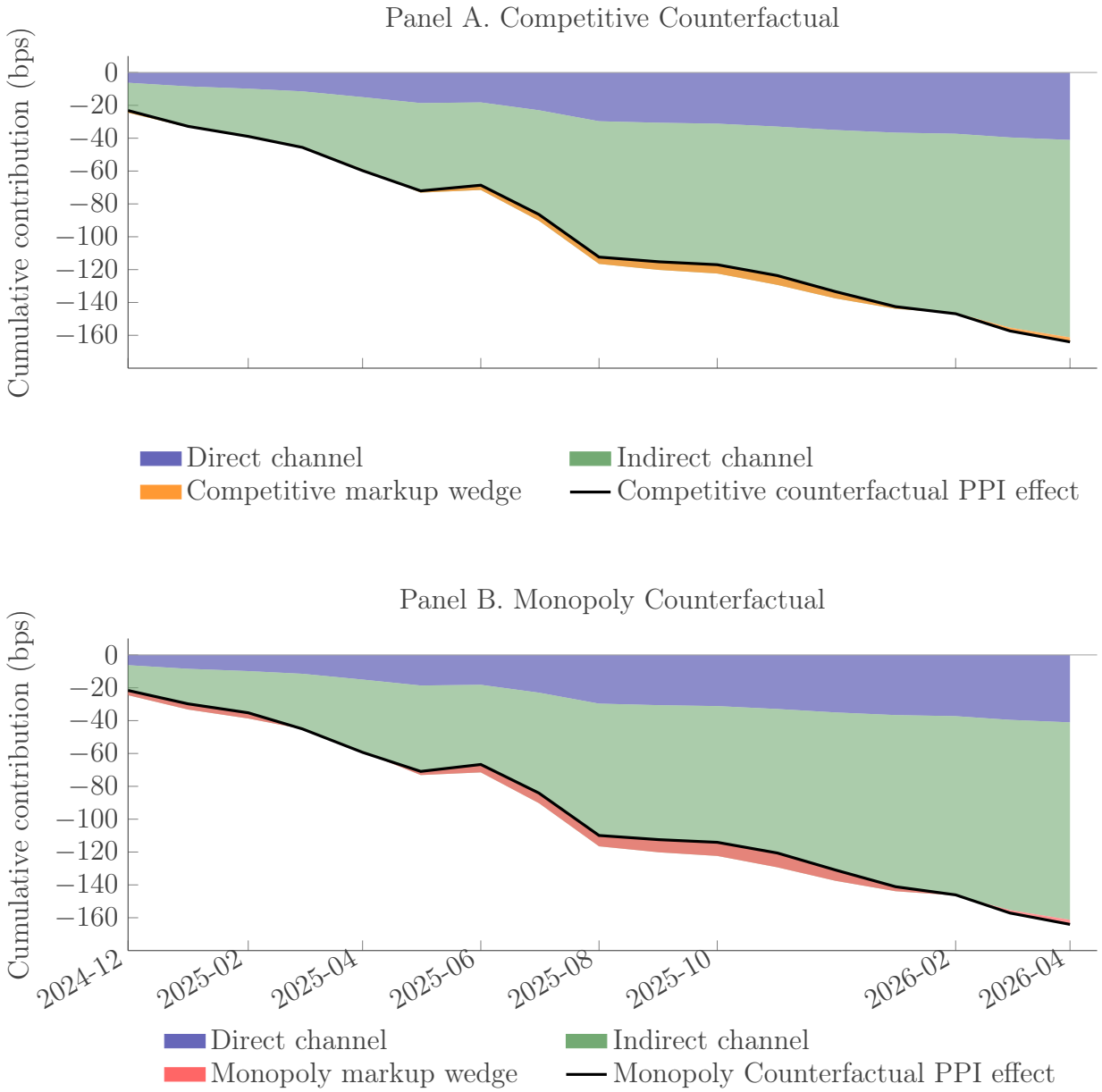
Since $\hat{\theta}_1 < 0$ (Table 5) and actual markups exceed unity, $\text{Wedge}_{s,t}^{\text{comp}} < 0$ when GPU costs fall indicates that competitive pricing would have transmitted more deflation downstream, while $\text{Wedge}_{s,t}^{\text{mono}} > 0$ indicates monopoly pricing would have transmitted 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 (37) yields economy-wide time series for the direct, indirect, and market-power

[11] $\bar{\mu}_{t-1}$ 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 (23), 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 is therefore a conservative upper bound on the market-power wedge that is credibly identified by the data.

wedge channels.

Figure 15 translates the micro pass-through result into economy-wide terms. Three results emerge. First, the indirect supply-chain channel accounts for the majority of the cumulative deflationary contribution in both panels, because 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 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: 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 15: 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 (41). 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 and shows that quality-adjusted AI prices fell by approximately 78 percent from early 2024 through early 2026. Estimating a structural demand system on OpenRouter token usage data and recovering markups via multi-product Bertrand inversion, I document that the AI services market is oligopolistic and that higher-markup providers absorb a larger share of upstream cost declines, attenuating the deflationary effect before it reaches downstream firms. Embedding the quality-adjusted price index in a Leontief input–output framework, I find a cumulative deflationary contribution of roughly 160 basis points to the PPI over 2024–2026, with indirect supply-chain linkages accounting for the majority of the aggregate effect; under competitive pricing, this contribution would have been appreciably larger. A welfare decomposition shows that the entry of new, higher-capability models, not price reductions on existing ones, accounts for most of the consumer surplus gains, suggesting a tension between compressing provider margins to improve cost pass-through and preserving the incentives for the product innovation that generates most downstream value. The findings also bear on price measurement: as AI input shares grow, statistical agencies that rely on nominal token prices will increasingly understate AI deflation, biasing estimates of real output growth and productivity.

References

Acemoglu, Daron, "The Simple Macroeconomics of AI," *Economic Policy*, 2025, *40* (121), 13–58.

_ **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.

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, "Importers, Exporters, and Exchange Rate Disconnect," *American Economic Review*, 2014, *104* (7), 1942–1978.

_ **,** _ **, and** _ , "International Shocks, Variable Markups, and Domestic Prices," *Review of Economic Studies*, 2019, *86* (6), 2356–2402.

Atkeson, Andrew and Ariel Burstein, "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 2008, *98* (5), 1998–2031.

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.

Auer, Raphael A. and Raphael S. Schoenle, "Market Structure and Exchange Rate Pass-Through," *Journal of International Economics*, 2016, *98*, 60–77.

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *Quarterly Journal of Economics*, 2020, *135* (2), 645–709.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, "Artificial Intelligence, Firm Growth, and Product Innovation," *Journal of Financial Economics*, 2024, *151*, 103745.

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.

Barkai, Simcha, "Declining Labor and Capital Shares," *Journal of Finance*, 2020, *75* (5), 2421–2463.

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.

Berman, Nicolas, Philippe Martin, and Thierry Mayer, "How do Different Exporters React to Exchange Rate Changes?," *Quarterly Journal of Economics*, 2012, *127* (1), 437–492.

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.

Bresnahan, Timothy F. and Manuel Trajtenberg, "General Purpose Technologies: "Engines of Growth"?," *Journal of Econometrics*, 1995, *65* (1), 83–108.

_ **and Robert J. Gordon, eds**, *The Economics of New Goods*, Vol. 58 of *NBER Studies in Income and Wealth*, University of Chicago Press, 1997.

Brynjolfsson, Erik, Avinash Collis, and Felix Eggers, "Using Massive Online Choice Experiments to Measure Changes in Well-Being," *Proceedings of the National Academy of Sciences*, 2019, *116* (15), 7250–7255.

Bustamante, Maria Cecilia and Andres Donangelo, "Product Market Competition and Industry Returns," *Review of Financial Studies*, 2017, *30* (12), 4216–4266.

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.

David, Paul A., "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox," *American Economic Review*, 1990, *80* (2), 355–361.

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?," in Carol Corrado, John Haltiwanger, and Daniel Sichel, eds., *Measuring Capital in the New Economy*, University of Chicago Press, 2005, pp. 323–362.

Dou, Winston Wei, Yan Ji, and Wei Wu, "Competition, Profitability, and Discount Rates," *Journal of Financial Economics*, 2021, *140* (2), 582–620.

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

Eisfeldt, Andrea L., Gregor Schubert, Miao Ben Zhang, and Bledi Taska, "Generative AI and Firm Values," *Journal of Finance*, 2023. Forthcoming.

Flamm, Kenneth, "Measurement of DRAM Prices: Technology and Market Structure," in Murray F. Foss, Marilyn E. Manser, and Allan H. Young, eds., *Price Measurements and Their Uses*, University of Chicago Press, 1993, pp. 157–206.

Gabaix, Xavier, "The Granular Origins of Aggregate Fluctuations," *Econometrica*, 2011, *79* (3), 733–772.

Gambacorta, Leonardo and Vatsala Shreeti, "The AI Supply Chain," BIS Papers 154, Bank for International Settlements 2025.

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.

Gofman, Michael, Gill Segal, and Youchang Wu, "Production Networks and Stock Returns: The Role of Vertical Creative Destruction," *Review of Financial Studies*, 2020, *33* (12), 5856–5905.

Goldman Sachs Global Institute, "Tracking Trillions: The Assumptions Shaping the Scale of the AI Build-Out," Technical Report, Goldman Sachs 5 2026.

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.

Grullon, Gustavo, Yelena Larkin, and Roni Michaely, "Are US Industries Becoming More Concentrated?," *Review of Finance*, 2019, *23* (4), 697–743.

Hamilton, James D., "Oil and the Macroeconomy since World War II," *Journal of Political Economy*, 1983, *91* (2), 228–248.

Hausman, Jerry, "Cellular Telephone, New Products, and the CPI," *Journal of Business & Economic Statistics*, 1999, *17* (2), 188–194.

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

Hou, Kewei and David T. Robinson, "Industry Concentration and Average Stock Returns," *Journal of Finance*, 2006, *61* (4), 1927–1956.

Korinek, Anton and Jai Vipra, "Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence," *Economic Policy*, 2025, *40* (121), 225–256.

Li, Bojie, "Incompressible Knowledge Probes: Estimating Black-Box LLM Parameter Counts via Factual Capacity," 2026. arXiv:2604.24827.

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.

NVIDIA Corporation, "NVIDIA Announces Financial Results for Fourth Quarter and Fiscal 2026," Form 8-K, U.S. Securities and Exchange Commission, February 25 2026.

Ozdagli, Ali K. and Michael Weber, "Monetary Policy Through Production Networks: Evidence from the Stock Market," *Review of Financial Studies*, 2025. Forthcoming.

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

Richmond, Robert J., "Trade Network Centrality and Currency Risk Premia," *Journal of Finance*, 2019, *74* (3), 1315–1361.

Rishabh, Kumar and Vatsala Shreeti, "The Geography of AI Firms," BIS Working Paper 1343, Bank for International Settlements April 2026.

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.

Weyl, E. Glen and Michal Fabinger, "Pass-Through as an Economic Tool: Principles of Incidence under Imperfect Competition," *Journal of Political Economy*, 2013, *121* (3), 528–583.