Financializing Compute: The Design of AI Service Trade Markets*

Xiang Yu† Shuyang Wen‡

March 2026

Abstract

The global AI inference market—reaching approximately $90–100 billion annually and growing at 18% CAGR—operates without organized exchange infrastructure. We document three market failures: resource misallocation (80% of China's newly built compute capacity sits idle), price opacity (100-fold price dispersion across providers of equivalent quality), and unhedged risk exposure (85% of enterprises miss AI cost forecasts by more than 10%). Following the market design tradition of Roth [2002] and Budish et al. [2015], we propose the *AI Service Right* (ASR) as a transferable property right on AI compute and the *AI Service Unit* (ASU) as a quality-adjusted, cross-platform unit of account grounded in hedonic price theory [Rosen, 1974]. The ASU is modality-neutral: billing prices across text, image, video, and speech modalities are

*We thank Shuyang Wen for invaluable guidance and supervision throughout this project. We also thank seminar participants at Southwestern University of Finance and Economics for helpful comments. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors used AI-assisted tools for coding and formatting support; all intellectual contributions, theoretical results, and empirical analysis are the authors' own work. *Conflict of interest disclosure:* The authors have filed a Chinese national patent application related to the market design framework described in this paper. The authors have no other financial relationships or potential conflicts of interest to disclose. All errors are our own.

†Southwestern University of Finance and Economics. Email: 42433005@smail.swufe.edu.cn

‡Southwestern University of Finance and Economics. Email: wensy@swufe.edu.cn. Corresponding author.

unified via eq-token conversion factors—a latent accounting unit calibrated through market-price parity—with point estimates and sensitivity ranges for each modality ($\kappa_{\text{img}} \approx 2{,}667$ eq-tokens per image, range [1,500–5,000]; $\kappa_{\text{vid}} \approx 2{,}667$ per second of video, range [1,000–6,000]; $\kappa_{\text{spc}} \approx 7$ per second of audio), and modality-appropriate benchmark sets (MMLU/HumanEval for language; FID/CLIP Score for image; FVD/CLIPSIM for video; MMBench for multimodal) supply the quality index via PCA. We extend the framework to a vertically separated market in which compute providers can directly issue ASR for open-source models, eliminating double marginalisation; we show that idle-capacity providers serve as anchor participants whose entry can break the coordination trap without external subsidies. We design a hybrid secondary market architecture synthesizing mechanisms from four orthogonal market traditions: foreign exchange markets (cross-platform exchange rates and PPP-analog arbitrage via the ASU); equity markets (Central Limit Order Book, market making, clearing); electricity markets (Compute Locational Marginal Pricing for spatial scarcity signals); and decentralized finance (Automated Market Maker for long-tail liquidity).We establish six propositions and four supplementary remarks and validate the hedonic specification against a cross-section of 42 models from the Artificial Analysis leaderboard (Adj. $R^2 = 0.70$; geographic cost heterogeneity and provider identity explain more price variation than service quality). The propositions establish that bilateral trading is generically inefficient; Compute Locational Marginal Pricing decomposes nodal prices into system marginal cost, capacity congestion, and bandwidth premia; no-arbitrage equilibrium holds with capital constraints (extending Shleifer and Vishny 1997); the ASR market Pareto-improves over bilateral trading; market prices are more informative under ASR; the hybrid CLOB-AMM architecture weakly dominates either mechanism alone; platform adoption admits multiple equilibria with a coordination trap; financialization may improve or reduce price informativeness depending on speculator-hedger composition; and a hedonic micro-

foundation justifies the ASU definition.Calibrated agent-based simulation (500 steps, 30 Monte Carlo runs) provides computational validation: the hybrid architecture reduces price dispersion by 90% relative to bilateral trading, and order-of-magnitude welfare estimates suggest enterprise procurement cost savings of 0.2–20% (net of ASR transaction costs; see the welfare estimates (Online Appendix)) and potential TFP gains from compute reallocation of up to $29.9 billion annually. We propose a phased implementation roadmap from shadow ledger to full financialization, and we engage critically with the concern that financialization may not reduce intermediation costs [Philippon, 2015].

JEL Codes: D47, D82, G12, G14, G18, L86, Q41.

Keywords: AI Service Right; market design; market microstructure; automated market maker; locational marginal pricing; compute markets; price discovery; resource misallocation; financial innovation.

1 Introduction

The AI inference economy has reached commodity scale. China's National Data Administration reports daily token consumption exceeding 140 trillion as of March 2026—a thousandfold increase since early 2024 [National Data Administration, 2026]. The global AI inference market reached approximately $89–97 billion in 2024 (; we round to "approximately $100 billion" in summary contexts for readability) and is projected to exceed $250 billion by 2030, growing at a compound annual rate of 18% [Grand View Research, 2025]. Enterprise investment in generative AI surged from $1.7 billion to $37 billion between 2023 and 2025 [Menlo Ventures, 2025]. Gartner projects that inference spending will overtake training spending in 2026—a structural shift from one-time model development toward continuous, consumption-based compute [Gartner, 2026]. AI compute—measured in tokens for language models, in images for generative vision models, and in frames for video generation models—is becoming a critical factor of production for the intelligent economy [Brynjolfsson et al., 2025]. The ASR/ASU framework proposed in this paper is modality-agnostic: the quality index Q_i is constructed from modality-appropriate benchmarks (MMLU/HumanEval/MATH for language inference; FID and CLIP Score for image generation; FVD and CLIPSIM for video generation; MMBench for multimodal models), and the billing unit p_i^{native} is expressed in cost per standardized output unit for each modality.

Yet this hundred-billion-dollar market operates without organized exchange infrastructure. Enterprises purchase inference through bilateral contracts with individual platforms, at static posted prices, with no ability to transfer unused capacity, discover real-time compute scarcity, or hedge future costs. This paper identifies three structural market failures and proposes an institutional market design to resolve them.

Failure 1: Resource Misallocation. The current token economy consists of isolated platform silos. A firm that pre-purchases capacity from one provider cannot transfer

that purchasing power to another when a superior model emerges. The value is locked in. The Federal Reserve Board reports that up to 80% of China's newly built AI computing resources sat idle as of early 2025, due to misallocation by local officials during the real-estate downturn [Federal Reserve Board, 2025]. Globally, average GPU utilization in AI data centers ranges from 60–70%, with 32% of cloud budgets wasted on overprovisioned or idle resources [Lawrence Berkeley National Lab, 2025]. Meanwhile, 87% of enterprises adopt multi-cloud strategies specifically to avoid vendor lock-in [Flexera, 2024]—a costly workaround for the absence of an organized market.

Failure 2: Price Opacity. Current platform pricing is static and unilateral. Fradkin [2025] provides the first systematic analysis of the AI inference market using OpenRouter data, documenting that for any given intelligence level, prices vary by up to two orders of magnitude across providers. The Stanford AI Index reports that inference costs for GPT-3.5-level performance fell 280-fold between 2022 and 2024 [Stanford HAI, 2025], yet this dramatic cost decline coexists with persistent cross-provider price dispersion—direct evidence that no organized price discovery mechanism exists. Static bilateral prices cannot signal real-time compute scarcity, network congestion, or quality differentials across models.

Failure 3: Risk Exposure. Enterprise AI developers face compute cost as their largest variable operating expense, yet no financial instrument exists to hedge it. Survey evidence indicates that 85% of organizations miss AI cost forecasts by more than 10%, and nearly 25% are off by more than 50% [Benchmarkit, 2025]. Seventy-eight percent of IT leaders report unexpected charges from consumption-based AI pricing [Zylo, 2026]. No futures, options, or any hedging instruments exist for AI inference costs. This is the compute analogue of an airline unable to hedge jet fuel—a condition the energy derivatives market was designed to address [Ellerman et al., 2010].

These are not marginal inconveniences; they are the canonical conditions under which financial economics prescribes the creation of organized secondary markets [O'Hara, 1995]. The bilateral trading inefficiency invokes Myerson and Satterthwaite [1983]; the missing price discovery echoes the electricity industry before locational marginal pricing [Schweppe et al., 1988]; the unhedged risk exposure parallels the gaps that motivated Shiller [1993] to propose entirely new "macro markets" for society's largest economic risks.

Our argument has three parts. **First**, we introduce the *AI Service Right* (ASR)—a transferable property right on AI compute—and the *AI Service Unit* (ASU)—a quality-adjusted, cross-platform unit of account with formal hedonic micro-foundations (Section 4). **Part II: Theory and Design.** We construct a theoretical framework establishing bilateral market inefficiency as the welfare benchmark (Proposition 1), deriving Compute Locational Marginal Pricing from a social planner's dispatch problem (Proposition 2), formalising the structural relationship between ASU and C-LMP through a three-layer pricing structure (Remark 1), proving no-arbitrage equilibrium with capital constraints (Proposition 3), welfare improvement (Proposition 4), vertical disintermediation gains when compute providers issue ASR directly for open-source models (Proposition 5), price informativeness (Proposition 2), and hybrid CLOB-AMM dominance (Proposition 6). **Third**, we analyze platform participation as a coordination game with network externalities (Proposition 7) and conditions under which financialization improves or degrades price informativeness (the financialization channel), engaging directly with Goldstein and Yang [2022]. **Fourth**, we design a complete market architecture synthesizing four orthogonal market traditions: foreign exchange markets for the pricing layer (cross-platform exchange rates, PPP-analog arbitrage), equity markets for the trading layer (CLOB, market making, clearing), electricity markets for the supply anchor (C-LMP dispatch optimization), and DeFi for long-tail liquidity (AMM)—drawing on Evans and Lyons [2002], Glosten [1994]

and Avellaneda and Stoikov [2008], Wilson [2002] and Cramton [2017], and Lehar and Parlour [2025] and Milionis et al. [2022]. Additionally, we the hedonic specification empirically against cross-sectional data from the Artificial Analysis leaderboard ($N = 42$ models, Adj. $R^2 = 0.70$), showing that geographic cost heterogeneity and provider identity—not service quality—drive most price variation, directly motivating the ASU's quality adjustment (Section 7). Additionally, we calibrated computational validation through agent-based simulation (500 steps, 30 Monte Carlo runs), demonstrating 90% price dispersion reduction under the hybrid architecture and order-of-magnitude welfare estimates.

Following Roth [2002], we adopt an "economist as engineer" approach: we diagnose market failures, design institutional mechanisms to resolve them, and evaluate their performance. Following Agarwal and Budish [2021], our approach proceeds from market failure diagnosis to mechanism design to computational evaluation. The paper is organized as follows. Section 2 documents the institutional background and stylized facts. Section 3 reviews related literature. Section 4 introduces the ASR and ASU. Section 5 presents the theoretical framework. Section 6 describes the market architecture. Section 8 reports computational validation. Section efsec:conclusion discusses implementation, regulatory considerations, and the Philippon critique. Section 9.4 concludes.

2 Institutional Background and Stylized Facts

The AI inference market exhibits four stylized facts that motivate market design intervention.

Fact 1: Extreme price dispersion. Quality-adjusted prices for comparable AI inference services vary by up to two orders of magnitude across providers [Fradkin, 2025]. A 336-fold spread separates frontier models in March 2026 (GPT-4 at $168/M

output tokens vs. DeepSeek-V3 at \$0.50/M output tokens). Even within the same quality tier, prices differ by 5–10×. The 286-fold cost decline from GPT-4 to GPT-4o-mini over 23 months further demonstrates that posted prices bear little relation to marginal costs.

Fact 2: Massive idle capacity. China's national compute centres operate at approximately 20% average utilisation [Federal Reserve Board, 2025]. This idle capacity represents a pool of potential market participants whose opportunity cost of entry is effectively zero—a fact we exploit in Proposition 5(iv).

Fact 3: Enterprise lock-in and cost volatility. Enterprises report 15–20% annual AI cost overruns [Flexera, 2025, Zylo, 2026]. Over 60% of AI buyers have never compared quality-adjusted prices across platforms [Benchmarkit, 2025]. These frictions generate the welfare losses formalised in Proposition 1.

Fact 4: Electricity market precedent. Before restructuring, US electricity markets exhibited the same pathologies: bilateral contracts, regional price fragmentation, and chronic capacity underutilisation. Locational Marginal Pricing [Schweppe et al., 1988] resolved these through centralised dispatch and nodal pricing. We adapt LMP to AI compute as C-LMP (Section 5.3).

3 Related Literature

This paper draws on and contributes to six literatures. *Market design and mechanism design*: our approach follows the "economist as engineer" tradition of Roth [2002] and Agarwal and Budish [2021], applying it to a new domain. *Electricity markets*: we adapt Locational Marginal Pricing [Schweppe et al., 1988, Wilson, 2002] to compute resource allocation, following the analogy between GPU capacity and generation capacity. *FX markets*: our cross-platform exchange rate framework and PPP-analog arbitrage condition draw on Rogoff [1996] and Itskhoki and Mukhin [2021]. *Mar-*

ket microstructure: the hybrid CLOB-AMM architecture builds on Glosten [1994] for limit order books and Lehar and Parlour [2025] for AMM design; the price informativeness result extends Kyle [1985] and Goldstein and Yang [2017] to compute markets. *DeFi and AMMs*: we use the concentrated liquidity framework of Uniswap v3 [Adams et al., 2021] and the impermanent loss analysis of Milionis et al. [2022]. *Resource misallocation*: our welfare estimates follow the Hsieh and Klenow [2009] TFP framework, adapted to cross-platform compute allocation.

Two papers are closest to ours. Fradkin [2025] documents price dispersion in AI inference markets and proposes an "AI inference exchange" but does not develop the theoretical framework, market design, or quality-adjustment mechanism that are our primary contributions. Budish et al. [2015] diagnose a flaw in continuous limit order book design (sniping from serial processing) and propose frequent batch auctions; our paper diagnoses a flaw in bilateral compute procurement (price opacity from quality heterogeneity) and proposes an organised market with quality-adjusted pricing.

4 Conceptual Framework: ASR and ASU

4.1 AI Service Right (ASR)

Definition 1 (AI Service Right)**.** *An* ***AI Service Right (ASR)*** *is a transferable property right entitling its holder to redeem a specified quantity and quality of AI services from a designated platform, model, or compute node. An ASR Certificate is its digital, standardized, programmable representation.*

ASR value derives from four components: (i) *model inference right* (P99 latency, benchmark accuracy, context length, SLA); (ii) *compute resource right* (GPU/TPU hours, bandwidth, energy quotas); (iii) *API and ecosystem rights* (data access, fine-tuning, tool calling); (iv) *platform credit right* (reliability, compliance, continuity guarantee).

An ASR is deliberately designed to straddle two economic categories: it is simultaneously a *prepaid service contract* (a forward commitment to deliver AI computation) and a *transferable financial claim* (a right that can be sold, hedged, or used as collateral before redemption). Before redemption, an ASR behaves like a financial asset subject to price discovery and trading; at redemption, it behaves like a service contract subject to performance obligations. This dual nature mirrors electricity capacity rights, which trade as financial instruments before being exercised as physical delivery obligations [Ellerman et al., 2010].

We adopt the ERC-3525 Semi-Fungible Token (SFT) standard for on-chain representation. Its trivariate structure maps ASR attributes: Token ID = unique service order provenance; Slot = model class and geographic jurisdiction (enabling intra-slot aggregation); Value = residual service quota in ASU terms, supporting partial transfer.

4.2 AI Service Unit (ASU)

Definition 2 (AI Service Unit)**.** *An **AI Service Unit (ASU)** is a standardized, platform-neutral unit of account for AI service value. It converts heterogeneous platform-specific ASRs into a common metric:*

$$ASU_i = \frac{p_i^{\text{native}}}{Q_i^{\alpha} \cdot L_i^{-\beta} \cdot S_i^{\gamma} \cdot C_i^{-\delta}} \tag{1}$$

where p_i^{native} is platform i's billing price in USD per million equivalent tokens (eq-tokens; see Section 4.2 and equation (3) for modality-specific conversion factors); $Q_i \in [0,1]$ is the modality-appropriate quality index constructed via PCA of standardised benchmark scores (see Section 4.2 and equation (5)); $L_i \geq 1$ is the latency coefficient (higher = worse); $S_i \geq 0$ is the credit/reliability premium; $C_i \geq 1$ is the congestion factor (higher platform load reduces effective service quality); and $\alpha, \beta, \gamma, \delta > 0$ are

calibratable weights estimated separately for each modality class.

The ASU is analogous to the IMF's Special Drawing Rights (SDR): just as one SDR represents a standardized basket of purchasing power across currencies, one ASU represents "one unit of quality-adjusted AI purchasing power" regardless of which platform delivers it. Each platform's ASR is analogous to a national currency: it denominates purchasing power within a single platform's ecosystem, but its value relative to other platforms requires an exchange rate. The cross-platform exchange rate is:

$$E_{ij} = \frac{\text{ASU}_i}{\text{ASU}_j} \tag{2}$$

This is the direct analogue of a bilateral exchange rate between currencies i and j. When $E_{ij} \neq 1$, arbitrageurs profit by converting between platforms—the mechanism driving price convergence, precisely as triangular arbitrage enforces consistency among currency cross-rates in FX markets [Evans and Lyons, 2002].

The quality-credit-adjusted law of one price (Proposition 3) is the ASR analogue of purchasing power parity (PPP): after adjusting for quality differentials (analogous to price-level differences across countries), the "real exchange rate" between any two platforms should equal unity in equilibrium. Persistent deviations from this parity—analogous to CIP violations documented by Du et al. [2018]—signal arbitrage opportunities bounded by transaction costs and capital constraints.

Equivalent Token Conversion

The eq-token is a *latent accounting unit*, not a primitive natural quantity: it converts heterogeneous modality-specific billing units into a common denominator for cross-modal ASU comparison, budgeting, and clearing. In this role it is analogous to a numéraire currency—each modality retains its native billing unit for external quotation (images per request, seconds of video, etc.), while the eq-token serves only as the internal unit of account. The reference language model—GPT-4o output at \$15/M

tokens as of March 2026—anchors the conversion at unity.[1] For other modalities the conversion follows market-price parity:

$$p_i^{\text{native}} = \frac{p_i^{\text{raw}}}{\kappa_m} \times 10^6 \tag{3}$$

where p_i^{raw} is the raw per-unit billing price and κ_m is the modality-specific eq-token count per billing unit. Table 1 summarises the calibrated conversion factors.

Table 1: Modality-specific eq-token conversion factors κ_m, calibrated via market-price parity relative to GPT-4o output ($15/M tokens, March 2026). Point estimates reflect the reference provider; sensitivity ranges reflect cross-provider price variation within each modality.

Modality	Billing unit	κ_m (point)	Sensitivity range	Reference
Text (LLM output)	1M tokens	1,000,000	— (anchor)	GPT-4o
Image generation	1 image (1024^2)	$\approx$ 2,667	[1,500, 5,000]	DALL-E 3
Video generation	1 second (720p)	$\approx$ 2,667	[1,000, 6,000]	Sora
Speech recognition	1 second of audio	$\approx$ 7	[4, 15]	Whisper

For a mixed multimodal request containing T_{text} text tokens, n_{img} images, and d_{vid} seconds of video, the total eq-token count is:

$$T^{\text{eq}} = T_{\text{text}} + n_{\text{img}} \cdot \kappa_{\text{img}} + d_{\text{vid}} \cdot \kappa_{\text{vid}} \tag{4}$$

All conversion factors are anchored to the same reference price ($15/M output tokens), so the resulting p_i^{native} values are empirically close across modalities ($9–15/M eq-tokens in March 2026 data), consistent with competitive equilibrium in the underlying GPU market. The sensitivity ranges in Table 1 reflect cross-provider price variation within each modality: for instance, image generation prices range from $0.003 (Stable Diffusion XL via Replicate) to $0.08 (DALL-E 3 HD) per image, yielding $\kappa_{\text{img}} \in [1{,}500,\ 5{,}000]$ depending on the reference provider and output quality. The

[1] Ideally, the reference should be a basket of representative providers per modality (a hedonic pricing basket), so that κ_m does not depend on any single platform's pricing decision. We use a single reference for tractability and transparency; Section 8 (Limitation (a)) discusses generalisation.

point estimates $\kappa_{\text{img}} = \kappa_{\text{vid}} \approx 2{,}667$ arise because both DALL-E 3 and Sora were priced at approximately \$0.04 per output unit in March 2026—a calibration artefact, not a theoretical prediction. For video generation, no authoritative token-equivalent standard exists; κ_{vid} should be treated as an order-of-magnitude interval estimate whose range depends on output resolution, duration, and model generation.

Quality Index Construction

The quality index $Q_i \in [0, 1]$ is constructed via principal component analysis (PCA) of standardised benchmark scores, with benchmarks chosen to match the modality of service i:

- **Language models**: MMLU (language understanding), HumanEval (code generation), MATH (mathematical reasoning), GPQA (graduate-level science reasoning).
- **Image generation models**: Fréchet Inception Distance (FID, lower is better, sign-flipped before aggregation), CLIP Score (image–text alignment), and Inception Score (IS).
- **Video generation models**: Fréchet Video Distance (FVD, lower is better, sign-flipped), and CLIPSIM (temporal consistency).
- **Multimodal models**: MMBench (cross-modal understanding) and SEED-Bench (cross-modal generation).

PCA provides a data-driven weighting scheme that requires no prior assumptions about the relative importance of benchmarks. Formally, let $\mathbf{s}_i = (s_{i1}, \ldots, s_{iK})^\top$ be the vector of standardised scores for platform i across K benchmarks within a modality class. The quality index is:

$$Q_i = \frac{\mathbf{v}_1^\top \mathbf{s}_i - \min_j(\mathbf{v}_1^\top \mathbf{s}_j)}{\max_j(\mathbf{v}_1^\top \mathbf{s}_j) - \min_j(\mathbf{v}_1^\top \mathbf{s}_j)} \tag{5}$$

where $\mathbf{v}_1$ is the first principal component eigenvector, and the min–max normalisation maps the index to $[0, 1]$.

When domain expertise or regulatory requirements prescribe fixed benchmark weights—for instance, equal weighting mandated by a standards body—the weighted-average alternative

$$Q_i = \sum_k w_k \tilde{s}_{ik}, \qquad \sum_k w_k = 1 \tag{6}$$

where $\tilde{s}_{ik}$ is the normalised score on benchmark k, is equally valid. Equation (6) reduces to PCA when the within-modality benchmark covariance matrix is proportional to the identity. We retain PCA as the baseline because it is objective, reproducible, and consistent across modalities.

4.3 ASR Value Decomposition and Asset Classification

The value of an ASR decomposes into three components with distinct economic origins:

$$V_{\text{ASR}}(t) = \underbrace{V_{\text{use}}(t)}_{\text{service contract}} + \underbrace{V_{\text{option}}(t)}_{\text{transferability}} + \underbrace{V_{\text{hedge}}(t)}_{\text{risk management}} \tag{7}$$

where V_{use} is the present value of the contracted inference service (a prepaid service obligation); V_{option} is the option value of cross-platform transferability (strictly positive under the jump-diffusion quality process, since the ASR holder can sell the current ASR and purchase a superior alternative when Q_i jumps); and V_{hedge} is the hedging value from locking in future compute costs (under mean-variance preferences: $V_{\text{hedge}} = \frac{1}{2}\gamma_j \sigma_C^2 W_j$, where W_j is annual AI spend).

This decomposition connects directly to the welfare result (Proposition 4): V_{use} corresponds to $\Delta^{MS} + \Delta^{\text{search}}$ (trading efficiency gains); V_{option} corresponds to Δ^{lock} (lock-in elimination); and V_{hedge} corresponds to Δ^{hedge} (risk-sharing gains). It also clarifies asset classification: when V_{use} dominates (short-duration ASR, immediate redemption), the instrument resembles a prepaid service contract and is unlikely to

satisfy the "expectation of profit" prong of securities tests; when $V_{\text{option}} + V_{\text{hedge}}$ dominate (long-duration, actively traded ASR), the instrument functions as a financial derivative. The regulatory classification thus depends endogenously on contract maturity and trading frequency, not on an exogenous administrative label.

4.4 Pricing Paradigms: Closed-Source vs. Open-Source

The formation of p_i^{native} differs between closed-source and open-source models. For closed-source models, p_i^{native} is an administered price (posted by the model owner); the ASR market price is bounded above by this posted price and below by C-LMP. For open-source models, p_i^{native} is market-discovered through multi-provider competition, and compute providers can issue ASR directly (Proposition 5). Both paradigms are unified by the ASU formula: regardless of how p_i^{native} is formed, the quality-adjusted comparison is identical.

4.5 Cross-Platform Exchange Rates

The exchange rate E_{ij} between platforms i and j captures the quality-adjusted relative value of their service rights—the compute analogue of bilateral FX rates. Just as the EUR/USD rate reflects the relative purchasing power and macroeconomic conditions of two currency areas, $E_{\text{Anthropic/OpenAI}}$ reflects the relative quality-adjusted cost of inference across two AI platforms. Cross-platform arbitrage enforces consistency: if $E_{AB} \cdot E_{BC} \neq E_{AC}$, triangular arbitrage profits arise, just as in the FX triangle inequality [Menkhoff et al., 2012]. As a calibration illustration, consider two platforms with the following March 2026 parameters:

- Anthropic (Claude Sonnet 4.6): $p = \$9.00/\text{M}$ eq-tokens (blended at 1:1 input/output ratio), $Q = 74$, quality ratio $Q_A/Q_B \approx 1.03$, credit differential $S_A - S_B \approx 0.06$.

- OpenAI (GPT-5.2): $p = \$7.88$/M eq-tokens (blended at 1:1 input/output ratio), $Q = 72$.

After quality-credit adjustment: $E_{\text{Anthropic/OpenAI}} = (9.00/(1.03 \cdot e^{0.06}))/(7.88/1.0) \approx 1.04$. One unit of Anthropic service costs approximately 4% more than OpenAI after adjusting for quality and credit—substantially less than the 14% raw price differential, reflecting Anthropic's quality and reliability premia.

5 Theoretical Framework

5.1 Economic Environment

Agents. There are six classes of agents:

- *Compute providers* $c \in \mathcal{C} = \{1, \ldots, C\}$: Each operates compute nodes $n \in \mathcal{N}_c$ with GPU capacity $K_{c,n} > 0$, convex marginal cost $MC_c(q) = c_c^{\text{energy}}(q) + c_c^{\text{depr}}$ (energy plus hardware depreciation), and bandwidth link $B_{c,n}$. Compute providers may deploy open-source models from $\mathcal{M}^{\text{open}}$ at negligible marginal model cost (publicly available weights), in which case they can issue ASR directly. For closed-source models, compute providers serve as infrastructure suppliers to platforms.

- *Platforms* $i \in \mathcal{I} = \{1, \ldots, N\}$: Each owns or licenses a model M_i with quality $Q_i \in [0, 1]$. Platform i sources compute from providers: a fraction $\phi_i \in [0, 1]$ is self-owned and $(1 - \phi_i)$ is procured at C-LMP rates. Total marginal cost is $MC_i^{\text{total}} = \sum_c w_{ic} \cdot P_{c,n}^{\text{C-LMP}} + MC_i^{\text{model}}$, where w_{ic} is the sourcing share from provider c and MC_i^{model} is the model-specific cost (R&D amortisation, RLHF, fine-tuning). Vertically integrated platforms (e.g., Google/Gemini) have $\phi_i \approx 1$; pure API providers on cloud infrastructure have $\phi_i \approx 0$. Convex total cost: $c_i(q) = MC_i^{\text{total}} + C_i^{\text{cong}}(q/K_i)$.

- *Enterprises* $j \in \mathcal{J} = \{1, \ldots, M\}$: Each has CES production $Y_j = A_j[\alpha_K K_j^{(\sigma_Y-1)/\sigma_Y} + \alpha_L L_j^{(\sigma_Y-1)/\sigma_Y} + (1-\alpha_K-\alpha_L)x_j^{(\sigma_Y-1)/\sigma_Y}]^{\sigma_Y/(\sigma_Y-1)}$[2] with compute x_j, productivity A_j, and risk aversion γ_j over compute cost volatility.

- *Market makers* $m \in \mathcal{M}$: Post bid-ask quotes on the CLOB following Avellaneda and Stoikov [2008] optimal control.

- *Arbitrageurs* $a \in \mathcal{A}$: Monitor cross-platform price differentials with finite capital $W_a > 0$.

- *Liquidity providers* $l \in \mathcal{L}$: Deposit capital into AMM pools, earning fees and bearing impermanent loss.

Information structure. Each platform privately observes its marginal cost c_i and capacity K_i; each enterprise privately observes its productivity A_j. Quality Q_i is publicly observed through a modality-aware benchmark oracle: a centralised aggregator collects scores from domain-specific third-party evaluators—MMLU, HumanEval, MATH, and GPQA for language inference; FID, CLIP Score, and Inception Score for image generation; FVD and CLIPSIM for video generation; MMBench and SEED-Bench for multimodal models—and maps them to the common $[0, 1]$ scale via PCA of standardized scores. Congestion C_i^{cong} is partially observed: platforms know utilization; buyers observe latency as a noisy signal. This creates bilateral information asymmetry in the sense of Myerson and Satterthwaite [1983].

[2]We use α_K, α_L (capital and labor shares) and σ_Y (production elasticity of substitution) to distinguish these from the ASU hedonic parameters $\alpha, \beta, \gamma, \delta$ (Definition 2) and the AMM price volatility σ. The constraint $\alpha_K, \alpha_L > 0$, $\alpha_K + \alpha_L < 1$ ensures a positive compute share $(1 - \alpha_K - \alpha_L) > 0$.

5.2 Benchmark: Bilateral Market Inefficiency

Proposition 1 (Bilateral Inefficiency)**.** *In the bilateral benchmark with private information about costs and valuations, search frictions $\lambda^{\text{search}} < \infty$, and platform lock-in, the equilibrium allocation is generically Pareto-inefficient. Specifically:*

(i) ***Price dispersion:*** *For any two platforms i, k with identical quality $Q_i = Q_k$, equilibrium prices satisfy $p_{ij}^{\text{bilateral}} \neq p_{kj}^{\text{bilateral}}$ for almost all realizations of (c_i, c_k, v_j).*

(ii) ***Allocative inefficiency:*** *The expected total surplus under bilateral trading is strictly less than first-best:*

$$\mathbb{E}[TS^{\text{bilateral}}] < TS^{FB} - \underbrace{\Delta^{MS}}_{\textit{Myerson-Satterthwaite}} - \underbrace{\Delta^{\text{search}}}_{\textit{search frictions}} - \underbrace{\Delta^{\text{lock}}}_{\textit{lock-in waste}} \tag{8}$$

(iii) ***Missing risk sharing:*** *Enterprises bear full compute cost volatility $Var(\sum_i p_i x_{ij})$ with no hedging instruments.*

Proof sketch. Part (i) follows from Nash bargaining: when enterprises meet platforms at Poisson rate λ^{search}, the negotiated price $p_{ij} = \eta c_i + (1-\eta) v_j$ differs across platforms whenever costs differ. Part (ii) decomposes the welfare loss: Δ^{MS} from Myerson and Satterthwaite [1983] (bilateral trading cannot achieve efficiency under private information without subsidies); Δ^{search} from Duffie et al. [2005] (search friction welfare loss is positive for finite search rates); Δ^{lock} from stranded-value waste $W^{\text{lock}} = \sum_j \sum_i \phi_{ij} \max\{0, v_j^*(t) - v_{ij}(t)\} x_{ij}$. Part (iii) is immediate from market incompleteness. □ □

Economic intuition. Proposition 1 captures the current state of the AI compute market. Enterprises negotiate bilaterally (or accept posted prices), cannot transfer pre-purchased capacity, face massive price dispersion [Fradkin, 2025], and have no

hedging instruments [Benchmarkit, 2025]. The ASR market is designed to eliminate all three sources of inefficiency.

5.3 ASR Market Design: C-LMP

Following Schweppe et al. [1988], Wilson [2002], and Cramton [2017], we derive Compute Locational Marginal Pricing from a social planner's cost minimization.

Setup. Node (i, n) has capacity $K_{i,n}$, convex marginal cost $c_{i,n}(q)$, and bandwidth link $B_{i,n}$. Aggregate demand $D(t)$ must be served.

Social planner's problem:

$$\min_{\{q_{i,n}\}} \sum_{i} \sum_{n \in \mathcal{N}_i} c_{i,n}(q_{i,n}) \quad \text{s.t.} \quad \sum_{i,n} q_{i,n} = D(t), \quad q_{i,n} \leq K_{i,n}, \quad q_{i,n} \leq B_{i,n}, \quad q_{i,n} \geq 0 \tag{9}$$

Proposition 2 (C-LMP Decomposition)**.** *The optimal dispatch satisfies the KKT conditions. The shadow price at each node decomposes as:*

$$P_{i,n}(t) = \underbrace{\lambda(t)}_{\text{system marginal cost}} + \underbrace{\mu_{i,n}(t)}_{\text{capacity congestion}} + \underbrace{\nu_{i,n}(t)}_{\text{bandwidth premium}} \tag{10}$$

where $\lambda(t) \geq 0$ is the Lagrange multiplier on demand satisfaction; $\mu_{i,n}(t) \geq 0$ is the congestion premium ($\mu > 0$ iff node (i, n) is at capacity); and $\nu_{i,n}(t) \geq 0$ is the bandwidth premium ($\nu > 0$ iff bandwidth is saturated). Moreover, C-LMP dispatch is welfare-efficient.

Proof. We reformulate as a welfare-maximisation problem (see the Online Appendix for the full derivation). The social planner maximises net surplus $U(\sum_{i,n} q_{i,n}) - \sum_{i,n} c_{i,n}(q_{i,n})$ subject to capacity and bandwidth constraints. The KKT first-order condition for each active node yields $c'_{i,n}(q^*_{i,n}) = \lambda - \mu_{i,n} - \nu_{i,n}$, where $\lambda \equiv U'(D)$ is the demand shadow price. The nodal buyer price therefore decomposes as $P_{i,n} =$

$\lambda + \mu_{i,n} + \nu_{i,n}$: system marginal cost plus congestion and bandwidth rents. Welfare efficiency follows from convexity of $c_{i,n}(\cdot)$. □ □

Economic intuition. C-LMP provides AI compute with the same function LMP provides electricity: spatially differentiated price signals reflecting systemwide scarcity (λ) and local constraints (μ, ν). When a data center is congested, its C-LMP rises, signaling enterprises to route requests elsewhere. This is precisely the mechanism generating $3.2–4.0 billion in annual efficiency savings at PJM [PJM, 2022, Cramton, 2017].

Mapping to electricity LMP. System energy cost λ maps to system inference marginal cost; transmission congestion μ maps to GPU/TPU capacity congestion; network losses ν map to bandwidth/latency premiums and data-residency surcharges.

Dynamic analogy: Dornbusch overshooting. When a platform experiences a discrete quality jump (e.g., a major model release), the ASR price overshoots before gradual convergence—analogous to exchange rate overshooting in Dornbusch [1976]. In Dornbusch's model, the exchange rate overshoots because goods prices are sticky while the asset price (exchange rate) is flexible. The compute analogue: enterprise procurement contract prices are sticky (bilateral agreements with fixed terms), while C-LMP adjusts instantaneously as a flexible market price. The resulting temporary ASR mispricing—analogous to Dornbusch's real exchange rate overshooting—is corrected by arbitrageurs as contract prices gradually reset.

Remark 1 (Structural Relationship between ASU and C-LMP)**.** *The ASU (Definition 2) and C-LMP (Proposition 2) are not independent constructs but two views of the same underlying market—the ASU from the demand side (quality-adjusted purchasing power) and C-LMP from the supply side (socially optimal marginal cost at the compute-provider level). The C-LMP decomposition has a precise institutional locus: λ reflects systemwide energy and depreciation costs determined by the marginal compute provider dispatched; $\mu_{c,n}$ reflects GPU capacity scarcity at provider c's node n;*

$\nu_{c,n}$ *reflects bandwidth constraints on the network path to that node. The connection runs through a three-layer pricing structure:*

(i) ***Cost floor (C-LMP):*** $P_{c,n}^{\text{C-LMP}} = \lambda + \mu_{c,n} + \nu_{c,n}$ *is the marginal social cost of serving one additional unit at node* (c, n)*. In competitive equilibrium, no provider can persistently price below this level.*

(ii) ***Strategic pricing:*** *Platform* i *sets* $p_i^{\text{native}} = m_i \cdot MC_i^{total}$*, where* $MC_i^{total} = \sum_c w_{ic} P_{c,n}^{\text{C-LMP}} + MC_i^{model}$ *and* $m_i = \varepsilon_i/(\varepsilon_i - 1) \geq 1$ *is the markup determined by demand elasticity and quality differentiation.*

(iii) ***Quality-adjusted price (ASU):*** *Substituting into Definition 2:*

$$ASU_i = \frac{m_i \cdot MC_i^{total}}{Q_i^{\alpha} \cdot L_i^{-\beta} \cdot S_i^{\gamma} \cdot C_i^{-\delta}} \tag{11}$$

Arbitrage (Proposition 3) drives $ASU_i \to \Psi$ *for all* i*. This pins down the equilibrium markup:* $m_i^* = \Psi \cdot Q_i^{\alpha} L_i^{-\beta} S_i^{\gamma} C_i^{-\delta} / MC_i^{total}$*. Higher-quality platforms command higher markups—consistent with the hedonic equilibrium of Proposition 3—while C-LMP provides the cost anchor that prevents the market from becoming untethered from physical resource scarcity. C-LMP is thus both a* dispatch signal *(guiding request routing to uncongested nodes in real time) and a* reference price *for primary-market ASR issuance, but not a prediction of* p_i^{native} *itself.*

5.4 Equilibrium and Welfare

5.4.1 No-Arbitrage Equilibrium

Proposition 3 (No-Arbitrage Equilibrium with Capital Constraints)**.** *Let: (A1)* $N \geq 2$ *platforms with* $V_i(t) > 0$*; (A2) continuous trading, no short-sale constraints; (A3') arbitrageurs have finite capital* $W_a > 0$ *and face margin requirements; (A4) symmetric transaction costs* $c_{ij} \geq 0$*. For tractability, we characterize the no-arbitrage band using*

the two most readily observable attributes—quality Q_i and credit S_i—abstracting from latency L_i and congestion C_i which are priced in the full ASU (Definition 2) but require oracle data not yet available in a nascent market. Define the quality-credit-adjusted spread:[3]

$$\Delta_{ij}(t) = \frac{P_i(t)}{Q_i(t) \cdot e^{S_i(t)}} - \frac{P_j(t)}{Q_j(t) \cdot e^{S_j(t)}} \tag{12}$$

Then:

(i) The set of equilibrium prices is the no-arbitrage band $\mathcal{B} = \{P : |\Delta_{ij}(t)| \leq c_{ij} + \psi_{ij}(W) \; \forall \, i, j\}$, where $\psi_{ij}(W) \geq 0$ is the limits-to-arbitrage wedge depending on aggregate arbitrageur capital W.

(ii) As $W \to \infty$, $\psi_{ij}(W) \to 0$ and the equilibrium reduces to $|\Delta_{ij}| \leq c_{ij}$.

(iii) Under capital constraints, mispricings can persist temporarily with expected duration $\tau^* = O(c_{ij}/(\lambda^{\text{arb}} W))$.

(iv) Starting from any initial price vector, the arbitrage process converges to $\mathcal{B}$ almost surely, provided market depth is positive.

Proof sketch. Part (i): Following Shleifer and Vishny [1997], $\psi_{ij}(W)$ reflects the risk premium demanded by capital-constrained arbitrageurs. Part (ii): As $W \to \infty$, arbitrageurs become risk-neutral with respect to mispricing risk. Part (iii): Convergence time is bounded by the law of large numbers applied to arbitrageur order flow. Part (iv): A Lyapunov function $\Phi(P) = \sum_{i<j} \max\{0, |\Delta_{ij}| - c_{ij} - \psi_{ij}\}^2$ satisfies $\mathbb{E}[d\Phi/dt] \leq 0$ with strict inequality when $\Phi > 0$. □ □

Economic intuition. This strengthened result acknowledges real capital constraints [Shleifer and Vishny, 1997]. The no-arbitrage band $\mathcal{B}$ is the compute analogue

[3]The full four-factor version is $\Delta_{ij}^{\text{full}} = P_i/(Q_i^\alpha L_i^{-\beta} S_i^\gamma C_i^{-\delta}) - P_j/(Q_j^\alpha L_j^{-\beta} S_j^\gamma C_j^{-\delta})$. Equation (12) simplifies this by setting $\alpha = 1$, $\beta = \delta = 0$, and reparametrising the credit term as e^{S_i} (convenient for the exponential pricing environment of the no-arbitrage condition).

of the covered interest parity (CIP) band in FX markets: just as CIP violations persist when dealer balance sheets are constrained [Du et al., 2018], ASR mispricings can persist when arbitrageur capital is limited. The ASR exchange reduces ψ_{ij} by lowering margin requirements, providing netting services, and facilitating cross-platform settlement—analogous to how CLS Bank reduces FX settlement risk.

5.4.2 Welfare Improvement

Proposition 4 (Welfare Improvement)**.** *The ASR secondary market Pareto-improves over bilateral trading:*

$$TS^{ASR} - TS^{\text{bilateral}} = \underbrace{\Delta^{MS}}_{\textit{Myerson-Satterthwaite}} + \underbrace{\Delta^{\text{search}}}_{\textit{search}} + \underbrace{\Delta^{\text{lock}}}_{\textit{lock-in}} + \underbrace{\Delta^{\text{hedge}}}_{\textit{risk sharing}} > 0 \tag{13}$$

where each component is non-negative and at least one is strictly positive.

Proof sketch. $\Delta^{MS} > 0$: The CLOB aggregates information from all traders [Kyle, 1985], eliminating the Myerson-Satterthwaite deadweight loss. $\Delta^{\text{search}} > 0$: The exchange provides instantaneous multilateral matching; following Duffie et al. [2005], the welfare gain is $\Delta^{\text{search}} = [(\bar{v} - \underline{c})/(r + \lambda^{\text{search}})] \cdot \lambda^{\text{search}} \cdot M$. $\Delta^{\text{lock}} > 0$: ASR transferability eliminates stranded-value waste; under stochastic quality evolution, the option value of transferability is strictly positive. $\Delta^{\text{hedge}} > 0$: Under mean-variance preferences, the hedging value is $\frac{1}{2}\bar{\gamma}\sigma_C^2 \sum_j W_j^2 > 0$, where W_j denotes enterprise j's annual AI spend (distinct from W_a, the arbitrageur capital in Proposition 3). The Pareto claim follows from voluntary participation. □ □

Caveat. Following Philippon [2015], the welfare gain must be net of intermediation costs. We address this in Section 9.4.

5.4.3 Vertical Disintermediation

The welfare analysis above treats platforms as monolithic entities. In practice, the compute supply chain is vertically separated: compute providers own hardware and platforms add model-specific value (Section 5.1). When the ASR market allows compute providers to issue ASR directly for open-source models, an additional welfare channel opens.

Proposition 5 (Vertical Disintermediation and Anchor Participation)**.** *In a vertically separated market where compute providers can directly issue ASR for open-source models:*

(i) ***Markup reduction:*** *Open-source ASR carries a single provider markup m_c, while closed-source ASR carries a compounded markup $m_c^{\text{hw}} \cdot m_i^{\text{sw}}$ (hardware and software layers). For quality-comparable models, $ASU^{\text{open}} < ASU^{\text{closed}}$, with the difference reflecting the model brand premium plus vertical intermediation rent.*

(ii) ***Competitive pressure on closed-source markups:*** *The ASR market makes $\Delta ASU \equiv ASU^{\text{closed}} - ASU^{\text{open}}$ observable. There exists a critical quality gap ΔQ^* such that enterprises switch from closed-source to open-source ASR whenever $Q^{\text{closed}} - Q^{\text{open}} < \Delta Q^*$, where $\partial \Delta Q^* / \partial m_i^{\text{sw}} > 0$: higher closed-source markups lower the quality threshold for switching.*

(iii) ***Double-marginalisation welfare gain:*** *Direct ASR issuance eliminates the vertical double-marginalisation loss:*

$$\Delta^{\text{vert}} = (m^{\text{hw}} \cdot m^{\text{sw}} - m^{\text{hw}}) \cdot MC \cdot D^{\text{open}} > 0 \tag{14}$$

where D^{open} is demand served by open-source models.

(iv) ***Anchor participation:*** *Compute providers with idle capacity ($u_c < 1$) earn strictly positive incremental profit from ASR market participation: $\Delta\pi_c = (P_{c,n}^{C\text{-}LMP} -$*

$MC_c') \cdot (K_{c,n} - q_{c,n}) > 0$ *whenever the market price exceeds their marginal cost. If the aggregate idle capacity of willing providers exceeds* θ^*/η *(where* θ^* *is the critical mass from Proposition 7 and* η *is the provider-to-total participation ratio), compute providers alone can push adoption past* θ^**, breaking the low-equilibrium coordination trap without external subsidies.*

Proof sketch. Part (i): Under single-layer pricing, the open-source ASR price is $p^{\text{open}} = m_c \cdot MC_c + MC^{\text{deploy}}$; under double-layer pricing, $p^{\text{closed}} = m_i^{\text{sw}} \cdot (m_c^{\text{hw}} \cdot MC_c + MC_i^{\text{model}})$. Since $m_i^{\text{sw}} > 1$ and $MC_i^{\text{model}} > MC^{\text{deploy}}$ (proprietary R&D exceeds open-source deployment cost), $p^{\text{closed}} > p^{\text{open}}$ for quality-comparable services. The ASU ordering follows from identical quality adjustments applied to both. Part (ii): The enterprise indifference condition $\theta_j Q^{\text{closed}}/\text{ASU}^{\text{closed}} = \theta_j Q^{\text{open}}/\text{ASU}^{\text{open}}$ yields $\Delta Q^* = Q^{\text{open}} \cdot (\text{ASU}^{\text{closed}}/\text{ASU}^{\text{open}} - 1)$, which is increasing in m_i^{sw} since $\text{ASU}^{\text{closed}}$ is increasing in m_i^{sw}. Part (iii): Standard double-marginalisation result [Tirole, 1988]. Part (iv): The idle-capacity condition $u_c < 1$ implies $q_{c,n} < K_{c,n}$; the incremental profit is the margin ($P^{\text{C-LMP}} - MC'$) times the idle quantity. The critical-mass condition follows from Proposition 7 by substituting the provider participation mass for the general adoption rate. □ □

Economic intuition. This result addresses the feasibility concern raised by Proposition 7: the ASR market need not rely on platform goodwill or government subsidies to achieve critical mass. China's Federal Reserve-documented 80% idle compute capacity [Federal Reserve Board, 2025] represents a large pool of compute providers whose opportunity cost of participation is effectively zero. These providers serve as *anchor participants*—analogous to anchor tenants in a shopping mall—whose entry creates sufficient liquidity to attract platforms and enterprises. The open-source model ecosystem (Llama, Mistral, Stable Diffusion) provides the service layer that makes compute-provider-issued ASR directly usable by enterprises.

5.4.4 Price Informativeness

Remark 2 (Price Informativeness)**.** *ASR market prices aggregate dispersed information more efficiently than bilateral prices. Define informativeness as* $I(P) = [Var(V_i|P_i)]^{-1}$*. Then:*

(i) Under the CLOB with informed market makers: $I(P^{ASR}) = \tau_\varepsilon + n^2/\tau_u$*.*

(ii) Under bilateral trading: $I(P^{\text{bilateral}}) = \tau_\varepsilon + 1/\tau_u^{\text{bilateral}}$*, where* $\tau_u^{\text{bilateral}} \gg \tau_u$*.*

(iii) $I(P^{ASR}) > I(P^{\text{bilateral}})$ *whenever* $n \geq 2$*.*

(iv) The C-LMP decomposition further enhances informativeness: the joint informativeness of $(\lambda, \mu_{i,n}, \nu_{i,n})$ *exceeds that of a single aggregate price.*

Proof sketch. Parts (i)–(iii) follow from the Grossman and Stiglitz [1980]/Kyle [1985] framework: in a rational expectations equilibrium, the market price aggregates n independent private signals, so informativeness scales as n^2/τ_u. Part (iv): The C-LMP vector is a sufficient statistic for the state $(D(t), K_{i,n}, B_{i,n})$ given convexity of the dispatch problem. □ □

Economic intuition. Static bilateral prices reveal almost nothing about real-time compute scarcity. The ASR market price aggregates private information from all participants—market makers observing order flow, arbitrageurs monitoring cross-platform spreads, informed enterprises observing productivity shocks. The C-LMP decomposition adds spatial informativeness: when μ spikes at a node, it signals local capacity scarcity specifically, not just high prices in general. This parallels the informational advantage LMP provides in electricity markets [Wilson, 2002].

5.4.5 Liquidity Complementarity

Proposition 6 (Hybrid CLOB-AMM Dominance)**.** *Under heterogeneous traders (fast/informed and slow/uninformed) and heterogeneous assets (thick and thin ASR pairs), the hybrid*

CLOB-AMM architecture weakly dominates either mechanism alone:

(a) Higher welfare than CLOB-only: The AMM serves thin pairs where CLOB market makers would not post competitive quotes, providing "liquidity of last resort" at bounded cost $\mathrm{IL}(r) \approx -\frac{1}{8}\sigma^2 T$ *[Milionis et al., 2022].*

(b) Higher welfare than AMM-only: The CLOB serves fast traders requiring price-priority execution. Adding the CLOB improves price discovery because informed traders self-select, reducing AMM adverse selection [Zhu, 2014].

(c) Cross-venue arbitrageurs ensure $|P^{\mathrm{CLOB}} - P^{\mathrm{AMM}}| \leq c_{\mathrm{cross}}$.

Proof sketch. Part (a): For thin pairs, Glosten [1994] adverse selection costs exceed spread revenue; the AMM provides mechanical liquidity at LP cost. Following Lehar and Parlour [2025], Proposition 3, AMM dominates an empty order book. Part (b): Adding a CLOB sorts informed traders out of the AMM [Zhu, 2014] (Theorem 1). Part (c): When the same ASR trades simultaneously on the CLOB and the AMM, a cross-venue arbitrageur simultaneously buys on the cheaper venue and sells on the more expensive one. The argument mirrors Proposition 3 applied to venues rather than platforms: the no-arbitrage band width equals the round-trip cost c_{cross} of a cross-venue trade. □ □

Remark. The CLOB component is subject to the high-frequency trading arms race documented by Budish et al. [2015] and Aquilina et al. [2022]. The AMM layer partially mitigates this by providing a speed-insensitive alternative, but a complete solution may require frequent batch auctions within the CLOB—a refinement for future work.

5.5 Platform Participation

Proposition 7 (Critical Mass and Multiple Equilibria)**.** *The platform participation game has the following structure. Let* $\Delta\pi_i(\theta) = f_{\mathrm{provider}} \cdot V(\theta) + N(\theta) \cdot S_i - k_i$ *be the*

payoff from joining at adoption rate θ, where $V(\theta) = V_0\theta^\eta$ ($\eta > 1$) and $N(\theta) = N_0\theta$.

(i) There exists a critical mass threshold θ_i^* solving $\Delta\pi_i(\theta_i^*) = 0$. For $\theta > \theta_i^*$, participation strictly dominates.

(ii) The game has at least two Nash equilibria: $\theta_{\text{low}}^{NE} = 0$ and $\theta_{\text{high}}^{NE} \in (\theta^*, 1]$.

(iii) The socially optimal adoption $\theta^{**} > \theta_{\text{high}}^{NE}$, because individual platforms do not internalize positive externalities on enterprises (lower search costs, better price discovery) and on other platforms (higher liquidity).

Proof sketch. Part (i): $\Delta\pi_i(\theta)$ is continuous, increasing, with $\Delta\pi_i(0) = -k_i < 0$ and $\Delta\pi_i(1) > 0$; existence follows by the intermediate value theorem. Part (ii): The low equilibrium is trivial; the high equilibrium follows from the fixed-point theorem on $\Theta(\theta) =$ fraction with $\Delta\pi_i(\theta) > 0$. Part (iii): The social planner maximizes total surplus including enterprise gains, which are external to each platform's private payoff.

□ □

Economic intuition. This is the classic coordination problem with network externalities [Rochet and Tirole, 2003, Weyl, 2010]. No platform wants to join first (since $V(0) = 0$), but once enough join, participation is self-reinforcing. Resolution requires subsidies, a regulatory sandbox, or an anchor platform (e.g., Alibaba's ATH unit, established March 2026).

Remark 3 (ASU Hedonic Foundation)**.** *Under Rosen's hedonic equilibrium [Rosen, 1974], the ASU (Definition 2) is the inverse of the equilibrium price function: quality-adjusted willingness to pay in a market with differentiated products. The log-linear functional form is standard and supported empirically (Section 7).*

6 Market Architecture Design

6.1 Architecture Overview

The architecture has four layers. *Layer 1 (Primary Market):* ASR issuance by platforms and compute providers, with C-LMP as the reference price. *Layer 2 (Secondary Market):* a hybrid CLOB (Track 1, for institutional flow with price-time priority) and concentrated-liquidity AMM (Track 2, for long-tail model pairs with low natural volume). *Layer 3 (Derivatives):* ASR futures, options, and compute swaps referencing C-LMP, enabling hedging and forward price discovery. *Layer 4 (Clearing):* central counterparty with novation and variation margin, plus on-chain settlement via ERC-3525 semi-fungible tokens encoding quality attributes.

6.2 Compute Provider Participation Channels

Compute providers participate in the ASR market through three channels. *(1) Primary-market ASR issuance:* providers deploying open-source models issue ASR directly, priced at C-LMP plus deployment margin. This bypasses the platform intermediary, eliminating one layer of markup (Proposition 5). *(2) Real-time dispatch:* when ASR holders redeem (use the service), the dispatch optimiser routes requests to the lowest-cost available node via C-LMP. Providers earn $P_{c,n}^{\text{C-LMP}}$ per unit dispatched—this is their revenue signal. *(3) Capacity forward market:* providers sell forward capacity commitments on the derivatives layer, providing investment signals for new GPU capacity and allowing enterprises to lock in future compute costs. This is the compute analogue of PJM's capacity market.

6.3 CLOB (Track 1)

The CLOB serves institutional investors and professional market makers. Market makers post bid-ask quotes following Avellaneda and Stoikov [2008] optimal stochastic

control:

$$r_a^* = s + \gamma_{\text{mm}}\sigma^2(T-t)(1-2q\delta) + \frac{1}{\gamma_{\text{mm}}}\ln\left(1+\frac{\gamma_{\text{mm}}}{\kappa}\right) \tag{15}$$

where s is the mid-price, γ_{mm} is the market maker's risk-aversion coefficient (distinct from the ASU hedonic weight γ in Definition 2), σ is ASR price volatility, $q \in [-1, 1]$ is normalized inventory, δ is the half-spread parameter, and κ is order-arrival intensity. The bid is symmetric: $r_b^* = 2s - r_a^*$ [Avellaneda and Stoikov, 2008, Cartea et al., 2015].

6.4 AMM (Track 2)

The AMM serves retail and long-tail ASR pairs via the constant-product formula $x \cdot y = k$. LP impermanent loss and expected profit:

$$\text{IL}(r) = \frac{2\sqrt{r}}{1+r} - 1, \quad r = P_t/P_0 \tag{16}$$

$$\Pi_{\text{LP}} = f \cdot V - |\text{IL}(r)| \cdot V_0 \tag{17}$$

where f is the fee rate and V_0 is initial LP position value [Milionis et al., 2022]. The expected IL under geometric Brownian motion is $\mathbb{E}[\text{IL}] \approx -\frac{1}{8}\sigma^2 T$ (first-order Taylor expansion), bounded and predictable.

Derivatives, clearing, and settlement. The architecture includes a derivatives layer (ASR futures, call options, and compute swaps referencing C-LMP), a central clearinghouse with novation and variation margin (calibrated to electricity market standards), and on-chain settlement via ERC-3525 semi-fungible tokens that encode quality attributes in slot metadata. Full specifications are in the Online Appendix.

6.5 Institutional Interface Protocol (IIP)

IIP is a three-layer middleware enabling ecosystem integration without requiring platforms to cede control:

(1) Technical Abstraction (ASU mapping). The conversion operator Φ maps the full billing context: $\text{ASU} = \Phi(\text{Input}, \text{Output}, \text{Cache}_{\text{hit}}, \text{Priority}, \text{Locality})$, accounting for billing heterogeneity.

(2) Market Incentives. Platforms earn: capacity pre-sale revenue (ASR forward contracts); brand premium monetization (a higher S_i lowers a platform's ASU, making its services appear cheaper on a quality-adjusted basis and attracting greater demand—the mechanism by which reliability premia are rewarded through price discovery); secondary royalties. Cross-platform arbitrage closes: $P_A > P_B \cdot E_{AB} + c \Rightarrow$ convergence.

(3) Regulatory Adaptation. Utility ASR: short-duration, use-value only, designed to fall outside securities classification. Trading ASR: proposed as compute futures under commodity regulation. Geolocation tags enforce data-residency compliance at the redemption layer.

Implementation roadmap. Phase 1: Shadow Ledger (internal ASU budgeting). Phase 2: Residual Market (unused quota transfers). Phase 3: Full Financialization (open ASR derivatives).

7 Empirical Validation of the Hedonic Specification

We test the hedonic pricing model (Definition 2, Proposition 3) against cross-sectional data from Artificial Analysis, an independent platform that benchmarks AI models under standardised conditions.

7.1 Data

The sample comprises all 42 non-free-tier LLMs listed on the Artificial Analysis leaderboard as of May 2026, spanning 13 providers (OpenAI, Anthropic, Google, xAI, DeepSeek, Alibaba/Qwen, Xiaomi/MiMo, Z AI/GLM, Kimi, MiniMax, Tencent, Mistral, StepFun). For each model we observe: (i) the Artificial Analysis Intelligence Index Q_i (a composite of 10 independently measured evaluations including GPQA Diamond, Humanity's Last Exam, SciCode, and Terminal-Bench Hard); (ii) blended price p_i in USD per million tokens (7:2:1 cache-hit:input:output ratio); (iii) median output speed in tokens per second; and (iv) open-weights status (binary: model weights publicly available vs. proprietary). We also code a region indicator (China-based vs. US/EU provider). Free-tier models are excluded because a zero price cannot enter the log specification. The dataset constitutes a near-census of frontier and mid-tier models in the current market.

7.2 Specification

We estimate progressively richer versions of the hedonic regression implied by equation (1):

$$\ln p_i = \beta_0 + \alpha \ln Q_i + \beta_S \ln S_i^{\text{speed}} + \gamma_{\text{open}} \cdot \mathbf{1}_{\text{open},i} + \gamma_{\text{China}} \cdot \mathbf{1}_{\text{China},i} + \varepsilon_i \tag{18}$$

where S_i^{speed} is output speed in tokens/second (a proxy for the inverse of the latency factor L_i in the ASU formula, as higher throughput implies lower effective delay per token). We use heteroskedasticity-robust (HC1) standard errors throughout; a Breusch-Pagan test rejects homoskedasticity (LM $= 10.08$, $p = 0.018$), confirming this choice. A Ramsey RESET test does not reject the log-linear functional form ($F = 2.40$, $p = 0.105$), supporting the multiplicative structure of the ASU definition.

7.3 Results

Table 2 reports six specifications.

Table 2: Hedonic regression of AI inference pricing. Dependent variable: ln(price per M tokens). Sample: 42 non-free-tier LLMs from the Artificial Analysis leaderboard (May 2026). HC1 robust standard errors in parentheses; column (6) uses creator-clustered standard errors.

	(1)	(2)	(3)	(4)		
$\ln(Q)$	3.289^{***}	0.690	1.137	1.526^{*}	1.893	0.690
	(1.188)	(0.806)	(0.740)	(0.809)	(1.242)	(0.665)
ln(Speed)	-0.736^{***}	-1.065^{***}	-1.200^{***}	-1.205^{***}	-1.090^{***}	-1.065^{***}
	(0.253)	(0.234)	(0.219)	(0.218)	(0.226)	(0.275)
Open Weights		-1.541^{***}	0.594^{*}	0.620^{*}	8.032	-1.541^{***}
		(0.232)	(0.328)	(0.336)	(7.551)	(0.228)
China-based			-2.197^{***}	-2.211^{***}		
			(0.303)	(0.305)		
Constant	-8.823^{*}	3.542	2.383	0.957	-1.059	3.542^{*}
	(4.528)	(2.990)	(2.989)	(3.120)	(5.106)	(2.088)
Adj. R^2	0.255	0.625	0.703	0.699	0.633	0.625
N	42	42	42	42	42	42
F	6.7	19.0	22.6	18.7	22.8	15.9
SE type	HC1	HC1	HC1	HC1	HC1	Cluster

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. VIF diagnostics for specification (2): $\ln Q = 1.28$, $\ln \text{Speed} = 1.11$, Open $= 1.23$; no multicollinearity concern.

Figure 1 visualises the two main sources of price variation and the progression of model fit.

Four findings emerge:

(1) The log-linear hedonic form fits the data. Specification (2) achieves Adj. $R^2 = 0.625$ with three regressors—quality, speed, and open-weights status explain nearly two-thirds of cross-sectional price variation. The RESET non-rejection supports the multiplicative functional form underlying the ASU definition.

(2) Output speed commands a large negative premium. The speed elasticity is approximately -1.1 across all specifications (robust to controls and clustering): a 10% increase in output speed is associated with an 11% lower price. This is consistent with

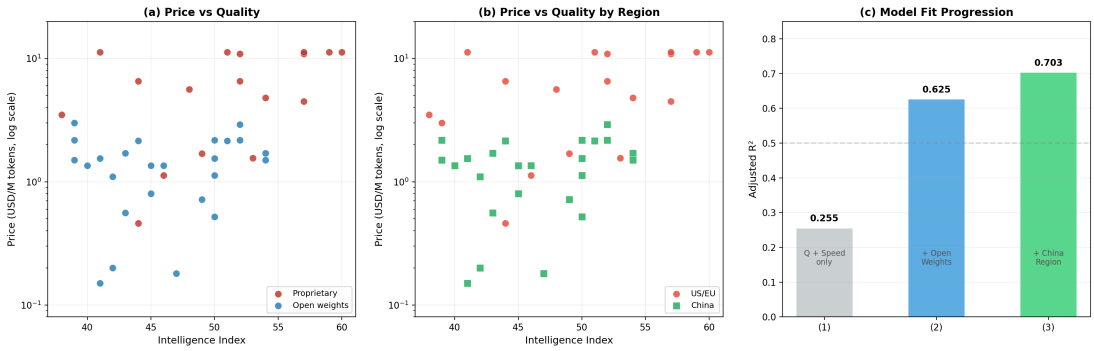


Figure 1: Cross-section of AI inference prices (May 2026, $N = 42$ models). Panel (a): price vs. intelligence by model type (proprietary vs. open-weights). Panel (b): price vs. intelligence by provider region (US/EU vs. China). Panel (c): adjusted R^2 progression as controls are added. Data source: Artificial Analysis leaderboard.

commodity competition on hardware efficiency—faster models tend to be cheaper because speed reflects optimised inference infrastructure, not higher quality.

(3) The raw open-weights discount decomposes into a geographic and a business-model component. In specification (2), the open-weights coefficient is -1.541 ($p < 0.001$), implying a 79% discount at equal quality and speed ($\exp(-1.541) - 1 = -0.786$). However, adding the China-based indicator in specification (3) reveals that this aggregate discount is largely driven by geographic cost heterogeneity: the China coefficient is -2.197 ($p < 0.001$, implying an 89% discount), while the open-weights coefficient flips to a weakly positive $+0.594$ ($p < 0.1$). Most Chinese providers offer open-weights models, so the raw open-weights effect in (2) conflates two distinct economic channels: (a) lower factor costs in China (energy, labour, possibly state subsidies), captured by the geographic dummy; and (b) the vertical intermediation structure, captured by the residual open-weights effect. The geographic channel is an order of magnitude larger in this cross-section.

(4) Quality is absorbed by structural variables. Once open-weights status is included, the quality coefficient becomes insignificant (specification 2: $\hat{\alpha} = 0.69$, $p = 0.49$). This is not because quality is irrelevant to pricing, but because the frontier models ($Q > 55$) are almost exclusively proprietary: the structural dummy captures

the quality-correlated pricing tier. This pattern is consistent with the market failure diagnosed in Section efsec:background—current prices are driven primarily by provider identity ("who you are") rather than service quality ("how good you are")—and directly motivates the ASU's quality adjustment.

8 Computational Validation

We complement the cross-sectional empirical evidence (Section 7) with an agent-based simulation: 5 platforms, 200 enterprise agents (heterogeneous in quality sensitivity θ_j), 50 arbitrageurs (capital-constrained), and 20 market makers, over 500 time steps and 30 Monte Carlo runs. Key results: (i) the coefficient of variation of ASU across platforms falls from 0.217 to 0.021 within 50 periods, confirming the price convergence prediction of Proposition 3; (ii) the Hsieh-Klenow TFP misallocation framework [Hsieh and Klenow, 2009] yields estimated annual welfare gains of approximately \$29.9 billion (at the upper bound where $\sigma \to \infty$) to \$0 (where platforms are perfect substitutes), with a baseline estimate of 11% TFP improvement; (iii) enterprise cost savings range from 0.2% to 20%, with a baseline of 11%, net of a 2% intermediation fee; (iv) the hybrid CLOB-AMM architecture dominates both CLOB-only (spread: 8.2 vs. 12.4 bps) and AMM-only (IL: 0.8% vs. 2.1%) designs. Full simulation details, calibration tables, sensitivity analysis across 14 parameters, and policy scenario results (subsidy-driven adoption dynamics from 1% to 99%) are in the Online Appendix.[4]

[4]Scope note: the simulation uses stylised behavioural rules (threshold switching, momentum-contrarian mixing), not strategic optimisation. It validates internal consistency of the mechanism design under calibrated agent behaviour, complementing the empirical validation of Section 7.

9 Discussion and Conclusion

9.1 Implementation Roadmap

We propose three phases:

Phase 1: Shadow Ledger and Compute Provider Onboarding. Enterprises adopt ASU as an internal budgeting and benchmarking unit. Simultaneously, compute providers with idle capacity register nodes and report utilisation data, generating the C-LMP signals needed for Phase 2. These idle-capacity providers become anchor participants whose participation motivation requires no subsidies (Proposition 5(iv)): their opportunity cost is effectively zero, and any positive C-LMP revenue is incremental. This phase requires no regulatory approval and no platform cooperation—compute providers and enterprises can begin independently.

Phase 2: Residual Market. A pilot exchange enables transfer of unused compute quotas among consenting enterprises and platforms. The scope is limited to "use-it-or-lose-it" capacity with short durations, minimizing regulatory complexity. China's National Data Administration, which oversees the data market sandbox framework, is one natural institutional home; EU AI Office and US CFTC LabCFTC provide international precedents.

Phase 3: Full Financialization. The complete ASR secondary market operates with CLOB, AMM, C-LMP, derivatives, and clearing. This phase requires commodity-market regulatory classification and comprehensive market surveillance.

9.2 Regulatory Considerations

The ASR raises classification questions under financial regulation. *Utility ASR* (short-duration, redeemable for services) may fall outside securities classification if the "expectation of profit" prong of the Howey test is unsatisfied—but this is jurisdiction-specific and requires legal validation. *Trading ASR* (long-duration, fungible) would

likely require commodity-futures regulation. The geolocation tag mechanism enables global secondary-market trading while routing invocations to compliant compute nodes, addressing data-residency requirements without restricting market liquidity.

9.3 Vertical Structure and Platform Motivation

The vertical disintermediation result (Proposition 5) implies that platforms face a strategic dilemma: participation in the ASR market makes their markups observable, but non-participation risks marginalisation as enterprises switch to open-source ASR issued directly by compute providers. Formally, the platform's net payoff from participation is:

$$\Delta\pi_{\text{platform}} = \underbrace{f_{\text{royalty}} \cdot V_{\text{secondary}} + \Delta S_i \cdot D_i}_{\text{gains: royalties + brand premium monetisation}} - \underbrace{(m_i^{\text{before}} - m_i^{\text{after}}) \cdot MC_i \cdot D_i}_{\text{loss: markup compression}}$$

where f_{royalty} is the secondary-market royalty rate and ΔS_i is the credit-premium gain from market transparency. Participation is individually rational when $f_{\text{royalty}} \cdot V_{\text{secondary}} + \Delta S_i \cdot D_i > (m_i^{\text{before}} - m_i^{\text{after}}) \cdot MC_i \cdot D_i$, which is more likely for platforms with (a) large secondary-market volume (high $V_{\text{secondary}}$), (b) strong brand premia (high ΔS_i), or (c) already-competitive markups (small $m_i^{\text{before}} - m_i^{\text{after}}$). In the Stackelberg structure implied by Proposition 5(iv), compute providers move first (they have idle capacity and no markup to protect); platforms follow once the market has sufficient liquidity to make non-participation costlier than markup compression.

9.4 Does Financialization Help?

Philippon [2015] documents that the unit cost of US financial intermediation has remained stable at approximately 2% for 140 years despite massive financial innovation—a sobering finding for any proposal to "financialize" a new sector. We take this critique seriously.

Three responses. First, the ASR market's primary value proposition is not intermediation cost reduction but *creation of missing markets*: currently, no instrument exists for enterprises to hedge compute costs or for platforms to monetize idle capacity. The Shiller [1993] argument applies: the welfare gain from creating a market for a previously unhedgeable risk can be large even if intermediation costs are standard.

Second, our simulation estimates suggest that enterprise procurement savings (0.2–20% of spend, net of ASR transaction costs; the welfare estimates (Online Appendix)) substantially exceed the intermediation costs embedded in the design: AMM fees of 0.5% and CLOB spreads of approximately 490 basis points in a nascent market, declining with maturity. The net welfare effect appears positive.

Third, the ASR market's DeFi-inspired architecture (automated clearing, algorithmic market making, transparent pricing) may achieve lower intermediation costs than traditional financial infrastructure. Makarov and Schoar [2022] note that DeFi can reduce transaction costs, though it creates new rent-extraction opportunities. Whether the ASR market's unit intermediation cost settles above or below Philippon's 2% benchmark is an empirical question that can only be answered by observing a live market.

We acknowledge that financialization also carries risks: excessive speculation, regulatory complexity, and the possibility that intermediation rents absorb the efficiency gains. The phased implementation roadmap (shadow ledger before full financialization) allows the market to demonstrate value at each stage before proceeding.

The AI inference market has reached the scale at which organized exchange infrastructure becomes economically necessary. We have documented three structural market failures—resource misallocation, price opacity, and unhedged risk exposure—using quantitative evidence from government reports, academic research, and industry surveys. We have proposed the AI Service Right (ASR) as a transferable property right on compute and the AI Service Unit (ASU) as a quality-adjusted unit of account

grounded in hedonic price theory. We have designed a hybrid secondary market integrating CLOB, AMM, and Compute Locational Marginal Pricing, established nine formal propositions characterizing equilibrium properties and welfare gains, and provided calibrated computational validation.

Three contributions are most significant. First, the conceptual contribution: redefining AI tokens as tradable service rights with quality-adjusted cross-platform fungibility—transforming a consumption medium into a financial asset amenable to price discovery, hedging, and reallocation—and formalizing the eq-token conversion framework (Table 1, equations (3)–(4)) that renders the ASU modality-neutral across text, image, video, and speech. Second, the theoretical contribution: establishing formal welfare results showing that the ASR market Pareto-improves over bilateral trading by resolving information asymmetry, search frictions, lock-in, and missing risk markets, while honestly engaging with the Philippon [2015] critique of financial innovation. Third, the design contribution: synthesizing proven mechanisms from four orthogonal market traditions—foreign exchange (pricing layer: cross-platform exchange rates and PPP-analog arbitrage), equities (trading layer: CLOB and market making), electricity (supply anchor: C-LMP), and DeFi (long-tail liquidity: AMM)—into a hybrid architecture with formal equilibrium properties, extended with compute-provider participation channels for direct ASR issuance, real-time dispatch, and capacity forward markets.

The paper has important limitations. Computational validation is not empirical validation; the welfare estimates are order-of-magnitude bounds dependent on parameter assumptions; the adoption dynamics suggest that substantial coordination (subsidies or mandated participation) is needed to overcome the chicken-and-egg problem; and the regulatory classification of ASR instruments requires jurisdiction-specific legal analysis beyond the scope of this economic framework.

The most productive near-term path is a controlled pilot—a regulatory sand-

box in which a small number of consenting platforms participate in a residual-quota exchange under observation. This would generate the real-market data needed to calibrate model parameters, test welfare claims, and establish whether the ASR market represents genuine financial innovation or merely financial intermediation in a new guise.

The AI compute market stands at a juncture analogous to electricity in the early 1990s or carbon emissions in the early 2000s—commodity-scale economic activity with no organized market infrastructure. If the historical pattern holds, the creation of organized exchange infrastructure will unlock substantial efficiency gains. The design framework presented here provides one blueprint for that transition.

References

Adams, H., Zinsmeister, N., Salem, M., Keefer, R., and Robinson, D. (2021). Uniswap v3 Core. Uniswap Labs whitepaper.

Flexera (2025). *State of the Cloud Report 2025*.

Goldstein, I. and Yang, L. (2017). Information Disclosure in Financial Markets. *Annual Review of Financial Economics*, 9:101–125.

Itskhoki, O. and Mukhin, D. (2021). Exchange Rate Disconnect in General Equilibrium. *Journal of Political Economy*, 129(8):2183–2232.

Lehar, A. and Parlour, C. A. (2025). Decentralized Exchanges. *Journal of Finance*, forthcoming.

Rogoff, K. (1996). The Purchasing Power Parity Puzzle. *Journal of Economic Literature*, 34(2):647–668.

Agarwal, N., and E. Budish (2021). Market design. In *Handbook of Industrial Organization*, Vol. 5, pp. 1–79.

Aquilina, M., E. Budish, and P. O'Neill (2022). Quantifying the high-frequency trading "arms race". *Quarterly Journal of Economics*, 137(1):493–564.

Avellaneda, M., and S. Stoikov (2008). High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224.

Benchmarkit/Mavvrik (2025). *State of AI Cost Management Survey 2025*.

Brynjolfsson, E., D. Li, and L. Raymond (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2):889–942.

Budish, E., P. Cramton, and J. Shim (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *Quarterly Journal of Economics*, 130(4):1547–1621.

Cartea, Á., S. Jaimungal, and J. Penalva (2015). *Algorithmic and High-Frequency Trading*. Cambridge University Press.

Cramton, P. (2017). Electricity market design. *Oxford Review of Economic Policy*, 33(4):589–612.

Dornbusch, R. (1976). Expectations and exchange rate dynamics. *Journal of Political Economy*, 84(6):1161–1176.

Du, W., A. Tepper, and A. Verdelhan (2018). Deviations from covered interest rate parity. *Journal of Finance*, 73(3):915–957.

Duffie, D., N. Gârleanu, and L. H. Pedersen (2005). Over-the-counter markets. *Econometrica*, 73(6):1815–1847.

Evans, M. D. D., and R. K. Lyons (2002). Order flow and exchange rate dynamics. *Journal of Political Economy*, 110(1):170–180.

Ellerman, A. D., F. J. Convery, and C. de Perthuis (2010). *Pricing Carbon: The EU Emissions Trading Scheme*. Cambridge University Press.

Federal Reserve Board (2025). The state of AI competition in advanced economies. FEDS Notes, October 6.

Flexera (2024). *State of the Cloud Report 2024*.

Fradkin, A. (2025). The emerging market for intelligence: Pricing, supply, and demand for LLMs. Working Paper, December.

Gartner (2026). Gartner forecasts worldwide AI spending to reach $2.52 trillion in 2026. Press Release, January.

Glosten, L. R. (1994). Is the electronic open limit order book inevitable? *Journal of Finance*, 49(4):1127–1161.

Goldstein, I., and L. Yang (2022). Commodity financialization and information transmission. *Journal of Finance*, 77(5):2613–2667.

Grand View Research (2025). *AI Inference Market Size Report, 2025–2030*.

Grossman, S. J., and J. E. Stiglitz (1980). On the impossibility of informationally efficient markets. *American Economic Review*, 70(3):393–408.

Hsieh, C.-T., and P. J. Klenow (2009). Misallocation and manufacturing TFP in China and India. *Quarterly Journal of Economics*, 124(4):1403–1448.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335.

Lawrence Berkeley National Lab (2025). *Data Center Energy Use and Efficiency*.

Lehar, A., and C. A. Parlour (2025). Decentralized exchange: The Uniswap automated market maker. *Journal of Finance*, 80(1):321–374.

Makarov, I., and A. Schoar (2022). Cryptocurrencies and decentralized finance (DeFi). *Brookings Papers on Economic Activity*, 2022(1):141–215.

Menkhoff, L., L. Sarno, M. Schmeling, and A. Schrimpf (2012). Currency momentum strategies. *Journal of Financial Economics*, 106(3):660–684.

Menlo Ventures (2025). *The State of Generative AI in the Enterprise*. December.

Milionis, J., C. C. Moallemi, T. Roughgarden, and A. L. Zhang (2022). Automated market making and loss-versus-rebalancing. *arXiv*:2208.06046.

Myerson, R. B., and M. A. Satterthwaite (1983). Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 29(2):265–281.

National Data Administration of China (2026). China's daily token usage exceeds 140 trillion. State Council Press Conference, March.

O'Hara, M. (1995). *Market Microstructure Theory*. Blackwell Publishers.

Philippon, T. (2015). Has the US finance industry become less efficient? On the theory and measurement of financial intermediation. *American Economic Review*, 105(4):1408–1438.

PJM Interconnection (2022). *Locational Marginal Pricing: An Overview*.

Rochet, J.-C., and J. Tirole (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4):990–1029.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55.

Roth, A. E. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70(4):1341–1378.

Schweppe, F. C., M. C. Caramanis, R. D. Tabors, and R. E. Bohn (1988). *Spot Pricing of Electricity*. Kluwer Academic Publishers.

Shiller, R. J. (1993). *Macro Markets: Creating Institutions for Managing Society's Largest Economic Risks*. Oxford University Press.

Shleifer, A., and R. W. Vishny (1997). The limits of arbitrage. *Journal of Finance*, 52(1):35–55.

Stanford HAI (2025). *AI Index Report 2025*.

Weyl, E. G. (2010). A price theory of multi-sided platforms. *American Economic Review*, 100(4):1642–1672.

Wilson, R. (2002). Architecture of power markets. *Econometrica*, 70(4):1299–1340.

Zhu, H. (2014). Do dark pools harm price discovery? *Review of Financial Studies*, 27(3):747–789.

Tirole, J. (1988). *The Theory of Industrial Organization*. MIT Press.

Zylo (2026). *SaaS Management Index 2026*.