

The Impact of AI Token Trading Replacing Traditional Currency in AI Technology Enterprises on Corporate Performance

Jingwen Yang

*Economics and Data Analysis, Hong Kong Baptist University, Hong Kong, China
22257829@life.hkbu.edu.hk*

Abstract. The intangible, instantly delivered, and continuously consumed nature of AI compute services exposes three inherent frictions in fiat-based settlement: temporal mismatches, coarse pricing granularity, and cross-border delays. This study investigates whether AI-Token—the atomic unit of model inference and training—can serve as a more efficient medium of exchange than fiat currency in B2B AI transactions, and how such a shift affects firm performance. Integrating transaction cost economics, agency theory, and corporate finance, it constructs a three-layer framework that traces how Token's roles as a means of payment, cost item, and asset class generate distinct causal pathways to revenue-side, cost-side, and asset-side performance. Employing theoretical synthesis and deductive analysis, this paper provides a unified conceptual architecture for future empirical research on programmable settlement media in the digital economy.

Keywords: AI-Token, medium of exchange, firm performance, transaction costs, intangible assets

1. Introduction

The core competitiveness of AI technology companies rests on large-scale computing infrastructure. With the rapid iteration of large language models and multimodal models, transactions such as compute leasing, model invocation, and data services between enterprises have grown into a rapidly expanding market. These transactions share a common feature: the subject matter is "computing power" and "data streams." They are intangible, delivered instantly, and consumed continuously rather than in a single delivery. When an AI company purchases 1,000 hours of GPU compute from a cloud provider, those 1,000 hours are not "delivered" at a single point in time but are drawn down continuously over a future period. Under these conditions, the fiat-based settlement system—which relies on interbank clearing, credit guarantees, and ex post contract enforcement—exposes three inherent tensions: a temporal mismatch between settlement cycles (T+1 or longer) and the instant delivery of compute, a disconnect between coarse pricing granularity (typically hourly or daily) and the continuous, per-second nature of compute consumption, and additional frictions arising from cross-border payment delays and exchange rate volatility. Whether a novel medium of exchange can mitigate these frictions constitutes the starting point of this study. Given that the use of AI-Tokens as

a B2B medium of exchange remains in its infancy, existing theories have yet to provide a complete answer, and systematic financial data are extremely scarce.

This paper presents a theoretical synthesis approach to develop a conceptual framework and testable hypotheses for future research in the AI-B2B context, laying a foundation for further analysis.

2. Methodology

2.1. The triple role of AI Token and the three dimensions of corporate performance

This study defines the AI-Token as the smallest unit of data processed during model training or inference, which also serves as the unit of pricing and settlement in B2B transactions between AI enterprises. It specifies its three roles with precision: as a means of payment, the Token prices and settles transactions by actual consumption in mid- and downstream B2B exchanges, replacing fixed-fee subscription models. As a cost item, it captures the computing power, electricity, and other resources consumed in generating or calling Tokens across all segments, benchmarked by per-Token cost. As an asset item, it refers to the expected Token output capacity of computing clusters held by upstream and midstream firms, serving as an economic resource that generates future economic benefits and qualifies for recognition as an intangible asset under accounting standards. The analysis examines the upstream, midstream, and downstream roles in the industry chain and their differential impacts on firm performance. Firm performance is decomposed into revenue-side, cost-side, and asset-side dimensions, based on the corporate finance theory that firm value equals the discounted value of expected future free cash flows. Revenue-side performance is measured by total revenue and asset turnover, influenced by the means-of-payment role, with theoretical foundations drawn from pricing theory [1]. Cost-side performance is measured by operating margin and administrative expense ratio, influenced by the cost-item role, with theoretical foundations drawn from agency theory and transaction cost economics [2]. Asset-side performance is measured by Tobin's Q and the degree of financing constraints, influenced by the asset-item role, with theoretical foundations drawn from intangible asset valuation theory and financing constraint theory [3, 4]. It is essential to clarify that the three roles are the "cause," and the three performance dimensions are the "effect"—this causal direction is irreversible.

2.2. Theoretical integration

This paper integrates the above theoretical strands into a three-layer progressive framework: transaction cost analysis, agency cost analysis, and the performance measurement interface. At the first layer, Williamson jointly point to measurement costs and execution costs inherent in B2B transactions [1, 5]. In AI compute trading, the high intangibility and heterogeneity of services render these two costs particularly salient. Token's programmability internalizes execution costs into code logic, while its standardized metrology transforms measurement costs from reliance on ex post contractual disputes into automated verification through technical parameters, making the Token not merely a medium of exchange but an endogenous component of the governance structure. At the second layer, extending the tripartite agency cost framework of Jensen and Meckling to inter-firm B2B transactions, Token's instant settlement eliminates credit risk, and its on-chain verifiability compresses the scope for disputes over service quality, jointly reducing inter-firm agency costs [2]. At the third layer, the projected reduction of these two cost categories relates to observable firm performance indicators. Fazzari noted that financing constraints affect investment-cash flow

sensitivity. If Token's instant settlement reduces inter-firm credit risk and transaction frictions, a decline in this sensitivity should be observed [4]. Peters et al. incorporated intangible assets into the investment-q model—if Token-generating capacity qualifies as a measurable intangible asset, an enhancement of Tobin's Q should be observed [3]. These two theoretical strands converge at the Token's three roles—means of payment, cost item, and asset item—each linked to one dimension of firm performance, forming a causal chain from the exchange medium to firm performance.

2.3. Mechanism derivation

Pathway one explores the relationship between payment and revenue-side performance. Traditional compute services employ fixed-rate or reserved-instance pricing, forcing customers to pay for idle capacity while lacking elastic scaling during demand surges. Token programmability enables smart contracts to automatically meter and charge by actual inference calls or training consumption, compressing pricing granularity from hourly to token-level. Li et al. identify elastic provisioning as a core economic attribute of cloud computing—Token's real-time continuous billing translates this technical possibility into economic reality [6]. First, finer pricing granularity allows service providers to implement differentiated pricing—charging a premium for high-frequency, low-latency clients and a discount for low-frequency, elastic-demand clients—thereby capturing consumer surplus lost under coarse pricing. Second, micropayment mechanisms make the cost per individual call extremely low, enabling small and medium enterprises and developer teams previously excluded by contractual thresholds to enter the market, opening up long-tail incremental demand. Consequently, Token's role in precise billing and pricing strategies enhances revenue generation and asset turnover.

Pathway 2 is from cost item to cost-side performance. Traditional fiat settlement involves delays and inefficiencies such as bank transfers and dispute resolution [1, 5]. In contrast, Token enables a clear, tamper-proof correlation between delivery and payment, effectively eliminating opportunities for rent capture due to measurement errors [1, 7]. At the same time, Token's instant settlement synchronizes payment and delivery, eliminating the accounts receivable cycle—sellers no longer bear credit risk, and buyers no longer worry about substandard service after prepayment, directly reducing the monitoring expenditures and residual loss in Jensen and Meckling's tripartite agency cost framework. Settlement automation further compresses non-productive activities such as manual reconciliation, auditing, and legal recourse, causing the administrative expense ratio to decline accordingly [2]. In sum, Token improves cost-side performance through the mechanism of "instant settlement compressing measurement and execution costs."

Pathway 3 links asset items to asset-side performance, highlighting that GPU clusters, classified as fixed assets, are undervalued under traditional accounting due to neglecting their true worth in generating Tokens. The continuous monitoring and forecasting of Token output rates fulfill the Financial Accounting Standards Board's asset definition, indicating future economic benefits. According to Peters et al., intangible assets capable of generating predictable cash flows are an important component of Tobin's Q—Token generation capacity falls precisely into this category, and its recognition by the market will push up firm valuation [3]. Meanwhile, the future cash flows of this asset are predictable and verifiable, enabling its use as collateral for financing, reducing information asymmetry in external financing, thereby alleviating the financing constraints in the sense of Fazzari, Hubbard, and Petersen—manifested as a decline in investment-cash flow sensitivity [4, 7]. Overall, Token generation capacity positively affects asset-side performance through the mechanism of intangible asset explication and collateralization.

3. Evaluation

Upstream computing infrastructure providers, such as NVIDIA, possess GPU clusters that constitute the material foundation of Token production—every unit of Token generated consumes measurable GPU computing power and electricity. At this segment of the supply chain, the core roles of the Token are the asset item and the cost item. NVIDIA's GPU clusters are traditionally classified as fixed assets and valued using the straight-line depreciation method. However, the true value of these clusters lies not in the physical equipment itself but in their capacity to continuously produce Tokens—that is, the "expected Token output stream." When such intangible assets are recognized in valuation, the firm's Tobin's Q rises: the market no longer prices the assets solely by their depreciated residual value but by their future Token output capacity [3]. NVIDIA's full-year FY2026 financial report shows that its data center business grew from approximately \$47.5 billion in FY2023 to \$115.2 billion in FY2025 (+142%) and further to \$193.7 billion in FY2026 (+68%), a more than fourfold increase over three years. Over the same period, the company's market capitalization rose from approximately \$360 billion to over \$3 trillion—a growth trajectory that traditional valuation models based on the book value of tangible assets can no longer explain. The market has evidently incorporated Token-generating capacity into its valuation premium. At the 2026 Annual Shareholders' Meeting, Jensen Huang explicitly stated that "every token is profit" and that "the core mission of the AI factory is to produce tokens—these monetizable units of intelligence are the raw materials for code, answers, designs, actions, and services." He further emphasized that "customers are not buying computing tools when they purchase NVIDIA systems, they are building AI factories that generate revenue directly." On the cost side, the continuous operation of GPU clusters consumes substantial electricity and cooling resources. NVIDIA officially identifies "cost per token" as a core metric for AI infrastructure TCO—the key to reducing Token cost lies in maximizing the delivered Token output. Blackwell's hourly GPU cost is only approximately twice that of Hopper (\$2.65 vs. \$1.41), yet the cost per million Tokens dropped from \$4.20 to \$0.12—a reduction by a factor of 35. This cost compression is directly reflected in profit margin expansion: within the same power budget, Blackwell delivers 50 times the Tokens per watt of Hopper, thereby maximizing the revenue of the AI factory. If the logic of Token assetization holds, the growth of NVIDIA's market capitalization should significantly exceed the growth of the book value of its physical assets, implying a rise in Tobin's Q. If the logic of Token cost optimization holds, the gross margin of NVIDIA's data center business should expand as the per-Token cost declines. The FY2026 financial report shows a data center gross margin of 71.1%. Despite a decline from the 75.0% recorded in FY2025, the FY2026 figure includes a \$4.5 billion inventory and purchase commitment charge related to H20 export controls, and the Token cost optimization effect in the core business has been fully reflected at the profit level. Midstream model and platform providers are exemplified by OpenAI, through which we illustrate the impact of the Token's three roles on firm performance. Beginning with the means-of-payment role, OpenAI shifted its pricing from a fixed subscription model (ChatGPT Plus at \$20 per month with unlimited usage) to per-Token consumption billing. The official API pricing page explicitly designates the Token as the unit of account, with separate pricing for input and output, and charges by actual Token consumption. The programmability of the Token enables automatic settlement based on actual inference counts, compressing pricing granularity from the user-per-month level to the Token-per-call level. This transformation has allowed OpenAI to implement differentiated pricing: different models form a price gradient. For example, GPT-5.6 Luna priced at \$0.20 per million input Tokens and \$1.20 per million output Tokens, while Terra is priced at \$2 per million input Tokens and \$12 per million

output Tokens—capturing the consumer surplus previously lost under fixed-rate pricing. Light users, who formerly overpaid relative to their usage under the fixed subscription, and heavy users, who previously paid less than the value they consumed, are now both aligned more closely with their true willingness to pay under Token-based billing. Verifiable performance data provide direct support for this mechanism: OpenAI's 2025 revenue reached \$13.07 billion, an increase of more than 2.5 times over the \$3.7 billion recorded in 2024, by year-end 2025, monthly revenue had reached \$2 billion, compared with quarterly revenue of only approximately \$1 billion at the end of 2024. Audited financial documents verified by the Financial Times indicate that OpenAI has become one of the fastest-growing enterprises in history. Holding assets constant, the refined pricing enabled by Token billing should manifest as an improvement in asset turnover, which is evident from the contrast between the revenue growth rate (+253%) and the asset base (approximately \$50 billion). On the cost side, OpenAI pays substantial compute leasing fees to upstream providers such as Microsoft Azure. If these payments were settled in Tokens, delivery and payment could theoretically be atomically executed, eliminating the receivables cycle and reconciliation costs inherent in fiat settlement, thereby reducing the monitoring expenditures and residual loss identified in Jensen and Meckling's framework [2]. However, verifiable performance data indicate that OpenAI currently remains predominantly on traditional fiat settlement, and the efficiency effects of the Token's cost role have yet to be fully realized: total costs and expenses for 2025 reached \$34 billion, including \$19.18 billion in R&D expenses, \$7.5 billion in cost of revenue, and \$1.57 billion in general and administrative expenses (an administrative expense ratio of approximately 12%), with an operating loss of \$20.92 billion—total expenditures exceeding 2.6 times revenue over the same period. On inference alone, spending approached \$5 billion in the first half of 2025, and cumulative inference costs from Q1 2024 to Q3 2025 exceeded \$12.4 billion. These data reveal the condition for its realization: the agency cost compression effect of Token settlement will only materialize once Tokens genuinely replace fiat settlement—this is precisely the core meaning of the paper's premise that AI-Tokens replace traditional currency. Regarding the Token's asset role, the expected inference Token output of the GPT series of models constitutes OpenAI's core intangible asset. Because both the price and the throughput of inference Tokens are observable, the expected Token output stream becomes a clearly bounded, measurable expected economic benefit, satisfying the definition of an asset under FASB Concepts Statement No. 8. Verifiable performance data are consistent with the predictions of Pathway Three: in October 2024, OpenAI completed a \$6.6 billion financing round at a post-money valuation of \$157 billion, and by the time of its 2026 financing, the valuation had further risen to \$730 billion. Its physical assets, totaling approximately \$50 billion at the end of 2025 with nearly half held in cash, are far below the valuation—the premium embedded in the market valuation is precisely the pricing of its Token-generating capacity as an intangible asset. OpenAI has entered into approximately \$1 trillion in computing capacity transactions, which are essentially advance commitments to future Token output capacity, consistent with the logic that intangible assets drive Tobin's Q. Taken together, the OpenAI case provides differentiated empirical evidence for the Token's three roles. As a means of payment, the Token enabled OpenAI to transition from fixed-rate to usage-based billing, with the 2.5-fold revenue growth providing direct evidence of the efficiency of refined pricing. As a cost item, the current data expose the fundamental predicament of the large-model business model: inference costs are escalating at a pace that far outstrips revenue, with cumulative inference costs of \$12.4 billion (Q1 2024–Q3 2025) already exceeding total revenue of approximately \$11.1 billion over the same period. However, this precisely illustrates the high level of agency costs in the absence of Token settlement, rather than falsifying Pathway Two. As an asset item, the market has accorded OpenAI a valuation far

exceeding the value of its physical assets—investors are betting on its future capacity to continuously produce Tokens, not on its existing servers and office equipment. The common thrust of the three roles is that the Token economy exhibits strong value-creation potential on the revenue and asset sides, but on the cost side—especially inference costs—it is consuming profits at an unprecedented pace. This contradiction highlights a key economic dynamic in the AI industry: Tokens drive remarkable revenue growth but also lead to significant losses. For downstream AI application developers, Token predominantly represent cost through midstream API calls while generating revenue by charging end customers based on the content volume produced. However, U.S. technology companies currently rarely disclose Token consumption costs associated with AI model calls separately in their financial reports, and no mature Token futures market exists for enterprises to conduct hedging operations. Under these constraints, this paper draws on Intuit's publicly documented practice in the stablecoin (USDC) payment context as an "observable proxy" for Token payments in the dimensions of settlement efficiency and cost transparency, to demonstrate the potential cost effects of Token as a means of payment. In December 2025, Intuit and Circle signed a multi-year strategic partnership to integrate USDC stablecoin infrastructure into core products including TurboTax, QuickBooks, Credit Karma, and Mailchimp. Intuit CEO Sasan Goodarzi explicitly stated that the partnership aims to "provide millions of consumers and businesses with faster, lower-cost, programmable money movement", CFO Sandeep Aujla further defined stablecoins as a new "'digital dollar' rail," noting that they "settle transactions nearly instantly, at materially lower cost." Circle, for its part, indicated that USDC provides Intuit with a "programmable, 24/7, low-friction money channel" that unlocks refund, remittance, and payment experiences "simply not possible on traditional rails." This practice provides an observable empirical anchor for the cost effects of Token payments. Traditional fiat settlement involves a series of non-productive procedures—bank transfer delays, invoice reconciliation, and dispute arbitration [1, 5]. The near-instant settlement of USDC eliminates the receivables cycle—sellers no longer bear credit risk, and buyers no longer worry about substandard service after prepayment—directly compressing the monitoring expenditures and residual loss. Its 24/7 availability eliminates holiday and cross-time-zone settlement delays, further reducing transaction frictions, programmability enables the atomic execution of payment and delivery, fundamentally eliminating the "public domain" for rent capture—when settlement and delivery correspond one-to-one and are tamper-proof, neither party can extract extra benefits from unpriced attributes [1]. The "low-friction" and "materially lower cost" emphasized by Intuit's management are direct mappings of the above mechanisms in managerial discourse. Extending this logic from USDC to the Token payment for AI enterprises allows developers to pay API fees using Tokens, synchronizing payments with inference delivery. This eliminates delays and reconciliation costs associated with fiat currency, compressing monitoring expenditures and residual loss while the standardized metrology of Tokens ensures accurate tamper-proof records between services rendered and costs incurred [1, 7]. Theoretically, this mechanism should manifest as a decline in the administrative expense ratio and an improvement in the operating margin. It should be emphasized that this analogy primarily applies to the logical chain of cost transparency and monitoring expenditures and does not extend to the dimensions of Token price volatility and hedging mechanisms. As a stablecoin pegged to the U.S. dollar, USDC does not possess price volatility and thus cannot serve as an analogy for the risk-hedging dimension of Token as a volatile cost item. Moreover, Intuit's USDC integration is in early stages, with no quantifiable financial performance data available for empirical testing. Improved industry disclosure standards and maturing AI markets will necessitate further empirical research.

4. Conclusion

This paper discusses three inherent frictions of fiat settlement in AI compute B2B transactions: temporal mismatch, coarse pricing granularity, and cross-border delays. Integrating search-theoretic monetary theory, transaction cost economics, agency theory, and corporate finance, this study constructs three causal mechanisms through which Tokens affect firm performance. As a means of payment, Token-based billing refines pricing to the token level, capturing consumer surplus and opening long-tail markets, thereby enhancing revenue-side performance. As a cost item, instant settlement eliminates the receivables cycle and, together with programmability, reduces measurement rent capture and agency costs, crucially, this efficiency gain materializes only when Tokens genuinely replace fiat, as evidenced by OpenAI's cumulative inference costs exceeding its revenue. As an asset item, the expected Token output stream is valued as a measurable intangible asset, raising Tobin's Q and alleviating financing constraints. Case evidence from NVIDIA, OpenAI, and Intuit corroborates the revenue- and asset-side effects. Limitations include the early-stage nature of Token adoption restricting data availability and external validity, the limited applicability of the USDC analogy to cost transparency rather than price volatility hedging, and the lack of uniform accounting recognition for Token-based intangible assets. Future research should conduct large-sample empirical tests using investment-q and investment-cash flow sensitivities and evaluate hedging strategies once Token derivatives market mature.

References

- [1] Barzel, Y. (1982). Measurement cost and the organization of markets. *The Journal of Law & Economics*, 25(1), 27–48. <https://doi.org/10.1086/467005>
- [2] Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305–360. [https://doi.org/10.1016/0304-405X\(76\)90026-X](https://doi.org/10.1016/0304-405X(76)90026-X)
- [3] Peters, R. H., & Taylor, L. A. (2017). Intangible capital and the investment-q relation. *Journal of Financial Economics*, 123(2), 251–272. <https://doi.org/10.1016/j.jfineco.2016.03.011>
- [4] Fazzari, S. M., R Glenn Hubbard, & Petersen, B. C. (1996). Financing constraints and corporate investment : Response to Kaplan and Zingales. National Bureau of Economic Research.
- [5] Williamson, O. E. (1979). Transaction-cost economics: The governance of contractual relations. *The Journal of Law & Economics*, 22(2), 233–261. <https://doi.org/10.1086/466942>
- [6] Li, H., Li, Q., Xu, Z., & Ye, X. (2024). Digital technologies. *Journal of Digital Economy*, 3, 240–248. <https://doi.org/10.1016/j.jdec.2025.02.001>
- [7] Xing, Y. (2026). AI token futures market: Commoditization of compute and derivatives contract design. *arXiv.Org*.