IFX AI Tocken Futures

A Cash-Settled Forward Market on Language-Model Inference Prices

June 2026 | IFX Research

Abstract. Enterprise spending on large language model (LLM) inference reached $8.4 billion annually by mid-2025 and is forecast to grow to $255 billion by 2030 [1]. Despite this scale, companies holding material token-cost exposure have no listed instrument to hedge it: GPU compute futures announced by ICE/Ornn [2] and CME/Silicon Data [3] reference hardware rental prices, which are structurally decoupled from the inference output prices set unilaterally by model creators. We describe IFX - a bilateral, cash-settled forward market on the published USD price per one million output tokens of a named language model. We document the natural hedging base spanning enterprise SaaS companies, AI startups, grant-funded research labs, and biotech firms using LLMs for drug discovery, and motivate a three-stage product roadmap that begins with pinned single-model OTC forwards and culminates in a volume-weighted token price index designed to hedge the aggregate AI cost exposure of investment funds holding diversified portfolios of AI companies.

1 Market Overview

1.1 The AI Infrastructure Capital Supercycle

The deployment of AI infrastructure is the largest synchronized capital investment cycle in technology history. The five largest hyperscalers - Amazon, Alphabet, Microsoft, Meta, and Oracle - plan aggregate capital expenditure exceeding $630 billion in 2026, a 36% increase over 2025, with approximately 75% of that sum directed toward AI servers, GPUs, and data center construction [4]. Goldman Sachs projects that cumulative hyperscaler capex from 2025 through 2027 will reach $1.15 trillion, more than double the $477 billion spent in the three preceding years [4]. Individual 2026 guidance stands at approximately $200 billion for Amazon, $175-185 billion for Alphabet, $115-135 billion for Meta, and $110-120 billion for Microsoft [4].

Nvidia's earnings provide the clearest real-time signal. Data Center segment revenue reached $30.8 billion in the quarter ending October 2025 (+112% year-on-year), and Blackwell GPU spot rental surged 48% from mid-February to mid-April 2026, rising from $2.75 to $4.08 per GPU-hour [5, 6]. The capital build is accelerating, not plateauing.

1.2 The Token Market: Scale, Growth, and Structure

The economic output of this capital stock is measured in tokens. Enterprise LLM API spend rose from $3.5 billion in Q3 2024 to $8.4 billion by mid-2025 - a 140% annualised growth rate - with the average company spending $85,500 per month on AI, up 36% year-on-year (CloudZero) [7]. OpenRouter served 100 trillion tokens in calendar 2025; Azure OpenAI processed 100 trillion tokens in Q2 2025 alone [7]. At the prevailing frontier output price of $8 per million tokens, those Azure tokens represent $800 million of quarterly inference expenditure on a single infrastructure layer.

Third-party forecasts project continued expansion at pace. MarketsandMarkets estimates the global AI inference market at $106 billion in 2025, growing to $255 billion by 2030 at a CAGR of 19.2% [1]. The Large Language Model API market is estimated at $22.7 billion in 2026 and expected to reach $137.7 billion by 2035 at a CAGR of 22.1% [8]. Gartner projects that AI spending will reach 8% of enterprise IT budgets by 2028, up from under 1% today, and forecasts that inference costs for a one-trillion-parameter model will fall a further 90% between 2025 and 2030 [9].

The supply side has grown in parallel. The number of distinct models on OpenRouter rose from 253 in January 2025 to 651 by December 2025, inference providers from 27 to 90, and model creators from 43 to 85 [7]. Despite this proliferation, market structure remains highly concentrated. Programming accounts for approximately 50% of all API calls, and the Herfindahl-Hirschman Index by use case reaches 8,000 for translation and 5,000 for programming - effectively monopoly and high-concentration levels [7].

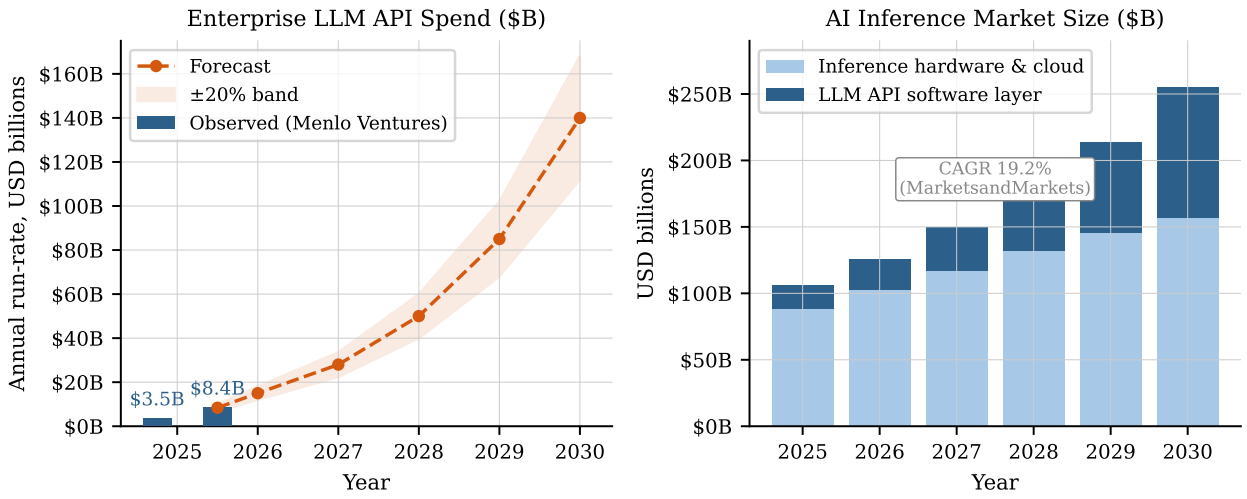

Figure 1: Left: enterprise LLM API spend annual run-rate with forward projection. Right: total AI inference market size and decomposition by segment (2025-2030). Sources: Menlo Ventures, MarketsandMarkets [1], BusinessResearch Insights [8].

1.3 Market Architecture

Demirer et al. [7] identify a three-layer architecture governing the flow of tokens from compute to end user (Figure 2). *Creators* train and maintain foundation models (OpenAI, Anthropic, Google DeepMind, Meta, Mistral, Cohere, xAI - 85 entities as of December 2025). *Inference providers* serve API calls, either co-located with the creator (direct API) or as independent re-hosters (Azure OpenAI, AWS Bedrock, Google Vertex, Together AI, Fireworks - 90 providers as of December 2025). *Aggregators* route traffic across providers with per-call price competition (OpenRouter, Portkey, LiteLLM, Requesty).

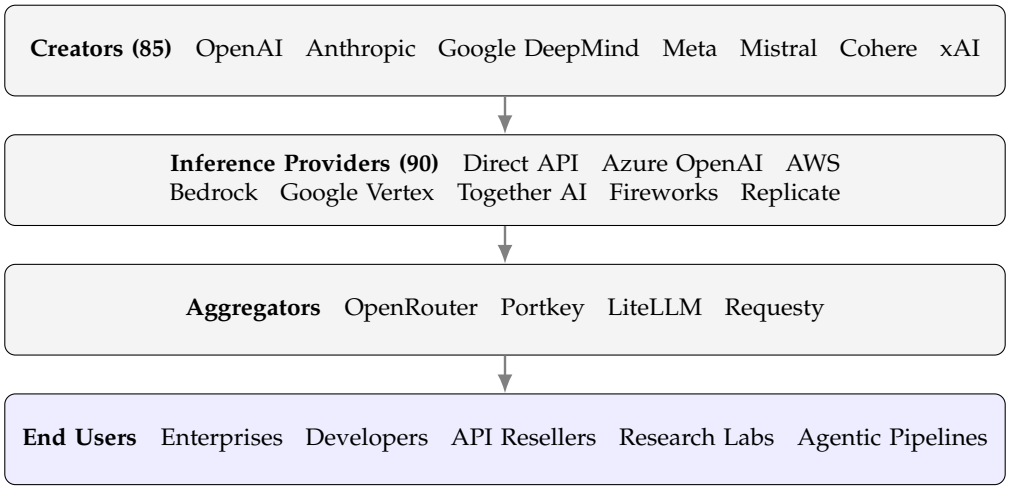

Figure 2: Three-layer inference market architecture (Demirer et al. [7]). Token price risk accumulates at the end-user layer.

The critical structural feature is vertical integration at the top: OpenAI and Anthropic are simultaneously creators and primary inference providers, setting administered prices that propagate through all downstream layers. This means that a company consuming `gpt-4.1` via Azure, via OpenRouter, or via the direct API faces the same underlying price risk - governed by a single administrative decision in San Francisco, not by market forces.

1.4 Token Pricing: Deflation and the Open-Closed Divide

The defining economic feature of the inference market is rapid, irregular, downward price movement. Demirer et al. [7] document a 1,000-fold decline in the price of GPT-4-class capability over two years. This is driven by compounding gains in hardware efficiency (A100 to H100 to H200 to B200) and algorithmic efficiency through distillation, quantization, and speculative decoding.

The pricing landscape divides sharply along the open/closed-source dimension. Proprietary frontier models carry a substantial capability premium: GPT-5.5 lists at $30 per million output tokens, Claude Opus 4.8 at $25, and Gemini 2.5 Ultra at $15. Open-weight models hosted on commodity providers carry an 87% discount to closed-source models at equivalent benchmark intelligence, a coefficient of -2.46 in log-terms from the price-intelligence regression in Demirer et al. [7]. DeepSeek V4 Flash is available via aggregators at $0.14 per million tokens; Meta's Llama 4 Scout at $0.18-0.59. The gap between commodity open-weight inference and frontier closed-source reasoning is structural and expected to widen as open-source capability converges while re-hosting remains unrestricted.

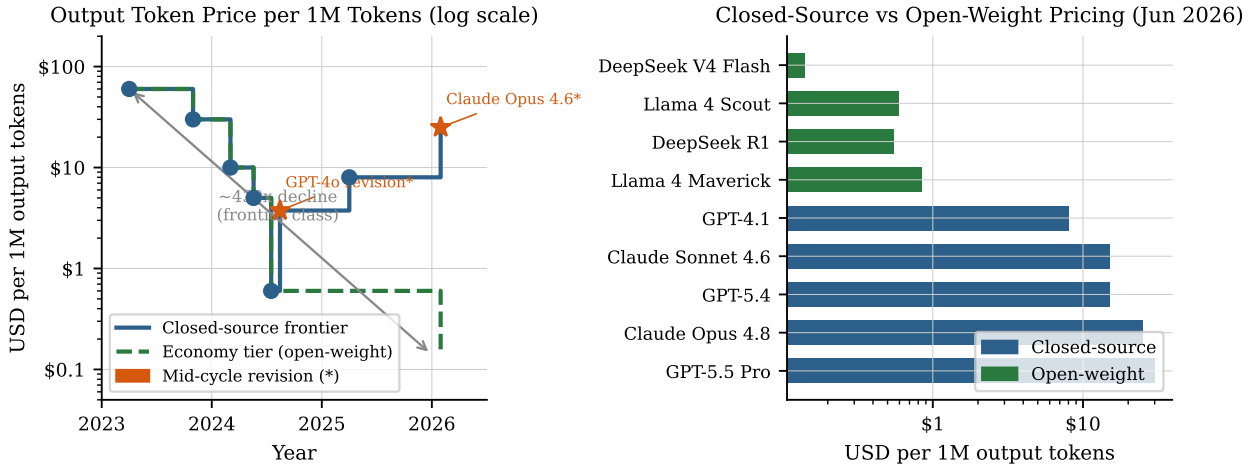

Figure 3: Left: closed-source frontier and economy-tier token price history on a log scale, 2023-2026. Asterisked events (*) are mid-cycle revisions of existing model IDs - the category relevant to pinned IFX forwards. Right: current open-weight versus closed-source pricing across representative models (output tokens, June 2026). Sources: OpenAI, Anthropic public pricing pages; Demirer et al. [7].

Closed-source creators have responded to competitive pressure through infrequent, large, downward price revisions of their existing model IDs. Table 1 summarises the documented events. Two patterns are immediate: all revisions are cuts, and the subset directly relevant to pinned forwards - mid-cycle revisions of an existing model ID - are rarer and larger on average than new-model launches.

1.5 Competing Financial Products and the Gap

Two derivatives initiatives were announced within days of each other in May 2026, signalling that institutional conviction in compute as a commodity asset class has reached a tipping point. ICE (owner of the NYSE) partnered with Ornn - an MIT-founded startup backed by a16z Crypto with $33 million in seed funding [10] - to launch cash-settled GPU compute futures referenced against the Ornn Compute Price Index (OCPI), live on Bloomberg since April 2026 [2, 6]. CME Group partnered with Silicon Data (founded by a former Bloomberg executive, backed by DRW and Jump Trading) on daily GPU benchmark futures distributed via Refinitiv [3]. The Shanghai Futures Exchange disclosed active research into AI token futures targeting the inference consumption layer, framed in Shanghai's municipal strategy to become a global AI financial hub [11].

Table 1: Closed-source frontier repricing events (2023-2026). Mid-cycle revisions (*) change the price of an existing model ID and are the direct risk exposure of a pinned IFX forward. New-model launches create new contract series without affecting outstanding positions.

Date	Model ID	Old price	New price	Change	Mid-cycle
Dec 2023	GPT-3.5 Turbo	\$2.00/1M	\$1.00/1M	−50%	
Mar 2024	GPT-4 Turbo	\$30.00	\$10.00	−67%	
May 2024	GPT-4o launch	\$15.00	\$5.00	−67%	
Jul 2024	GPT-4o mini	new	\$0.60	—	
Aug 2024	GPT-4o (rev.)	\$5.00	\$3.75	−25%	✓
Feb 2025	GPT-4.5	new	\$150.00	—	
Apr 2025	GPT-4.1	new	\$8.00	—	
Feb 2026	Claude Opus 4.6	\$75.00	\$25.00	−67%	✓

Venue	Partner	Status	Reference asset
ICE	Ornn / OCPI	Pending approval	GPU spot rental: H100/H200/B200/RTX 5090. Does not reference any model's output token price.
CME Group	Silicon Data	Pending approval	Daily GPU on-demand benchmark rates. Same structural gap: hardware cost, not inference price.
SHFE	Internal	Research stage	Explicitly targets the token consumption layer. No commercial launch date.
IFX	—	**Live**	**Published USD price per 1M output tokens of a named model ID. No other instrument covers this layer.**

A company spending \$500,000 per month on `gpt-4.1` API calls gains zero protection from GPU futures. Its token cost is set by OpenAI's published price, which is administratively determined and can change independently of hardware spot markets. The GPU-to-token cost transmission operates with a lag of three to eight months and a pass-through coefficient well below one: efficiency gains are partially captured as margin before reaching the end user. IFX addresses the layer where commercial token cost risk actually resides.

2 Token Price Economics

2.1 Academic Foundation

Three published works provide the theoretical and empirical substrate for IFX. Xing [12] (*arXiv:*2603.21690) delivers the first formal design of Standard Inference Token (SIT) futures, a supply-side cost model, a Poisson pricing framework, a Black (1986) condition audit, and Monte Carlo simulations showing 62-78% reduction in enterprise token-cost volatility. Demirer, Fradkin, Tadelis, and Peng [7] provide the empirical market structure study using a proprietary dataset from OpenRouter augmented with Azure volume data. Their outputs include the three-layer architecture, the price-intelligence regression, elasticity estimates, and firm concentration findings. Wu and Deng [13] contribute the Token Economics Impossibility Triangle, showing that granularity, real-time settlement, and optimality cannot all be achieved simultaneously in a token pricing system. IFX's 60-minute settlement window with a formula-determined median price resolves this triangle in favour of determinism.

2.2 Supply-Side Price Determinants

Xing [12] derives the marginal cost of token production from first principles. Token output capacity Q and marginal cost C_{marginal} are given by:

$$Q_{\text{token}} = \frac{\eta_H \cdot \eta_A}{C_E} \cdot K, \qquad C_{\text{marginal}} = \frac{C_E}{\eta_H \cdot \eta_A}, \qquad C_{\text{total}} = \frac{C_{\text{train}}}{N_{\text{lifetime}}} + C_{\text{marginal}}. \tag{1}$$

Notation. η_H is hardware efficiency measured in tokens generated per floating-point operation (tokens/FLOP); η_A is algorithmic efficiency, capturing how much benchmark capability is delivered per FLOP through post-training improvements; C_E is the energy cost per FLOP in USD; K is the total GPU capital stock deployed; C_{train} is the one-time training cost in USD; and N_{lifetime} is the expected number of tokens served over the model's commercial lifetime, so $C_{\text{train}}/N_{\text{lifetime}}$ represents the training cost amortized per output token.

The 1,000-fold price decline over two years [7] reflects compound gains in both η_H (hardware generations: A100, H100, H200, B200) and η_A (distillation, quantization, speculative decoding, mixture-of-experts routing). The training amortization floor $C_{\text{train}}/N_{\text{lifetime}}$ is a structural constant: as long as frontier model training costs tens of millions of GPU-hours, the marginal cost cannot reach zero. Gartner [9] projects a further 90% decline in inference costs for a one-trillion-parameter model by 2030 - consistent with continued compounding in η_H and η_A, but bounded from below by this floor.

2.3 Demand-Side: The Price-Intelligence Regression

Demirer et al. [7] estimate a log-log OLS regression of model price on benchmark quality:

$$\log P_{mt} = \beta \cdot \log(\text{Intelligence}_m) + \gamma \cdot \text{OpenSource}_m + \mathbf{X}'_{mt}\boldsymbol{\delta} + \varepsilon_{mt}. \tag{2}$$

Notation. P_{mt} is the published price per million output tokens of model m at time t; Intelligence_m is a composite benchmark score capturing model capability (aggregated from MMLU, HumanEval, and similar evaluations); OpenSource_m is a binary indicator equal to one for open-weight models; $\mathbf{X}_{mt}$ is a vector of controls including provider, date, and use-case fixed effects; and ε_{mt} is the residual.

The estimates $\hat{\beta} = +0.039$ and $\hat{\gamma} = -2.46$ (both at the 1% significance level) imply that a one-unit increase in benchmark intelligence is associated with a 3.9% higher price, and that open-source models trade at $e^{-2.46} \approx 8.5\%$ of the closed-source price at equivalent intelligence - an 87% structural discount. Prices are not arbitrary: they track compute investment, validating named model IDs as stable reference points for forward contracts.

2.4 Price Elasticity

Within-model cross-provider price elasticity of demand is estimated at -1.08 to -1.11 (Table 2 in Demirer et al. [7]). An elasticity near -1 is inconsistent with a Jevons Paradox: price cuts expand volume roughly in proportion, not faster. This means that as token prices fall, notional dollar demand remains roughly stable, so contract notional sizing does not need to account for explosive volume responses to price cuts.

3 Risk Identification

3.1 Nature of Administered Price Risk

Token prices for closed-source model IDs are set by fiat. The creator publishes a new price on its public pricing page, effective immediately, with no advance notice or market mechanism. The risk profile has three properties that distinguish it from any conventional commodity:

1. **Zero pre-trade price signal.** Changes are announced, not cleared. There is no order book, no futures strip, and no public forward curve from the creator.
2. **Discontinuous jumps.** Historical moves range from -25% to -67% in a single announcement. The piecewise-constant nature means intraday volatility is zero, but event-driven risk is large and sudden.

3. **No public hedge.** GPU futures hedge hardware rental cost, which can diverge from inference output price on an entirely different schedule. The transmission lag is three to eight months, and the coefficient is well below one.

3.2 Who Bears the Risk

The demand for IFX forwards arises from a structural mismatch: companies that commit to fixed-price output contracts (annual SaaS seats, API reseller agreements) while their largest variable input - tokens - can reprice without notice. For token buyers (natural longs), a surprise price cut generates a margin windfall that cannot be passed to locked-in customers until renewal. For inference providers and GPU cloud hosts (natural shorts), revenue per token falls with efficiency improvements, and a provider that pre-sold fixed-rate capacity at \$10 per million tokens faces a revenue shortfall when the market moves to \$5.

The critical finding from Demirer et al. [7] is that firms cannot diversify this exposure. Over 50% of companies on OpenRouter use a single model, and among multi-model users, over 90% of spend concentrates on one model. Brand loyalty reinforces this: Claude users substitute within Anthropic's lineup, not across to OpenAI. A company embedded in GPT-4.1 cannot hedge by switching to Gemini - it is structurally anchored to OpenAI's pricing decisions.

3.3 Exposure Magnitude

A company running one trillion tokens per year at GPT-4.1's \$8 per million incurs \$8 million in annual token spend. The average mid-cycle cut in the closed-source series is 46% (calibrated below). One such event creates a \$3.7 million annual COGS swing - a figure that cannot currently be hedged on any public market. At the average enterprise AI spend of \$85,500 per month (\$1.03 million per year), the same cut produces a \$474,000 annual variance relative to a budget set at the start of the year.

4 Forward Pricing

4.1 The Administered Price Process

The canonical price of a closed-source model ID follows a Poisson jump process: piecewise-constant between announcements, with negative jumps arriving at Poisson rate λ. This is the correct process for administered prices because between announcements the price is exactly constant (not Gaussian, not mean-reverting), and the arrival of the next announcement is unpredictable and memoryless.

Open-source models served across competing providers follow a different process: prices are market-determined, change continuously as providers compete, and follow competitive equilibrium dynamics rather than administrative jumps. Stage 1 IFX contracts are restricted to named closed-source model IDs precisely because only those have a uniquely defined, authoritative canonical price to pin.

4.2 Calibration and Repricing Probability

Of the eight documented repricing events in Table 1, seven are new-model launches that create new model IDs without altering pinned contracts. Only two events are mid-cycle revisions of existing IDs: the August 2024 GPT-4o revision (-25%) and the February 2026 Claude Opus 4.6 revision (-67%). These are the events that directly affect a holder of a pinned forward.

The Demirer et al. [7] dataset provides the empirical basis for calibrating the repricing frequency. Across the combined OpenAI and Anthropic frontier series, mid-cycle revisions occurred twice over 37 months of observation. For a single model ID at a single provider, the implied Poisson parameters are:

$$\hat{\lambda}_{\text{mid}} = \frac{1}{37} \text{ per month}, \qquad \hat{\delta}_{\text{mid}} = \frac{1}{2}(0.25 + 0.67) = 0.46. \tag{3}$$

Notation. $\hat{\lambda}_{\text{mid}}$ is the estimated arrival rate of mid-cycle repricing events in units of events per month per model ID; $\hat{\delta}_{\text{mid}}$ is the average magnitude of a mid-cycle price cut, computed as the equally weighted mean of the two observed cuts.

A broader calibration that includes new-model launches as signals of provider intent to commoditise gives:

$$\hat{\lambda}_{\text{series}} = \frac{8}{37} \text{ per month}, \qquad \hat{\delta}_{\text{series}} = 0.33. \tag{4}$$

The probability of at least one repricing event occurring before maturity at horizon τ months is:

$$p(\tau) = 1 - e^{-\hat{\lambda}\,\tau}. \tag{5}$$

Under the mid-cycle calibration, the repricing probability reaches 7.8% at three months, 15.0% at six months, and 27.9% at twelve months. These are not negligible probabilities for budget-cycle decisions. Importantly, the dispersion in cut size ($\hat{\delta}$ ranging from 0.25 to 0.67 across the two observed events) adds uncertainty beyond the Poisson rate estimate. A 67% cut like February 2026 would produce a payout three times larger than a 25% cut at the same notional and tenor, suggesting that buyers seeking full budget protection should target the 6M or 12M tenor rather than the 3M.

4.3 Indicative Forward Price

Under the Poisson model, the expected price at horizon τ is:

$$\mathbb{E}[P_\tau] = S_0 \cdot m(\tau), \qquad m(\tau) = 1 - \hat{\delta}\left(1 - e^{-\hat{\lambda}\tau}\right). \tag{6}$$

Notation. S_0 is the current spot price (the canonical published price per million output tokens at the time of contract inception); $m(\tau)$ is the forward price factor, which equals 1 when there is no repricing risk ($\hat{\delta} = 0$ or $\tau \to 0$) and falls toward $1 - \hat{\delta}$ as tenor grows large; τ is the contract tenor in months.

IFX publishes $F^{\text{ind}} = S_0 \cdot m(\tau)$ as a non-binding reference anchor. Counterparties set their own order prices. Table 2 shows the implied discounts under both calibrations.

Table 2: Indicative forward level as a fraction of spot S_0 under the two Poisson calibrations. $p(\tau)$: probability of at least one repricing event by maturity. *Mid-cycle* is the conservative base calibration targeting only events relevant to pinned contracts.

	Mid-cycle ($\hat{\lambda} = 1/37$, $\hat{\delta} = 0.46$)		Series ($\hat{\lambda} = 8/37$, $\hat{\delta} = 0.33$)	
Tenor	$p(\tau)$	$m(\tau)$	$p(\tau)$	$m(\tau)$
3M	7.8%	0.964	44.9%	0.851
6M	15.0%	0.931	63.9%	0.789
12M	27.9%	0.872	87.1%	0.713

4.4 Margin Framework for Stage 2

Because the price process is piecewise-constant, intraday variation is zero. For Stage 2 exchange-listed contracts, margin covers the jump event only:

$$M_{\text{init}} = \max\left(\alpha \cdot \hat{\delta} \cdot S_0 \cdot V_{\text{contract}},\ M_{\text{floor}}\right). \tag{7}$$

Notation. M_{init} is the required initial margin in USDT; α is a safety multiplier calibrated to cover the 99.5th-percentile loss conditional on a jump occurring (set at approximately 1.5 to account for the observed dispersion in $\hat{\delta}$); $\hat{\delta}$ is the jump magnitude parameter; S_0 is the spot price at inception; V_{contract} is the contract notional in token units; and M_{floor} is a minimum margin floor set by exchange rules. The initial margin target is 8-15% of notional, substantially lower than continuous-price commodity contracts because there is no Gaussian diffusion component.

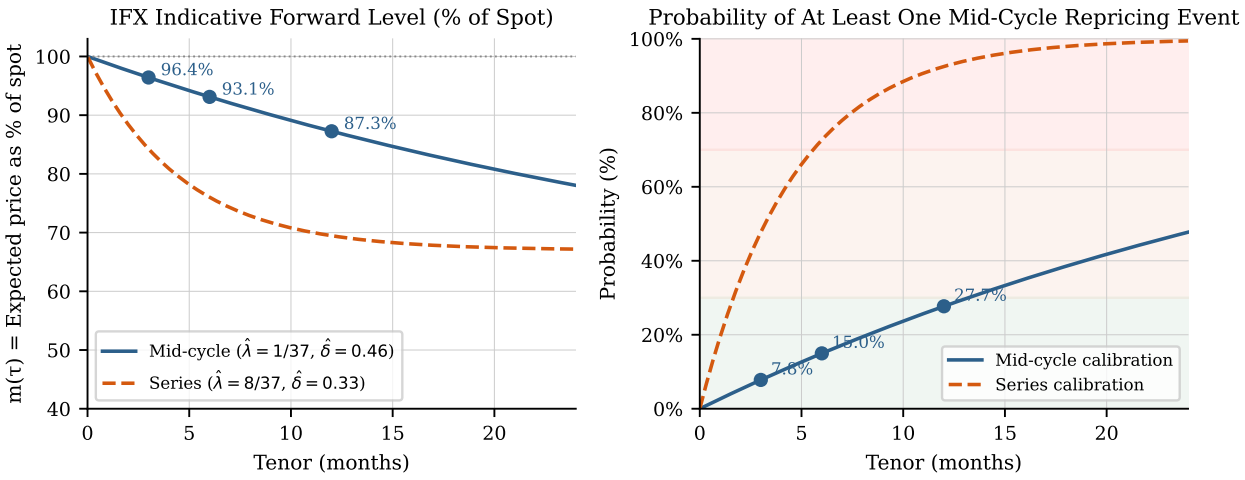

Figure 4: Left: indicative forward level $m(\tau)$ as a percentage of spot under mid-cycle and series calibrations. Right: probability of at least one mid-cycle repricing event by tenor. Coloured bands denote low risk (<30%), moderate (30-70%), and elevated (>70%) probability zones. Both calibrated from Table 1.

5 Demand Structure and Hedging Rationale

The natural demand base for IFX forwards is broader than the SaaS and API reseller segment typically discussed in API cost management literature. Five distinct participant archetypes face structurally different manifestations of token price risk, and Figure 5 classifies them by their position (long or short) and the mechanism through which they bear exposure.

Enterprise SaaS companies. A company selling annual software subscriptions that uses GPT-4.1 as its AI backbone is structurally short token price risk for the duration of each subscription cycle. Its output price is fixed at contract signing while its largest variable input can reprice without notice. A mid-cycle cut by OpenAI generates a margin windfall that cannot be captured until the next renewal; a price increase would directly compress margin. A 12-month IFX forward converts this open exposure to a fixed-cost commitment aligned with the subscription cycle. For a 10,000-seat product processing 800 million tokens per user per month, the hedge notional at one trillion tokens per year is $8 million - directly matching the API budget set at the start of the fiscal year.

AI startups. Seed- and Series A-stage companies building AI-native products operate on fundraising schedules that are independent of token price calendars. A startup that raises $5 million at $8 per million tokens and builds its burn model accordingly is financially exposed if OpenAI cuts to $5 per million before the next fundraise: the lower cost changes the competitive landscape (rivals can afford more features), and the startup's unit economics presentation to new investors is invalidated mid-cycle. A 3- to 6-month IFX forward provides a bridge from one fundraising milestone to the next, locking token COGS at the rate prevailing when the financial model was built.

Grant-funded research laboratories. Academic and government grants (NIH, NSF, DARPA, Wellcome Trust, ERC) are denominated in fixed dollar amounts awarded on multi-year cycles. A computational biology lab with a $2 million NIH grant awarded in January 2026 has a fixed token budget for the grant period. A mid-grant price increase directly reduces the number of experiments the lab can run within the token budget. A long forward converts a fixed-dollar grant award into a fixed-token budget, eliminating the translation risk between award date and experiment execution.

Biotech and drug discovery companies. This segment represents the highest per-token-dollar research intensity of any identified cohort. AI-assisted drug discovery pipelines consume LLM API calls for protein-structure simulation, molecular property prediction, literature synthesis, clinical-trial protocol generation, and regulatory document drafting. Token spend is directly proportional to the number of experimental cycles the pipeline can afford within a fixed research budget. Insilico Medicine's rentosertib program - the first drug where both target discovery and molecular design were performed entirely by generative AI - progressed

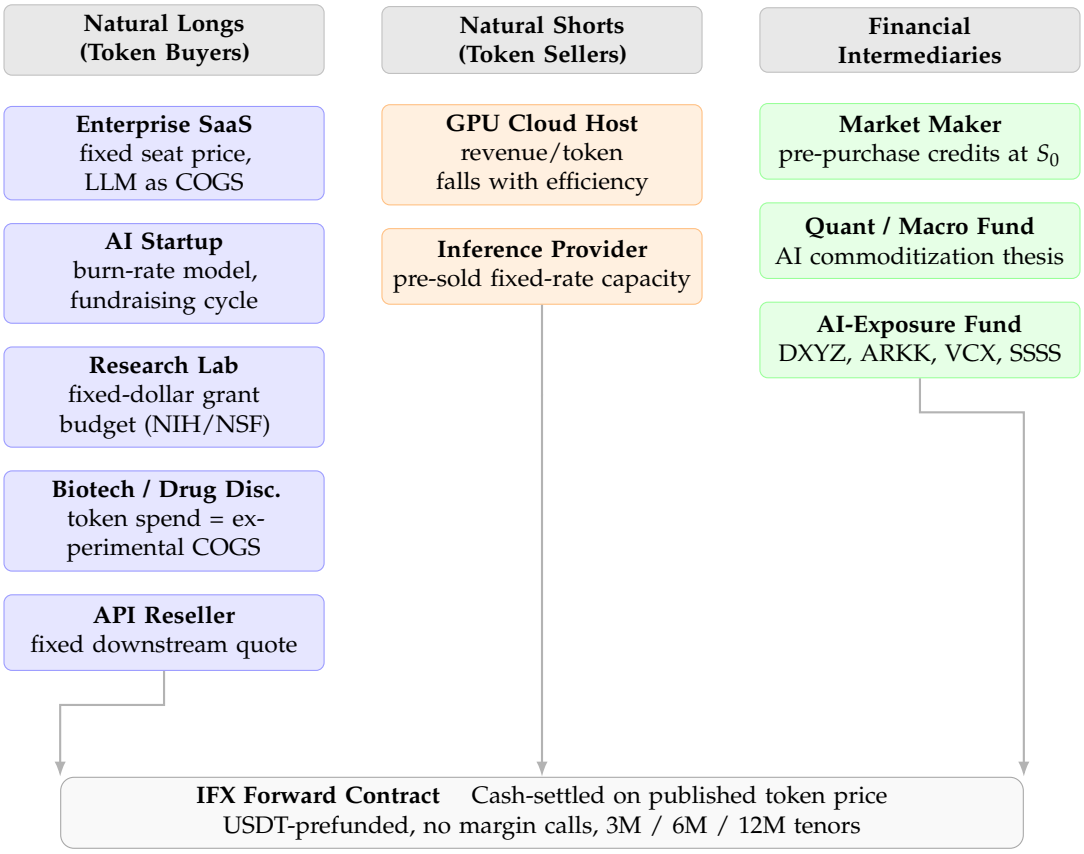

Figure 5: Market participant classification for IFX token forwards. Natural longs face the risk of price increases that compress margin on fixed-price output commitments. Natural shorts face the risk of price decreases that erode revenue on pre-sold token capacity. Financial intermediaries facilitate price discovery and provide speculative capital.

to Phase IIa positive results published in *Nature Medicine* after screening only 78 molecules in 18 months, at less than 10% of the average per-program cost [14]. For a biotech with $500,000 per quarter in AI API spend on a 24-month IND timeline, a 33% mid-cycle token price increase reduces the number of viable screening candidates by one-third. The sector's AI token spend is scaling rapidly: Eli Lilly's TuneLab platform (opened to external biotech partners in September 2025) and the subsequent $1 billion NVIDIA co-innovation lab (January 2026) signal institutional AI commitment at therapeutic scale [15].

API resellers and aggregators. Aggregators that quote fixed blended rates to downstream enterprise clients are structurally short the spread between the rate they quote and the model creator's canonical price. Any upward repricing after a quote is issued is a direct margin loss. A long forward or direct prepurchase via OpenRouter's prepaid credit facility locks the cost basis at the time of quoting.

Participant	Est. share†	Risk profile	IFX position
Enterprise SaaS	40%	Annual flat contract, LLM is COGS	Long 12M forward
Developer / startup	15%	Burn-rate model, fundraising cycle	Long 3-6M forward
Research lab	10%	Fixed-dollar grant budget	Long 6-12M forward
Biotech / drug disc.	5%	Per-experiment token budget	Long 12M forward
API reseller	10%	Fixed downstream quote	Long forward or prepurchase
VLA / physical AI	5%	Mission-critical budget certainty	Long 12M forward
GPU cloud / MM	15%	Revenue/token falls with efficiency	Short forward

† Demand tier estimates adapted from Xing [12].

Market-making backstop. Short counterparties can eliminate price risk by pre-purchasing OpenAI prepaid credits at S_0 and consuming them at expiry, making market-maker P&L equal to $(F - S_0) \times N$ - fixed and independent of F_{ref}. This makes Stage 1 viable without requiring speculative short interest from day one.

6 Contract Design

6.1 Single-Model Forwards: Rationale and Specification

The evidence from Demirer et al. [7] establishes that a forward on a single named model ID is not merely a sufficient hedge for most buyers - it is the necessary one. More than 50% of companies use exactly one model, and among multi-model users, more than 90% of spend concentrates on one. HHI values of 5,000-8,000 by use case indicate that substitution is not a realistic diversification strategy. An index forward would leave basis risk exceeding the hedge benefit for any company whose API spend is denominated exclusively in one model's tokens.

Underlying	Published USD price per 1M output tokens of a named model ID, pinned at inception. Oracle: IFX formula-driven observation at 00:00 UTC daily from the provider's canonical pricing page.
Settlement payoff	$\Pi_{\text{long}} = (F_{\text{ref}} - F) \times N$, paid in USDT at maturity. Capped at $\pm S_0 N$. F_{ref} is the median of IFX oracle observations over the 60-minute settlement window.
Models listed	GPT-4.1 ($8.00/1M out), GPT-5.4 ($15.00/1M out), GPT-5.5 ($30.00/1M out). Extended on client demand.
Notional sizes	100B, 500B, or 1T output tokens per contract.
Tenors	3M, 6M, 12M.
Prefunding	Both sides post $B = N \times S_0$ USDT into escrow at inception. No margin calls. Released at settlement.
Fee	1% of $F^{\text{ind}} \times N$ per matched position.
Settlement	Non-discretionary; formula-driven; no manual overrides permitted.

Worked example. GPT-4.1, 1T tokens, 6-month tenor, mid-cycle calibration. $S_0 = \$8.00$, $F^{\text{ind}} = \$8.00 \times 0.931 = \$7.45$, prefunding per side: \$8M USDT, fee: \$74,500. If GPT-4.1 is cut 46% to \$4.32, the long receives +\$3.13M. If no repricing occurs, zero transfer and both sides recover full prefunding.

6.2 Black (1986) Conditions

Xing [12] audits the Black (1986) checklist for futures market viability. All five conditions are satisfied for IFX token forwards. (1) *Homogeneous commodity*: a named model's output token is identical across consumers and

delivery channels. (2) *Price uncertainty*: mid-cycle cuts of 25-67% have occurred twice in the closed-source series, with no advance notice. (3) *Hedger demand*: $8.4 billion in annual enterprise API spend, with over 50% of firms structurally concentrated on a single model. (4) *Speculator supply*: GPU cloud hosts, inference providers, quant funds, and macro traders with AI commoditization views. (5) *Delivery feasibility*: cash settlement against the published canonical price removes any requirement for physical token delivery.

6.3 Stage 3: Volume-Weighted Index Contracts

Single-model forwards cannot serve a demand cohort that is growing at the fastest rate: investment funds with diversified exposure to the AI sector. Venture capital funds, growth-stage PE funds, and AI-focused ETFs (DXYZ, ARKK, ARKW, SSSS, VCX) hold equity across portfolios of AI-native companies. Their financial exposure to token repricing is not to any single model but to the portfolio-weighted average of token costs across all their investee companies. When OpenAI cuts GPT-4.1 by 33% and Anthropic simultaneously cuts Claude Opus by 67%, a fund holding equity in ten AI startups faces a blended exposure that no single-model forward can hedge.

The IFX Company AI Token Index captures this:

$$I_{c,t} = \sum_{h} \pi_{c,h,t} \cdot \sum_{m \in \mathcal{M}_{\text{active}}} w_{m,c,h,t} \cdot P^{\text{out}}_{m,c,h,t}. \tag{8}$$

Notation. $I_{c,t}$ is the composite token price index for company (or portfolio) c at time t; h indexes the procurement channel $h \in \{\text{direct API, cloud marketplace, aggregator}\}$; $\pi_{c,h,t}$ is the channel weight, representing the fraction of c's total token spend procured through channel h; $w_{m,c,h,t}$ is the model weight within channel h, representing model m's share of tokens consumed by c through that channel; $P^{\text{out}}_{m,c,h,t}$ is the observed output price per million tokens for model m in channel h; and $\mathcal{M}_{\text{active}}$ is the set of active model IDs in the current rebalancing period.

Channel weights and model weights are derived from observed spend data:

$$\pi_{c,h,t} = \text{Normalize}\left(\textstyle\sum_s r_s \cdot \text{Share}_{c,h,s,t}\right), \tag{9}$$

$$w_{m,c,h,t} = \text{Normalize}\left(\textstyle\sum_s r_s \cdot \text{Volume}_{m,c,h,s,t}\right). \tag{10}$$

Here, r_s is the reliability weight assigned to data source s (where sources include direct API usage logs, cloud marketplace invoices, and aggregator routing data), and $\text{Share}_{c,h,s,t}$ and $\text{Volume}_{m,c,h,s,t}$ are the spend fraction and token volume observed for company c in channel h from source s at time t. Rebalancing is scheduled monthly on the first business day and triggered on model deletion, rename, or a weight-threshold breach. Governance is non-discretionary throughout.

The Xing [12] Token Price Index (TPI) provides a simplified public reference:

$$\text{TPI}_t = \sum_i w_i \cdot P_{i,t}, \qquad w_i = \frac{V_i}{\sum_j V_j} \text{ (capped at 30\% per constituent)}, \tag{11}$$

where w_i is the volume-weighted share of model i in total observed token consumption (capped to prevent single-model dominance), and $P_{i,t}$ is the capability-adjusted price $P^{\text{raw}}_{i,t} \cdot S_{\text{SIT}} / S_i$ normalised to the Standard Inference Token (SIT) quality anchor, set to GPT-4-Turbo January 2024 benchmark scores.

7 Roadmap

The three-stage roadmap follows the natural sequencing of market liquidity formation and data infrastructure maturity.

Stage	Timeline	Product	Rationale and gate
1	Now - Q4 2026	Pinned OTC forwards on GPT-4.1/5.4/5.5. Full USDT prefunding. 3M/6M/12M tenors.	Serves the >50% of firms locked to a single model. No oracle data history or exchange partnership required on day one. Gate: 30 simultaneous positions, 5 completed settlements without dispute.
2	Q1 2027 - Q2 2028	Exchange-listed futures. Poisson-calibrated margin (8-15% initial). CCP or smart-contract clearing. Open order book.	Widens the participant set to hedge funds and discretionary macro traders. Aligns with the 2027-2028 optimal window identified by Xing [12]. Gate: 6 months daily price stability across two or more providers, audited oracle, documented cross-provider basis.
3	2028+	SIT-normalised TPI index contracts. IFX Company AI Token Index. Index governance committee. Exchange partnership.	Addresses the AI-exposure fund demand that single-model forwards cannot serve. VCs and AI-focused ETFs holding diversified portfolios need a single instrument to hedge the blended repricing risk across investee companies. Gate: 12 months cross-provider price history, stable normalisation methodology.

Stage 1 infrastructure. Bilateral ISDA 2002 or equivalent smart-contract escrow. Commodity forward exemption opinion under Dodd-Frank obtained before first trade. Minimum operating capital: $10 million USDT to warehouse book risk and fund escrow. Target clients: API aggregators and agentic companies with $100,000 or more per month in LLM spend.

References

[1] MarketsandMarkets. AI inference market—global forecast to 2030. Market research report, 2025. Market valued at USD 106.15B in 2025; projected USD 254.98B by 2030 at 19.2% CAGR.

[2] Intercontinental Exchange. ICE and Ornn to launch GPU compute futures contracts. Press release, May 2026.

[3] CME Group. CME group and Silicon Data partner to launch first compute futures. Press release, May 2026.

[4] MUFG Americas. Financing the AI supercycle: AI chart weekly. Research note, December 2025. Hyperscaler capex forecast exceeding $600B in 2026; Goldman Sachs $1.15T cumulative 2025–2027 projection.

[5] NVIDIA Corporation. Third quarter fiscal year 2026 results. Earnings release, November 2025. Quarter ending October 2025; Data Center revenue $30.8B, +112% YoY.

[6] Ornn. Ornn compute price index added to Bloomberg terminal. PR Newswire, April 2026.

[7] Mert Demirer, Andrey Fradkin, Steven Tadelis, and Linghao Peng. The emerging market for intelligence: Demand, supply, and market structure of LLM APIs. Working paper, 2025.

[8] Business Research Insights. Large language model (LLM) market size, share and forecast to 2035. Market research report, 2026. Market estimated at USD 22.74B in 2026; projected USD 137.66B by 2035 at 22.15% CAGR.

[9] Gartner. Gartner predicts that by 2030, performing inference on an LLM with 1 trillion parameters will cost GenAI providers over 90% less than in 2025. Press release, March 2026.

[10] Ornn. Ornn raises $33M in seed funding led by a16z crypto. Press release, June 2026.

[11] South China Morning Post. China plans compute futures in Shanghai as AI computing demand surges. News article, May 2026.

[12] Wei Xing. AI token futures market: Commoditization of compute and derivatives contract design. *arXiv*

preprint arXiv:2603.21690, 2026.

[13] Jian Wu and Hao Deng. Computational challenges in token economics: The impossibility triangle. Working paper, 2026.

[14] Insilico Medicine. Positive phase IIa results for rentosertib in idiopathic pulmonary fibrosis. Nature Medicine / Press release, 2025. First drug where both target discovery and molecular design were performed entirely by generative AI; 78 molecules screened in 18 months.

[15] Eli Lilly and Company. Lilly launches TuneLab platform to give biotechnology companies access to AI-enabled drug discovery models. Press release, September 2025. Over $1B in underlying AI drug discovery research investment; NVIDIA co-innovation lab announced January 2026 with combined $1B investment.