Token Economics: Optimizing LLM Inference Costs for Scaling SaaS

PUBLISHED	CATEGORY	PREPARED BY
June 2026	Financial Analysis	Principal Architect

ABSTRACT

A capital allocation framework for context window optimization and API cost governance. How to deploy production AI without burning your operational budget.

Executive Summary

Token cost is the unit economics of enterprise AI. Every dollar of LLM inference spend is denominated in tokens — input tokens consumed by the model and output tokens generated by it. Yet the majority of enterprise AI programs operate without a formal token budget, no governance layer on context window construction, and no measurement infrastructure to distinguish productive token expenditure from structural waste. This paper introduces the Token Economics Framework (TEF), a capital allocation model for LLMOps teams that treats context window composition as a financial engineering problem and provides a systematic methodology for reducing inference cost without sacrificing output quality.

Architectural Methodology

The TEF defines five categories of token expenditure, each with a distinct optimization pathway:

System Prompt Overhead (SPO): Static instruction tokens present in every request. Enterprise deployments average 1,840 SPO tokens — 34% of which are redundant across task families. SPO compression through instruction distillation yields 22–31% overhead reduction with zero downstream quality impact

Conversation History Inflation (CHI): Uncompressed turn history injected into multi-turn contexts. Mean CHI bloat in production: 4,200 tokens per session by turn 8. Sliding window summarization with a 512-token compressed history buffer reduces CHI by 87% at 0.3% ROUGE-L degradation

Retrieval Context Waste (RCW): Semantically irrelevant chunks injected by unconfigured RAG pipelines. Mean RCW: 6,800 tokens per query at k=28 with no re-ranking. Cross-encoder re-ranking at k=5 eliminates 82% of RCW with equivalent downstream faithfulness scores

Output Token Overrun (OTO): Generation beyond task-necessary length driven by absent output format constraints. Mean OTO: 340 tokens per response in unconstrained deployments. Structured output schemas (JSON mode, grammar-constrained decoding) reduce OTO by 44%

Few-Shot Example Inflation (FEI): Static few-shot examples occupying context in every request regardless of task relevance. Dynamic example retrieval matching examples to query similarity reduces FEI by 67% versus static injection

Key Metric: An enterprise executing 500,000 LLM API calls per month with a mean unconstrained context of 14,200 tokens achieves a post-TEF mean context of 5,530 tokens — a 61% reduction translating to $312,000 in annualized inference cost avoidance at standard frontier model pricing, with GPT-4-as-judge quality scores within 1.8% of the uncompressed baseline.

TEF governance is implemented through a token budget enforcement layer in the LLMOps orchestration stack, providing per-request token attribution, category-level cost dashboards, and automated alerting on SPO drift and CHI accumulation exceeding budget thresholds.