Solving the Token Yield Crisis: How Songlines Control® Delivers Execution Efficiency for Enterprise AI

Published by: Cetus AI

Date: June 2026

Reference: CETUS-WP-004

Audience: CIOs, CFOs, and Enterprise Architecture Leaders

Executive Summary

> *"The right question is not how many tokens a system consumes. It is what useful outcome the system produces per token consumed. In other words: token yield."*
>
> — Arvind Jain, CEO at Glean, June 2026

Enterprise AI has shifted from simple chat interfaces to long-running, multi-step agentic workflows. As this shift accelerates, a new operational constraint has emerged: token consumption is scaling exponentially, but business value is not scaling at the same rate. Deloitte's 2025 Tech Value Survey found that enterprises are allocating an average of 36% of their digital budgets to AI, yet organisations like Uber have reported burning through annual AI budgets in a matter of months.

The root cause is architectural. Token usage is rarely driven by the user's prompt alone — it is driven by the scaffolding around the prompt: context retrieval, tool schemas, intermediate reasoning, and execution traces. When this architecture is inefficient, enterprises pay frontier-model prices for routine operational work.

This white paper examines the four architectural levers determining token efficiency and details how **Songlines Control®** acts as the runtime governance and orchestration layer to solve the token yield crisis.

The Token Yield Crisis: An Architecture Problem

The signals are now hard to ignore. Ramp recently reported a 4x year-over-year increase in monthly enterprise AI spend. Deloitte found that more than half of enterprises allocate between 21% and 50% of their digital initiative budgets to AI. Yet in too many organisations, token consumption is rising quickly while business value is not rising at the same rate.

The prevailing instinct is to treat this as a model selection problem — to find a cheaper model or negotiate better API pricing. This instinct is wrong. Token usage is rarely driven by the model alone. It is shaped by the full system around the model: how context is retrieved, how tools are exposed, how work is decomposed, how models are routed, and how prior execution is reused.

If that architecture is inefficient, token spend climbs even when output quality does not. The waste is not in any single prompt. It is in the design of the system. That is why token yield is fundamentally an architecture question — and why solving it requires a governance and control layer that operates at runtime.

The Four Levers of Token Efficiency

Architectural Lever	The Inefficiency Problem	The Songlines Control® Solution
Context Quality	Models process whatever they are given. Poor retrieval forces models to spend token budgets reasoning over irrelevant data rather than solving the problem.	Inline payload optimisation filters noise and redacts PII before the payload reaches the model, ensuring only high-signal context consumes tokens.
Model Routing	Defaulting to frontier models for routine operational work (search, tool selection, validation) means paying premium prices for commoditised reasoning.	Intelligent Model Routing automatically routes requests based on task complexity, cost thresholds, and data residency rules — zero code changes required.
Continual Learning	Systems solve the same class of problem from scratch every time, paying the same exploratory token cost repeatedly.	Immutable execution telemetry captures every AI interaction, enabling identification of repeatable workflows and elimination of redundant reasoning loops.
Harness Design	Naive agent harnesses accumulate context endlessly, carrying unnecessary state forward at every step, leading to context bloat and degraded reliability.	Economic Control dashboards provide real-time token consumption visibility by user, team, and workflow — with hard consumption limits to prevent runaway spend.

Lever 1: Context Quality — Eliminating the Hidden Tax of Noise

The Problem

A short user instruction can trigger a massive token bill. In agentic systems, the visible prompt is often dwarfed by the system instructions, retrieved documents, and tool schemas injected into the context window. A prompt like "Analyse churn risk for these accounts and create follow-up

tasks" may appear small, but the actual token load includes system instructions, tool schemas, retrieved documents, intermediate reasoning, execution traces, and memory.

In many enterprise systems, most of the tokens are not typed by the user at all. They are generated by the scaffolding around the task. If the retrieval architecture is noisy, the model is forced to act as a filter rather than an engine of insight — spending its token budget reasoning over irrelevant or conflicting information instead of acting on the right signal.

> *"Weaker retrieval forced the system to compensate with more tool calls, more reasoning loops, and more over-fetching. That is the hidden tax of poor context architecture."*
>
> — Arvind Jain, CEO at Glean

The Songlines Control® Response

Songlines Control sits between the enterprise and the model. Before a request is processed, the platform inspects the payload and applies inline redaction for sensitive data — protecting compliance while simultaneously reducing payload size. Organisations can enforce payload size limits and content quality rules at the infrastructure layer, ensuring that only high-signal, compliant context reaches the model. The result is a direct reduction in the "noise tax" and a dramatic improvement in the ratio of useful output to tokens consumed.

Lever 2: Model Routing — Right-Sizing Intelligence

The Problem

A large share of enterprise AI work is operational: search, retrieval planning, tool selection, validation, and execution management. These steps are critical, but they do not require the reasoning capabilities of a frontier model. When every step in an agentic workflow defaults to GPT-4o or Claude 3.5 Sonnet, the enterprise is effectively paying frontier prices for routine work. As usage scales, this becomes one of the most significant drivers of AI cost inefficiency.

The Songlines Control® Response

Songlines Control features an Intelligent Model Routing engine that acts as a dynamic switchboard. It evaluates the incoming request against enterprise-defined policies and routes it to the most appropriate model. Routine summarisation or data extraction is routed to faster, cheaper models. Complex reasoning tasks are preserved for frontier models. Sensitive data is automatically routed to sovereign, locally-hosted instances.

Routing Rule Type	What It Does	Token Yield Impact
Cost-based routing	Routes requests below a complexity threshold to lower-cost models automatically	Average 40% cost reduction on mixed workloads
Latency-based routing	Routes time-sensitive requests to faster, lighter models	Reduces end-to-end workflow time without sacrificing quality
Sovereignty routing	Ensures regulated data never leaves Australian data residency	Eliminates compliance risk while maintaining operational efficiency
Budget threshold routing	Automatically falls back to cheaper models when monthly budget thresholds are reached	Prevents budget overruns without interrupting service delivery

All routing is executed entirely at the infrastructure layer — with zero changes to the underlying application code.

Lever 3: Continual Learning — Compounding Execution Efficiency

The Problem

Human workers do not solve the same problem from scratch every time. They document processes, reuse successful approaches, and build institutional knowledge. Enterprise AI systems, however, often pay the same exploratory token cost repeatedly for similar tasks. Every execution produces signal about how similar work should be done next time — which tools were useful, which retrieval path worked, which steps were unnecessary. If that signal is not captured and reused, the system keeps paying the same exploratory cost again and again.

The Songlines Control® Response

Through its Immutable Audit Trail, Songlines Control captures a cryptographically signed record of every AI interaction — including the prompt, the model used, the tokens consumed, the latency, the policy decision applied, and the output. This rich telemetry provides the foundation for continual learning. Enterprise architecture teams can analyse this data to identify highly

repeatable workflows, refine system prompts, and cache common responses — ensuring that the system compounds in efficiency over time rather than repeating costly exploratory work.

The best enterprise AI systems will compound in this way. Each completed task should improve the economics of the next related one. Songlines Control provides the data infrastructure to make that compounding possible.

Lever 4: Harness Design — Managing Context Bloat

The Problem

As agents take on longer-running, multi-step work, the harness — the system managing the agent's state — becomes a major determinant of both quality and cost. A naive harness keeps expanding the active context window. It carries more instructions, more tools, more state, and more intermediate outputs forward at every step. Cost grows as the workflow grows. Reliability usually degrades too.

The Songlines Control® Response

Songlines Control provides the visibility and control necessary to manage context bloat at the infrastructure layer. Through its Economic Control dashboard, organisations gain real-time, month-to-date visibility into token consumption at the user, team, and workflow level. When a specific agent or workflow begins exhibiting context bloat — characterised by exponentially rising token counts without corresponding output — administrators can intervene immediately: setting hard consumption limits, rerouting the workflow, or escalating for human review.

Economic Control Capability	What It Provides	Business Impact
Real-time MTD spend dashboard	Month-to-date token consumption and cost by model, user, and workflow	CFO-ready AI cost reporting without manual aggregation
Hard budget limits	Automatic enforcement of spend thresholds per team or workflow	Prevents the "Uber scenario" — burning through annual budgets in months
Cost attribution	Every token attributed to a specific user, workflow, and business unit	Enables chargeback and accurate ROI measurement by AI initiative
Anomaly detection	Alerts when token consumption patterns deviate from baseline	Early warning system for runaway agentic workflows before they become budget incidents

Conclusion: Execution Efficiency is the Real AI Moat

The CIOs and CFOs who succeed in the next phase of enterprise AI will not be those who simply deploy the most models. They will be those who master execution efficiency.

Token yield is fundamentally an architecture question. It requires a control plane that can manage context, route intelligently, capture execution telemetry, and enforce economic boundaries — operating at runtime, before the request reaches the model.

Songlines Control® delivers this control plane. It transforms AI from an unpredictable, opaque cost centre into a governed, highly efficient enterprise capability — one that compounds in efficiency over time, produces evidence for regulators and boards, and ensures that AI investment generates proportional business value.

The organisations that build this infrastructure now will not just reduce their AI costs. They will define how their enterprise competes in the AI era.

To see how Songlines Control® can address the token yield challenge in your organisation, contact us to request a 14-Day Control Baseline engagement.

Contact:
Cetus AI | Brookwater, QLD
contact@cetusai.com.au | www.cetusai.com.au