THE HIDDEN TOLL OF ARTIFICIAL INTELLIGENCE: COMMERCIAL, LEGAL, AND SOVEREIGN RISKS IN ENTERPRISE TOKEN ECONOMICS

Vidhi Patel[1], Symbiosis Law School, Nagpur

The End of Predictable SaaS, the Rise of the "Reasoning Tax," and the Legal Frameworks Required to Prevent Runaway Agent Liabilities

ABSTRACT

The transition of enterprise software procurement from predictable, flat-fee Software-as-a-Service ("SaaS") licensing to variable, consumption-based artificial intelligence ("AI") "token" billing represents one of the most consequential and least understood shifts in corporate technology governance. This paper examines three interlocking dimensions of that shift. First, it analyses the commercial mechanics of token economics, including the input–output cost asymmetry and the emergence of an invisible "reasoning tax" generated by chain-of-thought processing in frontier models, illustrated through documented enterprise incidents including Uber's exhaustion of its annual artificial intelligence budget within four months. Second, it interrogates the legal vacuum surrounding autonomous "agentic" artificial intelligence, arguing that traditional agency law and vendor indemnification structures are unequipped to allocate liability for runaway, self-compounding token consumption, and proposes that enterprise AI procurement be restructured along the lines of Engineering, Procurement and Construction ("EPC") contracts, incorporating hard spend caps, circuit breakers, liquidated damages, and compute escrow mechanisms. Third, it examines the data sovereignty and privacy risks created by "prompt caching," including timing side-channel vulnerabilities, the unresolved status of the right to erasure under the General Data Protection Regulation, and the specific compliance obligations arising under India's Digital Personal Data Protection Act, 2023. The paper concludes that corporate legal and financial leadership must treat artificial intelligence procurement as high-risk infrastructure rather than conventional software licensing, and offers a set of contractual mechanisms designed to close the liability gap before it manifests as a balance-sheet event.

Keywords: AI token billing; reasoning tax; agentic AI; vendor liability; EPC contracts; data sovereignty; prompt caching; Digital Personal Data Protection Act, 2023.

[1]* B.A. LL.B. (Hons.), Symbiosis Law School, Nagpur. All views expressed are personal.

I. INTRODUCTION

The global enterprise software market, long accustomed to the predictable rhythms of user-based subscription models, is undergoing a structural, irreversible upheaval. Valued at nearly $209.95 billion in 2024, the Software-as-a-Service ("SaaS") industry built its formidable foundations on predictable recurring revenue and flat-fee licensing. For decades, the fundamental economic premise of software was the zero marginal cost of distribution: once code was written, adding an additional user to a web-based dashboard cost the vendor practically nothing. Consequently, enterprise procurement teams enjoyed absolute budgetary certainty, and SaaS vendors enjoyed gross margins frequently exceeding eighty percent.

The integration of advanced Artificial Intelligence fundamentally fractures this model. In the era of autonomous agents and advanced reasoning models such as OpenAI's o3 series, GPT-5, and Anthropic's Claude architecture, compute is no longer a marginal, ignorable overhead — it is a highly volatile, strictly metered commodity. For Chief Financial Officers, Chief Information Officers, and corporate legal departments, the shift from predictable SaaS to variable, volatile "token" economics represents a multi-million-dollar exposure vector. Artificial intelligence deployment is no longer a standard software procurement exercise; it has become an infrastructure project, carrying the financial risks of large-scale supply-chain logistics and the legal liabilities of autonomous, fast-moving machinery.

This paper deconstructs the hidden commercial and legal risks of AI token billing. Part II examines the paradigm shift from predictable SaaS to tokenized software pricing. Part III analyses the insidious margin drain of the "reasoning tax." Part IV documents the unprecedented corporate liability generated by runaway autonomous agents. Part V interrogates the failure of traditional agency law to allocate that liability. Part VI proposes a legal firewall modelled on infrastructure contracting. Parts VII and VIII address M&A due diligence and the paradox of AI-assisted legal drafting. Parts IX and X address the silent compliance crisis emerging around data sovereignty, prompt caching, and India's Digital Personal Data Protection Act, 2023. Part XI situates these risks against the substantial cost advantages AI still offers when properly governed, and Part XII concludes with recommendations for enterprise leadership.

II. THE PARADIGM SHIFT: THE END OF PREDICTABLE SAAS AND THE TOKENIZATION OF SOFTWARE

Traditional SaaS pricing relies heavily on flat subscriptions or per-seat licenses, models in which customers pay a fixed fee tied directly to the number of human users interacting with the system. This strategy fails for AI products because the value customers derive, and the costs the vendor incurs, scale directly with computational usage in ways that flat pricing cannot absorb. When an AI product generates outputs — whether summarising financial reports, analysing legal contracts, generating code, or processing audio — it consumes substantial Graphics Processing Unit ("GPU") cycles. This consumption is measured and billed in "tokens," a mathematical representation of natural language roughly corresponding to parts of words, image patches, or seconds of audio.

A. The Input–Output Cost Asymmetry

The fundamental economic challenge of AI billing lies in the input and output cost asymmetry. Large Language Model providers universally charge different rates for reading data (input tokens) versus generating data (output tokens). Output tokens are priced significantly higher, owing to the intensive compute required for autoregressive generation, in which the model must calculate the probability of the next token iteratively. Anthropic's Claude Sonnet 4.5, for instance, charges $3.00 per million input tokens but $15.00 per million output tokens[2] — a five-fold gap that is rarely disclosed prominently in vendor sales materials, leaving enterprise procurement teams that benchmark against flat-rate SaaS pricing structurally disadvantaged from the outset.

A common and destructive pricing mistake is attempting to absorb these variable compute costs within a traditional flat-fee, per-seat licence. Where an enterprise user heavily utilises an AI copilot — for instance, a software engineer generating thousands of lines of code daily — the underlying API token costs routed to the foundational model provider can rapidly exceed the fixed monthly subscription fee paid to the SaaS vendor. This dynamic creates a margin problem that compounds precisely as a company's most engaged customers use the product more intensively. A deployment covering 5,000 knowledge workers averaging 65,000 tokens per workday can generate an indicative annual token spend between $700,000 and $1.5 million on the underlying infrastructure layer alone, distinct from application licensing costs.

[2]See Anthropic, Claude Pricing (2026), https://www.anthropic.com/pricing (on file with author) (setting out per-million-token rates for input and output tokens across the Claude model family).

B. The Rise of Hybrid and Usage-Based Billing Infrastructure

To survive this margin erosion, mature AI SaaS businesses are converging on hybrid pricing models: a predictable base fee coupled with a variable usage component or a capped "inference allowance." This transition has driven a rapid proliferation of specialised enterprise billing software designed to decouple metering logic from payment gateways, permitting the ingestion of billions of daily usage events and nuanced overage calculations. The table below summarises the principal categories of platform now competing in this space.

Billing Platform	**Primary Enterprise Persona**	**Architectural Advantage**	**Vendor Lock-in Risk**
Flexprice	Product, Finance, and RevOps	Open-source foundation; decoupled billing logic; transparent usage-aligned pricing.	Low
Solvimon	AI-native companies	Built specifically for token metering, credit systems, and hybrid contracts.	Low to Moderate
Stripe Billing	Early-stage / developer-first	Deep native integration with Stripe Payments; fast setup.	High
Orb	High-growth engineering teams	High-throughput event processing for complex overages and scaling.	Moderate
Maxio	Finance-led organisations	Integrates ASC 606 revenue compliance with real-time token metering.	Moderate

For enterprise procurement, this shift means that software contracts are no longer simple licensing agreements; they are complex consumption models requiring rigorous oversight, financial operations ("FinOps") strategy, and strict usage caps to prevent what the industry terms "bill shock."

III. THE "REASONING TAX": THE INVISIBLE COMPUTE METER

If variable token pricing introduces financial unpredictability into the enterprise software stack, the advent of "reasoning models" introduces an entirely hidden dimension of

compounding cost. Frontier foundational models — including OpenAI's o1, o3-mini and GPT-5 series, and advanced iterations of the Claude series — utilise internal "chain-of-thought" processes to solve complex algorithmic, scientific, and logic problems before presenting a final answer to the user. This architectural leap carries a severe commercial consequence that this paper terms the "Reasoning Tax."

A. Hidden Reasoning Tokens and Their Billing Mechanics

When a user prompts a standard model, the Application Programming Interface ("API") charges strictly for the input prompt and the visible text generated in response. Modern reasoning models, however, generate invisible tokens during an internal "thinking" phase. While never displayed to the end user, these tokens occupy space within the model's context window and, critically, are billed at the expensive output-token rate, categorised broadly as "completion tokens" on the billing dashboard.

Consider an enterprise agent reconciling a legal bill of materials against architectural drawings. The user inputs a prompt and context documents totalling 500 input tokens. The model spends forty-five seconds "thinking," generating an internal chain of thought to deduce that "PT" denotes pressure-treated wood across differing legends — an internal monologue totalling 4,000 reasoning tokens — before outputting a concise, accurate answer of 200 visible tokens. Under a non-reasoning architecture, the enterprise would be billed for 700 total tokens; using a reasoning model, it is billed for 4,700 tokens. The overwhelming majority of the financial cost is generated invisibly, behind a proprietary veil.

B. The Compounding Effect of Conversation History

This reasoning tax is aggressively compounded by the mechanics of context windows. AI models are inherently stateless and possess no intrinsic memory of a conversational thread; to maintain context, every sequential interaction requires the entire preceding conversation history to be resent as input tokens. At Turn 1, a 100-token prompt is billed as 100 input tokens. At Turn 2, the system must resend the first 100 tokens, and the enterprise is billed for 200 input tokens. By Turn 10, a fresh 100-token prompt requires resending 900 historical tokens, producing a bill of 1,000 input tokens. Combined with hidden reasoning costs, a single employee debugging a codebase or drafting a complex contract can unknowingly consume tens of thousands of tokens within minutes — a multiplier that, distributed across thousands of

knowledge workers, transforms apparently routine usage into a substantial financial liability.

C. Case Study: Uber and the Exhaustion of an Annual Artificial Intelligence Budget

The clearest documented illustration of the reasoning tax at enterprise scale is Uber's experience with Claude Code. In December 2025, Uber extended access to approximately 5,000 engineers; an internal leaderboard ranking engineers by AI usage volume drove adoption from thirty-two to eighty-four percent of the eligible workforce. By April 2026 — four months into the fiscal year — Uber had exhausted its entire annual AI budget, with monthly per-engineer spend ranging from $150 to $2,000 depending on usage intensity[3]. Uber's Chief Operating Officer publicly conceded that it was "very hard to draw a line" between rising AI costs and tangible customer-facing features, an admission later corroborated by the company's Chief Technology Officer. In direct response to incidents of this kind, Anthropic began billing agent tool usage at standard API rates separately from June 15, 2026.

The broader empirical picture confirms that Uber's experience was not anomalous. In 2025 alone, AI-native software spending nearly doubled, yet approximately eighty percent of enterprises missed their AI cost forecasts — a governance failure that ought to concern any board audit committee. Reasoning models can cost between five and twenty times more per task than standard models on account of internal thinking overhead alone; in a direct comparison of AI coding agents, one system consumed approximately 6.2 million tokens against a competitor's 1.5 million tokens for comparable tasks, a four-fold gap attributable almost entirely to reasoning verbosity.

IV. THE RUNAWAY AGENT NIGHTMARE: INFINITE LOOPS AND BILL SHOCK

While human employees generating unseen reasoning tokens present a serious budget-control problem, the deployment of autonomous "Agentic AI" introduces an existential financial and operational risk for which most corporate structures are entirely unprepared. Agentic AI refers to systems comprising multiple autonomous software agents capable of reasoning through complex goals, orchestrating multi-step plans, navigating external interfaces, executing transactions, and adapting workflows autonomously, without continuous human input. These agents operate at machine speed but, as systems architects have observed,

[3]See Fortune, Inside Uber's AI Budget Crisis (May 26, 2026); see also Fortune, "Tokenmaxxing" Is Over (May 28, 2026) (reporting enterprise disillusionment with undifferentiated growth in token consumption).

with "toddler judgment" — lacking human intuition, risk aversion, or an intrinsic understanding of financial consequence.

A. The Mechanics of Recursive Token Consumption

In traditional software engineering, an infinite loop is a bug that causes a program to run endlessly until it crashes or is terminated by the operating system. In the AI API economy, however, an infinite loop caused by an autonomous agent does not merely crash a local machine; it runs up a cloud-based token meter at thousands of transactions per second, generating expensive API calls continuously. Because an agent generates a fresh chain of thought for every retry and resends the entire history of its failed attempts, token consumption grows exponentially with each iteration. As modern AI APIs generally lack default, natively configured cost-linked rate limits, an agent trapped in a recursive loop over a weekend can generate six- or seven-figure billing overages before human oversight detects the anomaly on a Monday morning — and because cloud providers bill for consumed tokens regardless of whether the output was useful, the enterprise remains legally liable for the resulting invoice.

B. Documented Incidents of Runaway Billing

In November 2025, four agents in a research pipeline entered a recursive loop: an analyser and a verifier exchanged requests for eleven days, with the team assuming that escalating costs reflected organic growth. The final invoice was $47,000. No internal circuit breaker existed, and the loop was stopped only when a human noticed the anomaly on the billing dashboard[4]. Three months later, in February 2026, a data enrichment agent misread an API error code as an instruction to retry with different parameters and executed 2.3 million API calls over a single weekend; it was halted not by the enterprise's own controls, which did not exist, but by the external API provider's own rate limiter.

These incidents sit within a broader industry retrenchment. Google slashed Gemini API free-tier quotas by fifty to eighty percent in late 2025 and introduced enforced billing caps in April 2026, with Microsoft following shortly after — both moves amounting to an industry-wide acknowledgment that unmetered or lightly metered AI access was commercially

[4]See industry incident reporting compiled from enterprise FinOps case studies, Apr. 2026 (on file with author) (documenting an eleven-day recursive loop between two agents in a research pipeline resulting in a $47,000 invoice, and a separate February 2026 incident in which a misconfigured data-enrichment agent executed 2.3 million API calls over a single weekend).

unsustainable, though enterprises were given only days, rather than months, of notice before cost structures changed. By May 2026, commentators were openly describing the end of "tokenmaxxing" — the practice of maximising token volume as a proxy for AI productivity — as companies discovered that the anticipated return on large-scale AI consumption had failed to materialise.

V. THE LIABILITY SQUEEZE AND THE FAILURE OF TRADITIONAL AGENCY LAW

When human error results in financial loss or contractual breach, traditional corporate frameworks and employment law allocate responsibility efficiently. When an AI agent makes a commitment that triggers a massive financial liability or regulatory violation, the legal landscape is fraught with ambiguity — what legal commentators term the "AI Vendor Liability Squeeze." In traditional jurisprudence, "agency" refers to a fiduciary relationship in which one party is authorised to act on behalf of another to create legal relations with third parties; where a human agent goes rogue or exceeds the scope of actual or apparent authority, established doctrine provides a structured allocation of liability, often permitting the principal to pursue the individual agent for breach of fiduciary duty or fraud.

AI agents, however, are not recognised as legal entities. They cannot be sued, cannot hold insurance policies, and cannot bear fiduciary responsibility. When an autonomous software agent executes a faulty transaction, signs a disadvantageous contract, or causes a data breach by hallucinating a compliance approval, courts are increasingly applying the doctrine of vicarious liability and respondeat superior directly to the deploying organisation: the enterprise that deployed the agent, configured its permissions, and controlled its environment bears full responsibility for its actions.

Enterprises have historically relied on two safety nets when deploying enterprise software: vendor indemnification and comprehensive corporate liability insurance. Both are failing simultaneously in the AI era. First, AI platform providers are aggressively rewriting their terms of service and master service agreements to limit liability, disclaiming the accuracy of outputs and shifting the risk of hallucinations, intellectual property infringement, and runaway loops back to the customer; vendor liability is frequently capped at fees paid in the preceding twelve months, a mathematically inadequate shield where a runaway agent causes downstream damages worth tens of millions of dollars. Second, commercial insurance

underwriters, recognising that the risk profile of autonomous AI lacks historical actuarial data, are actively carving AI-related workflows out of standard liability policies[5]. As one industry analyst has put it: vendors will not accept liability, and insurers are pulling back from AI coverage, leaving businesses that deploy agents into high-stakes workflows unprotected by either their software vendor or their insurer.

For the corporate deployer, the reality is stark: every decision the AI makes is legally the corporation's decision, and every commitment it makes is the corporation's liability. An enterprise that relies solely on legacy vendor liability caps will inevitably find itself holding full responsibility for algorithmic failures it can neither control nor perfectly audit.

VI. THE LEGAL FIREWALL: PROCURING AI LIKE MASSIVE INFRASTRUCTURE

Given the scale of financial exposure created by token economics, and the liability void surrounding autonomous agents, traditional software procurement agreements are no longer sufficient. Corporate legal teams must restructure enterprise AI procurement by drawing on a different industry: heavy infrastructure and industrial construction. When a corporation builds a billion-dollar power plant or refinery, it does not use a standard service agreement; it utilises an Engineering, Procurement, and Construction ("EPC") contract, designed to manage high-complexity, massive-scale projects through structured risk allocation, guaranteed maximum prices, and rigorous performance metrics that transfer execution risk away from the project sponsor and onto the contractor. The procurement of enterprise AI compute at scale merits the same contractual severity.

A. Take-or-Pay Compute Contracts and Provisioned Throughput

In energy and infrastructure markets, a "take-or-pay" contract ensures that a buyer either takes a minimum guaranteed quantity of a resource or pays a predefined penalty if it does not, guaranteeing base revenue for the supplier while securing dedicated access for the buyer. As competition for enterprise AI compute intensifies, hyperscalers are moving away from simple pay-as-you-go models toward Provisioned Throughput Units and reserved-capacity agreements, under which enterprises pay for deployed, reserved GPU capacity rather than per-

[5]See remarks of industry risk analysts on AI vendor liability and insurance retrenchment, cited in enterprise AI governance commentary (2026) (on file with author).

token consumption. While this guarantees latency and uptime, it functions as a strict take-or-pay contract: if the enterprise fails to utilise the reserved capacity, it remains billed for the full deployed count — a multi-million-dollar commitment regardless of actual utilisation — and procurement teams must therefore subject Provisioned Capacity commitments to rigorous, real-world load testing before signature.

B. Contractual Hard Caps, Circuit Breakers, and Efficiency SLAs

To protect the enterprise against the runaway agent nightmare, corporate lawyers must negotiate three technically enforceable mechanisms directly into the Master Services Agreement. First, contractual hard caps: absolute financial ceilings which state explicitly that the vendor cannot legally bill the enterprise beyond a specified monthly amount without documented, executive-level authorisation, and which shift liability — rather than merely trigger a notification — once the threshold is breached. Second, platform-level circuit breakers: automated technical mechanisms, warranted in the contract itself, that sever API access when token burn rates exhibit anomalous spikes indicative of an infinite recursive loop; a contractual cap tested only after the first surprise invoice provides no operational margin and must instead be tested during the pilot phase. Third, "Efficiency SLAs," which legally require the vendor to pass down token efficiency savings and model optimisation benefits over the life of a multi-year contract, preventing the vendor from retaining margin improvements while the enterprise continues to pay legacy rates.

C. Further Legal Innovations: Compute Escrow, Liquidated Damages, Model-Freeze and Audit Rights, and AI Spend Insurance

Beyond the three core mechanisms above, a set of further contractual innovations — each borrowed directly from infrastructure and construction practice — merits inclusion in enterprise AI agreements. A compute escrow clause requires the vendor to maintain a pre-funded reserve of tokens purchased at agreed rates, which the enterprise draws down rather than paying on a post-consumption invoice; when the escrow depletes, service pauses, eliminating the post-facto billing-shock problem entirely. A liquidated damages clause for loop-caused overruns operates exactly as such clauses do in construction delay contexts: where a system-traced runaway loop is attributable to a model malfunction or API orchestration failure rather than user misconfiguration, the vendor pays a pre-agreed sum per million excess tokens billed. A model-freeze clause with version audit rights requires the vendor to provide

ninety days' written notice, together with token-efficiency impact estimates, before upgrading the underlying model powering a contracted service, and preserves the enterprise's right to remain on the prior model at the prior price for an agreed transition period. Token consumption audit rights, mirroring cost-record audit rights in large construction contracts, entitle the enterprise to a full token-consumption log itemised by agent, workflow, and task, with the vendor's failure to produce that log within seven business days constituting a presumption of billing error. Finally, an emerging market for AI spend insurance allows enterprises to negotiate vendor co-payment of cyber or AI insurance premiums covering runaway-compute events, aligning vendor incentives with the construction of better circuit breakers.

VII. M&A DUE DILIGENCE AND AI-SPECIFIC CONTRACT CLAUSES

As AI permeates the corporate environment, the acquisition of technology companies through mergers and acquisitions carries novel risks, and acquirers are increasingly demanding tailored representations and warranties to mitigate AI-related uncertainty. Indemnity reversals for intellectual property require enterprise buyers to push back against vendor attempts to shift liability for copyright infringement and data breaches onto the user, forcing the vendor instead to bear the financial burden of third-party claims stemming from its foundational training data or algorithmic bias. Human-in-the-loop mandates require that, for any workflow involving material financial consequence or sensitive protected classes, the vendor's architecture enforce mandatory human oversight — a documented defence against vicarious liability claims as courts expand direct accountability for discriminatory algorithmic outcomes. Data lineage and model-training restrictions must explicitly prohibit the vendor from using the enterprise's proprietary input data, prompts, or generated outputs to train or fine-tune future model iterations. Finally, audit rights and traceability provisions must guarantee immutable metadata logs tracing every agentic decision, without storing raw, sensitive content in violation of privacy law — a prerequisite to reconstructing, after the fact, exactly what an agent did and why.

VIII. THE PARADOX OF AI LEGAL TECHNOLOGY

A notable irony of the AI procurement challenge is that corporate legal departments are simultaneously attempting to use AI to manage the very contracts that govern AI. Tools powered by natural language processing and generative AI can summarise agreements, classify contract types, extract obligations, compare clauses against corporate playbooks, and draft

first-pass redlines within seconds, offering substantial theoretical efficiency for procurement analysts and in-house counsel overwhelmed by manual review cycles. Yet the industry is discovering that this efficiency is frequently illusory: enterprise workflows reliant on unstructured email threads, disparate PDF repositories, and version control maintained through disjointed file names were built for an analogue era, and AI cannot compensate for a chaotic, unstructured data environment. Generic models trained on the open internet may produce legally sound limitation-of-liability language that is nonetheless commercially disastrous for a specific, high-leverage deal, and the risk of hallucination — where the model invents obligations or misinterprets jurisdictional nuance — remains high. The core act of drafting therefore continues to require the same human-in-the-loop oversight demanded of engineering procurement; relying on AI to draft AI procurement contracts without strict playbook guardrails is a recursive risk the modern enterprise cannot afford.

IX. DATA SOVEREIGNTY IN THE AGE OF PROMPT CACHING

While token costs, agent liability, and contract structuring dominate the operational and financial risk landscape, a silent, highly technical crisis is unfolding within enterprise compliance departments regarding data privacy, centred on the optimisation technique known as "prompt caching." To reduce latency and the cost of repeated inference, LLM providers store the mathematical representation of a submitted context window — for example, a hundred-page master service agreement — as Key-Value tensors, which reside temporarily in a GPU's video memory or on GPU-local solid-state storage. Where a follow-up question is asked within a defined time-to-live window, ranging from five minutes to twenty-four hours, the model retrieves the cached tensors instantly, reducing both latency and token cost. This efficiency, however, introduces largely unresolved data-residency and security vulnerabilities.

A. Security Threats: Timing Side-Channel Attacks and PROMPTPEEK

Because cached prompts bypass the computational process of re-computation, cache-hit requests return answers significantly faster than uncached requests, and malicious actors sharing the same multi-tenant cloud infrastructure can exploit this differential through timing side-channel attacks to deduce what other enterprises are processing. A 2025 study conducted at Stanford University audited commercial LLM APIs and found that seven of eight caching-enabled providers shared caches globally across different users and organisations[6]. A related

[6]See Stanford University, Timing Side-Channel Vulnerabilities in Commercial LLM Prompt Caching (2025) (on file with author); see also the study describing the PROMPTPEEK attack, demonstrating token-level

study detailing the so-called PROMPTPEEK attack demonstrated that an attacker with standard, non-privileged API access could elevate such timing attacks from mere detection to complete input reconstruction, extracting a victim's cached prompt token-by-token with accuracy reported as high as ninety-nine percent, including sensitive personally identifiable information such as medical data.

B. The GDPR Article 17 Failure and the Residency Blind Spot

Beyond the immediate security threat, prompt caching creates an unmanageable regulatory blind spot. Cloud providers rarely specify the physical hardware location of ephemeral Key-Value cache tensors; an enterprise may contractually require processing within a specific sovereign zone, yet if the provider routes cache memory dynamically across shared regional nodes, sensitive corporate data may cross geopolitical borders without triggering any conventional audit log or alert. Under the European Union's General Data Protection Regulation, Article 17 guarantees a right to erasure[7], yet no major LLM provider currently offers an eviction application programming interface allowing enterprise users to trigger immediate, verifiable deletion of individual cached entries from GPU memory, leaving it unclear whether expired caches are promptly zeroed out or merely lazily overwritten — a state of affairs difficult to reconcile with the storage-limitation principle.

C. Zero-Retention and Sovereign Cloud Architectures

Reliance on standard vendor assurances that customer data is not used for model training is legally and technically insufficient. Even where data is not actively used for training, if it is cached, logged, buffered by middleware, or retained for abuse monitoring for a period of weeks, it remains susceptible to data breaches, side-channel attacks, and foreign law-enforcement process, including subpoenas compelled under the United States CLOUD Act. To mitigate this exposure, stringent enterprises are enforcing "zero-retention architectures," under which prompt and output content is processed entirely in volatile memory and discarded the moment a response is generated, with no conversation logs, content-level analytics, or training datasets written to disk. Compliance teams are moving further still, mandating sovereign cloud architectures operated entirely within a required jurisdiction with absolute physical isolation,

cached-prompt reconstruction via cache-hit timing analysis (2025–26).

[7]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation), art. 17, 2016 O.J. (L 119) 1 (right to erasure ("right to be forgotten")).

eliminating foreign administrative control and mitigating the risk of multi-tenant cache leakage across borders.

X. THE INDIAN CONTEXT: NAVIGATING THE DIGITAL PERSONAL DATA PROTECTION ACT, 2023

For multinational enterprises operating in India, the intersection of AI token processing, prompt caching vulnerabilities, and data sovereignty has gained acute legal urgency with the phased implementation of the Digital Personal Data Protection Act, 2023, finalised and brought into force under the Ministry of Electronics and Information Technology, with draft rules released in early 2025 and the Digital Personal Data Protection Rules, 2025 notified on 13 November 2025[8]. Historically, India lacked a comprehensive, standalone data protection framework, relying on the Information Technology Act, 2000; following the Supreme Court's recognition of privacy as a fundamental right, the Act introduces rigorous compliance mandates for the collection, processing, storage, and transfer of digital personal data, imposing immediate fiduciary duties regarding informed consent, data confidentiality, and system integrity on multinational companies.

Where an Indian enterprise deploys an AI agent that ingests customer data — an automated human-resources screening tool, a localised customer-service chatbot, or a financial advisory agent — that enterprise acts as a "Data Fiduciary" under the Act. If such sensitive data is sent to an external LLM provider, processed via an API, and potentially cached in a server farm located outside India, the enterprise faces three distinct regulatory hurdles.

A. Notice, Consent, and the Opacity Problem

The Act requires clear, standalone notices to Data Principals before processing personal data, itemising the personal data collected and the explicit purpose of processing. Where an AI agent generates invisible reasoning tokens or caches prompts containing personal data in unmapped data centres, the nature of that underlying processing must nonetheless remain transparent to the user — a substantial technical challenge where the LLM provider operates a proprietary, opaque model and data residency is fluid.

[8]The Digital Personal Data Protection Act, 2023, No. 22, Acts of Parliament, 2023 (India); Digital Personal Data Protection Rules, 2025 (India), notified Nov. 13, 2025.

B. Cross-Border Data Transfers under Rules 13 and 15

While the Act currently maintains a generally permissive approach to transferring personal data outside India, Rule 15 mandates strict compliance with conditions imposed by the Central Government, focused heavily on whether foreign states or intelligence agencies could access the transferred data, and Rule 13 grants the Government explicit authority to restrict cross-border transfers entirely for entities designated as Significant Data Fiduciaries, potentially banning transfers to blacklisted jurisdictions or requiring localised processing. An Indian enterprise routing AI prompts through a United States-based API endpoint subject to the CLOUD Act therefore risks sudden, material non-compliance the moment localisation mandates are tightened.

C. Security Safeguards, Sectoral Oversight, and Section 73 of the Indian Contract Act, 1872

The Act mandates that Data Fiduciaries adopt reasonable security safeguards, including encryption, access controls, and the prevention of unauthorised access. Given the vulnerabilities associated with prompt caching described above, utilising shared, multi-tenant AI platforms may be construed as a failure to maintain these mandated safeguards, exposing the Data Fiduciary to regulatory penalties and reputational harm. Entities regulated by the Securities and Exchange Board of India face an additional layer of obligation under SEBI's 2024 circular on the use of artificial intelligence and machine learning in regulated activities, which requires audit-trail maintenance and model explainability — requirements structurally difficult to reconcile with the opacity of reasoning-token generation[9]. Separately, under Section 73 of the Indian Contract Act, 1872, a party that suffers loss from a breach is entitled to compensation for losses naturally arising from that breach; where a vendor's API malfunction causes a runaway loop, and the vendor's terms of service disclaim liability for "unexpected usage patterns," Indian courts could look past such exclusion clauses if found unconscionable or contrary to public policy under Section 23 of the same Act. The absence of a reasonable circuit-breaker on the vendor's side may constitute a breach of the implied duty of care embedded in service contracts — an under-litigated proposition, but one unlikely to remain untested for long.

To comply with the Act, organisations deploying AI in India must structurally enforce

[9]Securities and Exchange Board of India, Circular on the Responsible Usage of Artificial Intelligence and Machine Learning in the Indian Securities Market (2024).

data residency. Data sovereignty is no longer a peripheral compliance task discharged by selecting an Indian data centre from a drop-down menu; it is a core design constraint that must be woven into the underlying AI architecture from inception, with prompt caching, reasoning-token generation, and all overarching processing confined strictly within national borders and governed by zero-retention principles and heavily audited infrastructure.

XI. THE HUMAN-VERSUS-AI COST PARADOX

Notwithstanding the risks catalogued above, AI agents remain, on average, between fifteen and three hundred times cheaper than human professionals for high-volume, repeatable tasks such as document processing, invoice reconciliation, or research briefing. A fully loaded human customer-service agent costs approximately $110,000 per year; an AI agent handling comparable volume costs a fraction of that figure, and one Stanford–Carnegie Mellon University study found AI agents completing tasks 88.3 percent faster and between 90.4 and 96.2 percent more cheaply than human professionals[10]. This is the paradox that most Chief Financial Officer budget models have yet to absorb: that cost advantage evaporates, and reverses, the instant an agent enters a runaway loop, deploys a reasoning model for a routine task, or operates without a contractual circuit breaker. A human lawyer billing at a conventional hourly rate is predictable; an AI agent reviewing the same contract may cost a few rupees if deployed correctly, or several lakhs if deployed carelessly with a reasoning model, a large context window, and no token cap. The savings narrative is genuine; the risk narrative is its shadow, and at Uber, the heaviest individual users generated monthly costs exceeding those of many human contractors performing comparable roles.

XII. CONCLUSION: A STRATEGIC IMPERATIVE FOR ENTERPRISE LEADERSHIP

The transition from predictable Software-as-a-Service to the variable, volatile economics of AI token billing represents one of the most significant shifts in enterprise technology procurement since the initial migration to cloud computing. The era of the flat-fee software licence is ending, replaced by a utility model in which every cognitive action taken by a machine draws directly from the corporate treasury. As foundational models become more sophisticated, the invisible reasoning tax will only increase, weighting operational expenditure

[10]See joint Stanford University–Carnegie Mellon University study on comparative task completion speed and cost between AI agents and human professionals (2025–26) (on file with author).

toward unseen, automated computation; simultaneously, the aggressive deployment of autonomous agents transfers unprecedented risk onto the balance sheet of the deploying organisation. The widening gap between aggressive vendor liability shields and the reluctance of commercial insurers to underwrite algorithmic error creates a genuine legal vacuum, in which a single runaway agent caught in an infinite loop can inflict devastating financial and reputational damage at machine speed.

To survive this structural transition, Chief Financial Officers, Chief Information Officers, and General Counsel must fundamentally change their procurement posture, ceasing to treat artificial intelligence as a standard software acquisition and instead recognising it as critical, high-risk digital infrastructure. Just as physical infrastructure projects require comprehensive EPC contracts with guaranteed maximum prices, liquidated damages, and uncompromising risk allocation, enterprise AI deployments require contractual hard caps, automated circuit breakers, strict human-in-the-loop mandates, and reversed indemnification structures. As regulatory frameworks including the General Data Protection Regulation and India's Digital Personal Data Protection Act, 2023 mature into active enforcement, the physical location, security, and retention mechanics of data processing — down to the ephemeral memory of a GPU cache — will increasingly dictate the outer boundary of corporate compliance. The integration of zero-retention architectures and physically isolated sovereign clouds is no longer an optional security upgrade; it is fast becoming a foundational legal requirement for operating in the modern digital economy. Those enterprises that build robust token-governance frameworks in 2026 will be writing the standard of care for the next decade of enterprise artificial intelligence; those that do not will spend that decade explaining their invoices to their boards, one runaway agent at a time.