



## AI Cost Governance

### Optimizing Costs to Prevent Enterprise AI Spend from Getting Out of Control

Unplanned AI expenses can grow quickly when AI adoption outpaces governance and cost management controls, creating significant enterprise risk; e.g.,

**A single compromised AI API key can result in over \$1 million in unauthorized charges within 48 hours.**

For information on the related topic of agentic AI costs and ROI, please refer to the [linked POV](#).

<sup>4</sup>Hillary Baron, *Autonomous but Not Controlled: AI Agent Incidents Now Common in Enterprises*, Cloud Security Alliance, 2026, p. 9.

<sup>5</sup>Supal Kanti Chowdhury, *Token optimization: The backbone of effective prompt engineering*, IBM Corporation, 16 September 2024.

<sup>6</sup>Cast AI, *2026 State of Kubernetes Optimization Report*, 2026, p. 5.

<sup>7</sup>CoreWeave, *Beyond the Hot Tier: Cut AI Storage Costs While Accelerating What Comes Next*, 2025, p. 2.

### Unplanned consumption costs can surge from unintended and uninformed AI misuse:



#### Agentic loops

Agents operating across systems and workflows can get stuck in logic loops, exhausting unpredictable budgets.



#### Inefficient prompting and API usage

Simple inefficiencies, such as excessive, polite phrasing (e.g., "please" or "thank you") in prompt engineering, can increase costs by 2–3x at scale.<sup>5</sup>



#### Model Routing

People often reach for premium models for straightforward tasks that smaller models could handle just as well, cheaper and faster.



#### GPU over-provisioning

Companies often lock into expensive hardware contracts before understanding actual needs, leading to massive, unused capacity, with GPU utilization across AI infrastructure averaging just 5%.<sup>6</sup>



#### Shadow infrastructure

82% of organizations have discovered AI agents or autonomous workflows created without the knowledge of security, IT, or governance teams, which can silently burn budget<sup>4</sup>



#### Hidden overhead

Storage costs from datasets, logs, and model checkpoints can add up quickly, with over 80% of data often going unused.<sup>7</sup>



#### API key exposure or misuse

Exploited GenAI API keys can result in rapid token consumption, triggering significant financial charges when usage exceeds intended levels.



#### Context window bloat

AI can leverage prior conversation and document context, but repeatedly receiving full histories in chats and API calls increases token usage and unnecessary cost.

### Services

#### 1 GenAI Incident Investigation & Response

Investigate potential and identified unplanned AI consumption incidents to assess immediate impact, perform root cause analysis, and implement stop-gap response and containment actions

#### 2 Training & Awareness

Provide training with persona-based messaging to promote responsible and efficient usage of tools and tokens and help users realize its value

#### 3 Governance Design & Implementation

Review current-state processes and implement controls to establish AI cost governance, including:

- › Real-time consumption reporting and observability
- › Built-in cost control and optimization through automated AI efficiency mechanisms (e.g., intelligent model routing, prompt optimization)
- › Data-driven guardrails and controls to align employee AI consumption with role and business expectations

#### 4 Monitoring & Anomaly Detection

Actively track costs and monitor AI consumption across personas throughout the organization and detect and alert anomalous consumption spend

### Value

#### 1 Multidisciplinary Team

Access to a multidisciplinary team of 200+ professionals with over 25 years of experience across relevant domains (e.g., digital forensics, cyber defense and resilience, AI governance, and tokenomics)

#### 2 Pre-Built, Customizable Accelerators

Established assets and accelerators to support effective AI consumption (i.e., risk and control library, training materials, governance frameworks, monitoring capabilities, cost control and efficiency metrics/KPIs, POVs on observability platforms, secrets managers, etc.) that can be tailored to organizational needs and processes

#### 3 Strategic Technology Alliances

Positioned to expedite and scale risk identification and mitigation through strategic alliances with leading technology and GenAI vendors

#### 4 Analytics and Proactive Threat Detection Capabilities

Experience implementing analytics and proactive threat detection capabilities to identify and mitigate unplanned consumption from inefficient AI use, AI misuse, and the exfiltration of sensitive data, such as credentials and GenAI API keys

### Contacts



**Cliff Goss, Ph.D.**  
Partner  
AI Risk Management  
Deloitte & Touche LLP  
[cgoss@deloitte.com](mailto:cgoss@deloitte.com)



**Arthur Ilyasov**  
Senior Manager  
AI Risk Management  
Deloitte & Touche LLP  
[arilyasov@deloitte.com](mailto:arilyasov@deloitte.com)



**Mike Weil**  
Managing Director  
AI & Digital Forensic Investigations  
Deloitte Financial Advisory Services LLP  
[miweil@deloitte.com](mailto:miweil@deloitte.com)



**Raj Thanabal**  
Senior Manager  
GenAI & AI Governance  
Deloitte Transactions and Business Analytics LLP  
[rthanabal@deloitte.com](mailto:rthanabal@deloitte.com)