Research Note 1/2026

MIB Trieste School of Management

The Invisible Bill: Governing the Hidden Costs and Compliance Risks of Agentic AI

Leonardo Felician

MIB Trieste School of Management

The Invisible Bill: Governing the Hidden Costs and Compliance Risks of Agentic AI

Leonardo Felician

Research Note 1/2026

July 2026

Felician, L. (2026). The Invisible Bill: Governing the Hidden Costs and Compliance Risks of Agentic AI IT: MIB Trieste School of Management. https://doi.org/10.69123/MIB.012026

Published by
MIB Trieste School of Management, Trieste, Italy

Abstract

This Executive Briefing is addressed to CEOs and Senior Management without assumed technical background to highlight three structural risks that the transition to Agentic AI is generating simultaneously: the **uncontrolled growth of costs** through token consumption that cannot be forecast or capped; an **auditability gap** that makes AI-assisted decisions structurally difficult to explain; and a **regulatory and litigation exposure** that transforms both gaps into legal liability.

Against these three risks, the article proposes a Proportional AI Architecture framework, built on a single counterintuitive principle. The most advanced AI architecture is not the one that uses AI everywhere. It is the one that masters workflow management discipline and deploys AI only where **uncertainty genuinely requires intelligence**, keeping everything else deterministic, auditable, and controllable. A new and constructive role for the Internal Audit function is also proposed within the same framework.

The challenge is not whether organizations should adopt AI. It is whether they can adopt it without creating invisible costs, opaque decision processes, and compliance liabilities that will become visible and expensive at the worst possible moment.

Contents

1. A revolution around the corner: from AI chatbots to AI agents

1.1 The birth of Generative AI

On November 30, 2022, **Generative AI** entered the mainstream, when OpenAI released ChatGPT to the general public. Within five days, it reached one million users; within two months, one hundred million (European Parliamentary Research Service, 2023). No consumer technology in history had reached that scale so quickly: not the smartphone, not social media, not the internet itself.

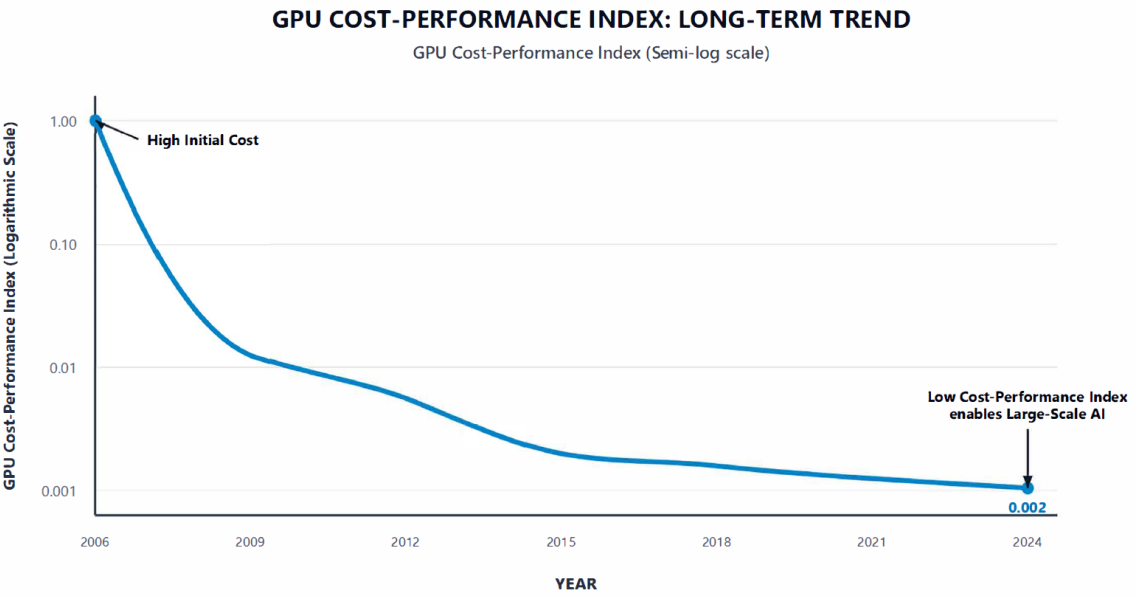
What followed was a gold rush. Google, Microsoft, Meta, Anthropic, and dozens of smaller players entered the market with competing AI systems, subsidized subscriptions and free tiers designed for rapid, unconstrained mass adoption at any cost. By 2024, Generative AI had become the fastest-growing software category on record, with enterprise deployments accelerating across every industry sector.

For most individual users, this remains a remarkably good deal: a tool of extraordinary capability, available for free or at trivial cost, accessible from any browser or smartphone. For organizations, the picture is fundamentally different, and the gap between these two realities is widening.

The distinction matters because AI is not standing still. The systems that captured public attention in 2022 were, by current standards, severely limited: they **answered questions through prompts,** generated text, and summarized documents. They were reactive. They waited to be asked. Their knowledge was frozen at a fixed training cutoff date, with no access to current information. In early 2023, the Retrieval Augmented Generation systems (RAG) began to move rapidly into mainstream AI applications, allowing these systems to query external databases and browse the internet in real time. This structural upgrade dissolved the boundaries of static training data, connecting the AI's reasoning engine with live, actionable information.

The economic landscape of AI is defined by a striking increase in computing power as shown in Figure 1. A Graphics Processing Unit (GPU) is a specialized processor originally designed for rendering graphics, but its highly parallel architecture makes it exceptionally efficient for the matrix operations that underpin the core architectures of modern AI. GPU price-performance has improved dramatically, with the cost of computing falling by over 99% since 2006, a trend that has fundamentally enabled the training and deployment of large-scale language models named Large Language Models (LLMs), that were economically prohibitive just a decade ago (Hobbhahn & Besiroglu, 2022). Over this same span, GPU processing performance has increased by approximately two orders of magnitude, with computational throughput doubling every 1.4 to 1.7 years.

Figure 1. GPU computation cost 2006-2024. Source: International Energy Agency (2025) | Stanford HAI (2025).



LLMs are based on neural networks: computational architectures loosely inspired by the biological brain, in which a large number of nodes are organized in interconnected layers. Each node acts as an artificial neuron, transmitting signals to the next. Between the input layer, where the prompt enters, and the output layer, where the response emerges, the model develops a series of hidden layers, invisible to the user, where its reasoning actually takes place.

1.2 Workflow Automation

Workflow Automation, often referred to as **Workflow Management**, emerged in the 1990s as a distinct discipline: not an AI discipline, but an organizational and operational one, built on the foundational contributions of Fernando Flores and Terry Winograd (Flores and Winograd, 1986). Its original objective was straightforward: to automate routine activities in manufacturing and service environments, improve efficiency and productivity, monitor process execution in real time, and increase operational accuracy and control.

Over time, workflow management evolved beyond its technological foundations (Georgakopoulos, 1995). Organizations realized that successful automation required more than software tools; it demanded a precise understanding of how decisions were made and how activities were executed across the enterprise. Workflow design therefore became closely linked to organizational knowledge, business rules, and managerial practice.

A workflow is simply a sequence of business activities whose execution path is known in advance and can therefore be audited, measured, and optimized. This **predictability** is not a technical detail; it is the foundation of operational governance.

More recently, AI capabilities have been integrated into these workflows, initially through **Machine Learning** and after 2022 through Generative AI. With Machine Learning, systems became capable of learning from historical data to detect patterns, classify information, and predict

outcomes. The subsequent introduction of generative AI further elevated this paradigm, enabling software to interpret unstructured context and generate text, code, and complex responses.

Far from being an obsolete discipline, workflow automation represents the **hidden competitive advantage** of organizations that have consistently invested in defining and optimizing their processes. In the age of AI this investment, as the chapters that follow will demonstrate, is more strategically valuable than ever.

Traditional software executes specific tasks. Workflow automation connects these tasks into rigid, automated sequences. AI-assisted workflows introduce probabilistic reasoning at key decision points within those sequences. **Agentic AI** does something qualitatively different: it pursues goals (Wang, 2024). Given an objective, an AI agent plans, acts, evaluates the results of its actions, and replans. It delegates sub-tasks, calls external tools and APIs, and loops autonomously until the objective is met, or until something goes wrong (Masterman, 2024).

By design, an Agentic system **dissolves the predictability** that workflow governance depends upon. The execution path is not defined in advance. It is constructed at runtime, by the agent, through a sequence of decisions that are not fully visible to the organization that deployed it.

As the complexity of agent tasks has grown, a further architectural shift has emerged: the **swarm-of-agents** paradigm (OpenAI, 2024), in which a complex objective is decomposed into multiple sub-tasks, each assigned to a specialized agent, with automatic coordination among them. This architecture is powerful. It is also the origin of a possible **Autonomous Agentic Loop**, the condition in which agents, interacting with one another to resolve ambiguity or clarify instructions, generate recursive cycles of activity that are invisible to the organization, but entirely visible on the invoice.

The mechanism is easier to grasp through analogy. Imagine two consultants, both billed by the hour, locked in a room with a problem to solve. You do not see their discussions, their iterations, their requests for clarification, or their failed attempts. They start debating behind your back, and you only find out when the results and the final bill arrive. Agentic AI operates in precisely this way: internal reasoning and inter-agent negotiation generate **hidden computational overhead**, what might be called unobserved orchestration and reasoning overhead, accumulating consumption that appeared in no specification document and in no business case.

As enterprise processes migrate toward Agentic architectures, the uncontrolled growth of costs represents the first critical risk this article addresses. A second, equally significant vulnerability, the breakdown of response reliability, will be fully explored in Chapter 3. The individual user, continuing to interact through a chat interface, will not notice this transition. The executive, on the other hand, cannot afford to ignore it.

The transition from traditional software to Agentic AI is therefore not merely a technological upgrade. It is a structural shift in how organizations incur costs, make decisions, and bear responsibility. The cost problem begins before the first token is generated: it begins with the economic dependency created by the infrastructure on which the workflows run.

2. The hidden economics of Agentic AI

2.1 The subsidized era

Before discussing technology, executives should ask a simpler question: are they in control of the infrastructure on which future business processes will depend? Most are not, and most do not yet know it.

The pricing that organizations have experimented with AI since 2022 does not reflect the actual cost of the underlying infrastructure. It reflects a deliberate act of commercial cross-subsidization, financed by investor capital on a scale without precedent in technology history. OpenAI, Anthropic, Google DeepMind, and their competitors have collectively absorbed hundreds of billions in losses to acquire users, establish API dependencies, and occupy organizational workflows. The goal was not profitability, it was position.

That era is not over, but it is changing (BenchLM 2026). Competitive dynamics are shifting, infrastructure costs are becoming visible on balance sheets, and the pricing assumptions under which organizations have built AI-dependent processes are no longer reliable guides to future costs. Many executives still think of AI cost as the cost of a subscription. The reality is that they are building operational processes on top of infrastructure whose prices do not reflect economic fundamentals and can change, materially and rapidly, in response to factors entirely outside their control.

This is the first risk: **strategic dependency**. An organization that has migrated core workflows to an Agentic architecture running on the APIs of a single large provider is not merely dependent on that provider's technology. It is dependent on their pricing decisions, and on their continued willingness to subsidize access.

Furthermore, when an organization builds Agentic workflows on top of proprietary APIs, it accumulates a second form of dependency that amplifies the first: **technical lock-in**. Switching providers is no longer merely a commercial decision. It becomes a full engineering migration, requiring redesign of integrations, re-validation of outputs, and retraining of operational staff. The more deeply Agentic AI is woven into enterprise processes, the higher that switching cost becomes. Dependency on pricing and dependency on architecture reinforce each other.

To understand why this matters, we must first look at how Generative-AI systems are priced. That pricing, in turn, is driven by what they actually consume, and why.

2.2 The Generative-AI pricing model

Unlike traditional enterprise software, which is priced by seat, by module, or by annual license, generative AI is priced by consumption: specifically, by the number of tokens processed. Any remaining per-seat pricing options are being abandoned by software providers, as this traditional approach fails to align subscription revenue with the actual cost of AI compute (OpenView, 2024).

A **token** is a sub-word fragment, the fundamental unit into which the model decomposes all text before processing it, explained in detail in §2.3. Every input sent to an AI model and every output it generates is metered and billed as a count of these fragments.

Furthermore, there is a fundamental price asymmetry between input and output tokens: generating a token requires a full computational pass through the model, while reading one does not, and providers charge accordingly, with output tokens consistently priced between three and ten times higher than input tokens. There is no fixed cost and **no predictable ceiling**, because consumption depends not on the length of a task, but on the complexity of the reasoning required to resolve it.

As of the time of writing, pricing across the major providers ranges from less than $1 to $30 per million output tokens, depending on model capability and provider. Figure 2 provides a representative comparison across flagship and mid-tier models from the three dominant US providers.

Figure 2. Indicative API Pricing for input (light color) and output tokens. Major Providers (mid-2025 to mid-2026)

Cost per million tokens, standard API rates. Input tokens are processed once; output tokens require full model generation and are billed at a consistently higher rate. Source: compiled from public provider pricing pages and independent comparison sources (costgoat.com, finout.io, benchlm.ai), June 2026

Even if flagship models deliver substantially higher accuracy and capability than mid-tier alternatives, two observations are essential. First, the differential between the most expensive flagship and the cheapest mid-tier model across providers can reach **50-to-1** on output tokens: an organization that deploys a flagship model where a mid-tier model would suffice is not making a quality decision, it is making an invisible cost decision. Second, the asymmetry between input and output pricing has profound implications for Agentic architectures: as agents reason, deliberate, invoke tools, and exchange messages, costs grow primarily on the generation side, making output volume a critical design variable. The more autonomous cycles an agent performs to resolve a task, the more heavily it hits the expensive side of this pricing asymmetry.

Nevertheless, the trend line of token prices, taken in isolation, appears decidedly favorable (Cottier, 2025). The inference cost for GPT-3.5-level performance fell from $20 per million tokens in November 2022 to $0.07 per million tokens by October 2024, a reduction of more than 280-fold in approximately 18 months (Stanford HAI, 2025). The OECD has documented a similar

trend: quality-adjusted AI prices declined by about 80% between 2023 and 2025, driven by both rapidly improving model capabilities and falling prices resulting from competition and efficiency gains (André et al., 2025).

This decline is real, well-documented, and genuinely significant. It is, however, for strategic planning purposes, an **insufficient basis for optimism**. Three dynamics operate simultaneously to undermine the conclusion that falling unit prices translate into controlled total costs. First, price reductions apply primarily to established models: each new frontier release resets pricing upward, and the reasoning models that power Agentic architectures have consistently maintained higher pricing than their non-reasoning predecessors, because organizations demonstrably pay for superior capability. Second, the volume of tokens consumed per task grows dramatically as systems move from conversational interactions to Agentic workflows: early chatbot exchange averages one to two thousand tokens per call, while multi-step Agentic tasks routinely consume fifty thousand or more (KPMG, 2025). **The unit price falls, but the quantity consumed rises faster**. Third, token consumption represents only part of the total cost of an Agentic deployment: workflow governance, security, compliance monitoring, and orchestration infrastructure add costs that remain stable or increase as deployments scale, independently of what happens to per-token pricing (KPMG, 2025; Gartner, 2025).

The net result is a dynamic that KPMG (2025) captures precisely in its title: the first token is free. The system that generates it is not. Executives who anchor their business case to the direction of unit prices are measuring the wrong variable. Organizations that built a business case on 2024 token costs and deployed at 2026 scale are already operating on assumptions that the market has moved past.

2.3 Tokens: the currency of AI

The key to understanding why costs in Agentic systems become structurally uncontrollable lies in the unit on which all consumption is measured. **That unit is the token, and it is not a word**.

A token is a sub-word fragment: the building block that a Large Language Model uses to read and generate text. Rather than processing full words, the model processes sequences of these fragments, each corresponding to a linguistically meaningful piece of a word: a common prefix or suffix, a root, or a high-frequency syllable combination. Tokenization algorithms are trained to find the most statistically efficient way to break down text. Consequently, tokenization can differ substantially across different languages, for instance English and Italian.

Figure 3. The anatomy of AI tokenization. Illustrative example: actual results may vary depending on the tokenizer used. The model decomposes each word into sub-word fragments before processing and bills for each one.

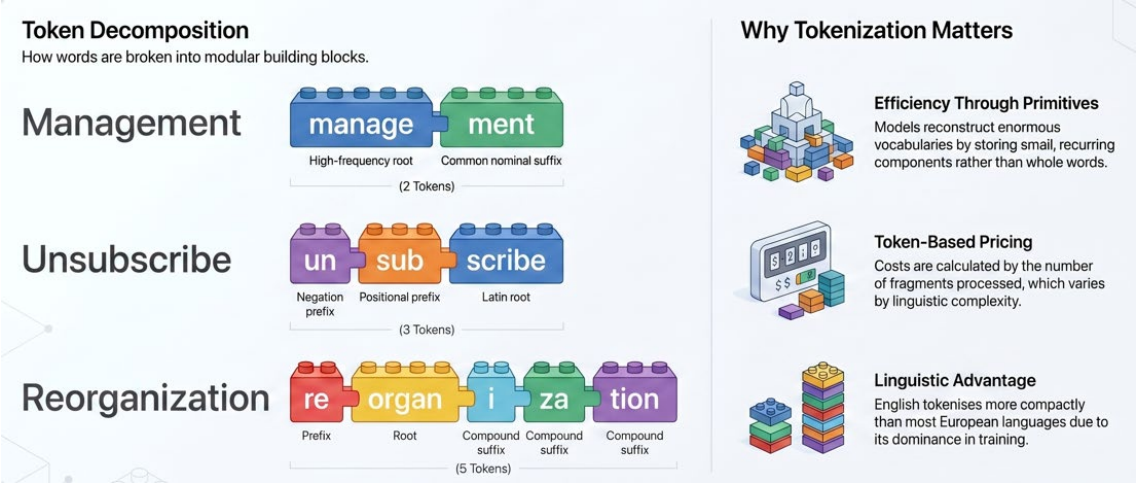


Figure 3 illustrates this with three familiar words. The logic is consistent across all three. The model does not memorize entire words; it memorizes the components from which words are constructed, because those components recur across thousands of different words. By storing *re-*, *-ment*, *sub-*, *-ize*, the model can reconstruct an enormous vocabulary from a compact set of primitives.

This is very efficient for the model. It means, however, that what an organization pays for is not the number of words it sends to the model, but the number of fragments those words are decomposed into: a number that varies with linguistic complexity, domain vocabulary, and, significantly, with the language used. English, being the dominant language in model training data, tokenizes more compactly than most European languages. An equivalent text in Italian will consume roughly 30 to 40 percent more tokens than the same content in English, a cost differential that most organizations never measure and few even know exists. Where AI can be invoked in English rather than in a language that tokenizes less efficiently, like Italian, doing so reduces input cost directly.

On average, 100 English words correspond to at least 130 tokens. For simple, single-turn interactions, this arithmetic is manageable and the cost is predictable. For Agentic systems, it is not. The reason requires understanding where most of the tokens in an Agentic workflow actually go.

Tokenization is the first and most consequential design decision in any LLM. Before the model begins to reason, its architects have already determined how language will be represented, and therefore a significant share of what the model can and cannot do. The statistical algorithms used for tokenization compress high-frequency strings into compact single tokens, while rare or technical strings are decomposed into multiple sub-units. An inefficient tokenizer uses more tokens to represent the same content: the context window shrinks, inference slows, costs rise, and meaning leaks at the edges. A more efficient one does the opposite: denser context, faster inference, lower cost, less information loss.

A useful analogy: in the US landline prefixes are all three digits (e.g. 212 for New York). Italian landline prefixes,by contrast, have been assigned by population size. Rome (06) and Milan (02) dial out on two digits, mid-sized cities on three, small towns on four. The system saves keystrokes at scale, billions of them, over decades. Tokenization works by a different mechanism, but the compression logic is the same: reserve the shortest codes for what appears most often.

These algorithms are among the most proprietary elements of any LLM: the details are generally not fully disclosed. The emergence of DeepSeek, a chinese competitor of American frontier models at substantially lower cost on less powerful hardware, is a reminder that algorithmic efficiency across the full processing pipeline can be as decisive as raw compute power.

2.4 The invisible meter of the Agentic Loop

To understand why Agentic AI token consumption is so difficult to predict and govern, it is necessary to understand the basics of how a Large Language Model processes information. As introduced in Chapter 1, an LLM is a **neural network**: Information enters through an input layer, propagates through a sequence of intermediate processing layers, and produces a result at the output layer. What matters for our purposes is not the mathematical detail of this propagation, but a single architectural fact: **the intermediate layers are invisible**. The user knows the input and the output. Every step that occurs between them, including internal reasoning, evaluation of alternatives, and progressive refinement of a response, involves computation that takes place entirely out of sight and generates tokens that appear on the invoice but not on the screen.

In a simple conversational interaction, this hidden processing is bounded and brief. Consider a manager asking an AI assistant: "*Summarize the key risks in this Non-Disclosure Agreement and flag any unusual clauses*." If the document is about one and a half A4 page long, about 600 words, the prompt might be 800 tokens; the response 400 tokens. At mid-tier pricing of $3 per million input tokens and $15 per million output tokens, the cost of that single exchange is a fraction of a cent: entirely calculable, entirely foreseeable and negligible in cost.

On models that disclose their internal reasoning, such as Claude with extended thinking, an additional 1,000 to 5,000 reasoning tokens are typically generated between input and output: invisible on screen, but present on the bill. Billed at rates comparable to output tokens, they can quietly double or triple the cost of a simple exchange. After any interaction, the platform records exactly how many reasoning tokens were used. In the OpenAI or Anthropic console, this appears in the usage logs for each call. For managers accessing AI through an enterprise tool, the same information is typically available in the billing or activity dashboard. Making it a habit to check this figure after a few benchmark queries gives an immediate, concrete sense of where costs actually land and how far they deviate from the visible input-output exchange.

The existence of this hidden reasoning process is also visible to individual users of some widespread chatbots. Models such as Gemini with Extended Thinking can display a side panel showing the reasoning steps generated by the model before producing the final answer. For relatively simple questions, the sequence may be brief. For more complex tasks involving multiple

alternatives, comparisons or evaluations, the chain of reasoning can become quite long, often comprising dozens of intermediate steps. While these displayed steps are not equivalent to reasoning tokens, they provide an intuitive illustration of the additional computation that takes place between the user's prompt and the model's final response.

But, unfortunately, **Agentic systems are structurally different**. When an AI agent is given a complex objective, it does not generate one response. It reasons through the problem across multiple steps: decomposing the task, generating candidate approaches, evaluating each, selecting the most promising, acting, observing the result, and replanning. Each of these reasoning steps generates tokens, not only in the visible exchange with the user, but in the internal deliberation that precedes each action. The organization bears the cost of every step of the agent's thinking, whether or not that thinking produces a visible output. In Agentic workflows, where the model plans, self-corrects, and iterates across multiple steps, that multiplier can reach ten times the base cost or even more (Introl 2025, Bertalanič 2026, Patil 2026).

This effect triggers a compounding cascade in multi-agent architecture where a complex task is decomposed and distributed across multiple specialized agents. When Agent A delegates a sub-task to Agent B, and Agent B encounters ambiguity and requests clarification, and Agent A interprets that request and responds, and Agent B re-evaluates and attempts the task again: each iteration of this loop generates further token consumption. The loop is invisible, the meter is running.

Figure 4. The Agentic Loop. The hidden cost of autonomy.

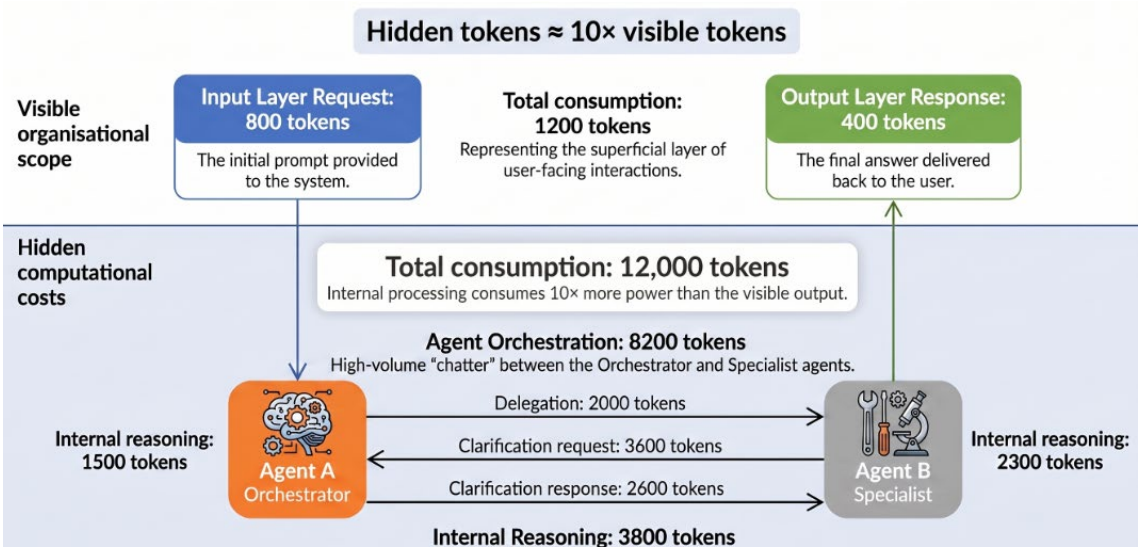


Figure 4 highlights the token cost structure of a single such exchange between two agents. The illustrative token count (12,000) for the inter-agent exchanges, and for internal reasoning represent only the internal hidden-layers consumption. Each iteration generates its own conventional input and output tokens as usual (1200), which are billed in addition at the proper rates. The hidden computational costs materialize separate billing events for three iterations of a single sub-task, none of which was visible to the organization that authorized the deployment.

Crucially, this internal overhead is not a marginal addition; empirical data shows that multi-agent communication topologies routinely multiply baseline input/output costs by a factor of 10x to nearly 12x (Zhang, 2025), driven by token duplication rates that can reach 86% in standard architectures (Parakhin, 2026).

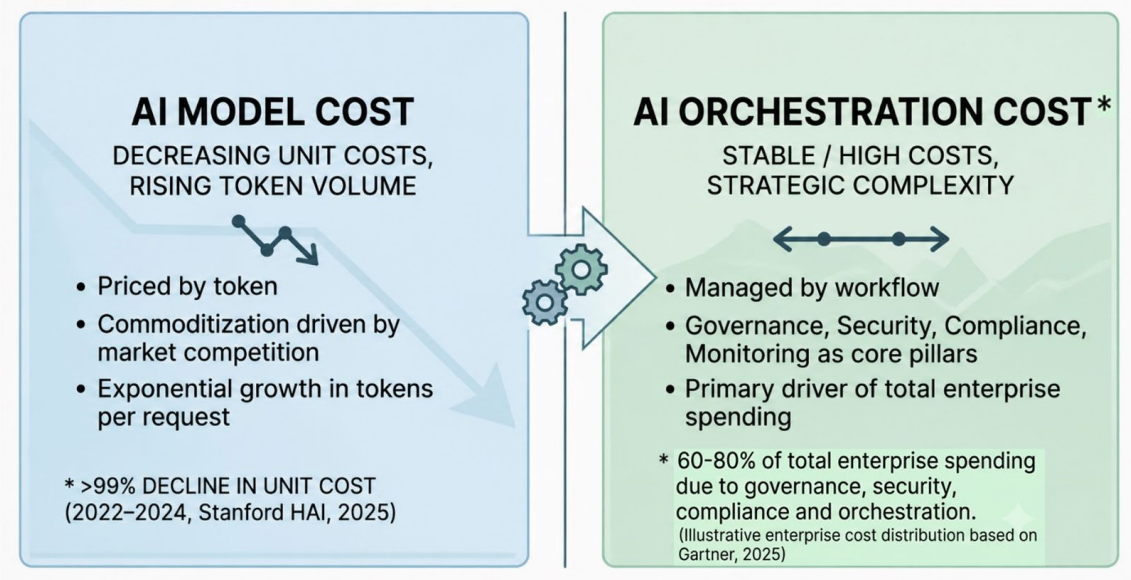
Consider now a concrete example: a mid-sized enterprise deploys an Agentic system to route unresolved queries from an AI-powered contact center (Tier 1) to human specialists (Tier 2). For each ticket, the orchestrating agent reads the full prior conversation history, queries a second agent to evaluate operator skills and current queue workloads, and ultimately produces an assignment recommendation. A representative prompt to the system might read: *"Ticket #4471 has been reopened by the customer after 24 hours. Prior conversation: [history]. Available specialists and current queue lengths: [data]. Assign to the most appropriate operator, with justification."* That prompt alone, including the conversation history and operator data, will typically run to 2,500 input tokens. The structured assignment response, including justification, will consume approximately 300 output tokens. At current mid-tier API rates of $3 per million input tokens and $15 per million output tokens, the visible cost per ticket is approximately $0.012. At 1,000 tickets per day and 220 working days per year, the annual token cost for this single operation approaches $2,600 for the visible exchange alone. However, once we factor in the ~10x hidden computational overhead—driven by the internal reasoning overhead, retry loops on ambiguous cases, and multi-agent orchestration—the realistic figure rises over $20,000 or beyond, before a single line of governance, monitoring, or compliance infrastructure is budgeted. Now multiply across the range of Agentic workflows a typical enterprise deploys in an initial wave of adoption: document classification, contract review, HR query routing, procurement matching, customer escalation triage. The aggregate token cost is no longer a line item. It is a structural operating expense that was never modelled, never approved, and remains largely invisible until the API invoice arrives, when the decisions are already locked in.

This is Agentic cost opacity, the first structural risk mentioned in the Abstract: it is not fraud, not negligence, but a structural feature of a technology whose resource consumption scales with the uncertainty it is asked to resolve. The more ambiguous the task, the more the agents deliberate. The more they deliberate, the longer the loop runs. The longer the loop runs, the higher the bill. And none of this is visible to the organization that authorized the deployment.

A summary is proposed in Figure 5. The real cost is not the AI models themselves, which are becoming increasingly cheaper and are rapidly turning into a commodity, but rather the growing number of tokens consumed per request. This increase is largely driven by the complexity of orchestration: that is, making multiple AI agents work together in a secure, controlled, reliable, and traceable manner. The more AI becomes *Agentic,* i.e., capable of acting autonomously, the higher this percentage rises, because every autonomous action multiplies the need for governance, security, and monitoring.

The answer is not to avoid Agentic AI. It is to understand, before deployment, which parts of a process genuinely require autonomous intelligence, and to govern everything else with deterministic, auditable, cost-controllable logic. That principle is the foundation of the framework proposed in this article.

Figure 5. The Economics of AI. Model Costs vs Orchestration Costs



3. The Cost of Opacity: Governance, Compliance and Legal Risks

3.1 The Auditability Gap

The previous chapter established that Agentic AI systems generate costs that are structurally difficult to observe and govern. This chapter addresses the second structural risk identified in the Abstract: **auditability gap**. It is a parallel problem, operating at a different level but sharing the same root cause: Agentic AI systems generate decisions that are structurally difficult to explain, audit, and defend. These two problems, cost opacity and decision opacity, are not independent. They are two expressions of the same architectural reality. To understand why, it is necessary to deepen something fundamental about how LLM learn.

Before proposing a solution, a conceptual distinction must be established, even though it is frequently collapsed in practice, with significant consequences. **Interpretability** refers to the capacity to understand what happens *inside* a model: for example why a specific pattern of weights produced a specific output, what internal representations drove a particular inference, etc. This remains, for current LLMs, an open research problem. **Auditability**, by contrast, refers to the capacity to demonstrate, after the fact and to an external party, that a system behaved in accordance with defined rules, that its decisions were traceable, and that accountability can be assigned. Auditability does not require interpretability. It does not require knowing what happened inside the model. It requires knowing what the system did, under what conditions, following what logic, with what human or automated oversight at each consequential step. This distinction is not merely academic: it means that the opacity of a neural network is not, in itself, a compliance barrier. The barrier arises when that opacity is allowed to extend to the *system as a whole,* when no logging, no deterministic guardrails, no decision checkpoints exist outside the model to reconstruct what occurred and why. Building an auditable system around an interpretability-opaque model is an engineering problem. It is a solvable one.

A neural network does not learn the correct answers. It learns the configuration of its internal parameters that minimizes a mathematical measure of error, called the **Loss Function**, calculated across billions of training examples. The model adjusts millions of numerical weights, iteration by iteration, in the direction that reduces this error. What emerges from this process is not a set of rules, not a decision tree, not an explicit reasoning chain. It is a statistical landscape: a vast, high-dimensional configuration of numerical relationships that encodes patterns extracted from training data. When the model produces an output, it is not following a retrievable logical path. It is traversing this landscape to minimize the loss function in a way that cannot be retraced after the fact.

This is at the very root of non-determinism, as anyone can observe simply by using an AI chatbot: it frequently provides slightly different responses to the exact same prompt. Given identical inputs on two separate occasions, a Large Language Model will not necessarily produce identical outputs, because the traversal of the statistical landscape involves sampling from probability distributions rather than executing deterministic instructions. This is not a malfunction. It is a direct

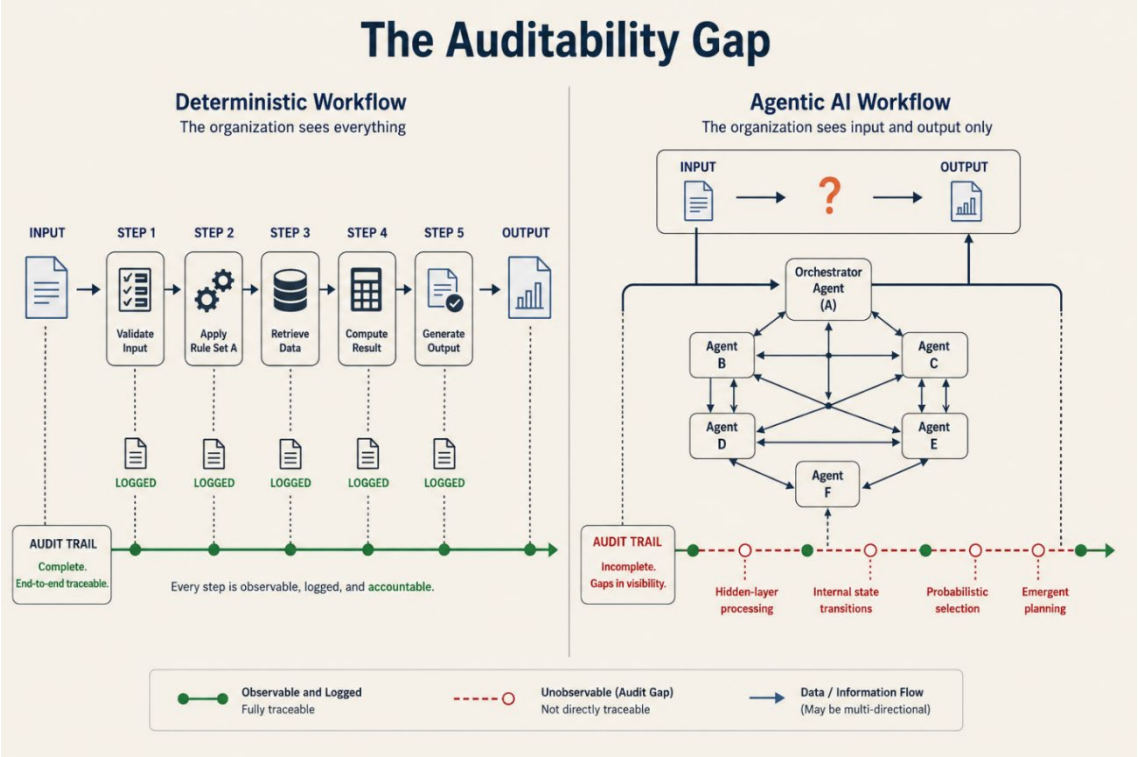
consequence of how the system was built. And it means that the concept of a reproducible, auditable decision path, which is the foundation of governance in every other domain of enterprise operations, does not exist in current AI architectures by default.

Hallucinations, the tendency of LLMs to produce confident, plausible, and entirely false outputs, are one of the well-documented failure modes associated with the probabilistic nature of LLM generation. Awareness of the problem is reasonably high among technical teams. But hallucinations, for all their notoriety, are not the central governance problem. An error can be detected, corrected, and attributed. What cannot be corrected is the absence of an explanation. The deeper issue is explainability, or rather, its structural limitation in current AI systems. When a Large Language Model produces an output, there is no retrievable record of the reasoning path that generated it. The model does not reason the way a human analyst reasons, moving through explicit steps that can be retraced. It operates through the interaction of billions of numerical parameters across representational spaces that have no direct correspondence to human-readable logic. The output exists. A human-readable, fully retrievable logical path to it generally does not.

For a single-turn interaction, a summary, a translation, a draft, this limitation is manageable. The output can be reviewed by a human before it produces consequences. In an Agentic architecture, the situation is categorically different. Agents act. They send communications, update records, approve transactions, reject applications, route cases, send money. By the time a human might review an outcome, the action has already been taken, often triggering downstream actions by other agents in the same pipeline. The decision path is not merely opaque: it is gone.

This is the **Auditability Gap** shown in Figure 6: the structural inability of Agentic AI systems to provide, after the fact, a coherent account of why a specific decision was made. It is an engineering reality, not a temporary limitation awaiting a software patch. And it has direct, material implications for governance, that the regulatory environment is now beginning to make explicit and legally enforceable.

Figure 6. The Auditabilty Gap. Difference in the audit trail introduced by Agentic AI.



3.2 The regulatory convergence: compliance is not a feature

Opacity is not only an engineering problem. In the European regulatory environment (European Commission, 2019), it is a legal liability, and an increasingly expensive one.

The **General Data Protection Regulation**, the ubiquitous GDPR (European Parliament and Council, 2016), established, years before Agentic AI existed, in Art. 22, in connection with Articles 13–15 (transparency obligations) and Recital 71, that individuals have the right to a meaningful explanation of automated decisions that affect them, even though this principle has seen limited and uneven enforcement in practice so far. The principle of *privacy by design*, introduced in Art. 25, requires that data protection be embedded in systems from the outset, not retrofitted as an afterthought. These obligations were written for a simpler world of rule-based workflow automation. They apply with full force, and considerably greater difficulty of compliance, to modern probabilistic, Agentic systems.

The EU AI Act (European Parliament and Council, 2024), the first comprehensive regulation worldwide, which entered into force in August 2024 and is now progressively becoming applicable across all Member States, goes further. Although the scope and the interpretation of this crucial provision are still evolving, it establishes a risk-based framework that imposes specific obligations on AI systems proportional to the severity of their potential impact. High-risk applications, such as those used in employment, credit assessment, access to essential services, and law enforcement, are subject to stringent requirements, including transparency, human oversight, auditability, and record-keeping, that current agentic architectures are, in their standard configurations, structurally unable to satisfy.

The key provision is not technical. It is conceptual. The EU AI Act proceeds from the premise that ignorance is not a valid technological defense. If an Agentic system rejects a loan application, declines a job candidate, or denies access to a service, the response that "*the algorithm decided*" is not an explanation. For Regulators, it is an admission of non-compliance. The organization that deployed the system remains responsible for its use and governance: not the model provider, not the technology vendor, not the complexity of the hidden layers. The organization is.

Human oversight is not optional under this framework. It is a legal requirement for systems that operate in sensitive domains. Meaningful human oversight becomes substantially more difficult without adequate auditability: human controllers cannot oversee a process whose decision logic is inaccessible to them. The two requirements are inseparable, and together they define the compliance problem that Agentic AI poses for any European enterprise.

The implication for system design is direct. A workflow governed by deterministic, auditable logic, traditional software, explicit rules, traceable code, can be audited, explained, and defended before a Regulator or a court. A workflow governed by an Agentic AI system, lacking specific architectural provisions for transparency, cannot. The gap between these two positions is not a matter of degree. It is a matter of category.

Workflow management, in this light, is not a step backward from innovation. It is the **organization's legal airbag**: the layer of deterministic, auditable process design that makes the use of AI, where genuinely necessary, governable and defensible. Organizations that have invested in defining their processes clearly before deploying AI are not behind the curve. They are structurally better positioned to comply, and to demonstrate compliance, than those that have layered Agentic systems onto undefined or poorly governed workflows.

The regulatory environment will tighten. The direction is not ambiguous. What remains for executive leadership to determine is not whether these obligations will apply to their organizations, but whether their current AI deployments can satisfy them.

3.3 The legal consequences: who is accountable?

The third structural risk identified in the Abstract moves from the regulatory domain into the judicial one. The EU AI Act is not merely establishing a compliance framework: as this section will show, it will also help assemble the evidentiary file for future AI litigation.

There is an aspect of the EU AI Act that has received remarkably little attention in executive discussions: its implications for civil litigation. Most commentary has focused on regulatory compliance, fines, and enforcement actions. The more material development, for organizations that deploy AI in operational processes, may be its impact on judicial proceedings.

The EU AI Act establishes a documentation chain that runs directly from system provider to deployer. Article 13 requires providers of high-risk AI systems to supply deployers with detailed technical instructions: the intended purpose of the system, its known limitations, its accuracy levels, foreseeable risks, required human oversight measures, and conditions of use. Article 26 requires deployers to use the system in conformity with those instructions. This appears, on first

reading, to be just a compliance obligation, a set of documents to be produced and filed. It is considerably more than that.

Italian law has moved quickly to translate this regulatory framework into litigation infrastructure. The implementing decrees enacted pursuant to Law 132 of 2025 (Italian Parliament, 2025), approved in preliminary examination in June 2026, remain subject to parliamentary review, potential amendments, and evolving legal interpretation before taking final effect, but introduce three provisions of direct relevance to any organization deploying AI in operational workflows.

The first is the right of access to technical documentation. A party who has suffered damage allegedly caused by an AI system may request disclosure of the technical documentation required under the AI Act. What Article 13 requires to exist becomes, in civil proceedings, discoverable evidence. The compliance file is also the litigation file.

The second is the presumption of causal link. When damage arises from a violation of AI Act obligations, a causal connection between the system's operation and the alleged harm is presumed and the burden of proof shifts to the deploying organization to disprove it. In a domain defined by structural opacity and the absence of retrievable decision paths, this reversal is not a procedural technicality. It is a fundamental shift in litigation exposure.

The third is the extension of liability beyond individuals. The new Article 437-bis of the Italian Penal Code punishes the omission or alteration of required safety measures for high-risk AI systems when the omission creates concrete and foreseeable harm. Equally significant, organizational liability under the well-known Legislative Decree 231/2001, Italy's framework for corporate administrative liability arising from criminal offenses, may attach in parallel, ensuring that accountability does not rest solely with individual employees but extends to the organization that deployed and benefited from the system.

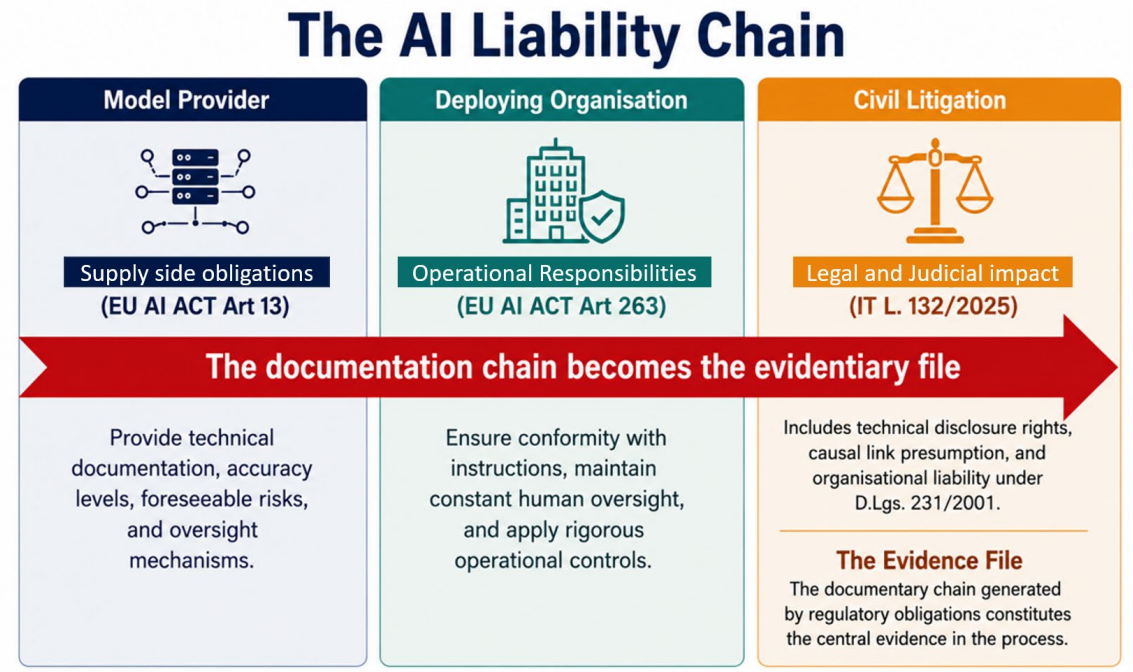
The aggregate effect is this: the EU AI Act is not merely establishing a compliance framework; it is, as authoritative legal commentary has observed, assembling the evidentiary file for future AI litigation (Il Sole 24 Ore, 2026). Every deployment decision, every instruction provided or omitted, every oversight measure implemented or bypassed, becomes a critical piece of evidence in future proceedings (Negri, 2026).

A very recent German ruling of the Regional Court of Munich I (Landgericht München I, 2026), issued a preliminary injunction prohibiting Google from disseminating false claims about two Munich-based publishing companies generated by its AI Overviews feature, confirms that this European-wide trend is already producing judicial consequences. Indeed, this pioneering judgment illustrates that European courts are already confronting questions of responsibility for AI-generated outputs and they tend to attribute the direct liability of AI agents to the corporations that deploy them. The court classified Google as a direct infringer. The ruling is a preliminary injunction, not a final judgment, and an appeal is expected, but its reasoning is already very significant: an organization that deploys an AI system producing autonomous statements cannot shelter behind the system's automated nature to escape accountability for what that system says.

As shown in Figure 7, AI governance, therefore, is not solely a regulatory obligation: it is a defensive strategy against the third risk mentioned in the Abstract. In the litigation of tomorrow,

what will matter is whether the organization can demonstrate exactly what instructions were provided, how the system was utilized, and what controls were genuinely in place.

Figure 7. AI Liability chain introduced by EU AI Act and other regulations mentioned in this chapter.



3.4 Compliance by Design: The Bridge to a Solution

With the distinction between interpretability and auditability established at the beginning of this chapter, the compliance picture that emerges from the rest of this chapter can be read with greater precision. The result is not comfortable, but it is clear. European organizations deploying Agentic AI in operational workflows face a convergence of three distinct but reinforcing pressures: an **engineering reality** that makes decision paths structurally opaque, a **regulatory framework** that requires those decision paths to be transparent and auditable, and an **emerging liti-gation environment** in which the absence of adequate documentation may increasingly weaken an organization's ability to demonstrate its compliance and defend its position. These three pressures do not operate sequentially. They operate simultaneously. An organization that deploys an Agentic system today is exposed to all three from the moment of deployment. The instinctive response, to add compliance measures after the fact, is structurally inadequate. *Auditability cannot be retrofitted onto an architecture that was not designed to produce it.* A logging layer imposed on top of an opaque Agentic pipeline does not make the pipeline auditable: it records that outputs were produced, not why. The Regulator requires the latter and probably also the court will demand the latter. The logging layer satisfies neither.

The only position that is simultaneously technically sound, regulatorily defensible, and legally robust is one in which auditability is *designed in from the outset*: where the boundary between what the AI system decides and what deterministic, traceable logic controls is explicit,

The only position that is simultaneously technically sound, regulatorily defensible, and legally robust is one in which auditability is designed in from the outset: where the boundary between what the AI system decides and what deterministic, traceable logic controls is explicit, documented, and enforceable. This is not a constraint on the use of AI. It is the condition under which AI can be used sustainably, at scale, in consequential enterprise processes. It would be intellectually dishonest, however, to present this regulatory architecture without acknowledging its competitive cost. The compliance burden described in this chapter protects European citizens, and that protection is legitimate.

Yet, protection and speed represent a direct trade-off. If the United States and leading Asian economies continue to deploy Agentic AI under lighter or absent regulatory frameworks, European organizations will face a structural disadvantage that is not merely administrative: it accelerates at the rate of the rapid expansion of AI capabilities. A system that is fully compliant but operationally slower and more expensive to build than its unregulated competitors is not a neutral outcome, it is a strategic liability. Whether the answer lies in reciprocal international regulatory convergence, in targeted exemptions for lower-risk applications, or in a deliberate European bet that trust and auditability will become competitive advantages in themselves, is a question that deserves a dedicated analysis. It falls outside the scope of this article, but it should not fall outside the attention of the executives and policymakers reading it.

However, the deliberate and proportional allocation of tasks between Agentic intelligence and deterministic governance is the foundation of the framework this article proposes. How that allocation is made, what criteria determine which parts of a process require autonomous AI and which require auditable logic, and how the resulting architecture can be governed, measured, and defended: that is the subject of the next chapter.

4. The Proportional AI Architecture Framework

4.1 The governing principle

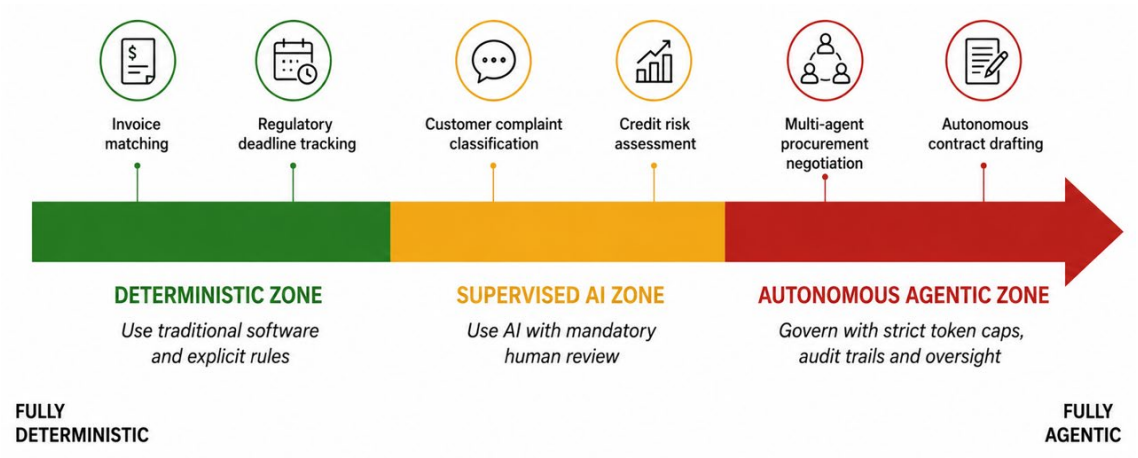
The challenge this article has described is not a reason to slow AI adoption. It is a reason to make it intelligent. The organizations that will extract sustainable value from Agentic AI are not those that deploy it most aggressively. They are those that deploy it most precisely: matching the capability of the technology to the genuine requirements of the task, and governing everything else with the simplest, most auditable tool available. This is the principle behind what this article proposes as Proportional AI Architecture.

The core insight is deliberately counterintuitive. The most advanced AI architecture is not the one that uses AI everywhere. It is the one that uses AI only where uncertainty genuinely requires intelligence. This reframes the strategic question executives have been asking. The question is not where AI can be used. Given current capability, the answer to that question is, increasingly, almost anywhere. The question that matters is where AI is actually needed: where its specific capacity to interpret ambiguity, synthesize unstructured information, and reason under uncertainty produces value that deterministic logic cannot. In most operational workflows, the answer is: less often than current deployment patterns suggest.

A significant share of business-process steps are deterministic, as shown in Figure 8. These steps have known inputs, known rules, and known outputs. Applying a Large Language Model to a deterministic step does not improve the outcome. It adds cost, adds latency, adds non-determinism, and adds an audit trail gap, in exchange for no measurable benefit. The proportion of a workflow that genuinely requires AI typically represents a minority of its steps: the share that involves ambiguity, interpretation, unstructured inputs, or contextual judgment.

Proportional AI Architecture is the discipline of identifying that minority precisely, governing it with appropriate AI capability and human oversight, and treating everything else as what it is: a software engineering problem with a software engineering solution, typically workflow management systems. The previous chapters have shown why this discipline is not optional. Cost opacity, decision opacity, and regulatory exposure are not risks an organization can choose to absorb indefinitely. They are the direct, structural consequence of applying Agentic intelligence to processes that did not require it. The four pillars that follow put this discipline into practice.

Figure 8. Where does AI actually matter? Proportional AI Architecture continuum shows some examples.



4.2 The four pillars of Proportional AI Architecture

Pillar 1: Deterministic first

When a rule, a condition, or an explicit algorithm can govern a process step reliably, use it. Traditional software is auditable, cost-controllable, and repeatable. It does not hallucinate, and it does not generate the hidden-layer token consumption described in Chapter 2. Default to deterministic logic, and depart from it **only when the task genuinely exceeds its capabilities**. The operational test is straightforward: can this step be fully specified as a finite set of conditions and outcomes, with a small, enumerable number of exceptions? If yes, it is a deterministic step, regardless of how the organization currently processes it. A significant share of tasks now routed through AI systems pass this test and were simply never re-engineered as rules once the AI alternative became available. That is not modernization. It is laziness: a form of technical debt accumulating under a new name.

Pillar 2: Structure before intelligence

Before introducing AI into any process, define and optimize the process itself. An AI system applied to a poorly designed workflow does not fix the workflow. It inherits its inefficiencies, amplifies their costs, and makes them harder to spot, because the model's apparent fluency masks the disorder underneath. The organizations with the strongest AI governance foundations tend to be those that had strong process governance before AI arrived.

The operational implication is sequencing, not philosophy: workflow redesign is not a prerequisite project to complete before AI is considered. It is the same project. Mapping a process precisely enough to deploy AI responsibly is, in most cases, the same exercise as mapping it precisely enough to know whether AI is needed at all.

Pillar 3: AI only at uncertainty points

Reserve AI capability for the steps in a workflow that require what AI genuinely provides: interpretation of unstructured inputs, reasoning under ambiguity, synthesis across heterogeneous

information sources. These points exist in almost every complex workflow. They are rarely the majority of it.

The operational diagnostic test for an uncertainty point: would two competent human professionals, given the same input, plausibly reach different but equally defensible conclusions? If the answer is yes, genuine judgment is involved, and AI assistance is appropriate, with human review proportional to the stakes. If the answer is no, the task is being treated as ambiguous only because no one has yet written down the rules that resolve it. Identifying the difference is an act of architectural discipline, not technological timidity.

Pillar 4: Human accountability remains external

The accountability for decisions made by AI-assisted processes cannot be delegated to the model. It remains with the organization, and therefore with the humans who designed, deployed, and oversee the system. This is not a philosophical position. It is a legal one, established by the regulatory framework now in force across European jurisdictions, and detailed in Chapter 3.

The operational implication is architectural, not procedural: consequential decisions require human review points, and those review points must be designed into the system before deployment, not appended after an incident. A review step added retroactively, once a regulator or a court has asked for it, demonstrates exactly the absence of foresight the framework is designed to penalize.

Together, these four pillars define an architecture in which AI is a precisely deployed capability within a deterministic, auditable governance structure, not a replacement for that structure. The AI does what only AI can do. The workflow does everything else. The human remains accountable for both.

This is not a conservative position. On the contrary, it is an innovative one. It is the condition under which AI can deliver its genuine economic and operational promise: at predictable cost, with adequate governance, and in compliance with the regulatory environment now taking shape around it.

4.3 From principle to practice: Governance and Internal Audit

Pillar 4 establishes that human accountability cannot be delegated to the model. What it does not yet specify is who, inside the organization, is best positioned to exercise that accountability in practice, and at what point in the system's lifecycle that function should be engaged. The answer reframes a function, that most organizations still treat as a control of last resort, into one of the principal beneficiaries of Proportional AI Architecture: Internal Audit.

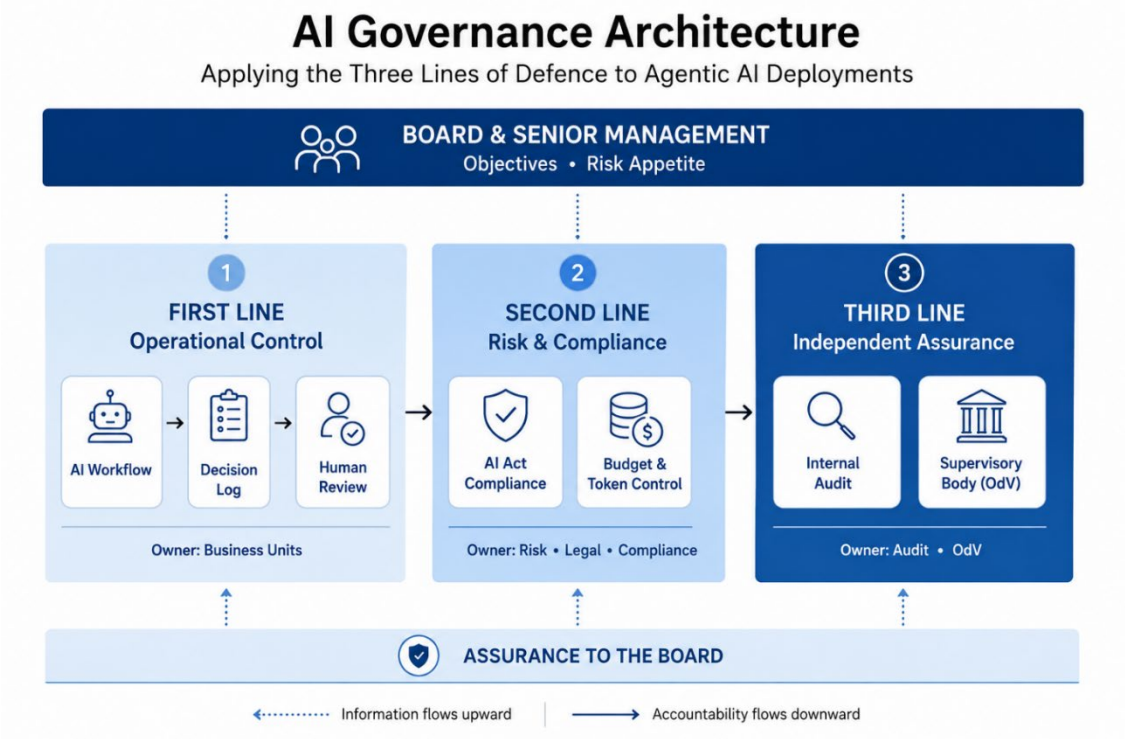
In most enterprises today, Internal Audit is consulted late, if at all, in AI deployment decisions. It reviews what has already been built, typically after an incident, a regulatory inquiry, or a scheduled audit cycle has made review unavoidable. Proportional AI Architecture inverts this sequence. Because the architecture requires, by design, that every AI-assisted process produce a traceable account of its own activity, it gives Internal Audit something it has rarely had in digital workflows:

a native, real-time, structured object to examine, rather than a reconstruction assembled after the fact.

In practice, for many workflows, the appropriate human oversight mechanism is not continuous human review, but a documented decision trail that enables human review, challenge and escalation of the actions taken by the AI system when required. Where experience demonstrates that this first-level review consistently surfaces no anomalies for a given workflow, the organization can graduate to **a more efficient second-level control: the AI system itself flags specific entries for manual review**, selected on the basis of detected anomalies or outlier characteristics. Internal Audit, in parallel, independently samples additional cases at random (cases the automated flagging did not select) to verify that the system's own anomaly detection is not itself developing blind spots.

This architecture maps directly onto the **Three Lines of Defence** model, the governance framework formalized by the Institute of Internal Auditors and now embedded in corporate control practice across regulated industries worldwide *(Institute of Internal Auditors, 2020).* Under this model, first-level controls are embedded in operational procedures; second-level controls are exercised by compliance and risk management functions; and the third line — independent assurance — is the responsibility of Internal Audit and, in the Italian context specifically, of the Supervisory Body (*Organismo di Vigilanza*) established under Legislative Decree 231/2001 *(Italian Parliament, 2001)*, working in concert with external audit. As shown in Figure 9, Proportional AI Architecture gives this independent assurance layer exactly what it has always needed and rarely had: a system designed from the outset to be examined, rather than one that must be reverse-engineered into examinability after deployment.

Figure 9 AI Governance architecture: the Three Lines of Model.



This is the modern reading of the aspirational principle underlying Law 231: that organizational accountability is most credible when it is built into the architecture of decision-making itself, not appended to it as a compliance formality. Internal Audit's role in Proportional AI Architecture, accordingly, should not be confined to the end of the deployment cycle. It is most valuable at the beginning, **defining what a workflow must be able to demonstrate** before it is approved, not auditing what it failed to demonstrate after the fact.

5. Strategic implementation and deployment

5.1 Make or buy?

The Proportional AI Architecture described in the former chapter becomes even more compelling when paired with a strategic reflection on *where* inference actually runs. The rise of high-quality **open-weight models,** Llama, Mistral, Qwen, and their derivatives, has fundamentally disrupted the assumption that large enterprises must purchase generative AI by the token from a **hyperscaler** like AWS, Microsoft Azure, Google Cloud Platform or Alibaba Cloud. *Hyperscalers* are cloud service providers managing computing architectures on a giant scale that have the capacity to expand and adapt processing power, memory, and storage in real time with virtually unlimited scalability. *Open-weight models*, on the other hand, are artificial intelligence models whose creators make publicly available and downloadable the *weights*, that is the parameters and mathematical connections learned during the training phase. *Weights* represent the condensed knowledge of the AI model. When a model is trained on huge volumes of data, it adjusts billions of internal numbers (the parameters or weights) until it learns to generate accurate text, images, or responses. Distributing the weights means allowing anyone to download the model file and run it on their own computer, on their own servers, or on a private cloud.

The main advantages of open-weight models are local execution and privacy, as the model can be run directly on company hardware without information or sensitive data having to pass through third-party APIs, as well as Fine-Tuning, because having the weights allows the model to be further trained on specific datasets (e.g., internal company documents, legal jargon, medicine) to adapt it to specific needs. However, the most important advantage is that the company does not have to pay a pay-per-use fee to an external platform for every request generated.

For organizations handling tens of thousands of contact center interactions per day, the cumulative API cost of even a well-optimized funnel can justify a serious make-or-buy analysis. Running a mid-size open-weight model on dedicated on-premise or private-cloud GPU infrastructure shifts the cost structure from variable (pay-per-token, unpredictable, scaling linearly with volume) to fixed (hardware amortization, energy, operations staffing). This shift should not be understood as a simple transition from variable pay-per-token costs to fixed in-house costs, but rather as a move from predominantly usage-based external expenditure toward a higher share of fixed and capacity-related costs. Beyond a certain interaction threshold, this trade-off can tilt decisively in favor of in-house deployment.

The decision, however, is rarely purely financial. Running open-weight models on an organization's own infrastructure can provide greater control over data location, access, and processing, while reducing dependence on external providers, a non-trivial advantage in regulated industries such as insurance, banking, or healthcare, where customer interaction data cannot easily transit third-party APIs, but it also demands organizational capabilities that many companies have yet to build: model evaluation, quantization, inference optimization, and ongoing fine-tuning as the deployed model drifts from production reality. The most pragmatic path for most enterprises is a **hybrid inference strategy**: proprietary frontier models via API for the residual

generative layer (low volume, high complexity, maximum quality required) and a fine-tuned open-weight model running in-house for high-volume, continuously recurring classification and clustering tasks at scale. This is not a one-time procurement decision; it is an architectural commitment that must be revisited as both model capabilities and hardware costs evolve.

5.2 From Model Routing to Cognitive Optimization

The emergence of agentic orchestration platforms introduces a potentially important new dimension of cost optimization. Today, these platforms are primarily being developed to coordinate agents, tools, workflows, and model calls. As the market matures, however, their role could evolve beyond simply routing a task to the appropriate agent or model. The key question is whether the orchestration layer will eventually become capable of optimizing not only *where* a task is sent, but also *how much cognitive capacity* should be allocated to resolve it.

The analogy with database query optimization is useful. A database management system does not execute every query in the same way. Its query optimizer evaluates the request against the characteristics of the available data and infrastructure, estimates the relative efficiency of alternative execution plans, and selects the strategy expected to deliver the best result at an acceptable cost. An agentic orchestration layer could evolve toward a comparable function. Rather than applying a fixed rule such as “task X always goes to model Y”, it could evaluate the specific task at runtime, estimate its cognitive complexity, and select the most cost-efficient model capable of handling it reliably.

This would represent a significant evolution beyond static API integrations and simple preconfigured model routing. The orchestrator could consider not only semantic complexity, but also the required context window, model latency, current availability, historical reliability on similar tasks, the cost of failure, and the price of alternative models. For example, a brand-compliance check on a slide deck, i.e.essentially a pattern-matching task against a documented ruleset, would not require frontier reasoning capacity. A runtime optimizer could route it to a lightweight, high-throughput model. A task requiring a multi-step legal reasoning across ambiguous contractual language could instead be escalated to a more capable model, because the additional cost would correspond to a genuine increase in cognitive complexity and potential value.

This capability should not be confused with a universal standard already present in today's orchestration platforms. The market is still developing, and the degree to which these systems will autonomously perform such evaluations remains an architectural choice and a likely area of future innovation. The proposition is that this is where a significant part of the know-how of Agentic AI may increasingly reside: not merely in connecting agents and models, but in dynamically allocating cognitive resources.

The implication is important. The cost of any individual Agentic interaction may remain difficult to predict precisely in advance. However, the aggregate cost of a multi-agent system could become actively managed through the intelligence of its orchestration layer, without requiring human intervention at runtime. Agentic AI, when properly designed, would not merely consume token budgets: it would allocate them.

The competitive advantage would therefore not necessarily belong to organizations using the most powerful models, but to those whose orchestration layer knows when the most powerful model is unnecessary. The strategic know-how of Agentic AI may ultimately lie not in accessing maximum intelligence, but in knowing precisely when to deploy it.

5.3 The CEO checklist

Translating the four pillars into practice begins well before any deployment decision is made. The first step is diagnostic: map your current AI-assisted workflows, and those under construction against the four tests described above. Which steps genuinely pass the Pillar 3 uncertainty test? Which are deterministic processes that have simply never been re-engineered? Where are the human review points, and were they designed in from the outset or added retrospectively? The answers to these questions, taken together, produce a clearer picture of where the organization's actual AI exposure lies than any technology audit conducted after deployment.

The second step is architectural selection. The market for Agentic AI solutions is large and growing rapidly, and not all platforms are equally governable. The framework described in this article points toward a specific class of solution: one that executes against explicit business rules, produces auditable logs of every action taken, and is designed from the ground up for traceability rather than retrofitted for it after the fact. Platforms of this kind, industrial-grade systems such as Kapto.ai (Kapto 2026), validate decisions against configurable business rules and maintain full traceability across operational workflows, thus embodying the Proportional AI Architecture principle in their design. This class of solution already on the market gives an answer to the right category of question: not "*what can this system do?*" but "*what can this system demonstrate, and to whom, and when?*"

Translating the four pillars into a deployment decision does not require a new governance department. It requires **a small set of questions**, asked by the CEO before approval, not after deployment. Does this step pass the Pillar 1 test; can it be fully specified as a finite set of conditions and outcomes? Has the underlying workflow been mapped and optimized, independent of the AI decision? Does this step pass the Pillar 3 test for genuine uncertainty? Where is the human review point, and was it designed in before deployment?

One question belongs also on Internal Audit's desk at the design stage: what would this workflow need to produce, in terms of traceable, searchable records, for third line internal audit assurance to be exercised without reconstruction after the fact? Two questions belong to the CTO before any Agentic workflow is approved: what is the token cost ceiling for this deployment, modeled against the hidden-layer multiplier described in Chapter 2? And how will the organization demonstrate auditability to a regulatory inspection, using the documentation chain described in Chapter 3? If the organization cannot answer any of these questions, the deployment should not be approved.

The board is pressing for AI everywhere. The CEO’s distinct value and the protection of the organization's EBITDA lies in knowing precisely where AI belongs, and where it does not.

Acknowledgements

It is a pleasure to thank prof. Andrea Tracogna, Francesco Venier and Roberto Pugliese for their comments on the draft of this article.

Bibliography

André, C., Bétin, M., Gal, P. and Peltier, P. (2025) Developments in Artificial Intelligence markets: New indicators based on model characteristics, prices and providers, OECD Artificial Intelligence Papers, No. 37, Paris: OECD Publishing. doi:10.1787/9302bf46-en, available at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2025/06/developments-in-artificial-intelligence-markets-new-indicators-based-on-model-characteristics-prices-and-providers_75e50b2a/9302bf46-en.pdf (Accessed: 18 June 2026).

BenchLM (2026) 'LLM API Pricing History: How AI Model Costs Have Changed Over Time', *BenchLM*. Available at: https://benchlm.ai/llm-pricing-trends (Accessed: 18 June 2026).

Bertalanič, B., & Fortuna, C. (2026). The cost of consensus: Isolated self-correction prevails over unguided homogeneous multi-agent debate. *arXiv preprint arXiv:2605.00914*.

Cottier, B., Snodin, B., Owen, D. and Adamczewski, T. (2025) 'LLM inference prices have fallen rapidly but unequally across tasks', *Epoch AI Data Insights*. Available at: https://epoch.ai/data-insights/llm-inference-price-trends (Accessed: 18 June 2026).

European Commission, High-Level Expert Group on Artificial Intelligence (2019) *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission. Available at: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Parliament and Council of the European Union (2016) *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. Official Journal of the European Union, L 119, 4 May 2016, pp. 1–88. Available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj

European Parliament and Council of the European Union (2024) *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union, L, 12 July 2024. Available at: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

European Parliamentary Research Service (2023), *Artificial intelligence, ChatGPT and generative AI: We need to talk*. European Parliament.

Flores, F. and Winograd, T., 1986. *Understanding computers and cognition: A new foundation for design*. Norwood, NJ: Ablex Publishing.

Gartner (2025) *Hidden Costs of Generative AI Make ROI an Inadequate Metric for CIOs in Banking*. Stamford, CT: Gartner. Available at: https://www.gartner.com (Accessed: 18 June 2026).

Gartner (2026) *Gartner predicts that by 2030, performing inference on an LLM with 1 trillion parameters will cost GenAI providers over 90% less than in 2025*. Gartner Newsroom, 25 March 2026. Available at: https://www.gartner.com/en/newsroom/press-releases/2026-03-25-gartner-predicts-that-by-2030-performing-inference-on-an-llm-with-1-trillion-parameters-will-cost-genai-providers-over-90-percent-less-than-in-2025

Georgakopoulos, D., Hornick, M. and Sheth, A., 1995. An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases*, 3(2), pp.119–153.

Hobbhahn, M. and Besiroglu, T. (2022) *Trends in GPU price-performance*. Epoch AI. Available at: https://epoch.ai/publications/trends-in-gpu-price-performance (Accessed: 18 June 2026).

International Energy Agency (2025) *Electricity 2025: Analysis and forecast to 2027*. Paris: IEA. Available at: https://www.iea.org/reports/electricity-2025 (Accessed: 18 June 2026).

Il Sole 24 Ore (2026) 'IA, responsabilità civile rafforzata e nuovo reato per i sistemi ad alto rischio', *Norme & Tributi Plus*, 11 June. Available at: https://ntplusdiritto.ilsole24ore.com (Accessed: 18 June 2026).

Institute of Internal Auditors (2020) *The IIA's Three Lines Model: An Update of the Three Lines of Defense*. Lake Mary, FL: The Institute of Internal Auditors. Available at: https://www.theiia.org/globalassets/documents/resources/the-iias-three-lines-model-an-update-of-the-three-lines-of-defense-july-2020/three-lines-model-updated-english.pdf

Introl (2025) *Inference Unit Economics: The True Cost Per Million Tokens*. Available at: https://introl.com/blog/inference-unit-economics-true-cost-per-million-tokens-guide (Accessed: 18 June 2026).

Italian Parliament (2001) *Legislative Decree No. 231 of 8 June 2001: Discipline of the Administrative Liability of Legal Persons, Companies and Associations, including those without Legal Personality*. Rome: Official Gazette of the Italian Republic No. 140 of 19 June 2001 Available at: https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2001-06-08;231

Italian Parliament (2025) 'Legge 23 settembre 2025, n. 132. Disposizioni e deleghe al Governo in materia di intelligenza artificiale', *Gazzetta Ufficiale della Repubblica Italiana*, Serie Generale n. 223 (25 September 2025).

Kapto (2026) *Kapto: Agentic AI with full traceability for insurance and operational workflows*. Available at: https://www.kapto.ai (Accessed: 18 June 2026).

KPMG (2025) *The First Token's Free: The Unknown Long-Term Costs of AI Adoption*. Available at: https://kpmg.com/us/en/articles/2024/first-token-free.html (Accessed: 18 June 2026).

Landgericht München I (2026) Judgement of *28.05.2026, case 26 O 869/26*.

Masterman, T., Besen, S., Sawtell, M. and Cheng, F. (2024) 'The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: a survey', *arXiv preprint*, arXiv:2404.11584. Available at: https://arxiv.org/abs/2404.11584

Negri, G. (2026) 'Obbligo di disclosure per i danni subiti dall'AI', *Il Sole 24 Ore*, 12 June. Available at: https://www.ilsole24ore.com (Accessed: 18 June 2026).

OpenAI, 2024. *Swarm: An educational framework exploring ergonomic, lightweight multi-agent orchestration*. GitHub repository.

OpenView Venture Partners (2024) *The State of SaaS Pricing*. Boston, MA: OpenView Venture Partners. Available at: https://openviewpartners.com (Accessed: 18 June 2026).

Parakhin, V. (2026). Token coherence: Adapting MESI cache protocols to minimize synchronization overhead in multi-agent LLM systems. *arXiv preprint arXiv:2603.15183*.

Patil, C. (2026) 'Beyond Per-Token Pricing: A Concurrency-Aware Methodology for LLM Infrastructure Cost Estimation', *arXiv*, 2606.11690. doi:10.48550/arXiv.2606.11690.

Stanford University Human-Centered Artificial Intelligence (2025) *Artificial Intelligence Index Report 2025*. Stanford, CA: Stanford HAI. Available at: https://aiindex.stanford.edu/report/ (Accessed: 18 June 2026).

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z. and Wen, J.-R. (2024) 'A survey on Large Language Model based autonomous agents', *Frontiers of Computer Science*, 18(6), article 186345. Available at: https://link.springer.com/article/10.1007/s11704-024-40231-1

Zhang, G., Yue, Y., Li, Z., Yun, S., Wan, G., Wang, K., Cheng, D., Yu, J.X. and Chen, T. (2025) 'Cut the crap: an economical communication pipeline for LLM-based multi-agent systems', in *Proceedings of the International Conference on Learning Representations (ICLR 2025)*. Available at: https://openreview.net/forum?id=LkzuPorQ5L (Accessed: 18 June 2026).