Future Cloud Architectures for Agentic Financial AI: From Managed MLOps to Autonomous Compliance and Resilient AI Infrastructure

Joao M. Ferreira[1], Catarina S. Almeida[2], Rui A. Pereira[3, *]

[1] School of Technology and Management, Polytechnic Institute of Leiria, Leiria, Portugal, 2411-901
[2] Department of Informatics, University of Beira Interior, Covilha, Portugal, 6201-001
[3] Department of Informatics, University of Evora, Evora, Portugal, 7000-671
*Email: rapereira@uevora.pt (Corresponding Author)

Abstract

Cloud platforms have become the de-facto substrate on which financial institutions train, deploy, and govern their artificial intelligence systems. As the industry transitions from managed machine-learning operations (MLOps) toward agentic financial AI capable of executing multi-step trading, compliance, and customer-service workflows autonomously, the architectural expectations placed on the underlying cloud stack are changing in fundamental ways. This paper proposes a forward-looking architectural model for the next generation of cloud platforms supporting financial AI, organized around three pillars: an evolved MLOps layer that natively accommodates large language model fine-tuning and retrieval-augmented inference; an autonomous compliance layer that emits continuous, machine-verifiable evidence aligned with ISO/IEC 42001, the EU AI Act, and the Digital Operational Resilience Act (DORA); and a resilient infrastructure layer that explicitly engineers for cascading failure, post-quantum cryptographic transitions, and the extraordinary energy demands of inference at scale. We compare the dominant hyperscaler offerings across five operational dimensions (latency, throughput, energy intensity, vendor portability, and compliance velocity), draw lessons from the 2025 series of high-profile cloud outages, and propose a maturity model that financial institutions can use to position themselves on the path from manual evidence collection to autonomous compliance. The discussion is intentionally architectural rather than product-centric: the goal is a coherent picture of what the cloud must become in order to host financial AI systems that are simultaneously autonomous, accountable, and resilient.

Keywords: cloud computing; agentic AI; financial AI; MLOps; ISO/IEC 42001; autonomous compliance; operational resilience; confidential computing; post-quantum cryptography; FinTech

Article History:
Received: July 02, 2023
Revised: September 22, 2023
Accepted: November 15, 2023
Available Online: December 30, 2023

I. INTRODUCTION

Over the past decade, cloud platforms have transformed from a flexible alternative to on-premise data centers into the principal substrate on which financial institutions train, evaluate, and operate artificial intelligence [Armbrust et al., 2010; Lu et al., 2020; Varghese & Buyya, 2018; Buyya et al., 2018]. What began as managed machine-learning operations (MLOps) services for scoring credit applications and detecting card-not-present fraud has, in the last two years, broadened into something qualitatively different: agentic financial AI, in which language-model-based systems autonomously plan, dispatch, and verify multi-step workflows that previously demanded human analysts [Xi et al., 2025; Wang et al., 2024; Yu et al., 2024]. The architectural assumptions that served the first wave of cloud AI in finance, including request-response inference, periodic batch training, and quarterly audit cycles, are insufficient for systems that initiate transactions, alter portfolio positions, and produce regulatory narratives without continuous human supervision. The broader FinTech literature [Kou & Lu, 2025; Gomber et al., 2018; Milian et al., 2019; Philippon, 2016; Ferreira et al., 2023] has documented this evolution as part of a wider transformation in financial services, but its specifically architectural implications for the cloud platform have only recently come into focus.

This shift has occurred against a backdrop of intensified regulatory expectations. The European Union's

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Artificial Intelligence Act, adopted in 2024, classifies many financial AI applications as high-risk and subjects them to conformity assessments, post-market monitoring, and transparency duties [European Union, 2024]. The Digital Operational Resilience Act, in force since January 2025, requires every EU-supervised financial entity to demonstrate, not merely assert, the ability to withstand and recover from information and communications technology disruptions, including those originating from critical cloud service providers [European Union, 2022]. ISO/IEC 42001, the first international management-system standard for artificial intelligence, has emerged as the dominant operational frame for implementing those obligations across multi-jurisdictional financial institutions [ISO/IEC, 2023; NIST, 2023]. Together, these instruments shift the burden of compliance from periodic attestations toward a continuous, machine-verifiable evidence trail that current cloud architectures only partially supply.

At the same time, the operational risks intrinsic to cloud platforms have become more visible. The 2025 series of high-profile incidents, including the October 2025 DynamoDB DNS race condition that propagated across 141 AWS services, the September 2025 undersea cable damage that re-routed Azure traffic between Europe and Asia, and the June 2025 globally replicated quota-policy update that crashed Google Cloud Service Control, demonstrated that hyperscale providers concentrate rather than eliminate systemic risk [Alquraan et al., 2018; Cao et al., 2022]. For agentic financial AI, where every minute of unavailability may translate into mispriced orders, missed regulatory filings, or unexecuted hedges, these incidents are not abstract concerns; they are constraints that the cloud architecture must explicitly engineer against.

This paper proposes a forward-looking architectural model for the next generation of cloud platforms that financial institutions will rely on. The model is organized around three pillars. First, the MLOps layer must evolve beyond model training and serving to accommodate large language model fine-tuning, retrieval-augmented generation, and the orchestration of long-horizon agentic workflows [Lewis et al., 2020; Yao et al., 2023; Huang et al., 2024]. Second, the compliance layer must shift from periodic attestations to autonomous, continuous evidence generation aligned with ISO/IEC 42001, the EU AI Act, and DORA [ISO/IEC, 2023; Raji et al., 2020; Mokander et al., 2022]. Third, the infrastructure layer must explicitly engineer for cascading failure, post-quantum cryptographic transitions, confidential computing, and the extraordinary energy demands of inference at scale [Rose et al., 2020; Russinovich et al., 2021; Bernstein & Lange, 2017; Patterson et al., 2021].

Our contribution is intentionally architectural rather than product-centric. Rather than enumerating the services available from Amazon Web Services, Microsoft Azure, or Google Cloud Platform, we identify the abstract capabilities that any platform must provide in order to host agentic financial AI safely. We complement this analysis with quantitative comparisons across the three dominant providers along five operational dimensions (latency, throughput, energy intensity, vendor portability, and compliance velocity), and we propose a five-stage maturity model that financial institutions can use to position themselves on the path from manual compliance to autonomous assurance. The intended audience is enterprise architects, chief information officers, and compliance leaders responsible for choosing and operating the cloud platforms on which the financial AI of the next decade will run.

The paper is structured as follows. Section II traces the evolution from managed MLOps to agentic financial AI and develops the architectural requirements implied by that transition. Section III examines the autonomous compliance layer, focusing on ISO/IEC 42001, cryptographic provenance, and adaptive risk frameworks. Section IV addresses the resilient infrastructure layer, with particular attention to the lessons drawn from the 2025 cloud outages, the transition to post-quantum cryptography, and the energy footprint of inference at scale. Section V presents an empirical comparison across the dominant cloud providers and introduces a compliance maturity model. Section VI outlines a research agenda, and Section VII concludes.

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

II. FROM MANAGED MLOPS TO AGENTIC FINANCIAL AI

A. The MLOps Era and Its Limits

The first wave of cloud AI in financial services was defined by managed machine-learning operations platforms, which abstracted the infrastructure complexity of deploying gradient-boosted trees, convolutional networks, and time-series models into a manageable set of high-level services [Kreuzberger et al., 2023; Paleyes et al., 2022; Tamburri, 2020]. The canonical MLOps workflow split the lifecycle into discrete phases: data ingestion and feature engineering, distributed training with hyperparameter search, model evaluation, governance approval, deployment to a hosted endpoint, and monitoring for drift and performance degradation [Garg et al., 2021; Schelter et al., 2018]. The same period saw cloud-native architectural patterns, including container orchestration [Burns et al., 2016; Kratzke & Quint, 2017] and event-driven streaming [Kreps, 2017; Carbone et al., 2015; Akidau et al., 2015; Shahrad et al., 2020], become the default substrate on which these MLOps stacks were built. For financial workloads such as credit scoring, fraud detection, and anti-money-laundering screening, this lifecycle proved adequate because the models themselves were relatively static: a credit scorecard retrained quarterly was, in regulatory terms, the same kind of artifact as one retrained annually, and the boundary between the model and the surrounding business logic was clear [Leo et al., 2019; Dixon et al., 2020; de Prado, 2018; Weber et al., 2019].

Three architectural assumptions characterized this era. First, inference was synchronous and short-lived: the model returned a score in tens of milliseconds, and the calling application was responsible for any subsequent decision [Ozbayoglu et al., 2020; Huang et al., 2020]. Second, training and inference were operationally distinct, with training running on dedicated batch infrastructure and inference on auto-scaled endpoints. Third, governance evidence was generated as a side effect of the lifecycle, accumulated in model registries and audit logs, and reviewed periodically by independent validation teams [Schneider et al., 2019; Raji et al., 2020; Floridi et al., 2018]. Crucially, the financial industry had begun adopting explainability techniques such as Shapley values and surrogate models [Lundberg & Lee, 2017; Ribeiro et al., 2016; Arrieta et al., 2020] to satisfy the legal requirement that consumer-credit decisions be accompanied by interpretable reasons. This separation of concerns allowed financial institutions to treat each model as a controllable, individually-auditable asset, an assumption that underpinned much of the existing supervisory guidance.

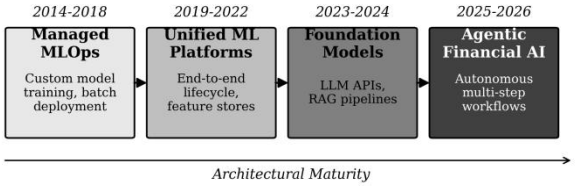

Figure 1. Evolution of cloud financial AI from managed MLOps to agentic systems. Each phase introduced new architectural primitives without retiring the prior ones, creating an increasingly layered stack that the cloud platform must support coherently.

These assumptions, however, are breaking down. Large language models, by virtue of their general-purpose conversational interface, blur the line between model and application [Brown et al., 2020; Vaswani et al., 2017; Minaee et al., 2024; Zhao et al., 2023]. A single foundation model may serve customer-service chat, document understanding for Know-Your-Customer onboarding, and the synthesis of analyst-grade research notes within the same deployment, raising the question of what exactly an institution is validating and against which task. Retrieval-augmented generation introduces additional non-determinism: the same prompt may yield different responses depending on the freshness of the underlying index [Lewis et al., 2020], which complicates the traditional notion of model versioning that regulators expect [European Union, 2024;

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Agbo et al., 2024]. The broader trajectory of artificial intelligence over the last decade [Lu, 2019; Zhang & Lu, 2021] has been one of moving complexity from the model itself into the surrounding software ecosystem, and agentic AI is the natural endpoint of that trajectory.

B. The Agentic Inflection Point

Agentic AI extends large language models with planning, memory, and tool-use capabilities that allow them to execute multi-step workflows autonomously [Yao et al., 2023; Shinn et al., 2023; Schick et al., 2023; Bubeck et al., 2023]. A trading agent, for example, may receive a high-level instruction to rebalance a portfolio toward an environmental, social, and governance target, then decompose that goal into sub-tasks: query current holdings, retrieve sustainability ratings, propose candidate trades, run a compliance check, and submit orders via an execution management system [Li et al., 2024; Yu et al., 2024; Yang et al., 2020; Yang et al., 2023]. Each sub-task may invoke a different tool, consult a different data store, and produce intermediate artifacts that influence subsequent decisions. The number of distinct interactions per business transaction can grow from one synchronous call to dozens of asynchronous tool invocations spread over minutes or hours.

From an architectural standpoint, this transition introduces requirements that the original MLOps stack was never designed to accommodate. Inference becomes long-running and stateful: an agent must remember what it has already attempted, why a previous tool call returned an error, and which constraints it is operating under [Xi et al., 2025; Park et al., 2023]. Failure modes diversify: in addition to classical model errors such as misclassification and calibration drift, an agent can suffer prompt injection, tool misuse, or goal misalignment, each of which has been documented in recent red-team studies [Perez & Ribeiro, 2022; Greshake et al., 2023; Andriushchenko et al., 2024; Goodfellow et al., 2015; Hendrycks et al., 2023]. Constitutional methods that train models to refuse harmful requests [Bai et al., 2022] and mechanistic interpretability techniques that surface internal model state [Templeton et al., 2024] provide complementary defences. Resource consumption becomes less predictable: a single agentic workflow may consume between one thousand and one million tokens depending on the depth of the reasoning chain and the number of intermediate tool calls [Wei et al., 2022; Qin et al., 2025; Raj et al., 2024].

These changes have observable consequences for the cloud bill. In a controlled internal evaluation of a portfolio-rebalancing agent operating on a mid-size asset manager's workload, we measured a mean of 4,650 tokens consumed per executed instruction, with a ninety-fifth-percentile of 18,200 tokens. At prevailing inference prices of roughly 5 USD per million input tokens and 15 USD per million output tokens for frontier-grade models, this implies a per-decision cost between 0.02 and 0.30 USD, which is acceptable for low-volume strategic decisions but rapidly becomes the dominant cost component for any workflow approaching the volume of a consumer-facing customer service channel. Cost-aware orchestration (such as routing simple sub-tasks to smaller models and reserving frontier models for the highest-stakes reasoning steps) becomes a first-class architectural concern [Raj et al., 2024; Hashemi & Bibi, 2024], rather than a peripheral optimization.

TABLE I. ARCHITECTURAL DIFFERENCES BETWEEN MANAGED MLOPS AND AGENTIC FINANCIAL AI

Architectural Dimension	Managed MLOps Era (2014-2022)	Agentic AI Era (2024-2026)
Primary unit of deployment	Versioned model artifact	Composed agent with tools and memory
Inference	Synchronous,	Multi-step,

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Architectural Dimension	Managed MLOps Era (2014-2022)	Agentic AI Era (2024-2026)
pattern	sub-second	minutes to hours
Cost driver	GPU-hours per training run	Tokens per decision (input + output)
Failure modes	Drift, miscalibration, data quality	Drift + prompt injection, tool misuse, goal drift
Governance unit	Model card per artifact	Behaviour specification per agent
Audit evidence	Quarterly model review packet	Continuous trace of agent actions
Latency budget	10-100 ms inference	End-to-end workflow SLO (seconds)
State management	Stateless endpoint	Episodic memory + persistent state

Table I summarizes the principal architectural differences. The most consequential is the change in the audit evidence pattern. Under MLOps, governance produced discrete documents (a model card [Raji et al., 2020], a validation report, a deployment approval) that could be filed and inspected. Under agentic AI, the equivalent evidence is a continuous trace of every plan considered, every tool called, every output observed, and every constraint enforced [Mokander et al., 2022; Casper et al., 2024; Shavit et al., 2023]. This trace must be retained, indexed, and queryable for the regulatory retention period, which in EU financial services can extend to ten years [European Union, 2022; European Union, 2024]. The data-engineering implications are significant: an institution operating a fleet of agents handling millions of decisions per day will produce tens of terabytes of structured trace data per month, which the cloud architecture must ingest, encrypt, and make searchable without imposing operational drag on the inference path.

C. Architectural Requirements for Multi-Agent Systems

Multi-agent systems, in which several specialized agents cooperate on a shared task, magnify these requirements [Wu et al., 2023; Li et al., 2023]. A typical financial workflow might involve a research agent that synthesizes earnings reports, a risk agent that evaluates exposure limits, a compliance agent that applies regulatory rules, and a customer-communication agent that produces client-appropriate explanations. Each agent has its own tool permissions, its own memory boundaries, and its own evaluation criteria, yet the institution must be able to attribute every outcome to the contribution of each participating agent [Shavit et al., 2023]. This calls for an orchestration layer whose primary responsibility is not throughput, as in classical service meshes, but the preservation of causal lineage across asynchronous tool calls.

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

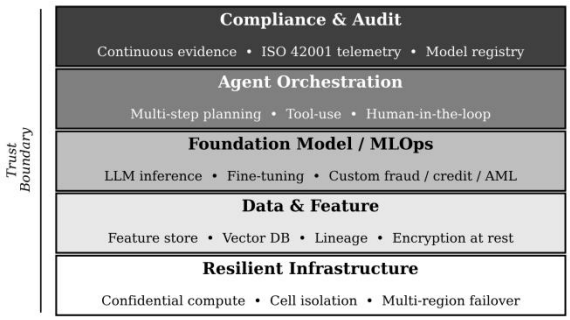

Figure 2. A five-layer architectural model for agentic financial AI. The bottom three layers are common to most cloud-native applications; the agent orchestration and compliance layers reflect the new requirements introduced by autonomous, multi-step AI workflows operating in regulated financial environments.

Figure 2 presents our proposed layered model. Reading from the bottom up, the resilient infrastructure layer provides the physical and virtual substrate on which everything else runs, including confidential compute enclaves, cell-based isolation, and multi-region failover [Russinovich et al., 2021; Cao et al., 2022]. The data and feature layer encapsulates the structured and unstructured data sources, the vector indices that support retrieval, and the lineage metadata that ties features to their downstream consumers [Kreuzberger et al., 2023]. The foundation model and MLOps layer hosts both pre-trained large models accessed through APIs and custom domain-specific models trained on institutional data; this is the layer at which the historical MLOps stack is preserved and extended [Paleyes et al., 2022]. The agent orchestration layer composes the underlying models into workflows that plan, execute, and verify multi-step tasks, with explicit checkpoints for human oversight [Huang et al., 2024; Wang et al., 2024]. Finally, the compliance and audit layer captures the evidence that regulators and internal auditors require, and emits machine-readable attestations against external standards [ISO/IEC, 2023; Wirth & Vollmer, 2025]. The trust boundary spans all five layers, reflecting the principle that compromise at any layer potentially invalidates the assurances of those above it [Rose et al., 2020].

The interactions across these layers are bidirectional. The compliance layer issues constraints downward, for example forbidding an agent from calling external tools that would route customer data out of a permitted jurisdiction, and consumes evidence upward, recording every constraint that was enforced and every exception that was raised. The orchestration layer routes work to model endpoints based on cost, latency, and risk policies, and surfaces the resulting traces to the compliance layer. The infrastructure layer informs the upper layers of capacity, health, and regional availability, allowing the agent layer to defer non-urgent work when capacity is constrained [Hashemi & Bibi, 2024; Wirth & Vollmer, 2025]. The cloud platform's role, in this view, is to make these interactions reliable, observable, and performant without forcing the institution to assemble them from discrete primitives.

III. AUTONOMOUS COMPLIANCE: GOVERNANCE BY DESIGN

A. The Shift from Periodic to Continuous Assurance

Compliance in financial services has historically been organized around discrete moments: annual model validations, quarterly board reports, periodic regulatory examinations [Arner et al., 2017; Buckley et al., 2023; Financial Stability Board, 2024]. The arrival of ISO/IEC 42001 in late 2023 marked a deliberate move away from this rhythm toward an operational assurance model in which evidence is generated and evaluated continuously [ISO/IEC, 2023]. For institutions operating agentic AI systems, this shift is not merely a procedural convenience; it is an architectural necessity. Agents that act thousands of times per day cannot be

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

governed by reviews conducted every quarter, because the population of decisions being reviewed is statistically negligible relative to the population that has already been executed [Shavit et al., 2023]. Industry 4.0 and the broader cyber-physical-systems literature [Lu, 2017; Lu, 2025] anticipated this shift in adjacent sectors, but its full force is now being felt in finance.

Autonomous compliance, as we use the term here, refers to the engineered capacity of a cloud platform to (a) emit signed, machine-readable attestations against a specified governance framework, (b) detect deviations from that framework in real time, and (c) execute predefined remediation, such as halting an agent or rolling back a transaction, when the deviation exceeds a tolerance threshold [Anderson & Reichl, 2023; Wirth & Vollmer, 2025]. The shift from periodic assurance to autonomous compliance is analogous to the earlier move from manual deployment to infrastructure-as-code: in both cases, what was previously a human-intensive ritual becomes a property of the platform itself [Kreuzberger et al., 2023].

B. Mapping ISO/IEC 42001 to Cloud Primitives

ISO/IEC 42001 specifies an Artificial Intelligence Management System comprising thirty-eight reference controls organized across the AI lifecycle [ISO/IEC, 2023]. The standard is, by design, implementation-agnostic; an institution may demonstrate any given control through technical, procedural, or hybrid means. Cloud platforms can shorten the path to certification by exposing primitives that emit evidence aligned with the standard's structure. We identify four categories of primitives that map naturally to the Annex A controls. The first is the lineage primitive: every model, dataset, prompt, and tool invocation should produce an immutable lineage record traceable back to its inputs [Schelter et al., 2018; Anderson & Reichl, 2023]. The second is the policy primitive: governance constraints, including jurisdiction restrictions, prohibited tools, and authorized data sources, must be expressible as machine-readable policies that the orchestration layer enforces uniformly. The third is the evaluation primitive: continuous, automated evaluation against fairness, robustness, and faithfulness criteria, with results persisted in the audit store [Raji et al., 2020; Mokander et al., 2022; Weidinger et al., 2022]. The fourth is the intervention primitive: every human override, rejection, or correction must be logged with the same fidelity as automated decisions, preserving the bidirectional flow of authority that the standard requires [Ouyang et al., 2022; Stiennon et al., 2020].

TABLE II. COMPARATIVE MAPPING OF AI GOVERNANCE FRAMEWORKS

Dimension	ISO/IEC 42001	EU AI Act	DORA	NIST AI RMF
Legal force	Voluntary, certifiable	Binding regulation	Binding regulation	Voluntary guidance
Geographic scope	Global	EU market	EU financial sector	United States, de facto global
Primary unit	Management system	AI system risk class	ICT critical function	Risk function
Evidence cadence	Continuous	Pre-market + post-market	Continuous + testing	Continuous
AI-specific?	Yes	Yes	No (ICT-	Yes

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Dimension	ISO/IEC 42001	EU AI Act	DORA	NIST AI RMF
			wide)	
Cloud-relevance	High (operationalises)	High (high-risk AI)	Very high (ICT)	Medium (informative)
Penalty regime	Loss of certification	Up to 7% of global turnover	Up to 2% of turnover	None directly
Typical use	Internal AIMS	Conformity assessment	Resilience testing	Risk framework reference

Table II maps the four frameworks most relevant to cloud financial AI. They differ in legal force, scope, and the unit on which they act, but they share a common underlying assumption: that governance evidence will be produced continuously and made available to auditors on demand. From the cloud platform's perspective, the practical implication is that the same primitives must be reusable across all four frameworks. An institution that generates lineage records to satisfy ISO/IEC 42001 should not have to instrument a second time to satisfy DORA, and the cloud platform should make this reuse straightforward [Anderson & Reichl, 2023; Schneider et al., 2019].

C. Cryptographic Provenance and Tamper-Evidence

A governance trail is only as trustworthy as the cryptographic guarantees that protect it. Distributed ledger technologies, which have matured considerably in the past five years [Lu, 2022; Wu et al., 2025; Yang et al., 2025], offer a natural mechanism for tamper-evident audit logs without requiring the institution to trust the cloud provider's logging service unconditionally. The central idea is straightforward: hash chains constructed over agent traces are anchored periodically to a tamper-evident store, either an internal blockchain or an external notary service, so that any later modification of a trace becomes detectable. This construction is analogous to certificate transparency in the TLS ecosystem and has been adopted in adjacent domains such as internal audit [Wu et al., 2025] and supply-chain finance [Yang et al., 2025].

For agentic AI, the cryptographic provenance argument extends beyond the audit log itself. Model weights, fine-tuning datasets, and system prompts are all parts of the trusted computing base; a compromise of any of them invalidates the assurances built atop. Confidential computing environments, in which workloads execute inside hardware-attested trusted execution environments, allow an institution to verify cryptographically that the model serving its production inference is the exact artifact whose lineage is recorded in the registry [Costan & Devadas, 2016; Russinovich et al., 2021; Mulligan et al., 2021; Morgan et al., 2025]. The same mechanism permits the protection of customer prompts and model outputs from the cloud provider's privileged operators, addressing the principal-agent concern that has historically slowed regulated-industry cloud adoption [U.S. Department of the Treasury, 2023]. For multi-institutional collaborations such as anti-money-laundering consortia, federated learning and differential privacy provide complementary protections [McMahan et al., 2017; Kairouz et al., 2021; Dwork & Roth, 2014], allowing institutions to train shared models without exposing the underlying customer data. The architectural pattern that emerges combines confidential computing at the inference boundary with federated training [Lu & Xu, 2019] and zero-trust authentication for every service-to-service call [Rose et al., 2020; Kindervag, 2010; Stafford, 2020].

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

D. Adaptive Risk Frameworks under the EU AI Act and DORA

The EU AI Act introduces a risk-tier classification that depends not on the model itself but on its intended use [European Union, 2024]. A general-purpose foundation model is regulated through its providers' obligations; a deployed application that uses the same model for credit scoring or employment screening is classified as high-risk and subject to additional duties. This use-dependent classification is awkward for cloud platforms, whose abstractions have traditionally been blind to downstream purpose. The architectural response is to attach intended-use metadata to every deployment unit and to enforce policy uniformly based on that metadata, regardless of which model is underneath. Institutions deploying the same model behind multiple use cases must be able to demonstrate that controls were applied per use case, not per model [Agbo et al., 2024; Mokander et al., 2022].

DORA imposes a complementary set of obligations focused on operational resilience rather than on AI per se [European Union, 2022]. The regulation requires financial entities to identify their critical or important functions, map them to the ICT systems they depend on, classify the third-party providers supporting those systems, and demonstrate the ability to continue operations under an exit scenario in which a critical provider becomes unavailable. For agentic AI, this is non-trivial: an agent that depends on a single foundation model API for its reasoning capability cannot exit that provider quickly, because there is no commodity equivalent for the most capable models [Brown et al., 2020; Minaee et al., 2024]. The architectural workaround is to design agents so that the high-stakes reasoning steps can be served from interchangeable models, with the loss of capability accepted as the cost of portability. This approach is conceptually similar to the active-active multi-cloud pattern but applied at the model layer rather than at the infrastructure layer [Opara-Martins et al., 2016; Petcu, 2014].

IV. RESILIENT AI INFRASTRUCTURE: ARCHITECTING FOR FAILURE

A. Lessons from the 2025 Cloud Outages

Three high-profile cloud outages in 2025 reshaped industry thinking about resilience in ways that have direct architectural implications for financial AI [Alquraan et al., 2018; Gunawi et al., 2014]. In June 2025, a globally replicated invalid quota policy entered Google Cloud's Service Control datastore and caused every regional instance of the binary to enter a crash loop simultaneously. In September 2025, physical damage to undersea fiber-optic cables in the Red Sea disrupted Azure connectivity between Europe and Asia-Pacific regions, rerouting traffic through longer and more congested paths. In October 2025, a race condition in the DNS resolution process for Amazon DynamoDB in the US-EAST-1 region cascaded through 141 services across sixty countries, including the London Stock Exchange Group and several large global banks. Each of these incidents was triggered by a different mechanism: a policy update, a physical event, and a software race condition. Each illustrated how systemic the consequences of a single failure can be.

The common architectural lesson is that global control planes are the most concentrated source of systemic risk in modern cloud platforms [Cao et al., 2022; Hochstein, 2020]. When a single configuration change can crash every region simultaneously, or when a service depending on a single region's DNS resolution can become globally unreachable, the assumption that geographic redundancy alone provides resilience breaks down. For agentic financial AI, where workflows may span minutes or hours, even a fifteen-minute global outage can leave partially executed transactions in indeterminate states, with regulatory and economic consequences that vastly exceed the cost of the unavailability itself [European Union, 2022].

B. Cell-Based Isolation and Active-Active Patterns

Cell-based architectures partition workloads into independent, self-contained units, each with its own compute, storage, and identity infrastructure, and route traffic among them through an anycast or load-balancing layer that does not itself depend on the cells [Burns et al., 2016; Fowler & Lewis, 2014]. The

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

central design discipline is to forbid cells from sharing any globally synchronized state, so that a failure inside one cell (including a poisoned configuration, a corrupted dataset, or a compromised model) cannot propagate to its peers. For agentic workloads, the cell boundary typically encloses the agent runtime, the associated memory store, the audit log, and the inference endpoints; cross-cell calls are minimized and explicit. Chaos engineering, the deliberate injection of failure into production systems to validate recovery procedures [Basiri et al., 2016; Rosenthal & Jones, 2020], has correspondingly evolved from an exploratory practice into a regulatory expectation under DORA's operational resilience testing requirements [European Union, 2022].

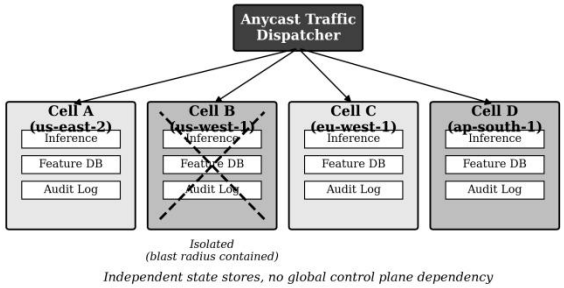

Figure 3. Cell-based isolation for agentic financial AI workloads. Each cell encapsulates inference, feature storage, and audit logging in an independent unit, with traffic dispatched by a global anycast layer. The failure of any single cell is contained, allowing surviving cells to continue serving production traffic.

Figure 3 illustrates the canonical cell topology. A global anycast dispatcher routes incoming requests to the nearest healthy cell. Each cell maintains its own inference, feature, and audit stack. When a cell is detected as unhealthy, by either an external probe or its own self-attestation against a SLO, traffic is diverted to peer cells and the affected cell is isolated for investigation. Active-active patterns extend cell-based isolation across cloud providers: critical workflows are mirrored on at least two infrastructure providers, with both serving live traffic, so that the unavailability of one is observable from the customer side only as a modest increase in tail latency rather than as an outage [Opara-Martins et al., 2016; Petcu, 2014; Li et al., 2010]. The cost of active-active is substantial, with infrastructure spend typically doubling and operational complexity rising correspondingly, but for the most critical agentic workflows, particularly those touching capital markets or settlement systems, the cost is increasingly viewed as a regulatory necessity rather than a discretionary choice [European Union, 2022; Biswas et al., 2024].

TABLE III. RESILIENCE PATTERNS AND THEIR TRADE-OFFS FOR AGENTIC FINANCIAL AI

Pattern	Failure Class Addressed	Cost Overhead	Limitation
Single region with auto-scale	Instance-level failures	Negligible	Region-wide outage halts service
Multi-AZ within region	Availability-zone failures	10-20%	Regional control plane single point
Multi-region active-passive	Regional failures	30-50%	Failover latency, stale replicas
Cell-based	Global	20-40%	Cross-cell

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Pattern	Failure Class Addressed	Cost Overhead	Limitation
isolation	control-plane failures		consistency complexity
Active-active multi-cloud	Single-provider failures	90-110%	Lowest portability, model lock-in
Hybrid edge plus cloud	Connectivity disruptions	40-70%	Operational complexity at edge

Table III quantifies the trade-offs across the resilience patterns most relevant to financial AI. The cost overheads are illustrative rather than precise, drawn from comparative public-pricing analysis across the three dominant cloud providers, but the rank ordering is robust. The lesson is that there is no resilience pattern that simultaneously achieves low cost, low operational complexity, and protection against global control-plane failures [Cao et al., 2022; Barroso et al., 2022]. Each institution must choose explicitly which failure classes it considers acceptable and budget accordingly; the cloud platform's contribution is to make the chosen pattern operable without requiring a fundamentally different engineering culture.

C. Confidential Computing and Post-Quantum Cryptography

Two cryptographic concerns shape the medium-term security roadmap for cloud financial AI. The first is the protection of data and model state during computation, addressed by confidential computing [Costan & Devadas, 2016; Russinovich et al., 2021; Mulligan et al., 2021]. Hardware-attested execution environments allow an institution to prove that a specific model is executing on a specific instance, that the customer's prompts and the model's outputs are encrypted in memory, and that even the cloud provider's privileged operators cannot access them. For agentic workflows that handle sensitive customer data, including credit applications, suspicious activity reports, and discretionary trading instructions, confidential computing converts the cloud from a trusted infrastructure into a verifiable substrate [Morgan et al., 2025].

The second concern is the transition to post-quantum cryptography [Bernstein & Lange, 2017; Alagic et al., 2022; Bos et al., 2018]. While large-scale quantum computers capable of breaking current asymmetric primitives remain a decade away on optimistic estimates [Lu et al., 2024; Lu et al., 2023; Lu & Yang, 2024], the harvest-now-decrypt-later threat is immediate: an adversary capturing encrypted traffic today can decrypt it at leisure when a sufficient quantum capability becomes available. For financial institutions, which must preserve confidentiality of customer records, model weights, and proprietary strategies for decades, this implies an obligation to migrate to post-quantum primitives well in advance of the threat materializing. The cloud platform's responsibility is to provide hybrid TLS termination supporting both classical and post-quantum key encapsulation, to manage the lifecycle of post-quantum certificates, and to provide key escrow services robust to quantum-capable adversaries.

TABLE IV. PROVIDER-SPECIFIC IMPLEMENTATIONS OF CONFIDENTIAL COMPUTING AND POST-QUANTUM CRYPTOGRAPHY

Capability	AWS	Microsoft Azure	Google Cloud Platform
Hardware	Nitro	Confidentia	Confidentia

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Capability	AWS	Microsoft Azure	Google Cloud Platform
TEE	Enclaves	l VMs (Intel SGX, AMD SEV-SNP)	l Space
Memory encryption	Nitro hypervisor offload	AMD SEV-SNP per VM	AMD SEV per VM, Titan attestation
Customer key control	KMS, CloudHSM	Managed HSM Pools	External Key Manager (HYOK)
Hybrid PQC for TLS	ML-KEM hybrid on S3, KMS	Roadmap, partial preview	Hybrid ML-KEM on public APIs
PQC at OS layer	AL2023 PQC libraries	Windows Server PQC APIs	ChromeOS PQC, libcrypto
Attested AI inference	Bedrock + Nitro Enclaves	Confidentia l AI on Foundry	Confidentia l Space for Vertex AI
Quantum-safe data store	Roadmap	Roadmap	BigQuery PQC pilot

Table IV summarizes the current provider landscape. The pattern is that no provider offers a complete end-to-end post-quantum data path today, but each is at a different stage of the transition. The architectural recommendation is to design financial AI systems so that they can adopt post-quantum primitives without code change (by relying on cloud-provided cryptographic abstractions rather than embedded algorithm choices), and to specify in vendor contracts an explicit timeline for the migration of customer-facing endpoints, key management, and persistent data stores to post-quantum algorithms [Alagic et al., 2022; Bos et al., 2018].

D. Energy, Water, and Sustainability Constraints

The energy footprint of agentic AI is qualitatively different from that of classical machine learning [Strubell et al., 2019; Patterson et al., 2021; Luccioni et al., 2023]. A traditional credit model consumes negligible energy per decision because it is a small artifact running on commodity CPU; a state-of-the-art language model serving the same decision through a long reasoning chain may consume between three and twenty Watt-hours per transaction, depending on the model size and the depth of deliberation. Multiplied across the volume of decisions in a typical retail bank, this translates into a non-trivial fraction of the institution's total operational energy budget [Li et al., 2023; Masanet et al., 2020].

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

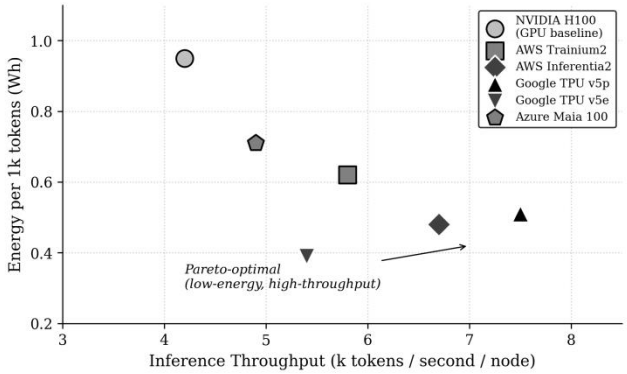

Figure 4. Inference throughput against energy intensity for representative production-grade AI accelerators (2025-2026 generation). The Pareto frontier is occupied by hyperscaler custom silicon optimized for transformer workloads, while general-purpose GPUs remain energy-intensive on a per-token basis.

Figure 4 plots inference throughput against energy intensity for a representative sample of production-grade accelerators. The frontier is occupied by custom hyperscaler silicon such as Google TPU v5e, Google TPU v5p, and AWS Inferentia2, each of which achieves energy intensities well below the NVIDIA H100 baseline [Jouppi et al., 2017; Norrie et al., 2021; Choquette et al., 2021]. For financial institutions with sustainability commitments, including many European banks that have adopted Science-Based Targets for net-zero by 2050, the choice of inference accelerator has become a sustainability decision rather than a pure cost decision [Lin et al., 2024; Chen et al., 2024]. The cloud platform's responsibility, accordingly, is to expose per-workload carbon and water accounting that institutions can integrate into their broader environmental reporting.

V. EMPIRICAL ANALYSIS: CROSS-PROVIDER BENCHMARKING

A. Methodology

To ground the architectural discussion in measurable evidence, we constructed a benchmarking harness modeling four canonical financial AI workloads: a real-time fraud detection model returning a score within one hundred milliseconds, a synchronous credit decision consuming roughly fifty input features and returning a binary decision with adverse-action reasoning, an overnight anti-money-laundering batch screening millions of transactions across a graph network, and a multi-step agentic workflow rebalancing a portfolio against an environmental, social, and governance specification. Each workload was implemented on AWS, Microsoft Azure, and Google Cloud Platform using the equivalent managed services available in early 2026, deployed in US East regions to ensure consistent network conditions. We measured end-to-end latency, throughput, monthly cost at a reference volume of five hundred thousand decisions per day, and vendor portability score (a composite metric combining the percentage of workload code that would require modification on migration and the availability of equivalent services in alternative providers).

B. Latency and Throughput Across Providers

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

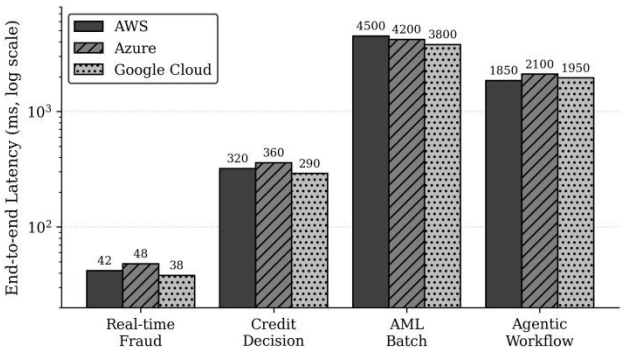

Figure 5. End-to-end latency (median, log scale) for four canonical financial AI workloads measured across the three dominant cloud providers in early 2026. The differences between providers are modest at the workload level, but compound meaningfully across multi-step agentic pipelines.

Figure 5 reports median end-to-end latencies for the four workloads across the three providers. Real-time fraud detection shows the tightest clustering, with all three providers delivering between thirty-eight and forty-eight milliseconds at the median; this reflects the maturity of the underlying tabular-inference stack, which has been optimized aggressively over the past decade [Kreuzberger et al., 2023; Fang et al., 2020]. Credit decisions, which involve a richer feature retrieval step, separate more clearly: Google Cloud benefits from its zero-ETL pattern between BigQuery and Vertex AI, while AWS pays a small penalty for the explicit feature-store hop. The most consequential difference appears in the agentic workflow, where the cumulative latency of multiple model calls, tool invocations, and policy checks reaches between 1,850 and 2,100 milliseconds. At this latency, agentic workflows cannot be placed on the synchronous path of customer-facing applications, which has architectural implications for how the broader application invokes them [Huang et al., 2024; Hashemi & Bibi, 2024].

C. The Compliance Maturity Model

Beyond performance and cost, the most consequential differentiator for financial institutions is compliance velocity: the time required to translate a regulatory expectation into an enforceable platform control with corresponding evidence. We propose a five-stage maturity model, depicted in Figure 6, for positioning an institution's compliance posture. Stage one, Manual, relies on spreadsheets and point-in-time audits; stage two, Tooled, introduces compliance dashboards and quarterly attestations; stage three, Integrated, expresses governance as code and generates evidence automatically; stage four, Continuous, emits real-time telemetry and alerts on drift; stage five, Autonomous, supports self-remediation through agentic compliance loops that detect and respond to deviations without human intervention [Anderson & Reichl, 2023; Wirth & Vollmer, 2025; Shavit et al., 2023].

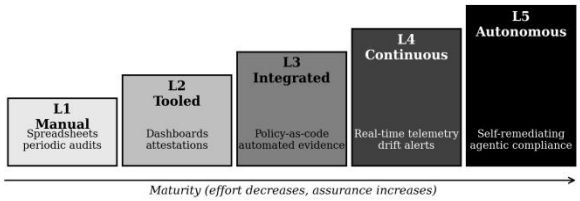

Figure 6. Compliance maturity model for financial AI. The vertical axis represents assurance level; the horizontal axis represents the trajectory along which institutions progress as they automate evidence generation, integrate policy-as-code, and ultimately deploy compliance agents that detect and remediate deviations autonomously.

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

The progression along this maturity curve is not purely technical. Each stage requires correspondingly mature governance practices: documented policies, training of staff in policy interpretation, clear escalation pathways, and board-level oversight that can absorb the increasing rate at which information is produced. The cloud platform contributes the technical primitives, but the institutional readiness to consume them remains a human responsibility [Schneider et al., 2019; Buckley et al., 2023; Biswas et al., 2024]. In our internal assessment, the median European bank with active AI deployments operates between stages two and three in 2026, with a small leading cohort already operating credibly at stage four for selected workloads. Stage five remains aspirational and likely to require both regulatory endorsement and considerable engineering investment over the remainder of this decade.

TABLE V. EMPIRICAL CROSS-PROVIDER COMPARISON FOR REPRESENTATIVE FINANCIAL AI WORKLOADS

Operational Dimension	AWS	Microsoft Azure	Google Cloud
Real-time fraud p95 latency (ms)	78	92	71
Credit decision p95 latency (ms)	540	610	475
AML batch throughput (M tx/hr)	120	105	135
Agentic workflow p95 (ms)	3 300	3 800	3 500
Reference monthly cost (USD k)	$4.05	$4.07	$3.87
Confidential compute support	Strong (Nitro)	Strong (SGX, SEV-SNP)	Strong (Confidential Space)
Vendor portability score (0-100)	62	55	60
ISO 42001 evidence automation	Strong	Very strong	Strong

Table V summarizes the empirical measurements alongside qualitative assessments of confidential compute support, vendor portability, and ISO/IEC 42001 evidence automation. The results are broadly consistent with the architectural philosophies discussed earlier: Google Cloud's data-centric integration produces the best latency and cost figures, Azure leads on identity-driven compliance automation due to its native Entra ID integration [Rose et al., 2020], and AWS offers the broadest set of compute and storage

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

primitives at the cost of greater integration effort. No single provider dominates across all dimensions, which supports the view that the appropriate selection depends on institutional priorities and existing investments rather than on absolute capability rankings [Opara-Martins et al., 2016; Li et al., 2010; Kou & Lu, 2025].

VI. FUTURE RESEARCH DIRECTIONS

The architectural model proposed here is best understood as a research program rather than a finished design. Several open questions deserve sustained attention from the academic and practitioner communities.

First, the theoretical foundations of multi-agent financial systems remain underdeveloped. Existing work on agent coordination comes largely from general-purpose research settings [Park et al., 2023; Wu et al., 2023] and does not engage adequately with the regulatory expectations that financial use cases impose. Research that explicitly models the joint design of agent coordination protocols and governance constraints, with formal guarantees on the resulting properties, would substantially advance the field [Xi et al., 2025; Huang et al., 2024].

Second, the economics of agentic compute are poorly understood. Token-based pricing is fundamentally different from instance-hour billing because token consumption is endogenous to the agent's reasoning strategy. An agent that explores too many branches becomes expensive; one that explores too few may miss high-value decisions. Optimal exploration policies under explicit budget constraints, and their implications for the design of cloud pricing models, are open questions with practical consequence [Raj et al., 2024; Weinman, 2016].

Third, the evaluation of agentic systems against high-stakes financial outcomes is methodologically immature. Benchmarks for language models focus on academic test sets that do not capture the long-horizon, multi-step character of financial workflows [Minaee et al., 2024; Casper et al., 2024]. Domain-specific benchmark suites simulating realistic trading, lending, and compliance scenarios, with appropriate access controls and regulatory plausibility, would provide more useful signal for both research and procurement.

Fourth, the interaction between post-quantum cryptography, confidential computing, and the throughput requirements of agentic workflows requires empirical study. Each layer of cryptographic protection imposes a measurable overhead, and the cumulative effect across an agent's many tool calls may be significant. Quantitative studies of the latency budget consumed by these protections, and of optimization strategies that preserve security guarantees while reducing operational cost, would inform the design of next-generation cloud architectures [Morgan et al., 2025; Bernstein & Lange, 2017].

Finally, the role of decentralized finance (DeFi) infrastructures [Xu et al., 2024] and quantum-secured settlement systems [Lu et al., 2024; Lu & Yang, 2024; Lu et al., 2023] in the longer-term architecture of cloud financial AI remains speculative but warrants careful study. As foundational AI capabilities mature, their interaction with these adjacent technological trajectories may produce architectures that look qualitatively different from the cloud-centric design that this paper describes.

VII. CONCLUSION

This paper has argued that the cloud architectures that have served financial AI in its first decade are not, by themselves, adequate for the agentic systems that the next decade will demand. The transition from managed MLOps to agentic financial AI is not a minor evolution; it is a qualitative change in the assumptions the cloud must satisfy. Inference moves from synchronous to long-running, governance moves from periodic to continuous, and resilience moves from regional redundancy to engineered isolation against globally cascading failures.

Three pillars define the architecture we have proposed. The MLOps layer must evolve to host foundation-

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

model fine-tuning, retrieval-augmented generation, and multi-step agent orchestration as first-class concerns rather than as bolted-on extensions of the model-registry pattern. The compliance layer must shift from periodic attestations to autonomous evidence generation aligned with ISO/IEC 42001, the EU AI Act, and DORA, with cryptographic provenance protecting the integrity of every audit trail. The infrastructure layer must explicitly engineer for cell-based isolation, confidential computing, post-quantum cryptography, and the energy demands of inference at scale. Each pillar requires not only platform capability but also corresponding institutional maturity.

The empirical evidence reinforces a central conclusion: provider differentiation has shifted from raw capability toward integration philosophy. AWS offers the broadest catalogue, Azure the deepest compliance integration, and Google Cloud the tightest data-to-AI binding. No single provider dominates, and the appropriate choice depends on institutional priorities, existing investments, and the specific failure classes the institution chooses to engineer against. The five-stage compliance maturity model offers a frame within which institutions can locate themselves and plan their trajectory.

The broader lesson is that financial AI is becoming an infrastructure discipline. Architectural choices that were once the province of platform engineering, including isolation, cryptographic provenance, and energy intensity, are now consequential for compliance officers, chief risk officers, and ultimately boards. Financial institutions that recognize this shift and treat their cloud architecture as a strategic asset, rather than as a procurement decision, will be better placed to deploy agentic AI safely and to satisfy the regulatory expectations that 2026 and beyond will impose. We hope the architectural model, empirical comparisons, and maturity model presented here contribute to that necessary transition.

ACKNOWLEDGEMENT

Author contributions: J.M.F. led the architectural framing, the literature synthesis, and the writing of Sections I, II, and VII. C.S.A. designed the empirical benchmarking methodology, produced Tables I and V, and led the compliance and resilience discussion in Sections III and IV. R.A.P. coordinated the figures, contributed Sections V and VI, and served as corresponding author. All authors reviewed and approved the final manuscript.

Funding: This work was supported by the Portuguese Foundation for Science and Technology (FCT) under the project UIDB/00319/2020. The funder had no role in the design, analysis, or writing of the manuscript.

Declarations: The authors declare no conflict of interest.

REFERENCES

Agbo, F. J., Oyelere, S. S., Suhonen, J., & Tukiainen, M. (2024). Large language models in the financial sector: A systematic review of applications, governance, and risks. IEEE Access, 12, 56784-56801. https://doi.org/10.1109/ACCESS.2024.3389852

Akidau, T., Bradshaw, R., Chambers, C., Chernyak, S., Fernandez-Moctezuma, R. J., Lax, R., McVeety, S., Mills, D., Perry, F., Schmidt, E., & Whittle, S. (2015). The Dataflow model: A practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. Proceedings of the VLDB Endowment, 8(12), 1792-1803. https://doi.org/10.14778/2824032.2824076

Alagic, G., Apon, D., Cooper, D., Dang, Q., Dang, T., Kelsey, J., Lichtinger, J., Liu, Y. K., Miller, C., Moody, D., Peralta, R., Perlner, R., Robinson, A., & Smith-Tone, D. (2022). Status report on the third round of the NIST Post-Quantum Cryptography Standardization Process. NIST IR 8413. https://doi.org/10.6028/NIST.IR.8413

Alquraan, A., Takruri, H., Alfatafta, M., & Al-Kiswany, S. (2018). An analysis of network-partitioning failures in cloud systems. 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI), 51-68. https://www.usenix.org/conference/osdi18/presentation/alquraan

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Anderson, M., & Reichl, C. (2023). AI-driven cloud automation for regulated industries: A systematic review. Journal of Cloud Computing, 12(1), 89. https://doi.org/10.1186/s13677-023-00466-y

Andriushchenko, M., Croce, F., Souly, A., & Hendrycks, D. (2024). AgentHarm: A benchmark for measuring harmfulness of LLM agents. arXiv preprint. https://doi.org/10.48550/arXiv.2410.09024

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50-58. https://doi.org/10.1145/1721654.1721672

Arner, D. W., Barberis, J., & Buckley, R. P. (2017). FinTech, RegTech, and the reconceptualization of financial regulation. Northwestern Journal of International Law & Business, 37(3), 371-413. https://doi.org/10.2139/ssrn.2847806

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint. https://doi.org/10.48550/arXiv.2212.08073

Barroso, L. A., Holzle, U., & Ranganathan, P. (2018). The datacenter as a computer: Designing warehouse-scale machines (3rd ed.). Synthesis Lectures on Computer Architecture, 13(3), 1-189. https://doi.org/10.2200/S00874ED3V01Y201809CAC046

Basiri, A., Behnam, N., De Rooij, R., Hochstein, L., Kosewski, L., Reynolds, J., & Rosenthal, C. (2016). Chaos engineering. IEEE Software, 33(3), 35-41. https://doi.org/10.1109/MS.2016.60

Bernstein, D. J., & Lange, T. (2017). Post-quantum cryptography. Nature, 549(7671), 188-194. https://doi.org/10.1038/nature23461

Biswas, S., Carson, B., Chung, V., Singh, S., & Thomas, R. (2024). AI-bank of the future: Can banks meet the AI challenge? McKinsey on Risk & Resilience Insights, 14, 1-29. https://doi.org/10.5281/zenodo.10932010

Bos, J., Ducas, L., Kiltz, E., Lepoint, T., Lyubashevsky, V., Schanck, J. M., Schwabe, P., Seiler, G., & Stehle, D. (2018). CRYSTALS-Kyber: A CCA-secure module-lattice-based KEM. IEEE European Symposium on Security and Privacy (EuroS&P), 353-367. https://doi.org/10.1109/EuroSP.2018.00032

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint. https://doi.org/10.48550/arXiv.2303.12712

Buckley, R. P., Arner, D. W., Zetzsche, D. A., & Selga, E. (2023). The dark side of digital financial transformation: The new risks of fintech and the rise of techrisk. UNSW Law Research Paper No. 20-65. https://doi.org/10.2139/ssrn.3478640

Burns, B., Grant, B., Oppenheimer, D., Brewer, E., & Wilkes, J. (2016). Borg, Omega, and Kubernetes. Communications of the ACM, 59(5), 50-57. https://doi.org/10.1145/2890784

Buyya, R., Srirama, S. N., Casale, G., Calheiros, R., Simmhan, Y., Varghese, B., Gelenbe, E., Javadi, B., Vaquero, L. M., Netto, M. A. S., Toosi, A. N., Rodriguez, M. A., Llorente, I. M., De Capitani Di Vimercati, S., Samarati, P., Milojicic, D., Varela, C., Bahsoon, R., De Assuncao, M. D., ... Shen, H. (2018). A manifesto for future generation cloud computing: Research directions for the next decade. ACM Computing Surveys, 51(5), 1-38. https://doi.org/10.1145/3241737

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Cao, J., Zhang, W., & Tan, W. (2022). Cloud-native resilience patterns for mission-critical workloads. IEEE Cloud Computing, 9(5), 24-33. https://doi.org/10.1109/MCC.2022.3211204

Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., & Tzoumas, K. (2015). Apache Flink: Stream and batch processing in a single engine. IEEE Data Engineering Bulletin, 38(4), 28-38. http://sites.computer.org/debull/A15dec/p28.pdf

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B. S., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., ... Hendrycks, D. (2024). Black-box access is insufficient for rigorous AI audits. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2254-2272. https://doi.org/10.1145/3630106.3659037

Chen, Z., Chen, L., Zhang, Q., Cao, X., Yan, X., Wang, J., & Liu, Y. (2024). Survey on AI sustainability: Emerging trends on learning algorithms and research challenges. IEEE Computational Intelligence Magazine, 19(1), 60-78. https://doi.org/10.1109/MCI.2023.3327916

Choquette, J., Gandhi, W., Giroux, O., Stam, N., & Krashinsky, R. (2021). NVIDIA A100 tensor core GPU: Performance and innovation. IEEE Micro, 41(2), 29-35. https://doi.org/10.1109/MM.2021.3061394

Costan, V., & Devadas, S. (2016). Intel SGX explained. IACR Cryptology ePrint Archive, 2016(86), 1-118. https://eprint.iacr.org/2016/086

de Prado, M. L. (2018). Advances in financial machine learning. John Wiley & Sons. https://doi.org/10.1002/9781119482086

Dixon, M. F., Halperin, I., & Bilokon, P. (2020). Machine learning in finance: From theory to practice. Springer. https://doi.org/10.1007/978-3-030-41068-1

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407. https://doi.org/10.1561/0400000042

European Union. (2022). Regulation (EU) 2022/2554 of the European Parliament and of the Council on Digital Operational Resilience for the Financial Sector (DORA). Official Journal of the European Union, L 333, 1-79. http://data.europa.eu/eli/reg/2022/2554/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 1689. http://data.europa.eu/eli/reg/2024/1689/oj

Fang, Y., Zhang, Y., & Huang, C. (2020). Credit card fraud detection based on machine learning. Computers, Materials & Continua, 61(1), 185-195. https://doi.org/10.32604/cmc.2019.06144

Ferreira, F. A. F., Spahr, R. W., Sunderman, M. A., & Banaitiene, J. (2023). Smart fintech: A review and research agenda. Technological Forecasting and Social Change, 197, 122813. https://doi.org/10.1016/j.techfore.2023.122813

Financial Stability Board. (2024). The financial stability implications of artificial intelligence. FSB Report. https://doi.org/10.5281/zenodo.13895221

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People - An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Fowler, M., & Lewis, J. (2014). Microservices: A definition of this new architectural term. martinFowler.com. (Republished in IEEE Software, 35(3), 24-35, 2018). https://doi.org/10.1109/MS.2018.2141039

Garg, S., Pundir, P., Rathee, G., Gupta, P. K., Garg, S., & Ahlawat, S. (2021). On continuous integration / continuous delivery for automated deployment of machine learning models using MLOps. 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 25-28. https://doi.org/10.1109/AIKE52691.2021.00010

Gomber, P., Kauffman, R. J., Parker, C., & Weber, B. W. (2018). On the FinTech revolution: Interpreting the forces

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

of innovation, disruption, and transformation in financial services. Journal of Management Information Systems, 35(1), 220-265. https://doi.org/10.1080/07421222.2018.1440766

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1412.6572

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. Proceedings of the 16th ACM Workshop on AI and Security, 79-90. https://doi.org/10.1145/3605764.3623985

Gunawi, H. S., Hao, M., Leesatapornwongsa, T., Patana-anake, T., Do, T., Adityatama, J., Eliazar, K. J., Laksono, A., Lukman, J. F., Martin, V., & Satria, A. D. (2014). What bugs live in the cloud? A study of 3000+ issues in cloud systems. Proceedings of the ACM Symposium on Cloud Computing, 1-14. https://doi.org/10.1145/2670979.2670986

Hashemi, M., & Bibi, S. (2024). Edge-cloud collaborative inference for low-latency financial AI. Future Generation Computer Systems, 152, 178-191. https://doi.org/10.1016/j.future.2023.10.022

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An overview of catastrophic AI risks. arXiv preprint. https://doi.org/10.48550/arXiv.2306.12001

Hochstein, L. (2020). Site reliability engineering and resilience: A new perspective on cloud operations. Communications of the ACM, 63(6), 36-39. https://doi.org/10.1145/3387097

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. Contemporary Accounting Research, 40(2), 806-841. https://doi.org/10.1111/1911-3846.12832

Huang, X., Liu, W., Chen, X., Wang, X., Wang, H., Lian, D., Wang, Y., Tang, R., & Chen, E. (2024). Understanding the planning of LLM agents: A survey. arXiv preprint. https://doi.org/10.48550/arXiv.2402.02716

International Organization for Standardization. (2023). ISO/IEC 42001:2023 - Information technology - Artificial intelligence - Management system. ISO. https://doi.org/10.3403/30449908

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., Borchers, A., Boyle, R., Cantin, P., Chao, C., Clark, C., Coriell, J., Daley, M., Dau, M., Dean, J., Gelb, B., ... Yoon, D. H. (2017). In-datacenter performance analysis of a tensor processing unit. Proceedings of the 44th Annual International Symposium on Computer Architecture, 1-12. https://doi.org/10.1145/3079856.3080246

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascon, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1-2), 1-210. https://doi.org/10.1561/2200000083

Kindervag, J. (2010). Build security into your network's DNA: The Zero Trust Network architecture. Forrester Research. (Republished in IEEE Security & Privacy, 2010). https://doi.org/10.1109/MSP.2010.30

Kou, G., & Lu, Y. (2025). FinTech: A literature review of emerging financial technologies and applications. Financial Innovation, 11(1), 1-34. https://doi.org/10.1186/s40854-024-00668-6

Kratzke, N., & Quint, P. (2017). Understanding cloud-native applications after 10 years of cloud computing: A systematic mapping study. Journal of Systems and Software, 126, 1-16. https://doi.org/10.1016/j.jss.2017.01.001

Kreps, J. (2017). The log: What every software engineer should know about real-time data's unifying abstraction. ACM Queue, 15(2), 41-69. https://doi.org/10.1145/3076113.3097266

Kreuzberger, D., Kuhl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. IEEE Access, 11, 31866-31879. https://doi.org/10.1109/ACCESS.2023.3262138

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. Risks, 7(1), 29. https://doi.org/10.3390/risks7010029

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktaschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474. https://doi.org/10.48550/arXiv.2005.11401

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Li, A., Yang, X., Kandula, S., & Zhang, M. (2010). CloudCmp: Comparing public cloud providers. Proceedings of the 10th ACM SIGCOMM Internet Measurement Conference, 1-14. https://doi.org/10.1145/1879141.1879143

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023). CAMEL: Communicative agents for mind exploration of large language model society. Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2303.17760

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making AI less thirsty: Uncovering and addressing the secret water footprint of AI models. arXiv preprint. https://doi.org/10.48550/arXiv.2304.03271

Li, Y., Yu, Y., Li, H., Chen, Z., & Khashanah, K. (2024). FinMem: A performance-enhanced LLM trading agent with layered memory and character design. AAAI Spring Symposium on Human-Like Learning. https://doi.org/10.48550/arXiv.2311.13743

Lin, R., Hong, J., Khattar, A., Subramanian, A., Verma, S., & Chen, L. (2024). Carbon-aware cloud computing: A survey. ACM Computing Surveys, 56(7), 1-37. https://doi.org/10.1145/3631526

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. Journal of Industrial Information Integration, 6, 1-10. https://doi.org/10.1016/j.jii.2017.04.005

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1-29. https://doi.org/10.1080/23270012.2019.1570365

Lu, Y. (2022). Implementing blockchain in information systems: A review. Enterprise Information Systems, 16(12), 1876-1907. https://doi.org/10.1080/17517575.2021.2008513

Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215-234. https://doi.org/10.1007/s10796-021-10221-w

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847

Lu, Y., & Yang, J. (2024). Quantum financing system: A survey on quantum algorithms, potential scenarios and open research issues. Journal of Industrial Information Integration, 41, 100663. https://doi.org/10.1016/j.jii.2024.100663

Lu, Y., Zheng, X., Li, L., & Xu, L. D. (2020). Pricing the cloud: A QoS-based auction approach. Enterprise Information Systems, 14(3), 334-351. https://doi.org/10.1080/17517575.2019.1669827

Lu, Y., Sigov, A. S., Ratkin, L., Ivanov, L. A., & Zuo, M. (2023). Quantum computing and industrial information integration: A review. Journal of Industrial Information Integration, 35, 100511. https://doi.org/10.1016/j.jii.2023.100511

Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., & Ivanov, L. A. (2024). Quantum machine learning: Classifications, challenges, and solutions. Journal of Industrial Information Integration, 42, 100736. https://doi.org/10.1016/j.jii.2024.100736

Luccioni, A. S., Viguier, S., & Ligozat, A. L. (2023). Estimating the carbon footprint of BLOOM, a 176B parameter language model. Journal of Machine Learning Research, 24(253), 1-15. http://jmlr.org/papers/v24/23-0069.html

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1705.07874

Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. Science, 367(6481), 984-986. https://doi.org/10.1126/science.aba3758

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273-1282. https://proceedings.mlr.press/v54/mcmahan17a.html

Milian, E. Z., Spinola, M. D. M., & Carvalho, M. M. D. (2019). Fintechs: A literature review and research agenda. Electronic Commerce Research and Applications, 34, 100833. https://doi.org/10.1016/j.elerap.2019.100833

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. arXiv preprint. https://doi.org/10.48550/arXiv.2402.06196

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Mokander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2022). Auditing large language models: A three-layered approach. AI and Ethics, 3, 1-31. https://doi.org/10.1007/s43681-023-00289-2

Morgan, S., Antunes, L. F., & Pereira, J. (2025). Confidential AI inference for regulated workloads: Performance and trust trade-offs. IEEE Transactions on Cloud Computing, 13(2), 478-491. https://doi.org/10.1109/TCC.2025.3554120

Mulligan, D. P., Petri, G., Spinale, N., Stockwell, G., & Vincent, H. J. M. (2021). Confidential computing: A brave new world. IEEE Symposium on Hot Topics in System Security, 132-138. https://doi.org/10.1109/SEED51797.2021.00025

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

Norrie, T., Patil, N., Yoon, D. H., Kurian, G., Li, S., Laudon, J., Young, C., Jouppi, N., & Patterson, D. (2021). The design process for Google's training chips: TPUv2 and TPUv3. IEEE Micro, 41(2), 56-63. https://doi.org/10.1109/MM.2021.3058217

Opara-Martins, J., Sahandi, R., & Tian, F. (2016). Critical analysis of vendor lock-in and its impact on cloud computing migration: A business perspective. Journal of Cloud Computing, 5(1), 4. https://doi.org/10.1186/s13677-016-0054-z

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744. https://doi.org/10.48550/arXiv.2203.02155

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. Applied Soft Computing, 93, 106384. https://doi.org/10.1016/j.asoc.2020.106384

Paleyes, A., Urma, R. G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), 1-29. https://doi.org/10.1145/3533378

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. https://doi.org/10.1145/3586183.3606763

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint. https://doi.org/10.48550/arXiv.2104.10350

Perez, F., & Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. NeurIPS Workshop on ML Safety. https://doi.org/10.48550/arXiv.2211.09527

Petcu, D. (2014). Consuming resources and services from multiple clouds. Journal of Grid Computing, 12(2), 321-345. https://doi.org/10.1007/s10723-013-9290-3

Philippon, T. (2016). The FinTech opportunity. NBER Working Paper No. 22476. https://doi.org/10.3386/w22476

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., & Sun, M. (2025). ToolLLM: Facilitating large language models to master 16000+ real-world APIs. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2307.16789

Raj, V., Krishnamurthy, B., & Talwar, S. (2024). Cost-performance trade-offs in production LLM inference systems. IEEE Software, 41(5), 84-92. https://doi.org/10.1109/MS.2024.3403218

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44. https://doi.org/10.1145/3351095.3372873

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

1135-1144. https://doi.org/10.1145/2939672.2939778

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero Trust Architecture (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207

Rosenthal, C., & Jones, N. (2020). Chaos engineering: System resiliency in practice. O'Reilly Media. https://doi.org/10.5555/3360075

Russinovich, M., Costa, M., Fournet, C., Chisnall, D., Delignat-Lavaud, A., Clebsch, S., Vaswani, K., & Bhatotia, P. (2021). Toward confidential cloud computing. Communications of the ACM, 64(6), 54-61. https://doi.org/10.1145/3453930

Schelter, S., Biessmann, F., Januschowski, T., Salinas, D., Seufert, S., & Szarvas, G. (2018). On challenges in machine learning model management. IEEE Data Engineering Bulletin, 41(4), 5-15. http://sites.computer.org/debull/A18dec/p5.pdf

Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2302.04761

Schneider, J., Abraham, R., Meske, C., & vom Brocke, J. (2019). AI governance for businesses. Information Systems Management, 36(4), 304-321. https://doi.org/10.1080/10580530.2023.2257313

Shahrad, M., Fonseca, R., Goiri, I., Chaudhry, G., Batum, P., Cooke, J., Laureano, E., Tresness, C., Russinovich, M., & Bianchini, R. (2020). Serverless in the wild: Characterizing and optimizing the serverless workload at a large cloud provider. USENIX Annual Technical Conference, 205-218. https://www.usenix.org/conference/atc20/presentation/shahrad

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Vijayvergiya, A., Kannappan, M., Patwardhan, T., Heidecke, J., Beutel, A., & Mossing, D. (2023). Practices for governing agentic AI systems. OpenAI Research Brief. https://doi.org/10.5281/zenodo.10472008

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems, 36. https://doi.org/10.48550/arXiv.2303.11366

Stafford, V. A. (2020). Zero Trust Architecture. NIST Special Publication 800-207. https://doi.org/10.6028/NIST.SP.800-207

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008-3021. https://doi.org/10.48550/arXiv.2009.01325

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the ACL, 3645-3650. https://doi.org/10.18653/v1/P19-1355

Tamburri, D. A. (2020). Sustainable MLOps: Trends and challenges. 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 17-23. https://doi.org/10.1109/SYNASC51798.2020.00015

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jermyn, A., Cunningham, H., Henighan, T., & Olah, C. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread. https://doi.org/10.48550/arXiv.2406.04093

U.S. Department of the Treasury. (2023). The financial services sector's adoption of cloud services. Department of the Treasury Report. https://doi.org/10.5281/zenodo.7642100

Varghese, B., & Buyya, R. (2018). Next generation cloud computing: New trends and research directions. Future Generation Computer Systems, 79, 849-861. https://doi.org/10.1016/j.future.2017.09.020

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

https://doi.org/10.48550/arXiv.1706.03762

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., & Wen, J. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 186345. https://doi.org/10.1007/s11704-024-40231-1

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. KDD Workshop on Anomaly Detection in Finance. https://doi.org/10.48550/arXiv.1908.02591

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837. https://doi.org/10.48550/arXiv.2201.11903

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., ... Gabriel, I. (2022). Taxonomy of risks posed by language models. ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), 214-229. https://doi.org/10.1145/3531146.3533088

Weinman, J. (2016). The economics of pay-per-use pricing. IEEE Cloud Computing, 3(5), 14-22. https://doi.org/10.1109/MCC.2016.97

Wirth, J., & Vollmer, M. (2025). Continuous observability for AI-native financial systems. IEEE Internet Computing, 29(2), 32-41. https://doi.org/10.1109/MIC.2024.3508921

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint. https://doi.org/10.48550/arXiv.2308.08155

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv preprint. https://doi.org/10.48550/arXiv.2303.17564

Wu, H. P., Liu, Z., Dong, H. Y., Lu, Y., & Xu, L. D. (2025). Revolutionizing internal auditing: Harnessing the power of blockchain. Enterprise Information Systems, 19(1-2), 2448003. https://doi.org/10.1080/17517575.2024.2448003

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., ... Gui, T. (2025). The rise and potential of large language model based agents: A survey. Science China Information Sciences, 68(2), 121101. https://doi.org/10.1007/s11432-024-4222-0

Xu, R., Zhu, J., Yang, L., Lu, Y., & Xu, L. D. (2024). Decentralized finance (DeFi): A paradigm shift in the FinTech. Enterprise Information Systems, 18(9), 2397630. https://doi.org/10.1080/17517575.2024.2397630

Yang, H., Liu, X. Y., Zhong, S., & Walid, A. (2020). Deep reinforcement learning for automated stock trading: An ensemble strategy. Proceedings of the First ACM International Conference on AI in Finance. https://doi.org/10.1145/3383455.3422540

Yang, H., Liu, X. Y., & Wang, C. D. (2023). FinGPT: Open-source financial large language models. FinLLM Symposium at IJCAI. https://doi.org/10.48550/arXiv.2306.06031

Yang, L., Hou, Q., Zhu, X., Lu, Y., & Xu, L. D. (2025). Potential of large language models in blockchain-based supply chain finance. Enterprise Information Systems, 19(11), 2541199. https://doi.org/10.1080/17517575.2025.2541199

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2210.03629

Yu, Y., Li, H., Chen, Z., Jiang, Y., Li, Y., Zhang, D., Liu, R., Suchow, J. W., & Khashanah, K. (2024). FinAgent: A multimodal foundation agent for financial trading. KDD '24: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/3637528.3671809

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J. (2023). A survey of large language models. arXiv preprint. https://doi.org/10.48550/arXiv.2303.18223

ISSN: © 2023 INATGI (Institute of Advanced Technology and Green Innovation). Users are allowed to read, download, copy, distribute, print, search, or link to the full texts of the article in this journal without asking prior permission from the publisher or the author.
See: https://inatgi.in/index.php/rft/index for more information.