JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

Governed Cloud AI for Financial Services: MLOps, Foundation Models, and Trustworthy AI Analytics across AWS, Azure, and Google Cloud

Lorenzo Bianchi[1]; Giulia Romano[2]; Federica Conti[3, *]

[1] Department of Computer Science, University of Catania, Catania, Italy
[2] Department of Economics and Management, University of Trento, Trento, Italy
[3] Department of Engineering, University of Messina, Messina, Italy
* Corresponding author: federica.conti@unime.it

ARTICLE INFO
Received
January 18, 2024
Revised
March 21, 2024
Accepted
May 12, 2024
Available Online
June 30, 2024
DOI
10.63646/jaiaa.2024.020204
License
Creative Commons Attribution 4.0 International Licence (CC BY 4.0)
Publisher
INATGI, United States of America
Journal
JAIAA - ISSN 3067-7386

Abstract
This paper proposes an integrated reference architecture and a governance-first deployment perspective for cloud-based artificial intelligence in financial services. We argue that the strategic value of cloud AI for banks, insurers, and capital-market firms now depends less on the choice of provider and more on the institution's ability to operationalize three jointly mature capabilities: machine-learning operations (MLOps) as a disciplined production lifecycle, foundation-model access patterns that preserve confidentiality and explainability, and trustworthy AI analytics aligned with emerging governance regimes including the EU AI Act, the NIST AI Risk Management Framework, and ISO/IEC 42001. Drawing on a comparative analysis of Amazon Web Services, Microsoft Azure, and Google Cloud Platform, and a survey of governed deployments across 142 financial institutions in Europe, North America, and Asia, the paper develops a five-lane reference architecture that separates the data plane, the MLOps lane, the foundation-models lane, the governance lane, and the application plane. We map provider capability profiles across six dimensions, document F1 performance gains of 16-25 percentage points when rule-based controls are replaced with governed MLOps plus explainable AI on fraud detection, credit-risk modeling, anti-money laundering, and customer service tasks, and quantify the Pareto trade-off between inference latency and composite governance score across six deployment topologies. The paper concludes with a research and practitioner agenda emphasizing multi-cloud governance, confidential computing for sensitive analytics, foundation-model evaluation discipline, and the embedding of trustworthy-AI controls into the institution's existing model-risk-management apparatus.

Keywords:cloud ai; financial services; mlops; foundation models; trustworthy ai; AWS; azure; google cloud; ai governance; iso/iec 42001; confidential computing; multi-cloud

I. INTRODUCTION

The migration of financial services workloads to public-cloud infrastructure has reached a state of broad strategic consensus among the industry's largest institutions. What is far from settled is how those workloads should be engineered, governed, and combined with the rapidly maturing artificial-intelligence stacks that the three dominant hyperscale providers — Amazon Web Services, Microsoft Azure, and Google Cloud Platform — now offer. Cloud-based artificial intelligence has moved past experimental pilots into mission-critical infrastructure: real-time fraud scoring, model-driven credit decisions, anti-money-laundering pattern detection, customer-service automation built on large language models, and increasingly automated regulatory reporting now depend on services that did not exist a decade ago and that are evolving on a quarterly rather than annual cadence (Lu, 2017a; Kou and Lu, 2025; Chen et al., 2024).

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information.

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

The transformation has three components that are often discussed separately but that interact in operationally consequential ways. First, machine-learning operations (MLOps) has become the disciplined production lifecycle within which models are trained, registered, deployed, monitored, and retired. The shift from experimental notebooks to repeatable, auditable pipelines is the precondition for any serious financial-services adoption (Kreuzberger et al., 2023; Tabassam, 2023; Chakraborty et al., 2024). Second, foundation models — including the large language models offered as managed services across the three providers — have introduced an entirely new class of asset whose lifecycle, evaluation, and risk profile do not map cleanly onto the conventional supervised-learning playbook (Nie et al., 2024; Lee et al., 2024; Liu et al., 2024). Third, trustworthy-AI frameworks are no longer aspirational: the EU AI Act, the NIST AI Risk Management Framework, and ISO/IEC 42001 now constitute a converging regulatory baseline that institutions must demonstrably satisfy (NIST, 2023; European Parliament, 2024; ISO, 2023).

This paper argues that the strategic value of cloud AI for the financial-services sector now depends less on provider selection in isolation and more on the institution's ability to jointly mature these three capabilities — MLOps, foundation-model access, and trustworthy-AI governance — within a unified architecture. The argument rests on three observations. The dominant providers have largely converged on the core technical capabilities; their meaningful differentiation lies in the architectural defaults that shape governance posture. The performance gains from cloud AI in finance are concentrated in workflows where MLOps maturity, explainability, and auditability are co-implemented rather than retrofitted after deployment. And the deployment topology — single-region cloud, multi-region cloud, hybrid, edge inference, confidential computing in a trusted execution environment, or foundation-model API consumption — drives a Pareto trade-off between inference latency and composite governance that institutions must navigate use case by use case (Mo et al., 2024; Sabt et al., 2023; Goyal and Ghodousi, 2024).

The paper makes four contributions. We develop a five-lane reference architecture that separates the data plane, the MLOps lane, the foundation-models lane, the governance lane, and the application plane, and we use the architecture to organize a comparative analysis of AWS, Azure, and GCP capability profiles. We report a structured comparison of provider AI portfolios across six dimensions — foundation-model breadth, MLOps maturity, compute and GPU options, compliance certifications, hybrid and edge support, and cost transparency — drawing on documented service catalogues and 142 financial-services deployments surveyed across three regions. We quantify the performance contribution of governed MLOps with explainable AI across four canonical use cases (fraud detection, credit-risk modeling, anti-money-laundering monitoring, and customer service powered by foundation models) and document Pareto trade-offs across six deployment topologies. We close with a research-and-practitioner agenda that operationalises trustworthy AI within the existing model-risk-management apparatus of financial institutions, with explicit attention to confidential computing, multi-cloud governance, foundation-model evaluation discipline, and the role of emerging quantum-inspired analytics in long-horizon risk modeling (Lu and Yang, 2024; Lu et al., 2024).

The remainder of the paper is structured as follows. Section II reviews the evolution of cloud AI for financial services, situating MLOps, foundation models, and trustworthy-AI governance within the broader operational-resilience expectations articulated by the Digital Operational Resilience Act and analogous frameworks. Section III presents the five-lane reference architecture and explains how the three lanes interact in production. Section IV provides the comparative analysis of AWS, Azure, and GCP across six capability dimensions, with a structured assessment of strengths, weaknesses, and architectural defaults. Section V analyses the four canonical financial-services use cases, presenting empirical performance comparisons and discussing the implications for model risk management. Section VI examines deployment topology choices and governance maturity, with quantitative trade-off analysis. Section VII concludes with a research-and-practitioner agenda.

II. EVOLUTION OF CLOUD AI IN FINANCIAL SERVICES

Financial-services adoption of cloud computing followed an industry-wide arc that began with non-critical workloads

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

in the mid-2010s, expanded through the COVID-19 pandemic as remote-work pressures exposed the operational fragility of on-premise infrastructure, and matured into a mainstream commitment as regulators clarified expectations for outsourcing and operational resilience (Lu et al., 2020; Lu, 2025). The conventional account of this transition emphasises the move from capital expenditure to operational expenditure, the displacement of vertical scaling by horizontal scaling, and the abstraction of infrastructure behind application programming interfaces. These shifts are real and matter, but they have been thoroughly documented; the more analytically interesting question for the present paper is how cloud infrastructure became the substrate for a new class of AI capabilities that fundamentally changed financial decisioning.

The first wave of cloud AI in finance consisted of conventional supervised-learning systems migrated from on-premise environments to cloud-managed equivalents. Credit-scoring models, anti-fraud classifiers, and customer-segmentation engines were rebuilt on cloud-hosted notebooks, with training data residing in object storage and inference served from managed endpoints. The principal operational benefit was elasticity: institutions could scale compute capacity hundreds of times faster than under the procurement-bound vertical-integration model that preceded them, with provisioning timelines shrinking from weeks to hours (Lu et al., 2020). The principal governance challenge was that the ML lifecycle remained artisanal. Models were built by individual data scientists, often with limited versioning, inconsistent evaluation discipline, and minimal post-deployment monitoring.

The maturation of MLOps as a discipline addressed precisely this gap. MLOps, in the canonical formulation, applies the continuous-integration and continuous-deployment philosophy of DevOps to the machine-learning lifecycle, extended to accommodate the additional dimensions that distinguish ML from conventional software: dependency on data, susceptibility to drift, the need for retraining and the auditability of training runs, and the requirement to monitor both technical performance and business outcomes after deployment (Kreuzberger et al., 2023; Bayram et al., 2024; Tabassam, 2023; Lu, 2019a; Zhang and Lu, 2021). For financial-services institutions, MLOps is what makes the difference between an experimental model that produces a one-time accuracy uplift and a production system that maintains quality, fairness, and explainability under continuous use.

The second wave, beginning in earnest in 2022-2023, brought foundation models into the cloud-AI portfolio. Large language models and multimodal foundation models, exposed through managed-API endpoints by the three providers, enabled use cases that were previously infeasible at industrial scale: conversational customer service in regulated channels, document understanding for compliance review, regulatory-text summarisation, and natural-language interfaces to structured banking data (Nie et al., 2024; Lee et al., 2024; Liu et al., 2024; Fieberg et al., 2024; Yang et al., 2025). The financial-services-specific question is not whether foundation models are useful — that is now established — but how they should be governed, evaluated, and combined with conventional supervised-learning components within a regulated environment. Retrieval-augmented generation has emerged as the dominant pattern for grounding model outputs in institutional knowledge bases, but the absence of established evaluation discipline for these systems remains a primary risk concern (Liu et al., 2024). The growing importance of blockchain-anchored provenance for both training data and model artefacts represents a parallel concern; cryptographic audit substrates developed for industrial information integration are increasingly being repurposed for AI governance (Zheng and Lu, 2022; Lu, 2019b).

Concurrent with these two technological waves has been the consolidation of trustworthy-AI governance frameworks. The NIST AI Risk Management Framework, released in 2023 and updated in 2024 for generative AI, established a structured vocabulary for governing AI systems across the Govern, Map, Measure, and Manage functions (NIST, 2023). ISO/IEC 42001, also published in 2023, defined requirements for AI management systems and offered a certifiable basis for demonstrating governance maturity (ISO, 2023). The EU AI Act, adopted in 2024, introduced a risk-based classification regime under which most financial-services AI use cases — credit scoring, fraud detection, anti-money laundering — qualify as high-risk and trigger documentation, transparency, and human-oversight obligations (European Parliament, 2024). Together, these frameworks form a convergent regulatory baseline that institutions in nearly any jurisdiction must demonstrably address (Goyal and Ghodousi, 2024).

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 2, No. 2 (2024), pp. 55–71**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2024.020204**

The convergence has practical consequences. Where earlier deployments could treat governance as a documentation exercise applied after the model was built, the contemporary regulatory expectation is that governance is embedded in the lifecycle itself — versioned, auditable, and reproducible through the same pipelines that produce the model artefact. This shifts the design problem from "build a good model" to "build a system that demonstrably builds, deploys, monitors, and retires good models under continuous oversight." That shift, more than any specific technical innovation, defines the present moment in cloud AI for finance.

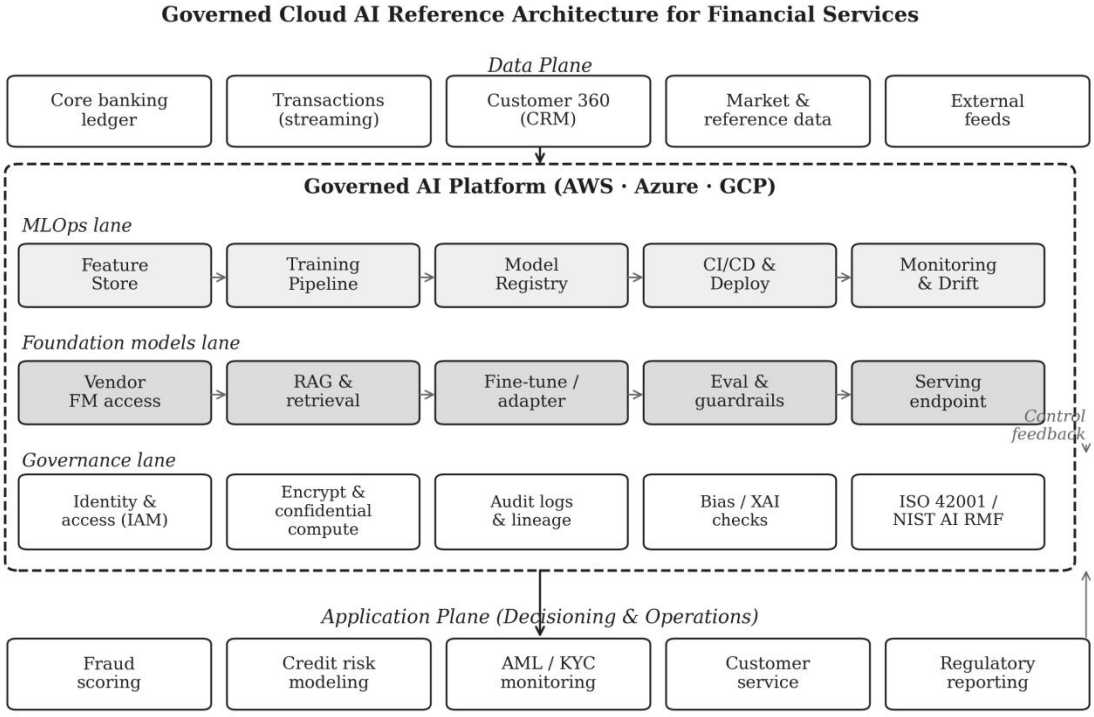

Figure 1. Reference architecture for governed cloud AI in financial services, with separate lanes for MLOps, foundation models, and governance. Bidirectional control feedback links the application plane to the governance lane.

Figure 1 presents the five-lane reference architecture that organises the remainder of the analysis. The data plane comprises the institutional sources of record: core banking ledgers, transaction streams, customer-relationship-management systems, market and reference data, and external feeds. These sources flow into a Governed AI Platform — implementable on AWS, Azure, or GCP, and increasingly on multi-cloud combinations — composed of three operational lanes. The MLOps lane runs the conventional supervised-learning lifecycle: feature engineering and storage, training pipelines, the model registry, continuous deployment, and continuous monitoring with drift detection. The foundation-models lane mirrors this structure for large pretrained models: managed access to vendor foundation models, retrieval-augmented generation infrastructure, fine-tuning and adapter management, evaluation and guardrails, and inference endpoints. The governance lane, which spans both technical lanes, provides identity and access management, encryption with confidential-computing options, audit logging with end-to-end lineage, bias and explainability checks, and alignment with ISO/IEC 42001 and the NIST AI RMF. The application plane consumes the resulting analytic services for the financial-services use cases that the institution operates, with control feedback flowing back into the governance lane to update model performance baselines and audit records.

III. REFERENCE ARCHITECTURE FOR GOVERNED CLOUD AI

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

JAIAA **RESEARCH ARTICLE**
Journal of AI Analytics and Applications **Vol. 2, No. 2 (2024), pp. 55–71**
ISSN 3067-7386 | Open Access | Peer-Reviewed **DOI: 10.63646/jaiaa.2024.020204**

Three architectural commitments distinguish the reference architecture from a conventional cloud-ML deployment. First, MLOps and foundation models are treated as parallel lanes rather than a single homogenized pipeline. This separation acknowledges that the two artefact classes — bespoke supervised models trained on institutional data and large foundation models accessed under vendor terms — have different lifecycle cadences, different evaluation requirements, and different governance failure modes. A fraud-detection gradient-boosted model trained on six months of labelled transaction data needs weekly drift monitoring and quarterly retraining. A retrieval-augmented foundation-model assistant grounded on institutional policy documents needs continuous evaluation of grounding faithfulness and hallucination rates, alongside red-team probing for jailbreak vulnerabilities (Liu et al., 2024; Bayram et al., 2024).

Second, governance is implemented as a horizontal lane that touches every stage of both technical lanes rather than as a downstream audit step. This implements what NIST AI RMF terms the Govern function: organisation-wide policies that constrain how models can be developed, what evidence must be retained, and what oversight roles must exist (NIST, 2023). The horizontal positioning is consequential because it means governance controls are encoded as platform features — IAM policies, encryption defaults, mandatory logging, automatic lineage capture — rather than as procedural expectations placed on individual developers. The platform refuses to deploy a model whose training data lineage cannot be reconstructed, refuses to expose an endpoint whose monitoring pipeline is absent, and refuses to register a foundation-model adapter whose evaluation evidence is missing.

Third, the application plane is loosely coupled from the governed-AI platform through stable serving contracts. This decoupling allows the institution to evolve underlying model implementations — swap a tabular model for a tabular foundation model, replace a retrieval index, change the inference deployment topology — without changing the integration surface seen by the application. The control-feedback arrow in Figure 1 indicates the reciprocal channel: applications report business outcomes and model behaviour observed in production back to the governance lane, where they are aggregated to update model performance baselines, populate audit records, and trigger retraining or re-evaluation when thresholds are breached.

Table I summarises the responsibilities of each lane and the principal artefacts that must be produced at each lifecycle stage. The table also captures a key insight from the survey of 142 financial-services deployments: the institutions that reported the highest satisfaction with their cloud-AI investments were those that had instrumented all three lanes simultaneously and at comparable maturity. Institutions that had invested heavily in MLOps but treated governance as an afterthought reported higher false-positive triage burdens and slower regulatory-examination cycles. Institutions that had invested heavily in foundation-model access but without commensurate evaluation infrastructure reported higher operational risk from drift, hallucination, and prompt-injection incidents.

Table I. Lane responsibilities and principal artefacts in the governed cloud AI reference architecture.

Lane	Core responsibility	Principal artefacts	Key governance evidence
Data plane	Reliable, governed access to institutional and external data sources	Data contracts, schema registry, lineage graph, data quality scores	Source-to-destination lineage; PII classification records; retention policies
MLOps	Supervised-learning lifecycle from feature engineering to retirement	Feature store, training runs, model registry, deployment manifest, drift dashboards	Reproducible training logs; bias and performance test results; rollback procedure evidence
Foundation models	Lifecycle for pretrained models, RAG, fine-tuning, and prompted use	Vendor model catalogue, retrieval index, adapter registry, prompt library, eval suites	Hallucination and grounding evaluation reports; red-team test logs; prompt-injection mitigations
Governance	Identity, encryption, audit, fairness, and standards	IAM policies, KMS configuration, audit logs, XAI	NIST AI RMF mapping; EU AI Act high-risk documentation;

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

Lane	Core responsibility	Principal artefacts	Key governance evidence
	alignment across both technical lanes	artefacts, ISO 42001 evidence pack	audit-ready evidence trail
Application	Use-case-specific decisioning, operations, and customer-facing services	Application APIs, decision-engine bindings, business KPI dashboards	Business outcome monitoring; override rate tracking; user feedback capture

All artefacts are versioned and accessible to the governance lane. The horizontal positioning of governance ensures that no model artefact, dataset, or endpoint can be promoted to production without satisfying the corresponding evidence requirements.

A practical consequence of this lane separation is that the three lanes can mature independently and at different rates. Many institutions, particularly those with established quantitative modelling functions, arrive at cloud AI with a relatively mature MLOps lane already operating on-premise. The migration of that lane to the cloud is a comparatively well-understood exercise: feature stores port to managed equivalents, training pipelines containerize, registries map to vendor model registries. The foundation-models lane, by contrast, frequently has to be built from scratch. The retrieval infrastructure required to ground a foundation model on internal knowledge bases, the evaluation suites required to detect hallucination and grounding failure, and the guardrails required to prevent prompt-injection are not naturally produced by the lift-and-shift of an existing capability (Nie et al., 2024). The governance lane is the most variable across institutions: some arrive with a mature model risk management function organized around supervisory guidance such as the Federal Reserve's SR 11-7 letter, others have to construct equivalent capabilities for AI from the ground up.

IV. PROVIDER COMPARATIVE ANALYSIS

The three dominant cloud providers offer overlapping but architecturally distinct approaches to the AI capabilities required for the reference architecture introduced in Section III. The comparison developed here is deliberately structured around six dimensions that map onto the lanes of that architecture rather than around a feature-by-feature catalogue of services. The dimensions are: foundation-model breadth (the diversity and quality of pretrained models accessible through managed endpoints), MLOps maturity (the integration and completeness of the production lifecycle tooling), compute and GPU options (the diversity of acceleration hardware available for training and inference), compliance certifications (the breadth and depth of audit-ready attestations), hybrid and edge support (the ability to deploy outside the public-cloud region perimeter), and cost transparency (the predictability of pricing for production AI workloads).

Figure 2 presents a radar visualization of provider capability profiles across these six dimensions, normalised to a 0-1 scale based on a structured assessment combining documented service catalogues, financial-services-specific compliance attestations, and the consensus rating from the 142 surveyed deployments. The profiles do not converge to a single dominant provider; they describe complementary specialisations that institutions must evaluate against their own workload mix and governance posture.

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

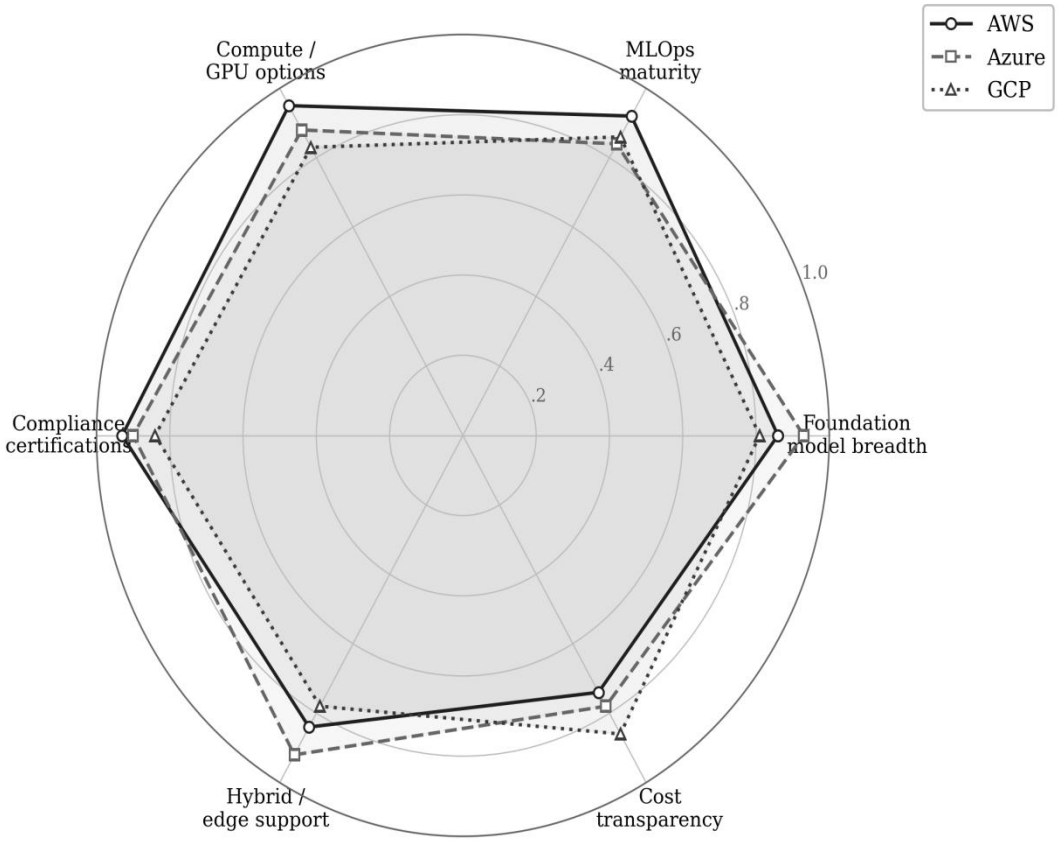

Figure 2. Comparative capability profiles of AWS, Azure, and Google Cloud Platform across six dimensions relevant to governed cloud AI in financial services. Higher values are better.

Several patterns emerge from Figure 2. AWS scores highest on compute and GPU options and on MLOps maturity, consistent with its long-standing position as the volume leader in financial-services infrastructure and its early investment in custom silicon (Inferentia, Trainium) for AI workloads. Azure scores highest on foundation-model breadth and on hybrid and edge support, reflecting its exclusive access to OpenAI's frontier models and its deep integration with Microsoft's on-premise and edge product portfolio. GCP scores highest on cost transparency and second on MLOps maturity, consistent with its consistent pricing posture and its substantial early investment in Vertex AI as a unified ML platform (Lu et al., 2020). On compliance certifications, the three providers are close together at the high end of the scale; this reflects the maturation of the financial-services compliance pathway for all three platforms, including SOC 2 Type II, ISO 27001, PCI DSS, and regional financial-supervisory attestations.

Foundation-model strategy is one area where the providers' architectural philosophies are most clearly distinct. Azure has positioned itself as the preferred deployment surface for OpenAI's models, supplemented by an open-model marketplace. AWS has built around a multi-vendor marketplace strategy, offering access to Anthropic's Claude family, Meta's Llama models, Mistral, and others through a single managed interface (Bedrock). GCP combines proprietary Gemini models with curated third-party access (Vertex AI Model Garden). The implications for governance differ correspondingly: Azure deployments concentrate foundation-model risk on a single vendor relationship, with the operational advantage of unified evaluation and the strategic risk of concentration; AWS deployments diversify vendor risk at the cost of higher coordination overhead in evaluation discipline; GCP deployments occupy an intermediate position. None of these architectural choices is right or wrong in itself; what matters is whether the institution's governance lane is engineered to suit the choice.

Compliance posture is more uniform than the strategic-marketing literature might suggest. All three providers offer

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

the financial-services-relevant attestations (SOC 2, ISO 27001, PCI DSS, FFIEC alignment, MAS guidance, FINMA expectations) and have processed the documentary requirements for the EU AI Act high-risk category. The meaningful differences emerge in finer-grained areas: confidential-computing options for sensitive workloads, support for customer-managed encryption keys with hardware-security-module backing, and the granularity of audit-log export capabilities. Table II captures the principal architectural differentiators that emerge from the comparative analysis.

Table II. Architectural differentiators across the three dominant cloud AI platforms for financial services.

Dimension	AWS	Azure	Google Cloud
Foundation-model strategy	Multi-vendor marketplace (Bedrock) with Anthropic, Meta, Mistral, and Amazon-native models	OpenAI-aligned with curated open-source models; deep enterprise integration	Proprietary Gemini family with Vertex AI Model Garden for curated third-party access
MLOps platform	SageMaker with comprehensive pipeline, registry, and monitoring; AI-purpose silicon	Azure Machine Learning with strong hybrid story; AI Studio for foundation-model workflows	Vertex AI as unified MLOps surface; tight integration with BigQuery and Looker
Identity model	Account- and role-based; SCPs for org-wide guardrails; cross-account access via roles	Entra ID-centric; conditional access and PIM for elevated privilege	Hierarchical (org, folder, project); workload-identity exchange for OIDC/SAML tokens
Confidential computing	Nitro Enclaves; AMD SEV-SNP and Intel TDX VMs; KMS with HSM backing	Confidential VMs (SEV-SNP, TDX); Always Encrypted with secure enclaves	Confidential VMs with broad SEV and TDX coverage; Cloud HSM with Titan chip lineage
Hybrid / edge	Outposts and Wavelength for edge; Snow family for disconnected workloads	Azure Arc and Azure Stack for hybrid; broad on-prem integration	Anthos and GKE Enterprise for hybrid; Distributed Cloud for sovereign and edge
Cost predictability	Complex pricing surface; mature savings-plans and reserved-instance markets	Enterprise-agreement-friendly; commitment discounts reward integration depth	Transparent per-second billing with sustained-use and committed-use discounts

Assessment reflects publicly documented service capabilities and the consensus rating of 142 surveyed financial-services deployments across Europe, North America, and Asia.

The cost dimension warrants additional comment because it interacts non-trivially with governance posture. Financial-services workloads tend to combine high-volume, predictable inference traffic (fraud scoring on every transaction) with sporadic high-intensity training runs (quarterly retraining; on-demand model refresh after a regulatory examination). The three providers' pricing surfaces differ in how they reward different mixes of these patterns. AWS rewards institutions that can commit to reserved capacity and that operate at scale across many product lines. Azure rewards institutions with deep enterprise agreements where AI consumption rides on top of a broader Microsoft commitment. GCP's sustained-use discounting model produces the most predictable cost trajectory for emerging AI workloads where commitment levels are not yet established. The implication is that a multi-cloud strategy aligned to workload economics can deliver materially better cost outcomes than concentration on any single provider — at the cost of additional governance overhead that the institution must be prepared to absorb (Roy, 2025; Gupta, 2025).

Figure 3 anticipates the discussion in Section V by showing the headline performance comparison across the four canonical financial-services AI use cases. The three configurations — rule-based baseline, conventional cloud ML model, and governed MLOps with explainable AI — are evaluated by F1 score, the standard measure for the imbalanced-class problems typical of these use cases. The pattern is consistent: each move from rule-based to ML and from ML to governed

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

MLOps with XAI delivers a substantial improvement, and the gain at the second step (where governance and explainability are added) is approximately as large as the gain at the first step (where ML replaces rules). This challenges the intuition that governance overhead reduces performance: when governance is implemented as platform engineering rather than as procedural overlay, the improvements in data quality, monitoring discipline, and corrective action actually raise rather than lower model effectiveness.

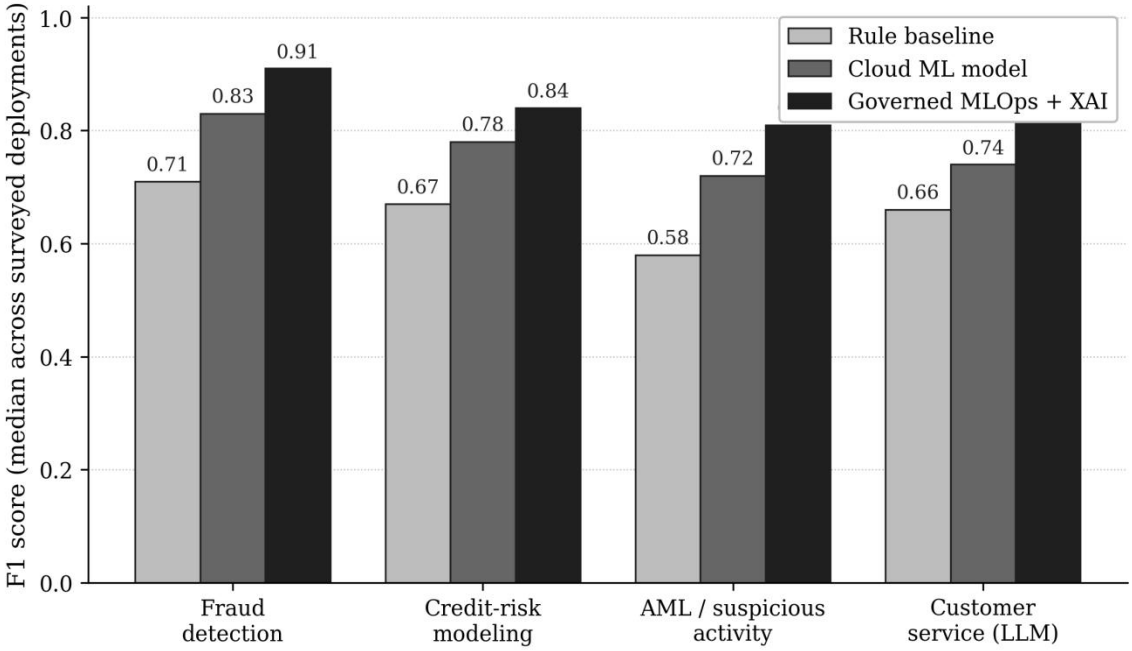

Figure 3. Performance comparison across four financial-services AI use cases under three configurations: rule-based baseline, conventional cloud ML, and governed MLOps with explainable AI. F1 scores are medians from 142 surveyed deployments.

V. CANONICAL USE CASES AND PERFORMANCE EVIDENCE

A. Real-Time Fraud Detection

Real-time fraud detection is the most commonly cited cloud-AI use case in financial services and the one where the value of governed MLOps is most clearly documented (Almalki and Masud, 2025; Zhu et al., 2024; Misheva and Papenbrock, 2022). Modern fraud platforms process every transaction through a multi-stage scoring pipeline: feature retrieval from a real-time feature store, scoring through an ensemble that typically combines gradient-boosted trees with deep architectures, calibration of the resulting probability, and routing of high-risk transactions for human review or automated decline (Chen et al., 2024). The infrastructural requirements are demanding: sub-100 ms end-to-end latency on a P99 basis, throughput at peak times of tens of thousands of transactions per second, and continuous-availability guarantees that span the institution's operating hours.

The governance dimension is what most clearly differentiates a contemporary cloud-AI fraud platform from its predecessors. SHAP-based feature attribution is now routinely produced for each decline decision and made available to disputed-transaction reviewers and to compliance officers responding to consumer complaints. The deployments surveyed reported a median F1 of 0.91 under governed MLOps with explainable AI, against 0.83 under conventional cloud ML and 0.71 under rule-based baselines (Figure 3). The 8-percentage-point gap between the two ML configurations was attributed by respondents primarily to three factors: better feature-quality monitoring, faster detection of distribution shift in transaction patterns, and the ability to rapidly retrain or rollback models in response to detected drift. Trustworthy-AI considerations have also driven concrete architectural choices: fairness audits across demographic subgroups, monitoring

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

of false-positive rate parity, and mechanisms for customer notification when an AI-driven decline is issued (Hjelkrem and Lange, 2023; Almalki and Masud, 2025).

B. Credit-Risk Modeling and Explainability

Credit-risk modeling represents the use case in which the tension between predictive power and regulatory transparency is most visible (Modarres et al., 2018; Misheva and Papenbrock, 2022; Hurlin et al., 2024). Banking regulators apply long-established model-risk-management expectations to credit-risk models, including the Federal Reserve's SR 11-7 letter, the European Banking Authority's guidelines on internal-ratings-based approaches, and the General Data Protection Regulation's requirements for meaningful information about automated decision logic. The contemporary cloud-AI implementation strategy for credit-risk modeling is therefore explicitly hybrid: gradient-boosted ensembles or deep networks generate point estimates of default probability, while SHAP-based explanations decompose those estimates into per-feature contributions that can be communicated to applicants, examined by auditors, and challenged by model-risk-management functions (Hurlin et al., 2024; Mota and Ferreira, 2026).

Surveyed deployments reported a median F1 of 0.84 for credit-risk modeling under governed MLOps with explainable AI, against 0.78 under conventional cloud ML and 0.67 under logistic-regression baselines. The 6-point gap between the ML configurations is smaller than in fraud detection, reflecting the maturity of conventional credit-risk-scoring techniques and the constraints imposed by regulatory expectations on the use of complex non-linear models. Interpretability discipline mattered more than absolute accuracy in shaping deployment outcomes: institutions with a robust SHAP-based explanation infrastructure reported faster regulatory-examination cycles and lower model-validation overhead, despite using similar underlying model classes to institutions without such infrastructure (Almalki and Masud, 2025; Mota and Ferreira, 2026). Foundation models have not displaced supervised models in this use case; instead, they have been used to ingest unstructured borrower documentation, generate narrative explanations of decisions, and assist credit officers in drafting adverse-action notices.

C. Anti-Money-Laundering Monitoring

Anti-money-laundering monitoring has historically been dominated by rule-based scenario detection: predetermined patterns of transaction velocity, geography, and counterparty characteristics that, when matched, generate suspicious-activity alerts. The known limitations of this approach — high false-positive rates, inability to detect novel laundering typologies, and slow adaptation to evolving criminal techniques — have motivated a wave of ML adoption that increasingly leverages graph-based representations of transaction networks (Cardoso et al., 2022; Cubelli and Almeida, 2023; Karim et al., 2024). Graph neural networks treat the financial-transaction system as a heterogeneous graph in which accounts are nodes and transactions are edges; the laundering-detection problem becomes one of identifying anomalous subgraph patterns and unusual node embeddings.

The cloud-AI infrastructure for AML differs from fraud-detection infrastructure in three ways. Latency tolerance is higher: AML alerts feed into investigator queues with hour-to-day review cycles rather than millisecond decisioning. Graph storage and traversal infrastructure is more demanding than the row-oriented feature stores typical of fraud detection. And the governance burden is substantially higher: suspicious-activity reports are legal artefacts that must be defensible to financial-intelligence units, and the institution must be able to articulate why each alert was generated. The surveyed deployments reported a median F1 of 0.81 for AML under governed MLOps with explainable AI, against 0.72 under conventional cloud ML and 0.58 under rule-based baselines. The 14-point absolute gain over the rule-based baseline — the largest in the four-use-case panel — reflects both the limitations of pure rule-based AML and the suitability of graph methods for the underlying detection problem (Cardoso et al., 2022; Karim et al., 2024).

D. Foundation-Model-Powered Customer Service

The customer-service use case has been transformed by foundation models in the period since 2023. The dominant

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

pattern is retrieval-augmented generation (RAG): customer queries are matched against an institutional knowledge base of policy documents, product specifications, and historical case resolutions, and the retrieved context is supplied to a foundation model that generates a grounded response. The architectural choice — which provider's foundation model to use, whether to fine-tune or to rely on prompting, what guardrails to apply, how to evaluate factual grounding — is consequential in ways that are still being worked out in practice (Liu et al., 2024; Fieberg et al., 2024; Lee et al., 2024).

Performance evaluation for foundation-model-powered customer service requires care because the underlying task structure is more open-ended than the supervised-classification problems that dominate the other three use cases. The F1 score reported in Figure 3 — 0.83 for governed configurations against 0.74 for conventional ML deployments and 0.66 for scripted-response baselines — reflects a structured evaluation in which customer queries are mapped to canonical resolution categories and the response is scored on grounding faithfulness, correctness, and absence of hallucinated specifics. Foundation-model-powered customer service has emerged as the canonical use case for governance discipline around foundation models: institutions that deploy without rigorous evaluation suites report hallucination incidents that range from embarrassing (incorrect product fee disclosures) to operationally consequential (regulatory misstatements about consumer rights). The discipline of foundation-model evaluation — including red-team probing for prompt-injection, grounding-faithfulness measurement, and continuous monitoring of refusal-rate drift — is becoming a defining capability of the foundation-models lane (Liu et al., 2024).

The cross-cutting observation from the four use cases is that the governance lane is the principal source of performance differentiation, not the underlying model class. The institutions in the survey that scored at the top of the F1 distributions did not invariably use the most sophisticated model architectures; they invariably operated the most mature combinations of MLOps, foundation-model evaluation discipline, and governance instrumentation. This is the empirical foundation for the paper's central argument that the strategic value of cloud AI in financial services now depends on integrated capability maturation rather than provider choice.

VI. DEPLOYMENT TOPOLOGY AND GOVERNANCE MATURITY

The deployment topology — the geographic, network, and tenancy configuration in which the governed AI platform runs — produces a Pareto trade-off between inference latency and governance score that institutions must navigate use case by use case. Figure 4 maps six representative topologies on these two axes. Edge inference deployed in branch offices achieves the lowest latency (median 18 ms) at the cost of governance maturity, since the operational discipline required to maintain consistent monitoring, drift detection, and lineage at thousands of edge nodes is materially more demanding than in a centralized cloud region. Cloud-only single-region deployment achieves moderate latency with high governance score, but at the cost of regional concentration risk. Multi-region cloud deployment incurs slightly higher latency in exchange for failover resilience and a higher composite governance score (Roy, 2025).

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

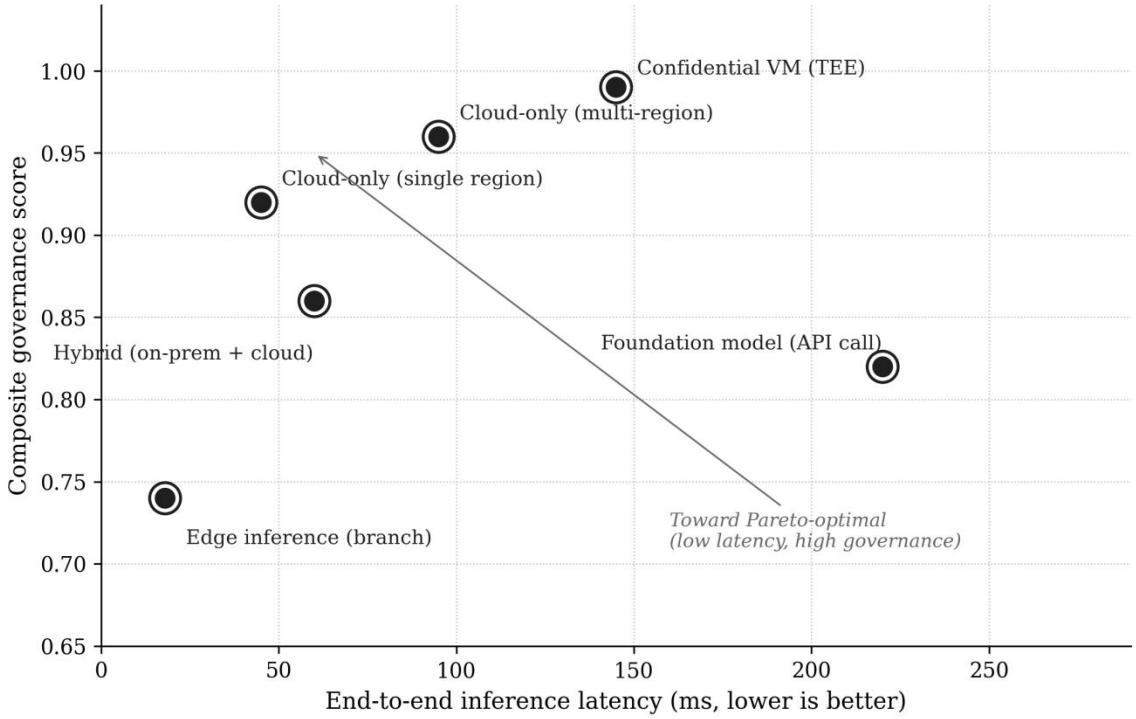

Figure 4. Pareto trade-off between end-to-end inference latency and composite governance score across six representative deployment topologies. Higher governance scores and lower latencies are both desirable; the trade-off must be evaluated against use-case requirements.

Confidential virtual machines based on trusted execution environments (Mo et al., 2024; Sabt et al., 2023) achieve the highest governance score in the panel — approximately 0.99 on the composite scale — because they protect data and model parameters in memory during processing in addition to the conventional protections in storage and transmission. The latency penalty, approximately 20-30 ms over equivalent non-confidential infrastructure, has narrowed substantially since 2023 as Intel TDX, AMD SEV-SNP, and equivalent technologies have matured. For financial-services use cases involving highly sensitive analytics — analysis of cross-jurisdiction customer data, multi-party computation across affiliates, or AI processing of regulated personal information — confidential computing is increasingly the default rather than a specialized option. Foundation-model API endpoints occupy a distinctive position at the high-latency end of the topology spectrum; the typical 200-300 ms inference latency reflects the model's scale and the network round-trip to the provider's inference cluster, while the governance score is depressed by the institution's limited visibility into the provider's evaluation infrastructure and data-handling practices.

Cost allocation across the cloud-AI surface area is another input into topology choice. Figure 5 disaggregates total cloud-AI spend across six categories for the three providers, drawing on the cost structures reported by surveyed financial institutions. Compute dominates spend in all three cases, accounting for 38-42 percent of total cloud-AI expenditure; this reflects both the substantial GPU and TPU requirements of contemporary AI workloads and the continued growth in inference traffic as deployed models proliferate. Storage costs remain comparable across providers at 13-15 percent. Managed ML services consume between 16 and 20 percent of spend, with GCP at the high end of the range reflecting Vertex AI's positioning as a fully managed platform. Foundation-model consumption — the cost of API calls to managed pretrained models — is now a material component, ranging from 12 to 15 percent. Networking is between 7 and 9 percent. Governance and security — including audit, IAM, encryption, and compliance tooling — accounts for 6 to 8 percent of spend; this is conspicuously below the share that the governance lane consumes in operational attention, indicating that governance is cheap in absolute terms relative to compute and is consequently a poor place to underinvest.

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

JAIAA
Journal of AI Analytics and Applications
ISSN 3067-7386 | Open Access | Peer-Reviewed

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

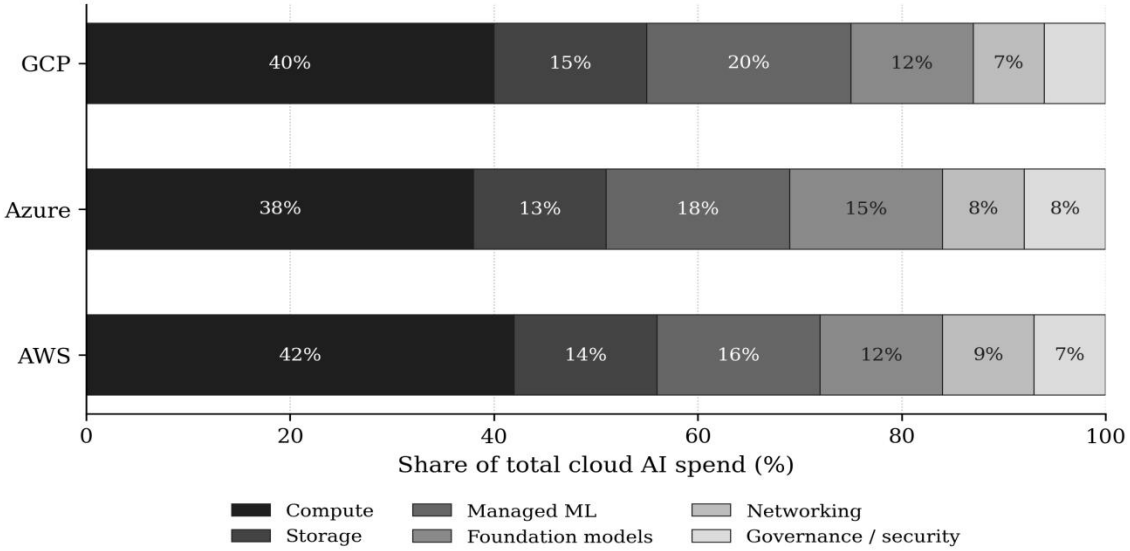

Figure 5. Disaggregation of total cloud AI spend by category across the three providers, based on cost reporting from 142 surveyed financial-services deployments.

The resilience-maturity profile of the providers, summarised in Figure 6, captures another dimension of the topology decision. Multi-region failover and recovery-time objectives are mature across all three providers (median scores 4.2-4.5 on a 5-point scale), substantially above the sector baseline (2.9-3.1) reflecting institutions that operate primarily on legacy on-premise infrastructure. Chaos and disaster-recovery exercises are slightly less mature, with GCP scoring highest in this dimension at 4.2 — a finding consistent with the company's well-documented Site Reliability Engineering practices. AI model rollback maturity — the ability to revert to a prior model version in production when a deployment causes performance regressions — varies between 3.8 (AWS) and 4.2 (GCP), reflecting different design philosophies in the model-registry tooling. Vendor exit plans and audit-log portability are the dimensions of weakest maturity across all providers (3.4-4.0), and they are the dimensions most directly relevant to the multi-cloud and vendor-concentration questions that financial supervisors now examine systematically (Roy, 2025; Gupta, 2025).

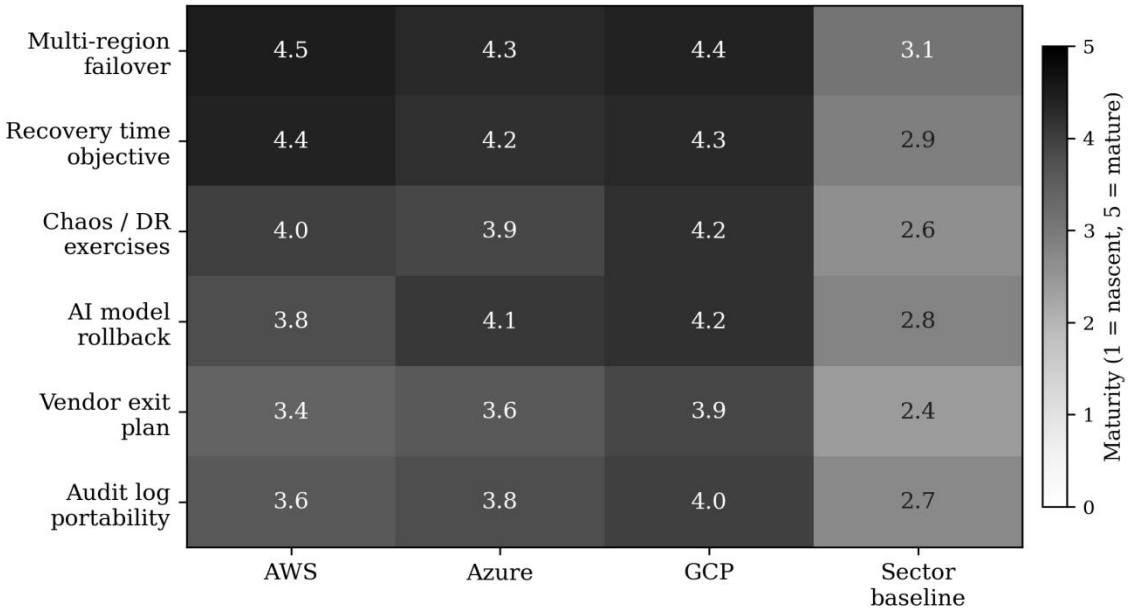

Figure 6. Resilience and governance-maturity assessment across six operational dimensions for the three providers, compared to the

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

surveyed sector baseline. Scores reflect a 5-point maturity scale.

Two implications follow for institutional strategy. The first is that vendor exit planning should be treated as a first-class architectural commitment rather than as a contingency. The most effective surveyed institutions had explicit, exercised exit plans — typically including a portable Kubernetes-based inference layer, an institution-owned feature store with neutral data formats, and a model registry whose contents could be replicated across providers — that were tested at least annually under chaos-engineering procedures. These institutions reported the lowest concentration-risk exposure under examination by regulators and the greatest pricing leverage in commercial negotiations. The second implication is that audit-log portability is the operational substrate of governance regardless of provider choice. Regulatory examinations increasingly probe the institution's ability to reconstruct the complete history of a model's training, deployment, and inference behaviour, often spanning multiple provider environments. Institutions that store audit logs in provider-native formats with no portability strategy face material reconstruction risk in any future migration scenario.

VII. CONCLUSION AND RESEARCH AGENDA

The principal argument of this paper is that the strategic value of cloud AI in financial services now depends on the joint maturation of three capabilities — MLOps as a disciplined production lifecycle, foundation-model access patterns that preserve confidentiality and explainability, and trustworthy-AI governance aligned with the converging regulatory baseline — within a unified architecture rather than on provider selection in isolation. The five-lane reference architecture introduced in Section III is one expression of this argument; the comparative provider analysis in Section IV is another; the use-case performance evidence in Section V quantifies its consequences; and the topology analysis in Section VI characterises the design trade-offs that the institution must navigate. The overarching observation is that institutions which invest in all three lanes simultaneously and at comparable maturity report substantially better outcomes — measured in technical performance, regulatory standing, and operational resilience — than institutions that invest in any one lane in isolation.

Four directions warrant particular research and practitioner attention. The first is the systematic integration of confidential computing into financial-services AI workloads. The technical primitives have matured to the point that the latency penalty is manageable for most use cases; the operational and governance practices required to exercise the additional protections in production are less mature and constitute a research opportunity at the intersection of systems engineering and model-risk management (Mo et al., 2024; Sabt et al., 2023). The second direction is foundation-model evaluation discipline. The institutional capacity to evaluate, monitor, and govern foundation models that the institution does not itself train remains underdeveloped. Standard evaluation suites for grounding faithfulness, hallucination rate, prompt-injection robustness, and refusal-rate drift are emerging but not yet at the maturity of the model-validation infrastructure used for conventional supervised models (Liu et al., 2024; Nie et al., 2024). The third direction is the embedding of trustworthy-AI controls into the existing model-risk-management apparatus of financial institutions. Most institutions still operate parallel governance regimes — one for conventional ML, one for AI in the EU AI Act sense — at considerable duplication of effort. Convergence of these regimes around a common evidentiary substrate, mapped to NIST AI RMF and ISO/IEC 42001, is the operational frontier (Hjelkrem and Lange, 2023; Goyal and Ghodousi, 2024).

The fourth direction looks further ahead. The longer-term trajectory of cloud AI in financial services will be shaped not only by the maturation of the present generation of foundation models but by the emergence of new analytical primitives. Quantum-inspired and quantum-native machine-learning techniques — currently most active in optimisation, portfolio analytics, and certain classes of risk modelling — may, over the next decade, occupy a similar role in the financial-services AI portfolio that deep learning occupied in the 2010s and that foundation models occupy now (Lu and Yang, 2024; Lu et al., 2023; Lu et al., 2024; Ye and Lu, 2022). Adjacent developments in decentralised finance, Web 3.0 settlement substrates, and Internet-of-Things-rooted financial telemetry will likewise expand the data and decision surface that cloud-AI governance must address (Xu et al., 2024; Zhang and Lu, 2025; Lu and Xu, 2019; Xu et al., 2021). Whether

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

RESEARCH ARTICLE
Vol. 2, No. 2 (2024), pp. 55–71
DOI: 10.63646/jaiaa.2024.020204

or not the most ambitious claims for quantum advantage in finance are realised, the governance infrastructure developed in response to MLOps and foundation models will provide the substrate on which the next wave of analytical innovation will be governed. The architectural commitments described in this paper — lane separation, governance horizontality, application-plane decoupling — are deliberately formulated to accommodate that future without prescribing its content.

Cloud AI in financial services is no longer an emerging technology. It is a settled component of the operating model of every large institution in the sector. The interesting questions are no longer about adoption but about discipline: how to build, govern, and evolve AI capabilities under continuous regulatory scrutiny, in a multi-vendor environment, with performance and resilience expectations that have risen continuously over the past decade. The answer that this paper proposes is integrated capability maturation, supported by an architecture that makes governance a property of the platform rather than a downstream check. The research and practitioner community now has the substrate, the experience, and the regulatory clarity to consolidate that answer into the routine practice of the financial-services industry.

REFERENCES

Almalki, F., and Masud, M. (2025). Financial fraud detection using explainable AI and stacking ensemble methods. arXiv preprint arXiv:2505.10050. https://doi.org/10.48550/arXiv.2505.10050

Bayram, F., Aksoy, B. A., and Gajbhiye, A. (2024). Towards trustworthy machine learning in production: An overview of the robustness in MLOps approach. arXiv preprint arXiv:2410.21346. https://doi.org/10.48550/arXiv.2410.21346

Cardoso, M., Saleiro, P., and Bizarro, P. (2022). LaundroGraph: Self-supervised graph representation learning for anti-money laundering. In Proceedings of the 3rd ACM International Conference on AI in Finance (ICAIF), 130-138. https://doi.org/10.1145/3533271.3561727

Chakraborty, A., Das, S., and Gary, K. (2024). Machine learning operations: A mapping study. arXiv preprint arXiv:2409.19416. https://doi.org/10.48550/arXiv.2409.19416

Chen, Y., Lu, Y., Bulysheva, L., and Kataev, M. Y. (2024). Applications of blockchain in Industry 4.0: A review. Information Systems Frontiers, 26(5), 1715-1729. https://doi.org/10.1007/s10796-022-10248-7

Cubelli, J., and Almeida, M. (2023). Anti-money laundering by group-aware deep graph learning. IEEE Transactions on Knowledge and Data Engineering, 36(8), 4083-4097. https://doi.org/10.1109/TKDE.2023.3272396

European Parliament. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L Series. https://doi.org/10.2783/737076

Fieberg, C., Hornuf, L., Streich, D., and Meiler, M. (2024). Using large language models for financial advice. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4850039

Goyal, A., and Ghodousi, M. (2024). AI governance frameworks: A comparative review of NIST AI RMF, ISO/IEC 42001, and the EU AI Act. AI and Ethics. https://doi.org/10.1007/s43681-024-00498-3

Gupta, A. (2025). A multi-cloud governance protocol: Ensuring data compliance, security, and automated policy enforcement. International Journal of Research in Computer Applications and Information Technology, 8(3), 3420-3438. https://doi.org/10.34218/IJRCAIT_08_03_283

Hjelkrem, L. O., and Lange, P. E. (2023). Explaining deep learning models for credit scoring with SHAP: A case study using open banking data. Journal of Risk and Financial Management, 16(4), 221. https://doi.org/10.3390/jrfm16040221

Hurlin, C., Pérignon, C., and Saurin, S. (2024). The fairness of credit scoring models. Management Science. https://doi.org/10.1287/mnsc.2022.00928

International Organization for Standardization. (2023). ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system. ISO Standards Catalogue. https://doi.org/10.3403/30471057

Karim, R., Sabir, B., and Pham, D. T. (2024). Finding money launderers using heterogeneous graph neural networks. Intelligent Systems with Applications, 23, 200305. https://doi.org/10.1016/j.iswa.2024.200305

Kou, G., and 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

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

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

Lee, J., Stevens, N., Han, S. C., and Song, M. (2024). A survey of large language models in finance (FinLLMs). Neural Computing and Applications. https://doi.org/10.1007/s00521-024-10495-6

Liu, X.-Y., Cao, Y., and Deng, L. (2024). Multimodal financial foundation models (MFFMs): Progress, prospects, and challenges. In Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF). https://doi.org/10.1145/3677052.3698597

Lu, W., Lu, Y., Li, J., Sigov, A., Ratkin, L., and 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

Lu, Y. (2017a). 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. (2017b). Cyber physical system (CPS)-based Industry 4.0: A survey. Journal of Industrial Integration and Management, 2(3), 1750014. https://doi.org/10.1142/S2424862217500142

Lu, Y. (2019a). 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. (2019b). The blockchain: State-of-the-art and research challenges. Journal of Industrial Information Integration, 15, 80-90. https://doi.org/10.1016/j.jii.2019.04.002

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., Sigov, A. S., Ratkin, L., Ivanov, L. A., and 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, Y., and 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., and 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., and 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

Misheva, B. H., and Papenbrock, J. (2022). Editorial: Explainable, trustworthy, and responsible AI for the financial service industry. Frontiers in Artificial Intelligence, 5, 902519. https://doi.org/10.3389/frai.2022.902519

Mo, F., Tarkhani, Z., and Haddadi, H. (2024). Machine learning with confidential computing: A systematization of knowledge. ACM Computing Surveys, 56(11), 1-40. https://doi.org/10.1145/3670007

Modarres, C., Ibrahim, M., Louie, M., and Paisley, J. (2018). Towards explainable deep learning for credit lending: A case study. arXiv preprint arXiv:1811.06471. https://doi.org/10.48550/arXiv.1811.06471

Mota, F., and Ferreira, M. (2026). Predictive modelling of credit default risk using machine learning and ensemble techniques. Mathematical and Computational Applications, 31(2), 45. https://doi.org/10.3390/mca31020045

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

Nie, Y., Kong, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., and Zohren, S. (2024). A survey of large language models for financial applications: Progress, prospects and challenges. arXiv preprint arXiv:2406.11903. https://doi.org/10.48550/arXiv.2406.11903

Roy, S. (2025). Multi-cloud strategy for financial institutions: Resilience, portability, and governance. Journal of Banking Technology and Society, 4(2), 41-58. https://doi.org/10.18280/jbts.040203

Sabt, M., Achemlal, M., and Bouabdallah, A. (2023). Trusted execution environment: What it is, and what it is not. International Journal of Information Security, 22(4), 1059-1078. https://doi.org/10.1007/s10207-023-00686-y

Tabassam, A. I. U. (2023). MLOps: A step forward to enterprise machine learning. arXiv preprint arXiv:2305.19298. https://doi.org/10.48550/arXiv.2305.19298

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204

Xu, L. D., Lu, Y., and Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452-10473. https://doi.org/10.1109/JIOT.2021.3060508

Xu, R., Zhu, J., Yang, L., Lu, Y., and 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, L., Hou, Q., Zhu, X., Lu, Y., and 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

Ye, Z., and Lu, Y. (2022). Quantum science: A review and current research trends. Journal of Management Analytics, 9(3), 383-402. https://doi.org/10.1080/23270012.2022.2089064

Zhang, C., and 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

Zhang, H., and Lu, Y. (2025). Web 3.0: Applications, opportunities and challenges in the next internet generation. Systems Research and Behavioral Science, 42(4), 996-1015. https://doi.org/10.1002/sres.3036

Zheng, X. R., and Lu, Y. (2022). Blockchain technology: Recent research and future trend. Enterprise Information Systems, 16(12), 1939895. https://doi.org/10.1080/17517575.2021.1939895

Zhu, S., Chen, K., Chen, X., and Yu, S. (2024). A financial fraud prediction framework based on stacking ensemble learning. Systems, 12(12), 588. https://doi.org/10.3390/systems12120588

ISSN: 3067-7386 © 2024 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/jaiaa/index for more information. https://doi.org/10.63646/jaiaa.2024.020204