NextGen Scientific Publication
American International Journal of Computer Science and Technology

Volume 6 Issue 1, Pg. No. 82-104, AIJCST-V6I1P109, 2024
https://doi.org/10.63282/3117-5481/AIJCST-V6I1P109

Original Article

Enterprise Architecture Framework for Secure Large Language Model Deployment in Financial Services

***Saad Khan**

Tech Lead, Senior Associate at JP Morgan Chase, USA.

Abstract:

The rapid advancement of Large Language Models (LLMs) has created new opportunities for financial institutions to enhance customer engagement, investment research, advisor productivity, and operational efficiency. However, enterprise adoption of LLMs remains challenging due to stringent regulatory requirements, data privacy concerns, cybersecurity risks, model governance complexities, and integration with existing enterprise systems. While existing research primarily focuses on LLM capabilities and model performance, limited attention has been given to a comprehensive enterprise architecture that enables secure, scalable, and governed deployment of LLMs within regulated financial environments. This paper proposes a novel Enterprise Architecture Framework for Secure Large Language Model Deployment in Financial Services to support the responsible adoption of generative AI across banking, asset and wealth management, and capital markets. The proposed framework consists of six vertically integrated architectural layers—Experience and Channels, AI Orchestration and Policy, LLM Services, Knowledge and Retrieval-Augmented Generation (RAG), Enterprise Integration, and Platform and Deployment—supported by four cross-cutting control domains: Security, Governance, Compliance, and Observability. These layers provide a scalable foundation for integrating LLM capabilities into enterprise applications while ensuring regulatory compliance, data protection, explainability, and operational resilience. A design science research methodology is employed to develop and validate the framework using a representative enterprise financial services case study. The evaluation demonstrates how the proposed architecture enables secure knowledge retrieval, intelligent research assistance, workflow automation, and decision support while maintaining identity-based access control, human oversight, auditability, and responsible AI governance. By combining cloud-native architecture, API-driven integration, zero-trust security, and continuous monitoring, the framework addresses critical enterprise requirements for secure LLM adoption. The proposed framework contributes a reusable enterprise reference architecture that bridges the gap between generative AI innovation and the operational, security, and governance demands of modern financial institutions, providing enterprise architects and technology leaders with practical guidance for implementing trustworthy and scalable LLM solutions.

Keywords:

Large Language Models, Enterprise Architecture, Financial Services, Retrieval-Augmented Generation, AI Governance, Responsible AI.

Article History:

Received: 26.11.2023

Revised: 28.12.2023

Accepted: 11.01.2024

Published: 22.01.2024

1. Introduction

The emergence of Generative Artificial Intelligence (Generative AI) and Large Language Models (LLMs) has fundamentally transformed the way organizations create, process, and utilize information. Recent advances in transformer-based language models have

enabled financial institutions to automate knowledge-intensive tasks, enhance customer engagement, improve investment research, streamline regulatory compliance, and support intelligent decision-making. As banks, asset and wealth managers, insurance companies, and capital market firms accelerate their digital transformation initiatives, LLM-powered applications are increasingly being integrated into enterprise workflows to improve operational efficiency, employee productivity, and customer experience. Despite these opportunities, deploying LLMs within enterprise financial environments presents significant technical and regulatory challenges. Financial institutions operate under strict regulatory frameworks that require robust data privacy, cybersecurity, auditability, model governance, and risk management. The integration of LLMs with sensitive customer information and mission-critical business systems raises concerns regarding data leakage, hallucinated responses, explainability, identity and access management, regulatory compliance, and responsible AI adoption. Traditional enterprise architectures were not designed to support the unique requirements of generative AI, creating the need for a secure, scalable, and governance-driven architectural approach.

Although existing research has extensively explored the capabilities of LLMs and their applications across various domains, relatively limited attention has been devoted to developing comprehensive enterprise architecture frameworks tailored to the security, governance, interoperability, and operational requirements of regulated financial institutions. Addressing these challenges requires an architecture that seamlessly integrates AI services with existing enterprise applications while maintaining compliance, resilience, and enterprise-grade security. This paper proposes a **novel** Enterprise Architecture Framework for Secure Large Language Model Deployment in Financial Services. The proposed framework introduces a layered architecture that integrates enterprise AI orchestration, Retrieval-Augmented Generation (RAG), enterprise integration, security and governance, and continuous monitoring to support the secure adoption of LLMs. The framework is evaluated through a representative enterprise financial services case study to demonstrate its practical applicability. The primary contributions of this research are: (1) the introduction of a reusable enterprise architecture framework for secure LLM deployment, (2) a comprehensive security and governance model for regulated financial institutions, and (3) practical implementation guidance for enterprise architects and technology leaders seeking to deploy trustworthy, scalable, and compliant generative AI solutions.

2. Related Work

Research on enterprise LLM deployment draws from several adjacent streams: foundation models, retrieval-augmented generation, financial-domain language models, responsible AI, cybersecurity, and enterprise architecture. These streams have advanced rapidly, but they are commonly treated as separate technical or governance concerns. This section synthesizes literature available through 2023 and identifies the architectural elements required to operationalize LLMs in regulated financial institutions.

2.1. Foundation Models and Enterprise Language Systems

Transformer architectures established the technical basis for contemporary language models by enabling scalable attention-based sequence processing [9]. Subsequent large-scale generative models demonstrated strong few-shot and in-context learning capabilities [10], while the foundation-model literature emphasized that model capabilities, downstream adaptation, and societal risks must be evaluated as an interconnected ecosystem rather than as isolated benchmarks [11]. Instruction tuning and reinforcement learning from human feedback improved the usability and alignment of general-purpose language models [50,51], but these approaches do not independently provide enterprise authorization, provenance, data residency, records management, or business-process controls. Consequently, a financial institution requires an architectural layer that governs how models are selected, invoked, monitored, and integrated with enterprise services.

2.2. Retrieval-Augmented Generation and Knowledge Grounding

Retrieval-Augmented Generation (RAG) combines parametric language-model knowledge with an external, updateable knowledge store and was introduced to improve knowledge-intensive generation and provenance [12]. Dense Passage Retrieval [13] and Fusion-in-Decoder [14] further advanced neural retrieval and evidence aggregation. Recent surveys describe RAG as a family of architectures spanning ingestion, indexing, retrieval, augmentation, and generation [15]. In regulated environments, however, retrieval quality alone is insufficient. Enterprise RAG must preserve source entitlements, document versions, retention classifications, geographic restrictions, and evidence citations. The present framework therefore extends standard RAG by treating authorization, provenance, and freshness as architectural controls rather than optional application features.

2.3. Tool Use, Orchestration, and Emerging Agentic Systems

ReAct demonstrated that language models can interleave reasoning and actions to interact with external environments [16], while Toolformer showed that models can learn when and how to invoke external tools [17]. These developments create opportunities for workflow automation, but they also introduce risks related to excessive agency, tool misuse, and non-deterministic execution. In financial services, model-generated suggestions must be separated from authorized transactions, and agent permissions must be constrained by deterministic policy engines, execution budgets, and human approval. This motivates the proposed AI Orchestration and Policy Layer.

2.4. Financial-Domain Language Models and Benchmarks

Domain-specific language modeling has shown that financial vocabulary, documents, and tasks benefit from specialized data and evaluation. FinBERT adapted transformer models for financial sentiment analysis [18]. BloombergGPT demonstrated a large-scale finance-oriented language model trained on mixed financial and general corpora [19], and FinGPT explored an open-source approach to financial LLM development [20]. FinanceBench subsequently showed that leading LLM configurations still performed poorly on realistic, evidence-based financial question answering, including when retrieval was used [7]. These findings reinforce the need for enterprise grounding, controlled use cases, risk-tiered evaluation, and human review rather than reliance on model capability claims alone.

2.5. Responsible AI, Explainability, and Documentation

Research on model risk has identified harms involving discrimination, misinformation, privacy, environmental impact, and misuse [35,36]. Model Cards [37] and Datasheets for Datasets [38] provide structured mechanisms for documenting intended use, limitations, and data provenance. Explainability approaches such as LIME [39] and SHAP [40] support local and feature-based interpretation, although their direct applicability to generative systems remains limited. For LLM applications, explainability must therefore be complemented by retrieval provenance, prompt and model versioning, policy decisions, and human approval records.

2.6. Security, Privacy, and Regulatory Governance

Enterprise LLM systems inherit conventional cloud and software-security risks while introducing prompt injection, sensitive-information disclosure, insecure tool use, model extraction, and supply-chain vulnerabilities. NIST Zero Trust Architecture [21] and the Secure Software Development Framework [22] provide foundational controls, while the OWASP Top 10 for LLM Applications identifies LLM-specific threats [23]. Research has demonstrated extraction of memorized training data [41], indirect prompt-injection attacks against integrated applications [42], adversarial attacks that bypass safety alignment [43], and the value of structured red-team methods [44]. Governance must also account for privacy and sector obligations, including GDPR [26], GLBA [27], PCI DSS [28], operational resilience guidance [29,30], and AI risk management under NIST AI RMF [8].

2.7. Synthesis

The literature provides strong building blocks for models, retrieval, tool use, security, documentation, and evaluation, but it does not offer a single enterprise reference architecture that integrates these capabilities with identity, legacy-system interoperability, model governance, compliance, observability, and human accountability. SF-LLM-EA addresses this gap by combining the technical execution path with cross-cutting institutional controls and evidence requirements.

3. Research Gap

The rapid evolution of Large Language Models (LLMs) has significantly advanced artificial intelligence applications across multiple industries, including financial services. Existing research has primarily focused on improving language model performance, prompt engineering, Retrieval-Augmented Generation (RAG), domain adaptation, and conversational AI capabilities. While these studies demonstrate the transformative potential of LLMs for customer service, investment research, fraud detection, and operational automation, they largely address individual technical components rather than the broader enterprise architecture required for production deployment in highly regulated financial environments.

Financial institutions operate under stringent regulatory and security requirements, including data privacy, identity and access management, cybersecurity, model governance, auditability, and compliance with industry regulations. Current literature provides limited guidance on how these requirements can be integrated into a unified enterprise architecture capable of supporting secure, scalable, and resilient LLM deployments. Most published approaches emphasize model-centric improvements without adequately addressing enterprise integration, governance, interoperability with legacy systems, cloud-native deployment, observability, and operational risk management.

Furthermore, existing enterprise AI frameworks often lack a holistic architectural approach that combines AI orchestration, enterprise data management, Retrieval-Augmented Generation (RAG), security controls, governance mechanisms, and continuous monitoring within a single reference architecture. As financial institutions increasingly adopt hybrid cloud environments and modern digital platforms, the absence of standardized enterprise architecture guidance presents challenges in ensuring scalability, regulatory compliance, explainability, and responsible AI adoption across multiple business functions.

To address these limitations, this research proposes a novel Enterprise Architecture Framework for Secure Large Language Model Deployment in Financial Services. Unlike existing model-focused studies, the proposed framework integrates six vertically integrated layers—Experience and Channels, AI Orchestration and Policy, LLM Services, Knowledge and Retrieval-Augmented Generation (RAG), Enterprise Integration, and Platform and Deployment—supported by the cross-cutting domains of Security, Governance, Compliance, and Observability—into a comprehensive reference architecture. The framework bridges the gap between advances in generative AI and the operational, security, and governance requirements of regulated financial institutions, providing enterprise architects and technology leaders with practical guidance for deploying trustworthy, scalable, and compliant LLM solutions

4. Proposed Enterprise Architecture Framework

The proposed framework, termed the Secure Financial-Services LLM Enterprise Architecture (SF-LLM-EA), is a vendor-neutral reference architecture for deploying generative artificial intelligence in regulated financial institutions. It addresses a central limitation of model-centric approaches: a high-performing language model is not, by itself, an enterprise system. Production deployment requires controlled user channels, policy-driven orchestration, approved model services, trusted knowledge retrieval, secure integration with systems of record, lifecycle governance, and continuous operational evidence. The framework therefore separates business interaction, AI decisioning, data grounding, enterprise connectivity and control functions while applying security, governance, compliance and observability across every layer.

4.1. Framework Objectives and Novelty

The framework is designed around five objectives: (1) prevent unauthorized disclosure or use of financial and personal information; (2) reduce unsupported model output through entitlement-aware retrieval and evidence-based responses; (3) permit controlled use of multiple models without locking business applications to one provider; (4) create auditable accountability for models, prompts, data, tools and decisions; and (5) scale from low-risk employee assistance to tightly governed client-facing and decision-support use cases. Its primary novelty is the integration of enterprise architecture, LLM-specific security, model-risk governance, RAG controls and financial-services operational resilience within one traceable architecture.

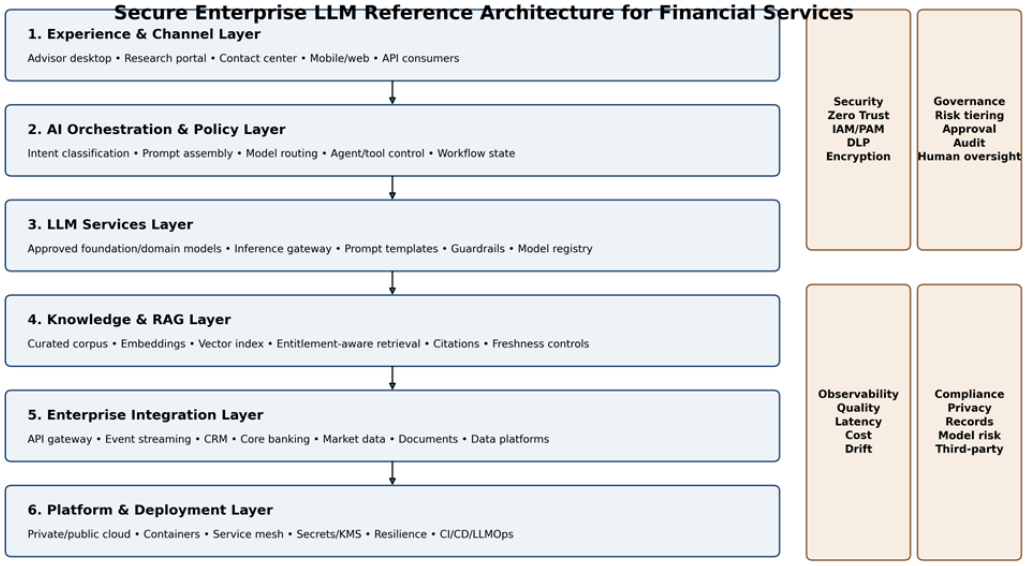

Figure 1. SF-LLM-EA Layered Reference Architecture with Cross-Cutting Security, Governance, Compliance and Observability

4.2. Experience and Channel Layer

The experience layer exposes approved LLM capabilities through business-controlled interfaces rather than direct access to foundation-model endpoints. Typical channels include advisor and relationship-manager workspaces, research portals, operations consoles, contact-center desktops, employee collaboration tools, mobile applications and machine-to-machine APIs.

Each channel transmits authenticated identity, business role, legal entity, jurisdiction, device posture and session context. The response contract should include provenance, confidence indicators, policy notices and an escalation path. For client-facing use, the layer enforces approved wording, accessibility, records-retention requirements and disclosure rules. High-risk actions are separated from conversational output and require explicit user confirmation or human approval.

4.3. AI Orchestration and Policy Layer

This layer is the control plane of the architecture. It classifies user intent, determines risk tier, assembles prompts, selects tools and knowledge sources, routes requests to approved models, manages conversation state, and applies policies before and after inference.

A policy decision point evaluates attributes such as user role, requested action, data classification, geography, client relationship, model approval status and purpose limitation. The orchestration service should support deterministic workflows for regulated processes and bounded agent behavior for exploratory tasks. Agents receive least-privilege tool permissions, execution limits, timeouts and transaction boundaries. No model is permitted to invoke a payment, trade, account change or client communication directly without a controlled business service and an approval gate.

4.4. LLM Services Layer

The LLM services layer provides a governed abstraction over internally hosted, private-cloud and externally managed models. An inference gateway exposes a consistent API while enforcing model allowlists, version pinning, quotas, data-handling restrictions, regional routing and fallback logic.

A model registry records the model owner, provider, intended uses, prohibited uses, evaluation results, training-data disclosures where available, security review, residual risks and approval expiration. Prompt templates and system instructions are versioned as controlled artifacts. Output guardrails detect restricted content, sensitive-data leakage, unsupported claims, policy violations and anomalous tool requests. Model selection is based on risk, capability, latency, cost and data residency—not solely benchmark performance.

4.5. Knowledge and Retrieval-Augmented Generation Layer

The knowledge and RAG layer grounds model responses in approved enterprise information. Sources may include policies, research, product documentation, procedures, contracts, client-approved records, market data and regulatory publications. Content passes ingestion controls for classification, malware scanning, duplication, quality, ownership, retention and freshness.

Retrieval is entitlement-aware: the vector index and document store must preserve source-level permissions rather than creating a universal semantic corpus. Query-time filtering applies user, account, geography, legal-entity and purpose constraints before any text is presented to the model. Retrieved passages include stable source identifiers, version dates and citations. The architecture separates retrieval confidence from model confidence and supports refusal when authoritative evidence is insufficient.

4.6. Enterprise Integration Layer

This layer connects generative AI with systems of record through governed APIs, events and integration services. Relevant platforms include CRM, core banking, wealth-management systems, payments, fraud and risk engines, market-data services, document repositories, master-data management, customer identity, data lakes and workflow platforms.

The integration layer converts model suggestions into typed business commands and validates them against existing authorization and business rules. It prevents an LLM from becoming a system of record. Read operations and write operations are separated; write operations use idempotency, transaction controls, approval and reconciliation. API gateways provide schema validation, throttling, threat protection and full traceability.

4.7. Platform and Deployment Layer

The platform layer supports hybrid and multi-cloud deployment patterns suitable for regulated institutions. It includes container platforms, service mesh, private networking, hardware-backed key management, secrets management, workload identity, policy-as-code, resilient queues, encrypted storage, CI/CD and LLMOps pipelines.

Sensitive prompts and retrieved content should remain within approved trust zones. External model access uses private connectivity where available, contractual no-training commitments, tenant isolation and egress controls. The design supports graceful degradation: if a model or retrieval service is unavailable, the application should fail safely, preserve records and avoid silently switching to an unapproved model.

4.8. Cross-Cutting Security, Governance and Compliance

Security and governance are not separate downstream review activities; they are architectural functions applied to every request and lifecycle stage. Controls include zero-trust access, phishing-resistant authentication, privileged-access management, encryption, tokenization, data-loss prevention, prompt-injection defenses, software-supply-chain controls, records management, privacy impact assessment, model-risk classification, third-party risk and human oversight.

The framework aligns governance activities with the NIST AI Risk Management Framework functions—Govern, Map, Measure and Manage—and incorporates LLM-specific threats identified by OWASP, including prompt injection, insecure output handling, sensitive-information disclosure, excessive agency, model denial of service and supply-chain vulnerabilities.

4.9. Monitoring and Observability Layer

Operational evidence is captured across user experience, orchestration, retrieval, model inference, integrations and business outcomes. Telemetry includes latency, availability, token use, cost, retrieval precision, citation coverage, groundedness, refusal rate, policy violations, prompt attacks, data leakage events, human overrides, user feedback and downstream errors.

Monitoring is segmented by use case and risk tier. Quality thresholds that are acceptable for employee brainstorming may be unacceptable for investment research or client communication. Continuous evaluation combines automated tests, curated challenge sets, red-team exercises, production sampling and incident review. Logs must support audit and investigation while minimizing unnecessary retention of sensitive prompts.

4.10. Secure Request and Response Flow

Figure 2 illustrates the runtime control sequence. Identity and context are established before intent classification. Data-loss prevention and policy controls are applied before retrieval or inference. RAG searches only authorized sources. The orchestration layer then routes the request to an approved model and tool chain. Output is validated for grounding, citations, sensitive content and business-rule compliance. High-risk responses or actions require human approval. Finally, the architecture records decision evidence, user feedback, quality measures and operational cost.

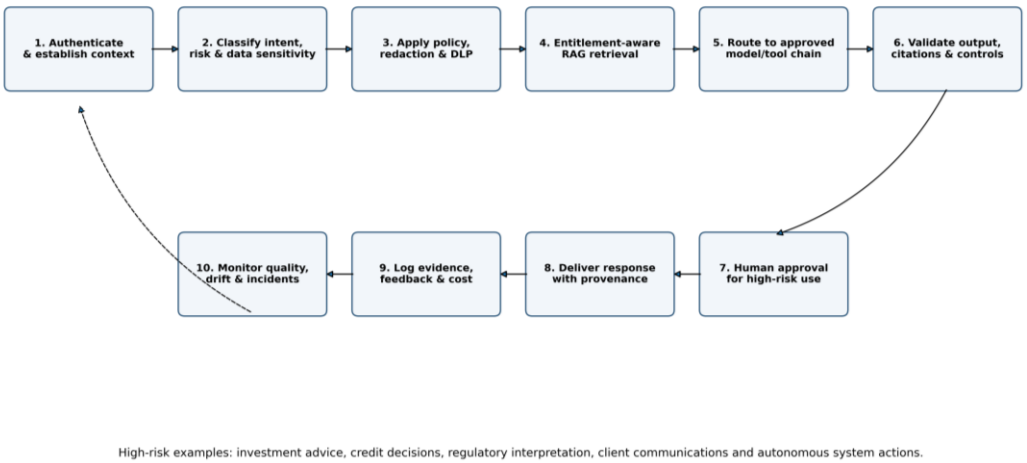

Figure 2. Secure End-To-End Request, Retrieval, Inference, Validation, Approval and Monitoring Flow.

4.11. Risk-Tiered Deployment Model

Table 1. Risk-Based AI Autonomy Tiers with Required Governance Controls and Evidence Requirements

Tier	Illustrative Use	Permitted Autonomy	Required Controls	Example Evidence
1 - Low	Summarization of public or non-sensitive material	Generate response	Authentication, content filters, logging	Usage logs, baseline quality tests
2 - Moderate	Internal research and knowledge assistance	Generate with approved RAG	Entitlements, citations, DLP, model allowlist	Retrieval tests, source provenance, user feedback
3 - High	Client communication or regulatory interpretation	Draft only; human approval required	Four-eyes review, records retention, legal/compliance controls	Approval record, final-versus-draft comparison
4 - Critical	Credit, trading, payments or account actions	No autonomous decision or execution	Independent decision engine, formal model-risk review, transaction controls	Decision trace, override records, reconciliation, incident tests

4.12. Governance Lifecycle

Every LLM use case is governed as a lifecycle rather than a one-time approval. A business owner and technology owner first define intended purpose, users, data, impact and prohibited uses. Architecture and control design follows the assigned risk tier. Models, prompts, RAG pipelines and integrations are evaluated together because system behavior emerges from their interaction. Approval creates a registered configuration baseline. Production monitoring, periodic testing, incident response and renewal ensure that material changes trigger reassessment.

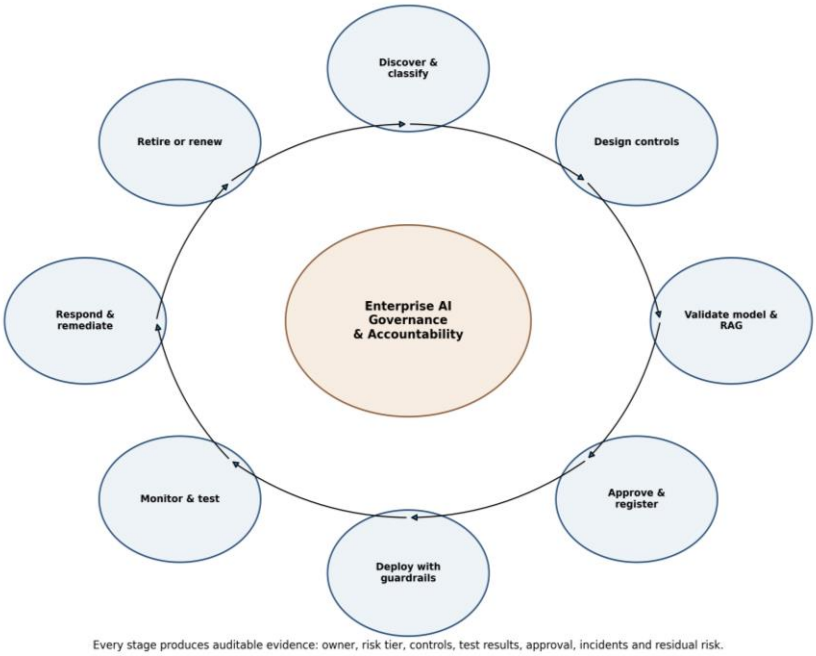

Figure 3. Lifecycle Governance Model for Accountable Deployment and Operation of Enterprise LLM Systems

4.13. Core Design Principles

- Security and privacy by design: Controls are applied before data reaches retrieval or model services, with least privilege and purpose limitation.
- Ground before generate: Authoritative enterprise knowledge and citations are preferred over unsupported parametric knowledge.
- Policy-driven orchestration: Risk and authorization decisions are externalized from prompts and enforced by deterministic services.
- Human accountability: LLMs assist qualified professionals; high-impact judgments and actions remain attributable to accountable humans.
- Model and provider portability: Applications depend on an inference contract and registry rather than proprietary provider-specific interfaces.
- Evidence-first governance: Approvals, tests, prompts, sources, outputs, overrides and incidents produce traceable records.
- Separation of suggestion and execution: Model output is never treated as an authorized transaction or system-of-record update.
- Resilience and safe failure: The system refuses, degrades or queues requests rather than bypassing controls during outages.
- Continuous evaluation: Security, quality, fairness, groundedness and business value are measured throughout operation.

4.14. Evaluation Measures

Table 2. Evaluation Dimensions, Performance Metrics, Target Interpretation, and Organizational Ownership for AI Governance

Dimension	Representative Measure	Target Interpretation	Control Owner
Groundedness	Percentage of material claims supported by retrieved evidence	Higher is better; threshold varies by risk tier	AI product and knowledge owner
Retrieval quality	Precision@k, recall@k and entitlement-filter accuracy	Must prove relevance and prevent unauthorized retrieval	Data/knowledge engineering
Security	Prompt-attack success rate and sensitive-data leakage rate	Zero tolerance for confirmed restricted-data disclosure	Cybersecurity
Reliability	Availability, p95 latency and safe-failure rate	Must meet use-case service objectives without bypassing controls	Platform engineering
Human oversight	Override, rejection and escalation rates	Used to detect low-quality output and automation bias	Business control owner
Compliance	Citation, disclosure and records-retention completeness	Must meet policy and jurisdictional requirements	Legal/compliance
Economics	Cost per completed task and productivity benefit	Value assessed after control and quality requirements are met	Product/finance

4.15. Architectural Implications for Financial Institutions

The framework enables institutions to introduce LLM capability incrementally. Low-risk internal assistance can use the same identity, orchestration, registry, knowledge and monitoring foundations later required for more sensitive use cases. This reduces duplicated controls and prevents isolated business units from creating incompatible AI stacks. The architecture also clarifies accountability: business owners define acceptable use and outcomes; data owners govern knowledge; model-risk and security teams define control thresholds; platform teams operate shared services; and qualified users remain responsible for high-impact decisions.

The framework is intentionally technology-neutral. Commercial APIs, open-weight models, private models and specialized financial models can coexist, provided each deployment satisfies the same control contract. This allows financial institutions to adapt as model capabilities, costs and regulatory expectations evolve without redesigning business applications or weakening governance.

5. Research Methodology

This study uses Design Science Research (DSR) to develop and evaluate SF-LLM-EA as an enterprise architecture artifact. DSR is appropriate because the research objective is to create a purposeful solution to a class of organizational and technical problems and to evaluate its utility, rigor, and limitations [31-33].

5.1. Research Objectives and Design Requirements

The research problem was translated into five design requirements: secure handling of regulated data; evidence-grounded generation; modular integration with models and systems of record; auditable governance and human accountability; and scalable operation across business lines and jurisdictions. Requirements were derived from peer-reviewed research, standards, regulatory guidance, and publicly documented enterprise practices available through 2023.

5.2. Artifact Development Process

The artifact was developed iteratively through six activities: problem identification; requirement definition; synthesis of architectural building blocks; construction of the layered reference architecture; demonstration through a representative Tier-1 financial-institution case; and evaluation against quality attributes and threat scenarios. The process follows established DSR guidance that links relevance to an organizational problem environment, rigor to the knowledge base, and evaluation to the artifact's intended purpose [31-33].

5.3. Evidence Base and Source-Selection Protocol

The literature review prioritized primary research papers, recognized technical standards, regulatory guidance, and authoritative industry frameworks. Sources were included when they addressed at least one of the following: LLM architecture, RAG, financial NLP, AI evaluation, cloud security, privacy, model governance, operational resilience, or enterprise architecture. Literature published after the early-2023 cutoff was excluded from the historical evidence base except where clearly identified as later policy guidance relevant to manuscript preparation.

5.4. Case-Study Construction

The case study is an illustrative design scenario rather than a report of a named institution's proprietary implementation. It synthesizes public characteristics of Tier-1 global banks, including multiple lines of business, hybrid-cloud infrastructure, geographically distributed data, complex identity entitlements, and stringent regulatory oversight. No confidential employer information, internal architecture, non-public metrics, or client data are used.

5.5. Evaluation Strategy

The framework is evaluated through scenario-based architecture analysis. Representative scenarios include investment research, advisor preparation, regulatory-policy assistance, operations knowledge retrieval, and customer-service drafting. Each scenario is assessed against functional suitability, security, reliability, performance efficiency, maintainability, compatibility, usability, and portability, drawing on ISO/IEC 25010 quality concepts [34]. Additional measures cover groundedness, citation correctness, entitlement enforcement, prompt-attack resistance, human override, and audit completeness.

5.6. Validation Logic and Traceability

Validation uses a requirements-to-controls traceability approach. Each design requirement is mapped to one or more architecture components, control mechanisms, evaluation measures, and evidence artifacts. For example, the requirement to prevent unauthorized disclosure maps to identity context, attribute-based authorization, entitlement-aware retrieval, data-loss prevention, encryption, and audit logging. This traceability supports review by enterprise architecture, cybersecurity, model-risk, legal, compliance, and business stakeholders.

5.7. Methodological Boundaries

The study does not claim experimental proof of performance or measured business outcomes. Evaluation findings are analytical and scenario based. Quantitative thresholds presented in the framework should be calibrated by each institution according to use-case risk, regulatory obligations, model characteristics, and operational service levels. These boundaries are addressed further in Section 10.

6. Case Study (Enterprise Financial Institution)

This case study demonstrates the application of the proposed Secure Financial-Services LLM Enterprise Architecture (SF-LLM-EA) in a representative large multinational financial institution. The scenario is illustrative and synthesizes publicly known enterprise banking practices rather than describing any confidential implementation. The institution operates retail banking, commercial banking, wealth management, capital markets, and asset management businesses across multiple jurisdictions.

6.1. Representative Institution Profile

The representative institution is modeled as a Tier-1 global financial group serving consumer, commercial, corporate, institutional, and wealth-management clients across North America, Europe, and Asia-Pacific. Its technology estate includes thousands of applications, multiple regulated legal entities, hybrid public- and private-cloud platforms, regional data-residency constraints, centralized and federated identity services, and extensive third-party data and software dependencies. The institution employs a large population of advisors, analysts, operations specialists, risk professionals, software engineers, and control functions. This scale creates a realistic environment in which an LLM platform must support heterogeneous use cases without creating fragmented model access, inconsistent controls, or uncontrolled copies of sensitive data.

6.2. Business Challenges and Risk Context

Key challenges include fragmented knowledge repositories, inconsistent customer service, manual research processes, long document review cycles, increasing regulatory obligations, and the need to protect sensitive financial information while improving employee productivity.

6.3. Existing Technology and Data Landscape

Before adoption of SF-LLM-EA, knowledge is distributed across document-management platforms, CRM records, research repositories, policy libraries, collaboration sites, market-data systems, ticketing tools, data lakes, and legacy shared drives. Authorization models differ by platform, and document metadata is frequently incomplete or inconsistent. Existing search tools rely heavily on keywords and may not preserve the business context needed for advisor, research, or compliance workflows. Point solutions have begun experimenting with commercial LLM APIs, creating concerns about duplicated integration, uncontrolled prompts, provider lock-in, inconsistent logging, and data being transmitted outside approved trust zones. The target architecture must therefore introduce a shared control plane while allowing business-specific experiences and knowledge collections.

6.4. Framework Implementation

The institution adopts the proposed six-layer architecture with cross-cutting control domains. Relationship managers and analysts access a secure AI assistant through existing digital channels. An orchestration service classifies requests, applies policy, retrieves only authorized documents using entitlement-aware RAG, routes prompts to approved LLMs, validates responses with citations, and records audit evidence. Enterprise integration connects CRM, document management, market-data platforms, identity services, workflow systems, and data platforms through governed APIs.

6.5. Phased Deployment and Operating Model

Deployment is organized into four phases. Phase 1 establishes shared identity, model gateway, prompt registry, logging, and a low-risk internal knowledge assistant using public or non-sensitive content. Phase 2 introduces entitlement-aware RAG for internal research and operations knowledge, with source citations and feedback capture. Phase 3 adds high-risk drafting use cases, including client communications and regulatory analysis, subject to four-eyes approval, records retention, and legal or compliance review. Phase 4 permits tightly bounded agentic workflows in which tools can prepare typed commands, but final execution remains with authorized deterministic services and accountable human users. A federated operating model assigns platform ownership to a central AI engineering function while business owners, data owners, cybersecurity, model risk, privacy, and compliance retain control over intended use, data access, approval thresholds, and ongoing monitoring.

6.5.1. End-To-End Advisor Research Scenario

An authenticated wealth advisor requests a briefing for an upcoming client meeting. The channel supplies the advisor's identity, client relationship, jurisdiction, device posture, and purpose. The orchestration layer classifies the request as moderate risk, applies prompt and data-loss-prevention rules, and invokes RAG only against repositories the advisor is authorized to access. Retrieved passages are filtered by client, legal entity, geography, retention status, and effective date. The model gateway selects an approved model based

on risk, data residency, latency, and cost. The generated briefing includes citations and clearly separates source-grounded facts from suggested talking points. Output validation checks unsupported claims, restricted information, and policy wording. The advisor reviews the response; no CRM update or client communication occurs automatically. The full evidence chain—identity, policy decision, retrieval sources, model and prompt versions, output, user action, and feedback—is retained according to records policy.

6.6. Representative Use Cases

Table 3. Representative Enterprise AI Use Cases with Associated Business Value, Risk Levels, and Primary Governance Controls

Use Case	Business Value	Risk Level	Primary Controls
Investment research assistant	Accelerates research synthesis and knowledge discovery	Moderate	RAG, citations, human review
Advisor client preparation	Improves relationship management and meeting preparation	Moderate	Entitlement-aware retrieval, audit logs
Regulatory policy assistant	Speeds compliance analysis	High	Legal approval, version-controlled knowledge
Operations knowledge assistant	Reduces manual search and repetitive work	Low	Role-based access, monitoring
Customer service draft responses	Improves consistency and productivity	High	Human approval, content filtering

6.7. Case-Study Acceptance Criteria

The deployment is considered architecturally acceptable only when it demonstrates: (1) no retrieval outside the requester's entitlements; (2) traceable citations for material factual claims; (3) approved model and prompt versions; (4) safe failure when models, retrieval services, or policy engines are unavailable; (5) human approval for high-risk outputs; (6) complete audit evidence; (7) measurable service-level performance; and (8) an incident-response path for prompt attacks, data leakage, harmful output, or provider failure. These criteria prevent a successful demonstration from being mistaken for production readiness.

6.8. Expected Architectural Benefits and Implementation Lessons

The framework demonstrates improved knowledge accessibility, consistent governance, reusable integration patterns, centralized model management, stronger auditability, and simplified compliance oversight. It also enables incremental AI adoption by supporting multiple business domains through a shared enterprise platform. The most important benefit is not unrestricted automation but reuse of common controls across business units. A central inference gateway reduces provider-specific integration, while the policy and retrieval layers prevent each application from independently implementing authorization and data handling. The case also reveals key implementation lessons: knowledge quality and entitlement metadata are prerequisites for useful RAG; model performance must be evaluated within the complete system; human reviewers require clear provenance and escalation paths; and production monitoring must connect technical quality to downstream business outcomes. Any quantified benefits would require controlled measurement and are not asserted by this illustrative case.

7. Framework Evaluation

The framework is evaluated qualitatively using recognized enterprise architecture quality attributes and secure AI deployment principles. The objective is to assess whether the proposed architecture satisfies the operational and governance requirements of regulated financial institutions.

7.1. Evaluation Method

The evaluation combines architecture quality analysis, control-coverage analysis, and scenario walkthroughs. Rather than scoring the framework solely on model accuracy, the analysis examines whether the complete socio-technical system can deliver useful output while maintaining security, governance, compliance, resilience, and human accountability. Evidence is drawn from the architecture's explicit components, request flow, risk-tier model, governance lifecycle, and case-study controls.

7.2. Quality-Attribute Analysis

Quality attributes are interpreted as design claims that must be supported by mechanisms and measurable evidence. Security is supported by identity context, Zero Trust enforcement, DLP, encryption, prompt defenses, and separation of suggestion from execution. Scalability is supported by stateless orchestration, model abstraction, horizontally scalable retrieval, and cloud-native deployment. Interoperability is supported by API contracts, event integration, and typed commands. Maintainability is supported by separation of concerns and versioned artifacts. Reliability is supported by fallback policies, circuit breaking, queues, safe failure, and observability. Responsible AI is supported by provenance, refusal behavior, human oversight, and lifecycle governance.

7.3. Security and Threat-Scenario Evaluation

Threat scenarios include direct and indirect prompt injection, unauthorized retrieval, sensitive-data leakage, malicious documents in the RAG corpus, insecure tool invocation, model or dependency compromise, denial of service, and inappropriate autonomous action. For each scenario, the framework applies prevention, detection, containment, and recovery controls. For example, indirect prompt injection is mitigated through content classification, separation of instructions from retrieved data, tool allowlists, constrained execution, output validation, and human approval. No single control is treated as sufficient; defense in depth is required because LLM behavior remains probabilistic.

Table 4. Evaluation of Enterprise AI Framework Support across Key Quality Attributes

Quality Attribute	Evaluation	Framework Support	Assessment
Security	Zero Trust, IAM, DLP, encryption, prompt protection	Cross-cutting security layer	Strong
Scalability	Cloud-native orchestration and modular services	Layered architecture	Strong
Governance	Model registry, approvals, audit trails	Governance lifecycle	Strong
Compliance	Traceability, records, explainability	Compliance controls	Strong
Interoperability	API-first integration	Enterprise Integration Layer	Strong
Maintainability	Loose coupling and reusable services	Modular layers	Strong
Reliability	Monitoring, fail-safe routing, resilience	Observability layer	Strong
Responsible AI	Human oversight, citations, policy enforcement	Governance + orchestration	Strong

7.3.1. Quantitative Evaluation Plan

A future empirical implementation should report, at minimum, retrieval precision and recall, entitlement-filter accuracy, groundedness, citation correctness, hallucination rate, prompt-attack success rate, restricted-data leakage, p50 and p95 latency, availability, cost per completed task, human override rate, approval turnaround, incident rate, and task-level productivity or service-quality outcomes. Metrics should be segmented by risk tier, use case, model, jurisdiction, user role, and document type. Statistical confidence intervals and baseline comparisons are necessary before claiming improvement.

7.4. Comparative Analysis

Compared with model-centric deployment approaches, the proposed framework explicitly integrates enterprise architecture, retrieval-augmented generation, security-by-design, governance, operational monitoring, and financial-services compliance into a unified reference architecture. This improves architectural completeness and makes the framework suitable for enterprise-scale deployment rather than isolated proof-of-concept implementations.

7.5. Architecture Trade-Offs

The framework introduces deliberate trade-offs. Strong entitlement filtering and output validation increase latency; human approval limits automation but preserves accountability; multi-model portability adds operational complexity; extensive logging improves auditability but creates privacy and retention concerns; and highly curated knowledge improves groundedness but requires sustained data stewardship. These trade-offs should be resolved according to risk tier rather than through one uniform enterprise configuration.

7.6. Interpretation of Evaluation

The evaluation indicates that separating orchestration, model services, enterprise knowledge, integration, and governance reduces architectural coupling while improving portability across LLM providers. Embedding policy enforcement and observability

throughout the lifecycle enhances trustworthiness and operational resilience. Although the evaluation is scenario-based rather than empirical, it demonstrates that the framework addresses the principal architectural concerns reported in enterprise AI literature through 2023.

8. Security, Governance and Regulatory Compliance

8.1. Introduction

The adoption of Large Language Models (LLMs) within financial institutions introduces significant opportunities for intelligent automation and enterprise decision support. However, unlike conventional AI systems, LLMs interact with highly sensitive enterprise information through natural language, increasing the risk of unauthorized disclosure, hallucinated responses, prompt injection, model misuse, and regulatory non-compliance. Financial institutions therefore require an enterprise security architecture that extends beyond traditional cybersecurity controls by integrating AI governance, identity management, privacy protection, human oversight, and continuous monitoring throughout the model lifecycle. The proposed Secure Financial-Services LLM Enterprise Architecture (SF-LLM-EA) embeds security and governance as cross-cutting architectural capabilities applied consistently across every layer of the enterprise AI ecosystem.

8.2. Identity and Access Management (IAM)

Identity management serves as the first line of defense for enterprise AI systems. Every interaction with the LLM platform must be authenticated, authorized, and continuously evaluated based on user identity, business role, geographic location, device posture, and organizational context.

The framework incorporates:

- Single Sign-On (SSO)
- Multi-Factor Authentication (MFA)
- OAuth 2.0/OpenID Connect
- Role-Based Access Control (RBAC)
- Attribute-Based Access Control (ABAC)
- Privileged Access Management (PAM)
- Just-in-Time privileged access

Instead of allowing unrestricted access to enterprise knowledge, the orchestration layer dynamically evaluates user entitlements before retrieval operations are executed. This prevents unauthorized exposure of confidential client records, research documents, and regulatory information.

8.3. Zero Trust Security Architecture

Traditional perimeter-based security models are insufficient for AI-enabled enterprise systems. The proposed framework adopts Zero Trust principles where no user, application, model, or service is inherently trusted.

Core Zero Trust principles include:

- Verify every request
- Least privilege access
- Continuous authentication
- Micro-segmentation
- Device trust validation
- Policy-based authorization
- Continuous risk evaluation

Every LLM request undergoes identity verification, policy evaluation, and data classification before any prompt reaches the model.

8.4. Encryption and Data Protection

Enterprise AI systems process highly confidential financial information.
The framework applies encryption across three states:

8.4.1. Data at Rest

- AES-256
- Encrypted vector databases
- Encrypted document repositories

8.4.2. Data in Transit

- TLS 1.3
- Mutual TLS
- API Gateway encryption

8.4.3. Data in Use

- Secure execution environments
- Confidential computing
- Tokenization
- Dynamic masking

Personally identifiable information (PII) and sensitive financial data are masked or tokenized before inference whenever appropriate.

8.5. Model Governance

Enterprise deployment of LLMs requires comprehensive governance extending beyond model selection.

The proposed governance framework includes:

- Model Registry
- Model Version Control
- Approval Workflow
- Risk Classification
- Model Documentation
- Performance Benchmarking
- Bias Assessment
- Explainability Assessment
- Model Retirement Policy

Each model receives a documented risk rating before production deployment.

8.6. Prompt Security

Prompt injection represents one of the most significant threats to enterprise LLM systems.

The proposed architecture introduces multiple defensive layers:

- Prompt validation
- Prompt sanitization
- Prompt isolation
- System prompt protection
- Context window validation
- Tool permission restrictions
- Output filtering

These controls significantly reduce the risk of prompt manipulation and unauthorized tool execution.

8.7. Data Privacy

Financial institutions must comply with stringent privacy requirements.
The framework supports compliance with:

- GDPR
- CCPA
- PCI DSS
- GLBA
- SOX
- Regional banking regulations

Privacy controls include:

- Purpose limitation
- Data minimization
- Consent management
- Right-to-delete support
- Data retention policies
- Secure audit trails

8.8. Auditability and Explainability

Enterprise AI must produce decisions that are transparent and auditable.

The proposed framework records:

- User identity
- Prompt version
- Retrieved documents
- Model version
- Generated response
- Human approval
- Policy decisions
- Confidence indicators
- Timestamp

This evidence supports both internal governance and regulatory audits.

8.9. Human Oversight

LLMs are positioned as decision-support systems rather than autonomous decision-makers.

High-risk activities—including investment recommendations, credit decisions, payment execution, regulatory interpretation, and client communications—require human review before action.

Human oversight reduces automation bias while ensuring accountability for critical business decisions.

8.10. AI Risk Management

The framework aligns AI risk management with the NIST AI Risk Management Framework (AI RMF) and addresses common LLM threats identified by the OWASP Top 10 for LLM Applications**.**

Risk categories include:

- Prompt injection
- Hallucinations
- Sensitive data leakage
- Excessive agent autonomy
- Model denial of service
- Supply-chain vulnerabilities
- Bias and fairness concerns
- Third-party model risks

Continuous monitoring, periodic testing, red-team exercises, and incident response processes are incorporated to maintain ongoing operational resilience.

8.11. Regulatory Alignment

The proposed architecture supports compliance with widely adopted standards and guidance, including:

Table 5. Standard Security, Privacy, and AI Governance Frameworks Supporting Secure AI System Development

Framework	Purpose
NIST AI RMF 1.0	AI governance and risk management
NIST SP 800-207	Zero Trust Architecture
NIST SP 800-218	Secure Software Development Framework
OWASP Top 10 for LLM Applications	LLM-specific security risks
ISO/IEC 27001	Information Security Management
ISO/IEC 42001	Artificial Intelligence Management Systems
PCI DSS	Payment card data protection
GDPR	Data privacy and protection

9. Discussion

The rapid adoption of Large Language Models (LLMs) has fundamentally reshaped enterprise artificial intelligence by enabling natural language interaction, knowledge synthesis, intelligent automation, and decision support across multiple business domains. Within the financial services industry, however, the deployment of LLMs extends beyond technical implementation and requires careful consideration of regulatory compliance, enterprise security, operational resilience, and governance. Unlike many existing studies that primarily focus on model performance or natural language processing capabilities, this research approaches LLM adoption from an enterprise architecture perspective, emphasizing the organizational, technological, and governance foundations required for secure deployment in highly regulated financial environments.

The proposed Secure Financial-Services Large Language Model Enterprise Architecture (SF-LLM-EA) contributes to the existing body of knowledge by introducing a comprehensive, layered reference architecture that integrates enterprise AI orchestration, Retrieval-Augmented Generation (RAG), enterprise integration, governance, security, and continuous monitoring within a single architectural framework. Rather than treating LLMs as standalone AI services, the framework positions them as enterprise capabilities operating within established business processes, identity systems, and governance structures. This architectural perspective enables organizations to systematically incorporate generative AI into existing technology ecosystems while maintaining operational control and regulatory compliance.

A key contribution of this research is the explicit separation of architectural responsibilities into distinct but interconnected layers. The Experience Layer provides controlled user interaction through enterprise-approved digital channels, while the AI Orchestration Layer centralizes prompt management, policy enforcement, workflow execution, and model selection. The LLM Services Layer abstracts model providers through standardized interfaces, enabling organizations to adopt multiple commercial or open-source models without introducing vendor lock-in. The Knowledge and Retrieval-Augmented Generation Layer further strengthens response reliability by grounding model outputs in authoritative enterprise knowledge rather than relying solely on parametric model memory. This separation of concerns improves architectural modularity, maintainability, scalability, and long-term adaptability as AI technologies continue to evolve.

Another significant contribution is the integration of security, governance, and compliance as cross-cutting architectural capabilities rather than isolated technical controls. Financial institutions operate within complex regulatory environments that demand comprehensive auditability, data privacy, explainability, and risk management. The proposed framework embeds Zero Trust principles, identity and access management, encryption, prompt security, model governance, continuous monitoring, and human oversight across every stage of the AI lifecycle. By integrating these controls directly into the enterprise architecture, organizations can reduce operational risk while improving trust in AI-assisted decision-making.

The incorporation of Retrieval-Augmented Generation (RAG) represents another important architectural advancement. Enterprise knowledge is often distributed across document repositories, CRM platforms, policy libraries, market research databases, and internal collaboration systems. Traditional LLM deployments frequently generate responses based solely on pre-trained knowledge, increasing the likelihood of outdated or unsupported information. The proposed framework addresses this limitation through entitlement-aware retrieval, semantic search, document version control, and citation generation. These capabilities improve response transparency while supporting regulatory requirements for evidence-based decision support and knowledge traceability.

From an enterprise architecture perspective, the framework also promotes technology neutrality. Financial institutions increasingly adopt hybrid cloud environments while integrating commercial AI services, internally hosted foundation models, and domain-specific language models. The proposed architecture deliberately abstracts model providers behind standardized orchestration services, allowing organizations to replace or extend LLM technologies without redesigning enterprise applications. Such flexibility is particularly valuable as model capabilities, regulatory expectations, and deployment strategies continue to evolve.

The practical implications of this research extend beyond technology implementation. For enterprise architects, the framework provides a reusable reference architecture that can guide strategic planning, solution design, and enterprise AI modernization initiatives. By separating orchestration, integration, governance, and knowledge management into modular architectural layers, enterprise architects can reduce technical complexity while improving long-term maintainability and interoperability across business domains.

For technology leaders, including Chief Information Officers (CIOs), Chief Technology Officers (CTOs), Chief Data Officers (CDOs), and Chief Information Security Officers (CISOs), the framework offers practical guidance for balancing innovation with enterprise risk management. Rather than evaluating AI solely through the lens of model performance, the proposed architecture encourages organizations to assess governance maturity, operational resilience, security controls, regulatory readiness, and business integration as equally important success factors. This broader perspective supports more sustainable enterprise AI adoption strategies.

The framework also provides meaningful implications for financial institutions pursuing digital transformation. Banks, wealth management firms, insurance providers, and capital market organizations increasingly seek to enhance employee productivity, improve customer experiences, accelerate knowledge discovery, and automate repetitive business processes through generative AI. The proposed architecture demonstrates how these objectives can be achieved while maintaining enterprise-grade security, governance, and compliance. Furthermore, the layered architecture enables incremental implementation, allowing organizations to introduce low-risk internal use cases before expanding toward more sophisticated enterprise-wide deployments.

Despite these contributions, several limitations should be acknowledged. First, the proposed framework represents a **conceptual enterprise architecture** developed using a Design Science Research methodology rather than an empirical evaluation based on operational deployment data. Although the representative case study demonstrates architectural feasibility, future research should validate the framework through multiple real-world implementations across different financial institutions and technology environments.

Second, the rapid evolution of foundation models, AI agents, multimodal systems, and enterprise AI platforms may require periodic refinement of architectural components. While the proposed framework is intentionally technology-neutral, future advancements in model architectures, autonomous AI systems, and regulatory requirements may introduce additional considerations that extend beyond the scope of this study.

Finally, although the framework integrates recognized enterprise security, governance, and compliance principles, implementation outcomes will inevitably depend on organizational maturity, data quality, enterprise architecture capabilities, and institutional governance practices. Consequently, the framework should be viewed as a reference architecture that supports informed enterprise decision-making rather than a prescriptive implementation methodology.

Overall, this research demonstrates that successful enterprise adoption of Large Language Models requires substantially more than selecting high-performing foundation models. Sustainable deployment within regulated financial environments depends upon the integration of enterprise architecture, AI governance, cybersecurity, knowledge management, regulatory compliance, and continuous operational monitoring into a unified architectural strategy. By providing a comprehensive and technology-neutral reference

architecture, this study contributes both to academic research and to enterprise practice, offering a practical foundation for secure, scalable, and responsible deployment of generative AI across modern financial institutions.

10. Threats to Validity

The study's validity is constrained by its design-science and scenario-based character. Construct validity may be affected by the use of broad concepts such as trustworthiness, groundedness, and governance, which can be measured differently across institutions. This risk is reduced by defining representative measures and mapping requirements to specific controls and evidence.

10.1. Internal Validity

Because the framework is not evaluated through a controlled production deployment, causal claims regarding productivity, cost, risk reduction, or customer outcomes cannot be made. The case study demonstrates logical feasibility and control coverage, not measured effectiveness.

10.2. External Validity

The representative Tier-1 institution reflects common global banking characteristics, but smaller institutions, fintech firms, insurers, and institutions operating under different regulatory regimes may require simplified or additional components. Multi-institution validation is needed before generalizing implementation outcomes.

10.3. Conclusion Validity

The qualitative assessment may be influenced by the selected quality attributes and threat scenarios. Future studies should use independent evaluators, predefined scoring rubrics, quantitative benchmarks, and sensitivity analysis to test whether conclusions remain stable under different assumptions.

10.4. Temporal Validity

LLM capabilities, attack techniques, vendor offerings, and regulatory expectations evolve quickly. The framework is intentionally vendor-neutral, but specific controls and benchmark thresholds require periodic reassessment. The literature cutoff and framework version should therefore be stated whenever the artifact is applied.

11. Conclusion

This research presented the Secure Financial-Services Large Language Model Enterprise Architecture (SF-LLM-EA), a comprehensive enterprise architecture framework designed to support the secure, scalable, and governed deployment of Large Language Models (LLMs) within regulated financial institutions. Recognizing that successful enterprise AI adoption extends beyond model selection, the proposed framework integrates six complementary architectural layers—Experience and Channels, AI Orchestration and Policy, LLM Services, Knowledge and Retrieval-Augmented Generation (RAG), Enterprise Integration, and Platform and Deployment—supported by Security, Governance, Compliance, and Observability—into a unified reference architecture that aligns business objectives with technology, security, and regulatory requirements.

The study contributes to enterprise architecture and financial technology research by providing a vendor-neutral, technology-agnostic framework that addresses critical challenges associated with enterprise LLM adoption, including identity and access management, Zero Trust security, data privacy, model governance, prompt security, auditability, regulatory compliance, and responsible AI. Unlike model-centric approaches, the proposed framework positions generative AI as an enterprise capability embedded within existing business processes, governance structures, and operational controls, thereby enabling organizations to integrate AI solutions while maintaining resilience, transparency, and accountability.

A representative enterprise financial institution case study demonstrated how the framework can support knowledge management, investment research, advisor productivity, customer service, and regulatory assistance through secure orchestration, entitlement-aware RAG, controlled enterprise integration, and human oversight. The qualitative evaluation further indicated that the framework satisfies key enterprise architecture quality attributes—including security, scalability, interoperability, governance, maintainability, reliability, and responsible AI—making it suitable as a reusable reference architecture for financial institutions pursuing enterprise-wide AI modernization.

12. Future Work

Future directions for secure, governed, and measurable enterprise LLM adoption in financial services.

The proposed Secure Financial-Services Large Language Model Enterprise Architecture (SF-LLM-EA) establishes a foundation for controlled enterprise adoption of generative artificial intelligence; however, the technology and regulatory environment will continue to evolve. Future research should therefore extend the framework beyond conversational assistance toward adaptive, multimodal, distributed, and increasingly autonomous enterprise AI systems. These extensions should preserve the framework's central principles of least privilege, evidence-based generation, human accountability, policy-driven orchestration, and continuous evaluation. The following research agenda identifies six priority areas.

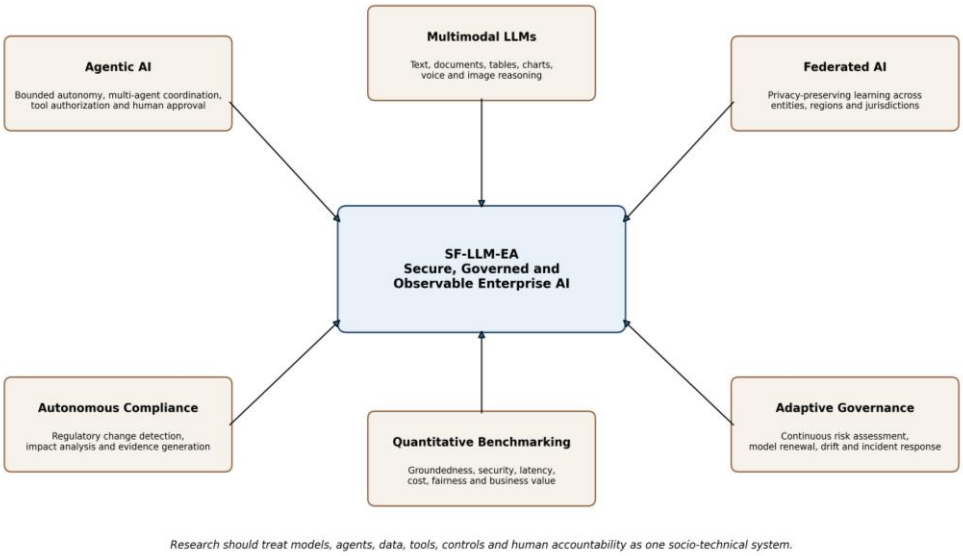

Figure 4. Future Research Roadmap for Extending SF-LLM-EA.

12.1. Agentic AI and Controlled Multi-Agent Systems

Future work should investigate how the framework can support agentic AI systems that plan, reason, interact with tools, and coordinate multiple specialized agents. Early agent benchmarks show that autonomous LLM systems remain vulnerable to failures in long-horizon reasoning, instruction following, and decision-making [1]. In financial institutions, these limitations are especially important because agents may interact with research repositories, CRM platforms, workflow systems, payment services, or compliance tools. Research should define bounded autonomy models in which agents receive task-specific identities, least-privilege tool permissions, execution budgets, transaction limits, and explicit termination conditions. Multi-agent collaboration should be evaluated for role separation, conflict resolution, shared memory, provenance, and resistance to cascading errors. Human approval should remain mandatory for client communication, investment recommendations, credit decisions, payments, trades, regulatory interpretation, and changes to systems of record.

12.2. Multimodal Llms for Financial Documents and Interactions

Multimodal LLMs can combine text with tables, charts, scanned documents, images, audio, and other data modalities [2,3]. This creates opportunities for integrated analysis of annual reports, investment presentations, account statements, contracts, handwritten forms, call-center recordings, and market visualizations. Future research should extend the Knowledge and RAG Layer with multimodal ingestion, modality-specific embeddings, cross-modal retrieval, document-layout understanding, and evidence linking between generated statements and their original visual or audio sources. Evaluation must address optical and speech-recognition errors, conflicting evidence across modalities, accessibility, privacy of recorded conversations, and the risk that visually plausible but incorrect interpretations may be accepted without verification.

12.3. Federated and Privacy-Preserving Enterprise AI

Large financial institutions frequently operate across subsidiaries, legal entities, and jurisdictions that cannot freely centralize sensitive information. Federated learning and federated instruction tuning offer a potential path for improving models across distributed

environments while keeping source data local. Prior federated-learning research has demonstrated that decentralized training and secure aggregation can improve shared models while keeping source data distributed [4]. Future work should investigate federated domain adaptation, secure aggregation, differential privacy, confidential computing, data-poisoning defenses, cross-border governance, and local model personalization. The architecture should also distinguish between federated model learning and federated retrieval, because sharing model updates may reduce raw-data movement but does not automatically resolve legal, privacy, intellectual-property, or model-inversion risks.

12.4. Autonomous Compliance and Regulatory Intelligence

A significant future direction is the development of compliance agents that continuously monitor regulatory publications, identify changes, map obligations to policies and systems, generate impact assessments, and assemble audit evidence. These capabilities could reduce manual monitoring and improve the traceability of regulatory change management. Nevertheless, autonomous compliance should be treated as assisted interpretation rather than legal authority. Research should establish authoritative source hierarchies, jurisdiction-aware retrieval, effective-date management, conflict detection, legal-review workflows, and immutable evidence trails. Benchmarks such as LegalBench illustrate the value of domain-specific evaluation and also demonstrate that legal reasoning comprises multiple distinct capabilities rather than one generalized accuracy score [5].

12.5. Quantitative Benchmarking and Empirical Validation

The current framework is evaluated primarily through design-science reasoning and representative scenarios. Future studies should conduct empirical evaluation across multiple financial institutions and use cases. A suitable benchmark should combine model-level, system-level, security, compliance, and business metrics. HELM demonstrated the importance of multi-metric evaluation across accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency [6]. FinanceBench further showed that financial question answering requires evidence-based evaluation and that strong general-purpose models can still perform poorly on realistic financial documents [7]. Future evaluation should measure retrieval precision and recall, groundedness, citation correctness, hallucination rate, entitlement-filter accuracy, prompt-attack success, sensitive-data leakage, latency, availability, cost per completed task, human override rate, workflow completion, and measurable productivity or service-quality improvement. Results should be segmented by risk tier, user population, jurisdiction, model version, and document type.

12.6. Adaptive Governance and Continuous Assurance

Traditional governance processes rely heavily on periodic reviews, while enterprise LLM systems can change whenever a model, prompt, tool, policy, retrieval corpus, or provider configuration changes. Future work should therefore develop continuous assurance mechanisms that automatically detect material changes, trigger targeted regression tests, update risk assessments, and suspend deployments when controls fall below approved thresholds. The NIST AI Risk Management Framework organizes risk activities around Govern, Map, Measure, and Manage, providing a useful foundation for such adaptive governance [8]. Future research should translate these functions into machine-readable control policies, continuous control monitoring, model and prompt attestations, automated evidence collection, and risk-based renewal workflows. This would enable governance to operate at the speed of AI delivery without reducing human accountability.

12.7. Prioritized Research Agenda

Table 6. Comparative Analysis of Research Directions, Key Measures, Risks, and Expected Contributions in AI Governance

Research Direction	Core Research Question	Key Measures	Primary Risks	Expected Contribution
Agentic AI	How much autonomy can be granted without weakening accountability?	Task success, unsafe-action rate, approval rate	Excessive agency, tool misuse, cascading error	Bounded-agent reference architecture
Multimodal LLMs	How can mixed financial evidence be retrieved and verified?	Cross-modal accuracy, citation fidelity, privacy	Misinterpretation, OCR/audio errors	Multimodal RAG and provenance controls

Federated AI	Can institutions collaborate without centralizing sensitive data?	Utility, privacy loss, communication cost	Poisoning, inversion, jurisdictional conflict	Federated governance and deployment model
Autonomous compliance	Can AI accelerate regulatory change while preserving legal accountability?	Coverage, traceability, reviewer acceptance	Incorrect interpretation, stale regulations	Human-governed compliance-agent framework
Benchmarking	Which metrics predict production trustworthiness and value?	Groundedness, leakage, latency, cost, outcomes	Benchmark overfitting, weak external validity	Financial-services evaluation suite
Adaptive governance	How can controls continuously respond to model and data change?	Control coverage, detection time, remediation time	Silent drift, incomplete evidence	Continuous assurance architecture

12.8. Summary

Future development of enterprise LLM systems should not be evaluated solely by increases in model capability. The more consequential research challenge is to determine how advanced capabilities can be introduced while preserving authorization, evidence, privacy, resilience, regulatory compliance, and accountable human control. Agentic AI, multimodal reasoning, federated learning, autonomous compliance, quantitative benchmarking, and adaptive governance represent complementary extensions of SF-LLM-EA. Together, they provide a research pathway from secure conversational assistance toward trustworthy enterprise intelligence capable of operating across complex financial workflows without converting probabilistic model output into unaccountable institutional action.

Statements and Declarations

Funding

The author received no specific funding for this work.

Competing Interests

The author declares no competing financial or non-financial interests directly related to this manuscript.

Data Availability

No new empirical dataset was generated or analyzed. The case study is illustrative and is based exclusively on publicly describable enterprise patterns and the cited literature.

Ethics Approval and Consent to Participate

Not applicable. The study did not involve human participants, personal data, or animals.

Author Contributions

Saad Khan conceived the study, developed the framework, performed the literature synthesis and architecture analysis, and prepared and reviewed the manuscript.

Use of Generative Artificial Intelligence

Generative AI tools were used to assist with drafting support, language refinement, structural editing, and formatting. The author reviewed, revised, fact-checked, and accepts full responsibility for the accuracy, originality, citations, analysis, and conclusions of the final manuscript.

Acknowledgements

The author acknowledges professional experiences in enterprise architecture and financial-services technology that informed the problem context. No confidential information, proprietary architecture, internal metrics, or unpublished client data from any employer were used.

References

[1] Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., et al.: AgentBench: Evaluating LLMs as agents. arXiv:2308.03688 (2023).
[2] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. arXiv:2306.13549 (2023).
[3] Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of the International Conference on Machine Learning, 19730-19742 (2023).
[4] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for privacy-preserving machine learning. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 1175-1191 (2017).
[5] Guha, N., Nyarko, J., Ho, D.E., Ré, C., Chilton, A., Narayana, A., et al.: LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in Neural Information Processing Systems 36 (2023).
[6] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., et al.: Holistic evaluation of language models. Transactions on Machine Learning Research (2023).
[7] Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., Vidgen, B.: FinanceBench: A new benchmark for financial question answering. arXiv:2311.11944 (2023).
[8] Tabassi, E.: Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology (2023). https://doi.org/10.6028/NIST.AI.100-1
[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems 30 (2017).
[10] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems 33, 1877-1901 (2020).
[11] Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al.: On the opportunities and risks of foundation models. arXiv:2108.07258 (2021).
[12] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems 33, 9459-9474 (2020).
[13] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.-t.: Dense passage retrieval for open-domain question answering. Proceedings of EMNLP, 6769-6781 (2020).
[14] Izacard, G., Grave, E.: Leveraging passage retrieval with generative models for open-domain question answering. Proceedings of EACL, 874-880 (2021).
[15] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., et al.: Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997 (2023).
[16] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations (2023).
[17] Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., et al.: Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems 36 (2023).
[18] Araci, D.: FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063 (2019).
[19] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., et al.: BloombergGPT: A large language model for finance. arXiv:2303.17564 (2023).
[20] Yang, H., Liu, X.-Y., Wang, C.D.: FinGPT: Open-source financial large language models. arXiv:2306.06031 (2023).
[21] Rose, S., Borchert, O., Mitchell, S., Connelly, S.: Zero Trust Architecture. NIST Special Publication 800-207. National Institute of Standards and Technology (2020). https://doi.org/10.6028/NIST.SP.800-207
[22] Souppaya, M., Scarfone, K., Dodson, D.: Secure Software Development Framework (SSDF), Version 1.1. NIST Special Publication 800-218. National Institute of Standards and Technology (2022). https://doi.org/10.6028/NIST.SP.800-218
[23] OWASP Foundation: OWASP Top 10 for Large Language Model Applications, Version 1.0 (2023).
[24] International Organization for Standardization: ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection - Information security management systems - Requirements. ISO, Geneva (2022).
[25] International Organization for Standardization: ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system. ISO, Geneva (2023).
[26] European Parliament and Council: Regulation (EU) 2016/679, General Data Protection Regulation. Official Journal of the European Union (2016).
[27] United States Congress: Gramm-Leach-Bliley Act, Pub. L. 106-102, 113 Stat. 1338 (1999).
[28] PCI Security Standards Council: Payment Card Industry Data Security Standard, Version 4.0 (2022).
[29] Financial Stability Board: Artificial intelligence and machine learning in financial services: Market developments and financial stability implications. FSB (2017).
[30] Basel Committee on Banking Supervision: Principles for operational resilience. Bank for International Settlements (2021).
[31] Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Quarterly 28(1), 75-105 (2004).
[32] Peffers, K., Tuunanen, T., Rothenberger, M.A., Chatterjee, S.: A design science research methodology for information systems research. Journal of Management Information Systems 24(3), 45-77 (2007).
[33] Gregor, S., Hevner, A.R.: Positioning and presenting design science research for maximum impact. MIS Quarterly 37(2), 337-355 (2013).

[34] International Organization for Standardization: ISO/IEC 25010:2011 Systems and software engineering - Systems and software quality requirements and evaluation - System and software quality models. ISO, Geneva (2011).
[35] Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., et al.: Ethical and social risks of harm from language models. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 214-229 (2022).
[36] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 610-623 (2021).
[37] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., et al.: Model cards for model reporting. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 220-229 (2019).
[38] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K.: Datasheets for datasets. Communications of the ACM 64(12), 86-92 (2021).
[39] Ribeiro, M.T., Singh, S., Guestrin, C.: 'Why should I trust you?': Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD Conference, 1135-1144 (2016).
[40] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30 (2017).
[41] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., et al.: Extracting training data from large language models. 30th USENIX Security Symposium, 2633-2650 (2021).
[42] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M.: Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. Proceedings of the ACM Workshop on Artificial Intelligence and Security, 79-90 (2023).
[43] Zou, A., Wang, Z., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models. arXiv:2307.15043 (2023).
[44] Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., et al.: Red teaming language models with language models. Proceedings of EMNLP, 3419-3448 (2022).
[45] Manakul, P., Liusie, A., Gales, M.J.F.: SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. Proceedings of EMNLP, 9004-9017 (2023).
[46] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P.W., Iyyer, M., Zettlemoyer, L., Hajishirzi, H.: FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. Proceedings of EMNLP, 12076-12100 (2023).
[47] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al.: Survey of hallucination in natural language generation. ACM Computing Surveys 55(12), 1-38 (2023).
[48] Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., et al.: Augmented language models: A survey. Transactions on Machine Learning Research (2023).
[49] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (2022).
[50] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems 35, 27730-27744 (2022).
[51] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., et al.: Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073 (2022).
[52] OpenAI: GPT-4 technical report. arXiv:2303.08774 (2023).
[53] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288 (2023).
[54] McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. Proceedings of AISTATS, 1273-1282 (2017).
[55] Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., et al.: Advances and open problems in federated learning. Foundations and Trends in Machine Learning 14(1-2), 1-210 (2021).
[56] European Commission High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI. European Commission (2019).
[57] Organisation for Economic Co-operation and Development: Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449 (2019).
[58] National Institute of Standards and Technology: NIST Privacy Framework: A tool for improving privacy through enterprise risk management, Version 1.0 (2020).
[59] National Institute of Standards and Technology: Cybersecurity Framework, Version 1.1. NIST (2018).
[60] Financial Stability Board: Enhancing third-party risk management and oversight: A toolkit for financial institutions and financial authorities. FSB (2023).