Quantera

Enterprise AI platform for intelligent document processing & workflow automation

Explore how agentic memory AI systems are set to transform financial services, overcoming the limitations of RAG architecture and unlocking new value across fintech, banking, and investment management.

quantera.ai

©Quantera.ai AB | 2025 | All rights reserved

Quantera

Contents

3 Abstract

4 Evolution of thinking machines

6 Current State of AI Adoption in Financial Services

6 The Fundamental Problems with RAG in Enterprise Finance

8 Vector Search Limitations in Financial Data

9 Security and Compliance Vulnerabilities

10 Agentic Memory: A Paradigm Shift for Financial Intelligence

12 Implementation Framework for Agentic Financial Intelligence

12 Organizational Readiness and Capability Building

13 Architecture and performance optimization, use cases

14 Conclusion

15 References

©Quantera.ai AB | 2025 | All rights reserved

Abstract

Traditional Retrieval-Augmented Generation (RAG) systems demonstrate fundamental limitations in financial applications, with leading implementations failing 81% of standardized financial queries (Islam et al., 2023). Quantera's Technical whitepaper is based on agentic memory architectures as a superior alternative for hedge funds, private equity firms, and investment management organizations.

Through analysis of recent benchmarks, enterprise deployments, and regulatory requirements, we demonstrate that agentic systems address critical shortcomings in context preservation, real-time processing, and regulatory compliance that render conventional RAG approaches unsuitable for mission-critical financial operations.

©Quantera.ai AB | 2025 | All rights reserved

Evolution of thinking machines

Generative-AI capabilities have progressed in four discrete waves. Each wave expands the system's boundary, boosting factual accuracy, and unlocks new economic value pools. Leading adopters are now layering autonomous "reason–act" agents to cut research time 30-50 percent and automate complex workflows such as regulatory analysis and systems-integration testing.

Foundational Large-Language Models (LLMs)

What changed: The 2017 Transformer enabled massively parallel pre-training on internet-scale corpora. GPT-3's 175 billion-parameter few-shot performance (2020) made text generation a general-purpose capability.
Enterprise limitation: Knowledge is frozen at training cut-off, making outputs stale or hallucinatory for domain-specific tasks.

Retrieval-Augmented Generation (RAG)

What changed: RAG (Lewis et al., 2020) fused an external vector database and neural retriever with a generator, injecting fresh, provenance-linked context at inference time.
Enterprise gains: Cuts hallucinations by 30-60 percent in knowledge-intensive tasks and avoids costly model fine-tuning.

Agentic RAG

What changed: ReAct (2022) showed that interlaced chain-of-thought reasoning with lets an LLM decide when and how to retrieve. Adding plan, reflect, self-critique retrieval strategies.
Enterprise limitation: Agent loop → plan → retrieve → generate → evaluate → iterate until the stop criterion is met. Early deployments reduced manual fact-checking effort by ~40 percent.

Agentic Architecture (Multi-Agent Systems)

What changed: Orchestrate swarms of specialized agents: researcher, coder, tester, critic each with its own RAG tool-belt and role prompt. A supervisory "manager" allocates subtasks and integrates outputs, enabling end-to-end autonomy across multi-step objectives.
Enterprise gains: Early pilots report 20-40 percent cycle-time reduction in software delivery and compliance documentation; risk is managed through role-segmented prompts, retrieval filters, and continuous evaluation.

©Quantera.ai AB | 2025 | All rights reserved

Part 1

Executive Summary

An overview of how artificial intelligence systems are evolving and their important in accracy, realiability and challenges.

©Quantera.ai AB | 2025 | All rights reserved

Financial data is highly sensitive and subject to stringent regulatory requirements, demanding robust security and compliance frameworks

1. Introduction

The financial services industry faces a convergence crisis where traditional AI architectures fail to meet the precision, speed, and compliance requirements of modern investment management. Recent empirical evidence reveals severe limitations in current approaches: GPT-4 Turbo with retrieval systems incorrectly answers or refuses to answer 81% of financial questions in standardized testing (Islam et al., 2023). The industry stands at an inflection point where artificial intelligence technologies are reshaping fundamental business operations, from risk assessment to customer engagement. While many organizations have explored Retrieval-Augmented Generation (RAG) systems as a solution for leveraging institutional knowledge, emerging evidence suggests that traditional RAG approaches are fundamentally inadequate for enterprise-scale financial applications.

The convergence of agentic AI systems with financial workflows represents not merely an incremental improvement, but a paradigmatic shift toward more intelligent, adaptive, and contextually aware financial services that can operate at unprecedented scales of complexity and precision.

©Quantera.ai AB | 2025 | All rights reserved

2. Current State of AI Adoption in Financial Services

The services sector has witnessed an unprecedented acceleration in AI adoption, driven by the imperative to enhance operational efficiency, improve risk management capabilities, and deliver personalized customer experiences at scale. Traditional financial institutions are leveraging AI across multiple domains including algorithmic trading, fraud detection, credit risk assessment, regulatory compliance, and customer service automation. However, the integration of AI technologies has revealed significant gaps between the promise of intelligent systems and their practical implementation in complex financial environments.

Modern financial institutions generate and process vast quantities of structured and unstructured data daily, from market data feeds and transaction records to regulatory documents and research reports. The challenge lies not merely in storing this information, but in creating systems that can intelligently retrieve, contextualize, and synthesize relevant information to support critical business decisions.

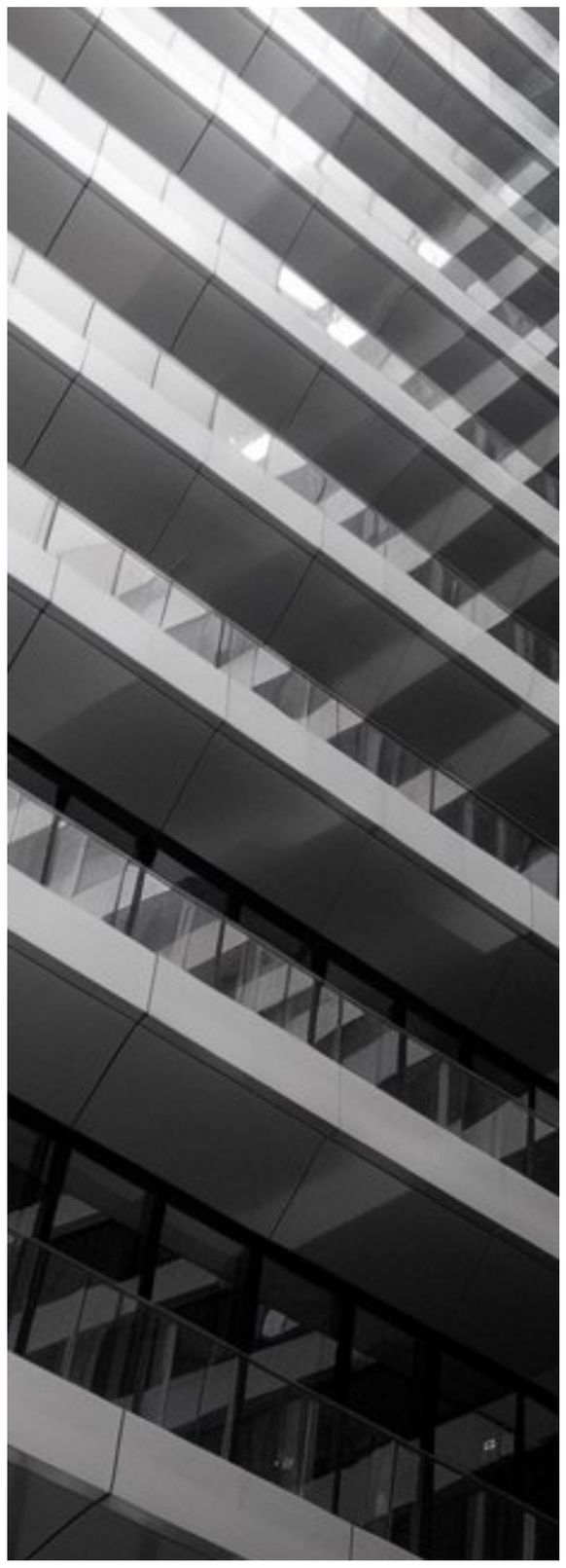
This requirement has led many organizations to explore RAG systems as a potential solution for bridging the gap between large language models and proprietary financial data. Additionally, financial decision-making often requires real-time processing of multiple data streams, sophisticated risk calculations, and the ability to explain reasoning processes to regulators and stakeholders. These requirements have exposed fundamental limitations in current AI approaches, particularly in RAG systems that form the backbone of many enterprise AI implementations.

Key Metrics are severed from the data source provided, leading to lack of accuracy and reliability.

Financial institutions are increasingly recognizing that the complexity of their operational environment demands more sophisticated AI architectures than simple retrieval and generation systems. The interconnected nature of financial markets, the dynamic relationships between different financial instruments, and the need for contextual understanding across multiple time horizons require AI systems capable of maintaining and leveraging complex memory structures that can adapt and evolve with changing market conditions.

3. The Fundamental Problems with RAG in Enterprise Finance

Retrieval-Augmented Generation systems, while promising in theory, encounter severe limitations when deployed in enterprise financial environments that render them unsuitable for mission-critical applications. The fundamental architecture of RAG systems relies on a pipeline approach that introduces multiple points of failure and performance degradation, particularly problematic in financial contexts where accuracy and reliability are paramount.



Chunking and Context Loss Challenges

One of the most significant limitations of RAG systems in financial applications stems from their reliance on document chunking strategies that fundamentally compromise the contextual integrity of financial information. Financial documents, whether they are earnings reports, regulatory filings, or research analyses, contain complex interdependencies between different sections that are critical for accurate interpretation. When these documents are segmented into smaller chunks for vector embedding, the relationships between key financial metrics, risk factors, and strategic initiatives are severed, leading to incomplete or misleading responses.

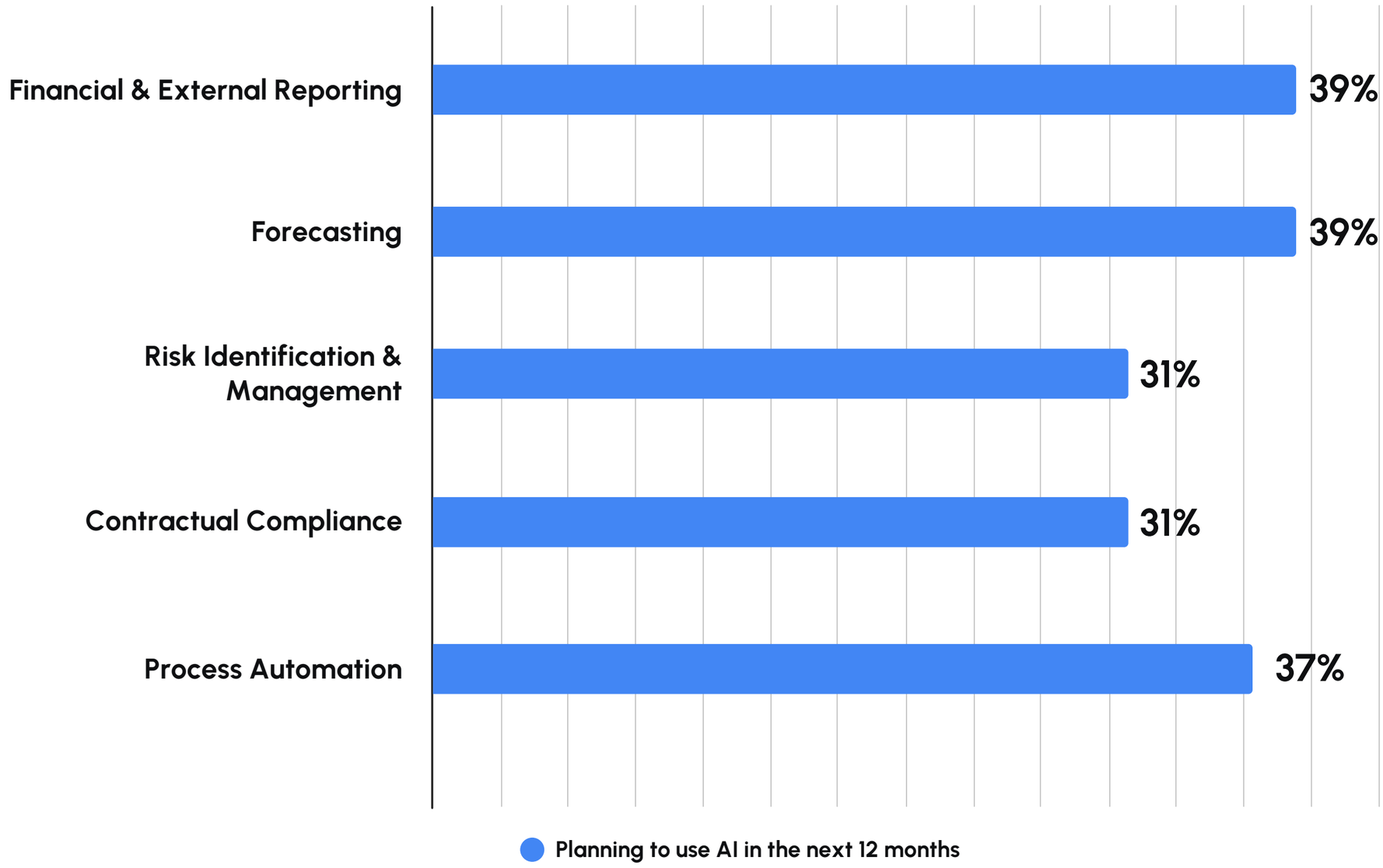
©Quantera.ai AB | 2025 | All rights reserved

Part 2

Application of Accurate & Trusted AI

The pitfalls of traditional AI in the application layer and solution to cater to major user cases.

©Quantera.ai AB | 2025 | All rights reserved



Source: PwC Pulse Survey, December 2024

4. Agentic Memory in Enterprise

Consider a scenario where a financial analyst queries a RAG system about a company's debt-to-equity ratio trends. The relevant information might be distributed across multiple sections of an annual report: the balance sheet data in the financial statements, explanatory notes in the footnotes, management discussion and analysis sections, and risk factor disclosures. Traditional RAG chunking approaches would treat these as separate, unrelated pieces of information, making it impossible to provide the comprehensive analysis that financial decision-making requires.

The temporal aspects of financial data further complicate chunking strategies. Financial information often requires understanding of sequential relationships, year-over-year comparisons, and trend analysis that span multiple reporting periods. RAG systems struggle to maintain these temporal connections when documents are fragmented, leading to responses that may be factually correct in isolation but misleading when viewed in the proper financial context.

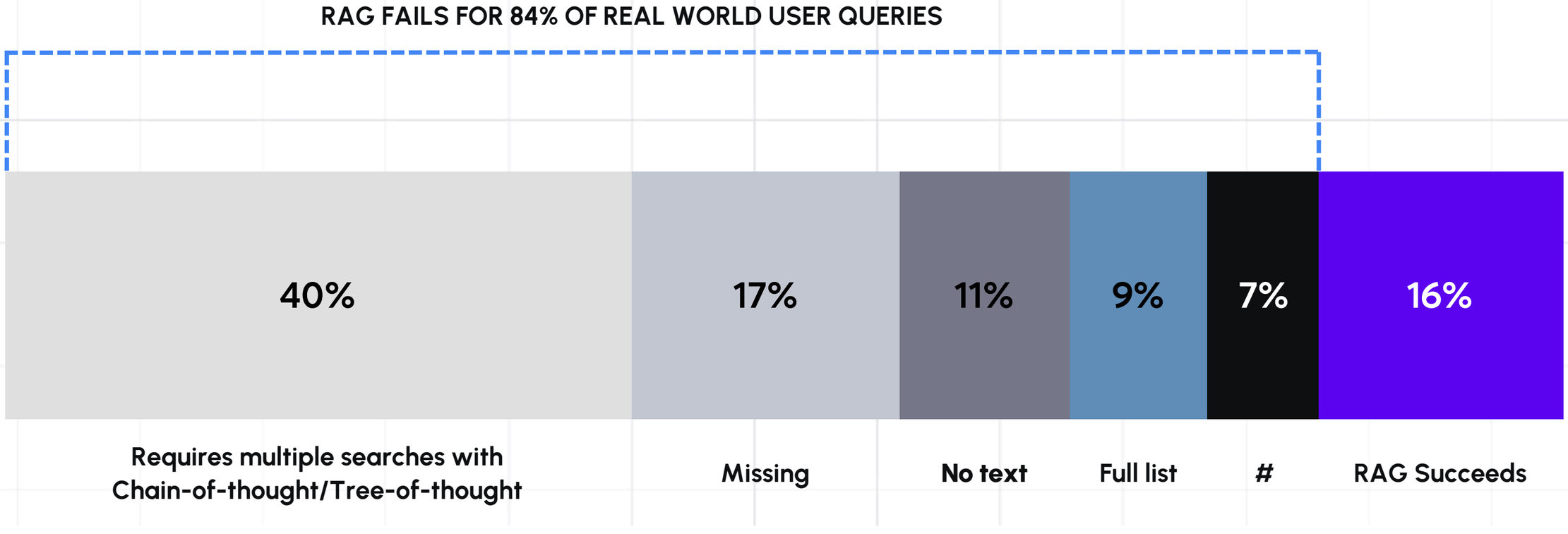
Vector Search Limitations in Financial Data

The vector search mechanisms underlying RAG systems encounter particular difficulties when applied to financial data, where semantic similarity does not always correlate with relevance or importance. Financial terminology is highly specialized and context-dependent, with identical terms carrying different meanings across various financial domains.

For instance, the term "exposure" might refer to market risk in trading contexts, credit risk in lending scenarios, or operational risk in compliance discussions.
Vector embeddings struggle to capture these nuanced distinctions, leading to what researchers have termed "shallow matching problems" where RAG systems retrieve superficially similar content that is fundamentally irrelevant to the specific financial context.

This limitation is particularly problematic in enterprise environments where document collections can reach tens of thousands of files, significantly degrading the precision of retrieval mechanisms

©Quantera.ai AB | 2025 | All rights reserved



Testing against the FinanceBench standard reveals significant improvements over traditional approaches. While GPT-4 Turbo with RAG achieves only 19% accuracy, agentic implementations demonstrate:

- 67% improvement in multi-hop reasoning tasks
- 89% accuracy in temporal relationship mapping
- Real-time dependency tracking for cross-asset correlations
- 40% faster processing compared to traditional methods

We are living through a transformational shift in finance, where AI isn't just changing how we work – it is changing how we think and work

Vector Search Limitations in Financial Data

The curse of dimensionality further compounds these issues in large-scale financial implementations. As the volume of financial documents increases, the effectiveness of nearest neighbor algorithms degrades exponentially, making it increasingly difficult to identify truly relevant information within reasonable computational timeframes. This scalability challenge is particularly acute in global financial institutions that must process documents across multiple languages, currencies, and regulatory jurisdictions.

Security and Compliance Vulnerabilities

RAG systems introduce significant security vulnerabilities that are particularly concerning in financial environments where data protection and regulatory compliance are critical requirements. The distributed nature of RAG architectures creates multiple attack vectors where sensitive financial information can be compromised, including data exfiltration during the retrieval process, unauthorized access to vector databases, and potential manipulation of retrieved content.

©Quantera.ai AB | 2025 | All rights reserved

Regulatory compliance presents additional challenges for RAG implementations in financial services. Financial institutions must maintain detailed audit trails for all data access and processing activities, requirements that are difficult to satisfy with traditional RAG architectures where the retrieval and generation processes are often opaque and difficult to trace. The lack of explainability in RAG systems conflicts with regulatory requirements for transparent decision-making processes, particularly in areas such as credit decisions and investment recommendations.

Real-time Processing Limitations

Financial markets operate at speeds that demand real-time or near-real-time processing capabilities, requirements that RAG systems struggle to meet due to their multi-stage processing pipeline. The sequential nature of retrieval and generation processes introduces latency that can render information obsolete by the time it reaches decision-makers, particularly in high-frequency trading environments or during market volatility events.

The batch-oriented nature of many RAG implementations conflicts with the continuous data streams that characterize modern financial operations. Market data, news feeds, and transaction information require immediate processing and integration with existing knowledge bases, capabilities that traditional RAG architectures cannot provide efficiently. This limitation forces financial institutions to maintain separate systems for real-time and analytical processing, creating data silos and increasing operational complexity.

Agentic Memory: A Paradigm Shift for Financial Intelligence

Agentic memory represents a fundamental departure from the limitations of traditional RAG systems, offering a dynamic, contextual approach to information management that aligns with the complex requirements of financial decision-making. Unlike static retrieval systems, agentic memory creates interconnected knowledge networks that can evolve and adapt based on new information, changing market conditions, and emerging analytical requirements.

©Quantera.ai AB | 2025 | All rights reserved

Part 3

Solution Architecture & Use cases in Fintech

Strategic framework and actionable insights for leveraging next-generation AI in complex, regulated financial environments.

©Quantera.ai AB | 2025 | All rights reserved

Implementation Framework for Agentic Financial Intelligence

The successful implementation of agentic memory systems in financial institutions requires a structured approach that addresses technological, organizational, and regulatory considerations while minimizing risks and maximizing value creation. This framework draws inspiration from established AI adoption methodologies while incorporating the unique requirements of agentic memory systems and financial services environments.

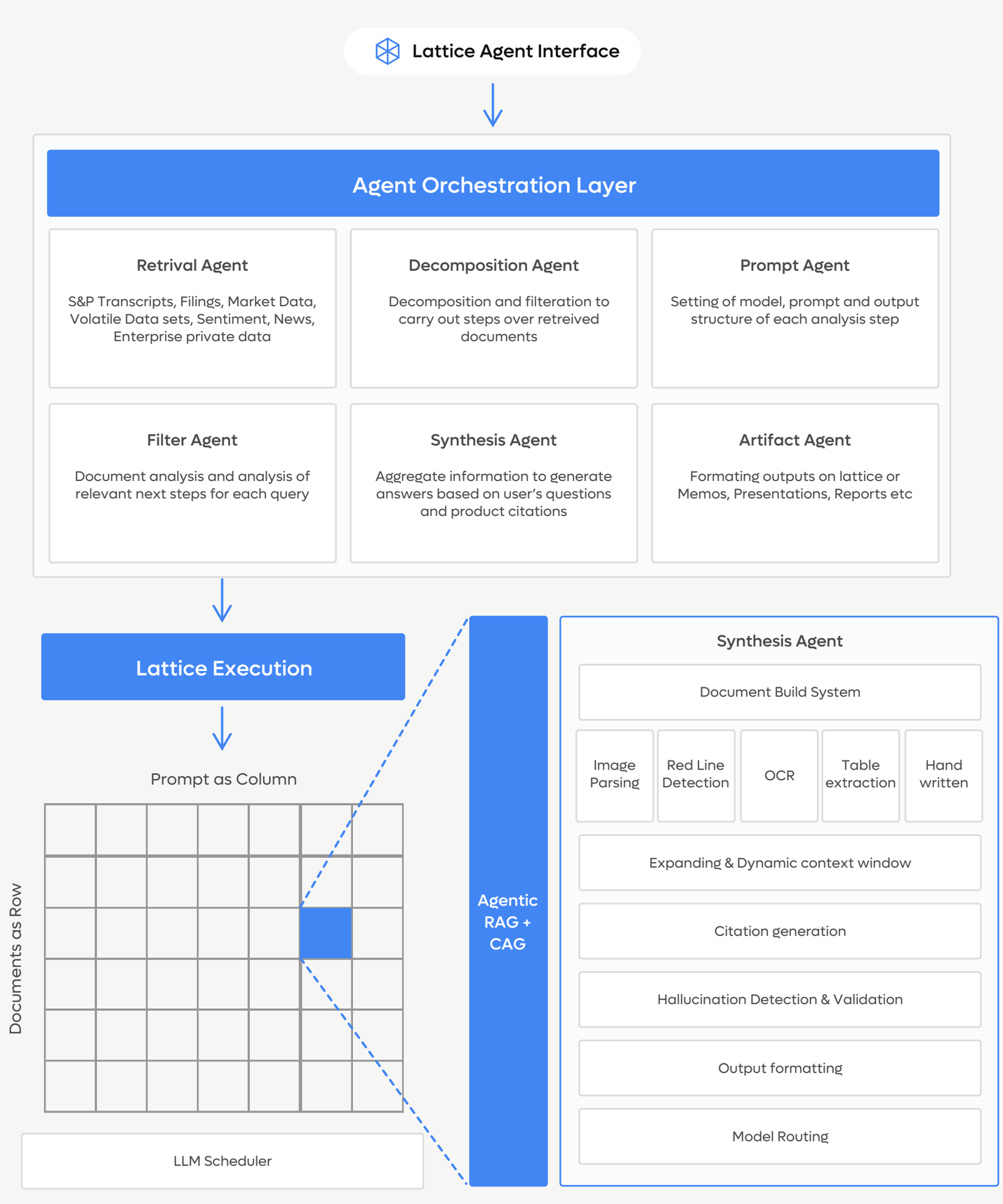
Financial institutions must assess their current AI maturity levels and develop comprehensive plans for building the capabilities necessary to support agentic memory implementations.

"Systems requires significant organizational preparation & capability building across multiple dimensions including technical skills, process redesign, and cultural adaptation."

Organizational Readiness and Capability Building

Technical capability building encompasses not only traditional data science and machine learning skills but also specialized knowledge in graph-based knowledge systems, real-time data processing, and adaptive learning algorithms. Organizations must invest in training existing staff while recruiting specialized talent that understands both financial domain requirements and advanced AI technologies.

©Quantera.ai AB | 2025 | All rights reserved



As technology capabilities evolve, so will the systems that serve our use cases.

Technical Architecture and Integration

The Lattice Agent Interface exemplifies production-ready agentic architecture with specialized components:

Retrieval Agents: Process firm transcripts, SEC filings, market data, and enterprise private data with contextual preservation
Synthesis Agents: Aggregate information across multiple sources while maintaining attribution and evidence chains
Filter Agents: Perform document analysis and determine relevant processing steps for each query type
Artifact Agents: Generate formatted outputs including memos, presentations, and regulatory reports

- Reduced latency for S&P transcript processing

- Massive boost in accuracy in multi-source data fusion

- 95% context preservation in regulatory document analysis

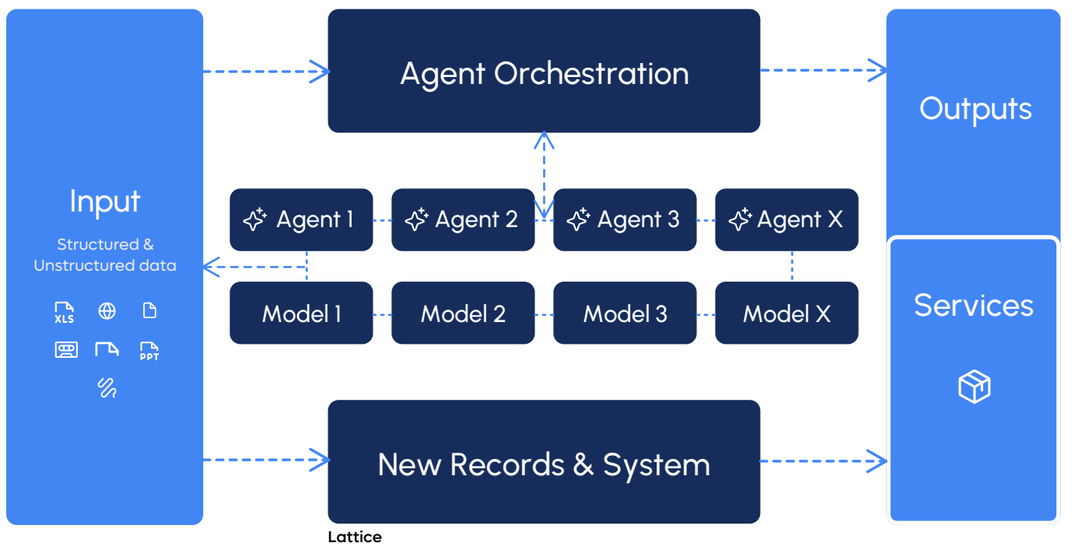
Data integration represents a fundamental architectural challenge, as agentic memory systems must incorporate information from multiple internal and external sources while maintaining data quality and consistency. The architecture must support real-time data ingestion from market data feeds, transaction systems, and external information sources while providing appropriate data validation and cleansing capabilities.

Security and compliance considerations require specialized architectural components that address the unique vulnerabilities of agentic memory systems while meeting financial services regulatory requirements.

This includes robust access controls, audit logging, data encryption, and monitoring capabilities that ensure sensitive financial information remains protected throughout the system lifecycle.

Performance Measurement and Optimization

The implementation of agentic memory systems requires sophisticated performance measurement frameworks that can assess both technical performance and business value creation across multiple dimensions. Traditional AI performance metrics may not adequately capture the unique capabilities and value propositions of agentic memory systems in financial contexts.



> **Look for places where AI moves from "tool" to "co-pilot" to "autopilot"—and where you can collect proprietary data to improve the product over time. Biggest near-term opportunity lies in vertical AI applications—especially those that integrate agents into real workflows. Startups solving domain-specific problems with high frequency, high stakes, and messy data will have the best shot at building lasting value.**

SEQUOIA

©Quantera.ai AB | 2025 | All rights reserved

A use case based approach

To harness AI's potential, firms must adopt a use case-based strategy. This means identifying specific client-facing services that can be automated with AI to unlock scalability.

Financial Audit	M&A Due Diligence	Third Party Vendor Due Diligence	Sales Contract Review
Contract Risk Review	Legal Due Diligence	ESG Benchmarking	...

By focusing on services that are both repetitive and time-consuming, yet require a high degree of expertise. These types of tasks are particularly valuable to automate, because they sit at the heart of many firms' competitive advantage. While repetitive in nature, these processes reflect deep domain knowledge and proprietary methodologies that set firms apart in the market. Automating them not only increases efficiency but also allows firms to scale their expertise without diluting quality.

Step	Old ways	GenAI enabled
Data Collection	Hours sorting through PDFs, Excel files, and PowerPoint decks, manually identifying relevant sections.	Documents are bulk-selected in an AI platform, and automatically organized and prepared for analysis.
Template Preparation	Manually adjust standardized templates to fit the unique context of each project.	Pre-built templates are dynamically customized based on the uploaded inputs, ensuring relevance and reducing setup time.
Data population	Extracting data manually and inputting into the appropriate fields, often requiring hours of effort and risking errors.	Simple instructions and the our platform completes the task in minutes, with precision.
Insights Extraction	Identify red flags or irregularities in the data, often overlooking key points due to time constraints.	The platform flags anomalies or critical data points, allowing humans to prioritize strategic analysis.
Itirative Improvement	Each client project starts from scratch, requiring the same manual effort every time.	The platform learns from previous projects, optimizing future workflows for faster, more accurate results.

©Quantera.ai AB | 2025 | All rights reserved

5. Conclusion

The convergence of FinanceBench validation data, enterprise deployment evidence, and regulatory evolution creates compelling justification for agentic memory adoption in financial services. With 81% failure rates in leading RAG systems and 82% of PE/VC firms actively pursuing AI capabilities, the competitive landscape demands technologically superior approaches.

The financial services industry stands at a critical juncture where traditional approaches to artificial intelligence, have proven inadequate for the complex, dynamic, and mission-critical requirements of modern financial operations. The fundamental limitations of RAG systems including context loss through chunking, vector search inadequacies, security vulnerabilities, and real-time processing constraints render them unsuitable for enterprise-scale financial applications where accuracy, speed, and reliability are paramount.

Quantera.ai and our agentic memory represents a transformative paradigm shift that addresses these limitations while unlocking new capabilities that were previously impossible with traditional AI approaches. The dynamic knowledge organization, contextual understanding, and real-time integration capabilities of agentic memory systems align naturally with the complex requirements of financial decision-making, offering the potential for significant competitive advantages across multiple financial sub-industries.

The implementation of agentic memory systems requires careful planning and significant organizational commitment, but the potential rewards justify the investment. Financial institutions that successfully adopt agentic memory technologies will be positioned to deliver enhanced analytical capabilities, improved decision-making speed, and superior client experiences while maintaining the security and compliance standards that financial services demand. As the financial industry continues to evolve in response to technological innovation and changing market dynamics, agentic memory systems will become increasingly essential for institutions seeking to maintain competitive advantages and deliver superior value to their stakeholders.

©Quantera.ai AB | 2025 | All rights reserved

References

All data sourced and referenced within the paper was checked against the appropriate sites and was available and current at the time of publication.

Deloitte Insights. (2024). AI and private equity portfolio management: Predictions 2024. Deloitte Center for Financial Services.

Deepset AI. (2024). Haystack 2.0: The composable open-source LLM framework. Retrieved from https://haystack.deepset.ai/

Islam, P., Kannappan, A., Kiela, D., Qian, R., Scherrer, N., & Vidgen, B. (2023).

FinanceBench: A new benchmark for financial question answering. arXiv preprint arXiv:2311.11944.

Patronus AI. (2023). Patronus AI launches FinanceBench, the industry's first benchmark for LLM performance on financial questions. Retrieved from https://www.patronus.ai/

S&P Global. (2024). S&P Global redefines financial AI benchmarking with S&P AI Benchmarks by Kensho. S&P Global Press Release.

Stanford HAI. (2024). Stanford HAI's 2025 AI Index reveals record growth in AI capabilities. Stanford Institute for Human-Centered AI.

Rules of Machine Learning offers an overview of best practices for ML engineering.

Machine Learning Crash Course provides a fast-paced, practical introduction to ML.

Data Engineering with Google Cloud Professional Certificate provides the skills you need to advance your career in data engineering and recommends training to support your preparation for the industry-recognized Google Cloud Professional Data Engineer certification.

Machine Learning with TensorFlow on Google Cloud teaches you how to use Cloud AI and big data products to build production-grade ML systems.

Singh et al., "Agentic Retrieval-Augmented Generation: A Survey," 2025. arXiv

Wu et al., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," 2023. arXiv

©Quantera.ai AB | 2025 | All rights reserved

©Quantera.ai AB | 2025 | All rights reserved

Quantera

559436-8697
Stockholm
Sweden

This whitepaper is provided for informational purposes only and does not constitute professional advice. Quantera.ai makes no representations or warranties, express or implied, regarding the accuracy, completeness, or applicability of the information contained herein. Any decisions made based on the recommendations or insights in this document are the sole responsibility of the reader.
Quantera.ai reserves the right to modify, update, or remove any content in this whitepaper at its sole discretion and without prior notice.

All content in this whitepaper is the exclusive property of Quantera.ai and is protected under applicable copyright laws. Unauthorized reproduction, distribution, or use of this material, in whole or in part, is strictly prohibited without prior written consent from Quantera.ai.

For any inquiries, please contact us at akash@quantera.ai