TAZI

AI Architecture Playbook for Financial Services

09.05.2025

+ 1 408 290 5491 www.tazi.ai I info@tazi.ai

Table of Contents

Summary

The artificial intelligence landscape is evolving at a breakneck pace, moving beyond the initial dominance of monolithic Large Language Models (LLMs). A new era of architectural specialization is dawning, where novel frameworks are designed to solve specific, complex problems with unprecedented efficiency and power. For the financial services industry, this shift presents a transformative opportunity.

This paper provides a strategic overview of the modern AI spectrum—from established Transformers to cutting-edge Graph Neural Networks and State Space Models. It then maps these powerful architectures to 12 high-impact use cases across capital markets, risk management, and customer strategy, offering a clear playbook for unlocking the next wave of competitive advantage.

Part 1: The Spectrum of Modern AI Architectures

Understanding the expanding toolkit of AI architectures is the first step toward strategic implementation. While Transformers and LLMs are incredibly versatile, specialized models often provide superior performance, efficiency, and reliability for specific tasks. Maturity of an architecture is also quite important when it comes to setting expectations. Many of the approaches may have been around for a while, they have existed in research domains and not real-life, particularly enterprise-style, implementations. Whether an architecture requires a particular data type, amount or preparation needs is another factor.

Comparative Overview of Key Architectures

Architecture	Core Idea	Primary Use Cases	Key Strength
Transformers/LLMs	Learning deep contextual relationships in sequential data via self-attention.	NLP, content generation, conversational AI.	Unmatched versatility on language tasks.
Hierarchical Reasoning Model (HRM)	Decomposing problems with a dual-level, "fast and slow thinking" architecture.	Complex reasoning, strategic planning, audits.	Deep, verifiable reasoning with little data.
State Space Models (SSMs)	Processing sequences with linear efficiency and parallel training.	Ultra-long sequence analysis (market data).	Drastic speed/cost improvements for long context.
Diffusion Models	Generating data by progressively de-noising a random signal based on a	High-fidelity media generation, synthetic data.	State-of-the-art quality in generative tasks.

	prompt.		
Graph Neural Networks (GNNs)	Operating on graph-structured data (nodes and edges) to model relationships.	Network fraud detection, recommendation engines.	Uniquely leverages network structure.
Hierarchical Temporal Memory (HTM)	A biologically inspired model for online pattern recognition in data streams.	Real-time anomaly detection, time-series forecasting.	Continuous, unsupervised learning.
Joint-Embedding Predictive Arch. (JEPA)	Self-supervised learning by predicting abstract representations of masked inputs.	Computer vision, learning "common sense" models.	Efficient learning of meaningful features.
Memory-Augmented NNs (MANN)	Enhancing a network with an explicit, external memory it can read from and write to.	One-shot learning, knowledge-intensive Q&A.	Rapidly assimilates and retrieves new info.
Physics-Informed NNs (PINNs)	Integrating the laws of physics (as equations) into the network's learning.	Scientific simulation, derivatives pricing.	High accuracy with less data; plausible outputs.

Part 2: Mapping Architectures to High-Impact Use Cases

Applying the right architecture to the right business problem is critical for success. This section provides a detailed analysis of 12 key financial services use cases, identifying the optimal AI tool for each job.

Use Case & Architecture Master Guide

Domain	Use Case	Recommended Architecture
Risk & Compliance	Insurance Claims Fraud	Graph Neural Networks (GNNs)
	Anti-Money Laundering (AML)	Graph Neural Networks (GNNs)
	Insurance Claims Audit	Hierarchical Reasoning Model (HRM)
	Market Scenario Generation	Diffusion Models
Customer Strategy	Voice of Customer (VoC)	Transformers/LLMs
	Hyper-personalized Communications	Transformers/LLMs + Diffusion Models
	Customer Churn/Acquisition	Traditional Machine Learning
Lending & Underwriting	Complex Underwriting & Credit	Hierarchical Reasoning Model (HRM)

Capital Markets	Algorithmic Trading (HFT)	State Space Models (SSMs)
Capital Markets	Exotic Derivatives Pricing	Physics-Informed NNs (PINNs)
Investment Management	AI-Powered Research & Alpha Gen	Transformers/LLMs
Investment Management	One-Shot Learning for Portfolio Mgmt	Memory-Augmented NNs (MANN)

Detailed Use Case Analysis

Risk Management & Compliance

- **Insurance Claims & AML Fraud Detection**
  - **Problem:** Fraud and money laundering are network activities.[2] Criminals use complex webs of accounts, providers, and shell companies to hide their tracks.[3]
  - **Architecture: Graph Neural Networks (GNNs)**
  - **Rationale:** GNNs are the state-of-the-art for these problems.[4] They analyze the entire network of relationships at once, detecting suspicious topologies (e.g., fraud rings, money laundering circles) that are invisible to models that inspect transactions individually.
- **Insurance Claims Audit & Compliance**
  - **Problem:** Auditing a claim requires a systematic, multi-step verification against a complex policy document.
  - **Architecture: Hierarchical Reasoning Model (HRM)**
  - **Rationale:** HRM is designed for deep, rule-based logic.[5] Its planner/worker structure can perfectly mirror an audit checklist, ensuring every step is followed reliably and providing a fully auditable reasoning trail for its decision.
- **Market Scenario Generation & Stress Testing**
  - **Problem:** Portfolios must be tested against crises that haven't happened yet. Historical data is insufficient for modeling true "black swan" events.

- **Architecture: Diffusion Models**
- **Rationale:** Diffusion models can generate high-fidelity, realistic, yet novel synthetic market data.[6] Risk managers can prompt the model to create specific scenarios (e.g., "stagflationary shock") to test portfolio resilience against a much wider range of plausible futures.

Customer Strategy & Operations

- **Voice of Customer (VoC) Analysis**
  - **Problem:** Gleaning actionable insights from vast amounts of unstructured customer feedback (calls, emails, reviews).
  - **Architecture: Transformers/LLMs**
  - **Rationale:** This is a core strength of LLMs. They excel at sentiment analysis, topic extraction, and summarization, allowing firms to understand customer pain points and priorities at scale.
- **Hyper-personalized Communications**
  - **Problem:** Moving beyond "Dear [First Name]" to create genuinely unique and context-aware communications.
  - **Architecture: Transformers/LLMs + Diffusion Models**
  - **Rationale:** LLMs generate the dynamic, personalized text.[7] Diffusion models can add a layer of visual personalization, such as creating a unique image for a marketing email that reflects a customer's recent activity or interests.

- **Customer Churn & Acquisition Modeling**
  - **Problem:** Predicting which customers will leave and which prospects will convert based on historical, structured data.
  - **Architecture:** Traditional Machine Learning (e.g., Gradient Boosting)
  - **Rationale:** For this classic prediction task, traditional models remain the most efficient, effective, and interpretable solution. The complexity of deep learning is often not necessary.

Lending & Underwriting

- **Complex Underwriting & Credit Decisioning**
  - **Problem:** Automating complex underwriting decisions for commercial loans or insurance requires following a deep, multi-stage rulebook, not just calculating a simple score.
  - **Architecture: Hierarchical Reasoning Model (HRM)**
  - **Rationale:** HRM can codify the entire underwriting policy. It methodically steps through the

required checks, providing a robust and transparent decision-making process that aligns with regulatory demands for explainability.

Capital Markets & Investment Management

- **Algorithmic Trading & Microstructure Analysis**
  - **Problem:** Finding predictive signals in high-frequency market tick data, which consists of millions of data points per day.
  - **Architecture: State Space Models (SSMs)**
  - **Rationale:** Transformers are too slow and expensive for this task. The linear-time performance of SSMs allows them to efficiently process these ultra-long sequences and model the market's microstructure to uncover subtle, short-term trading opportunities.
- **Exotic Derivatives Pricing & Hedging**
  - **Problem:** Complex derivatives are defined by partial differential equations (PDEs) that are slow and difficult to solve with traditional numerical methods.
  - **Architecture: Physics-Informed Neural Networks (PINNs)**
  - **Rationale:** A PINN can learn to solve these governing financial equations.[8] It acts as an ultra-fast pricing engine, allowing traders to price and hedge complex instruments in real-time, a significant competitive advantage.

- **AI-Powered Research & Alpha Generation**
  - **Problem:** Analysts are drowning in information. They need to rapidly synthesize insights from thousands of pages of news, filings, and reports.
  - **Architecture: Transformers/LLMs Digests**
  - **Rationale:** A specialized financial LLM can act as a powerful research associate, summarizing documents, identifying risks, tracking competitor strategies, and surfacing thematic trends (alpha) from unstructured text.[9]

- **One-Shot Learning for Portfolio Management**
  - **Problem:** A sudden geopolitical event or market shock occurs. A portfolio manager needs their AI assistant to immediately understand this new reality and its implications without needing to be retrained.
  - **Architecture: Memory-Augmented Neural Networks (MANN)**
  - **Rationale:** MANNs can assimilate new facts instantly by writing them to their external memory. A PM could "tell" the MANN about the event, and the model could immediately use that new piece of knowledge to re-evaluate portfolio risks or answer questions about its impact.

Conclusion: Specialized Architectures for Different Use Cases

The era of a one-size-fits-all approach to artificial intelligence is over. While Large Language Models (LLMs) will continue to be a foundational technology, sustainable competitive advantage in financial services will be driven by the strategic deployment of specialized architectures.

Firms that move beyond generic AI and learn to leverage the right tool for the right job—using GNNs for network fraud, HRMs for complex underwriting, and SSMs for market data—will be the ones to build more intelligent, efficient, and resilient operations for the future.