PAYMENTS · DIGITAL ASSETS · MARKET EXPANSION

AI in Financial Services

A Practical Primer for Bank and Fintech Leaders

clarinovapartners.com

Why This Matters Now

Every bank and fintech board in the region is being asked the same question: what is our AI strategy? The honest answer usually starts with a vocabulary gap — the terms get thrown around in vendor decks and press coverage faster than most leadership teams can absorb them. This guide is a plain-English primer: the handful of concepts that actually matter for a financial services decision-maker, what they cost, and the questions worth asking a vendor before signing anything.

It is deliberately non-technical. You will not need to write a line of code to use it — only to hold your own in the next conversation with a technology partner, regulator, or board member.

The Building Blocks, in Plain English

Large Language Model (LLM)
The AI system itself — the thing that reads a question and writes an answer. Claude, ChatGPT, and Gemini are examples. Think of it as a very well-read, very fast junior analyst who has read an enormous amount of text but has never actually worked inside your bank.

Tokens
The units an LLM reads and bills by — roughly three-quarters of a word each. Every AI vendor's pricing is quoted per million tokens processed. This is the metre that's running whenever the system is in use.

Embeddings
A way of turning text into a numerical 'fingerprint' so a computer can judge whether two pieces of text mean the same thing — even if they use completely different words. This is what lets a customer-service assistant recognise that 'overdraft charge' and 'insufficient funds fee' are the same issue, without a human ever building that mapping by hand.

RAG (Retrieval-Augmented Generation)
The standard pattern for a compliant, auditable enterprise AI assistant. Rather than retraining the model on your data, the system fetches the relevant passages from your own policy manuals, product terms, or regulatory filings at the moment of the question, and has the model answer strictly from that material.

> *This is also where the bank's proprietary data gets ringfenced: your documents stay inside your own environment, and only the relevant retrieved passages are sent to the LLM at query time — the model is never trained or improved on your data.*

Hallucination
The single biggest risk in this category: a fluent, confident, entirely wrong answer. It's the primary reason no serious deployment in a regulated environment goes live without RAG grounding and a human-in-the-loop review step for anything customer-facing.

Agents / Agentic AI
The shift the industry made in 2025–26: instead of just answering questions, the system can take actions — checking a balance, opening a ticket, escalating a case — by calling other systems directly. This is where most of the real productivity case now sits, and where governance controls matter most.

Where the Value Actually Shows Up

Across GCC and South Asian institutions, the deployments that move past pilot stage tend to cluster around three patterns:

Customer & staff assistants (RAG)	A chat interface grounded in your own policy documents, product terms, and FAQs — reducing call-centre load and first-line query time, with every answer traceable to a source document rather than the model's own memory.
Pattern-matching at scale (embeddings)	Flagging similar-looking transactions, complaints, or documents for fraud review, AML screening, or complaint triage — finding conceptual matches, not just keyword matches, across millions of records.
Workflow automation (agentic AI)	KYC document review, onboarding checklist completion, and internal reporting drafts — where the system doesn't just answer a question but completes a multi-step task, with a human sign-off gate before anything customer-facing goes out.

What It Actually Costs

Vendor rate cards quote a per-token price, but that is rarely where the budget actually goes. In practice, cost sits across four layers:

- **Model usage — the per-query AI cost itself. This can vary significantly depending on which model is used: costs across leading vendors can differ by 5–10x for comparable workloads, so model selection is itself a cost decision, not just a capability one.**
- **Data infrastructure — the ongoing cost of storing and searching your own documents so the AI can retrieve them. Most institutions run this in parallel across development, evaluation/testing, UAT, and production environments, and each typically needs its own storage and usage allowance — one of the most commonly underestimated cost items, since budgets are often set for a single environment rather than four.**
- **Governance and testing — compliance review, accuracy testing, monitoring, and human oversight workflows. For AI specifically, a dedicated evaluation stage — testing model behaviour, hallucination rate, and edge cases with production-like data volume — is worth keeping distinct from generic functional QA. Industry experience consistently shows this overall category is the largest cost, not the model itself.**
- **Integration — connecting the system to your core banking, CRM, or ticketing platforms.**

A Note on Vendor Choice

Not every AI vendor is priced or governed the same way. Lower-cost alternatives exist and are worth piloting for internal, non-regulated workloads. But for anything touching customer data or a regulated process, price is the wrong first filter — data residency, auditability, and the vendor's own regulatory standing in your jurisdiction should be settled before cost is. This is precisely the kind of trade-off worth working through with an advisor before it becomes a procurement decision.

Questions Every Financial Services CEO Should Ask

Before any AI deployment moves past pilot stage, these are the questions worth putting to the technology team or vendor directly — not because the answers are always simple, but because a vague answer to any of them is itself the signal to slow down.

Where will our data reside?
Confirm exact processing/storage locations and subprocessors — and check that against your regulator's residency requirements, not the vendor's marketing claim.

How are AI answers audited?
Every answer should be traceable back to a specific source document and version, with the original prompt, retrieved documents, and response retained in an audit log. If an answer can't be explained, it shouldn't be relied upon.

What happens when the model is wrong?
There should be a defined correction workflow. Customer-facing responses should have appropriate human oversight, errors should be captured and analysed, and knowledge sources updated so the same mistake is less likely to recur.

Where is human approval required?
Anywhere a decision is customer-facing, financially material, irreversible or regulatory. Examples include credit decisions, transactions, regulatory filings and personalised financial advice.

Can we switch LLM vendors without rebuilding everything?
Only if the solution has been designed for portability from the outset. The data, retrieval layer and application logic should remain independent of any single model provider, making it easier to change vendors as technology, pricing or regulation evolves.

What is the total cost of ownership — not just token pricing?
Model usage is often the smallest component. Total cost typically includes data infrastructure, integration, governance, testing, monitoring, security and ongoing operational support. Procurement decisions should consider the full operating model, not just the AI bill.

Which use cases should we deliberately avoid?
Avoid high-risk decisions without appropriate human oversight, particularly where errors could affect customers, regulatory compliance or financial outcomes. Governance should determine where AI assists people — not where it replaces accountability.

Where ClariNova Partners Helps

We work with banks, fintechs, and investors across the India–GCC–ASEAN corridor to translate this kind of technology shift into a board-ready strategy: vendor and architecture selection, regulatory positioning, operating model design, and go-to-market for digitally-enabled financial products.

Engagement models: Strategic Advisory | Fractional Leadership | Project Delivery | Executive Workshops

Bobby Kakar

Managing Partner, ClariNova Partners

bobby@clarinovapartners.com | clarinovapartners.com