INDUSTRY REPORT · AGENTIC AI × BANKING

Not every process *needs an agent*

A bank's guide to choosing the right pain points, the right stack, and the architecture that survives scale and audit. Because the hardest agentic AI decision is not how, and it is not how fast. **It is where, and whether.**

SERIES
Responsible AI in Financial Services

FOCUS
Selection · Architecture · Governance

PUBLICATION
Clarium · 2026



Photo: Unsplash

EXECUTIVE SUMMARY

Most agent programs fail on selection, not capability. *Choose like an underwriter.*

The industry's most cited statistic on enterprise generative AI is not about models. MIT researchers found that 95% of organizations studied saw no measurable profit-and-loss impact from their pilots, and the report's title, the GenAI Divide, blames learning and integration gaps, not model quality. In banking the pattern is sharper: pilots are bolted onto processes that should never have been candidates. This report is a selection guide. It gives you a filter for choosing pain points, a five-layer stack built for scale and resilience, and four go/no-go gates derived from the 4Ds of AI fluency.

95%	200k	4	2011
Enterprise generative AI pilots with no measurable P&L impact, per MIT's 2025 GenAI Divide study	Users JPMorgan's LLM Suite onboarded in eight months: scale as a governed copilot, not an autonomous agent	Gates every candidate must pass: Delegation, Description, Discernment, Diligence. Fail one, and the use case waits	The year of SR 11-7, the model-risk guidance regulators read as covering your agents today

01 **Selection beats capability.** The 95% statistic is a selection statistic. High-volume, low-discretion processes with clean data are the reliable first moves; anything requiring judgment belongs to a copilot or a human.

02 **The stack that survives audit is boring by design.** Five layers: data, integration, model, harness, oversight. Model choice gets the headlines; the harness and oversight layers decide whether anything ships.

03 **Scale and resilience are architecture decisions.** Deterministic fallbacks, idempotent actions, circuit breakers, canary rollouts and cost-per-decision economics are designed before the first pilot, not after the first outage.

04 **The 4Ds are your gate system.** If a candidate cannot pass Delegation, Description, Discernment and Diligence checks, it does not matter how good the demo was. It waits.



Judgment stays human. Volume goes to the machine. Photo: Unsplash

A filter before a pilot: *volume and discretion* decide where agents belong

Two questions sort every candidate process faster than any vendor deck. **How much volume does it carry, and how much judgment does it require?** Get those answers honestly and the deployment mode chooses itself.

EXHIBIT 1 The agent-fit matrix: two axes, four verdicts

LOW DISCRETION ←	DISCRETION AND JUDGMENT →
LOW VOLUME · LOW DISCRETION **Plain software** Existing rules, workflows and scripts already do this. Adding a model adds cost and failure modes without adding capability. *Don't buy a model to do a regex's job.*	LOW VOLUME · HIGH DISCRETION **Human judgment** Rare, high-stakes, context-heavy. The economics of an agent never clear, and the risk never justifies autonomy. *Restructurings, complex complaints, board-level analysis.*
HIGH VOLUME · LOW DISCRETION **Automate now** The reliable first move. Structured inputs, checkable outputs, measurable error rates. Agents compound value quickly here. *Reconciliation, data extraction, document classification.*	HIGH VOLUME · HIGH DISCRETION **Agent with harness** The interesting middle. Agents draft, chase and monitor; humans verify and sign. Autonomy expands only as the evidence grows. *KYC refresh, credit memo drafting, exception triage.*

Framework: Clarium. Volume axis measured in transactions or cases per period; discretion axis measured by the cost of a wrong output reaching a customer.

MIT's Project NANDA put a number on what bankers suspect: 95% of organizations studied saw no measurable P&L impact from their generative AI pilots. The report attributes the divide to learning and integration gaps, and in banking the largest integration gap is upstream of the technology. **Pilots are attached to processes that were never good candidates.**

The matrix also explains why early wins look unglamorous. Nobody wins a board presentation with invoice reconciliation, yet that is precisely where agents earn trust: every output can be checked against a source system, every error is visible, and the fallback is the workflow you never dismantled.

SO WHAT

Resist the demo that automates your most impressive process. **Score every candidate on volume and discretion first,** and let the matrix, not the vendor, pick your first three deployments.

SECTION 02 · THE RIGHT STACK

The stack that survives audit is *boring by design*

An agent in a bank is not a chatbot with keys. It is a five-layer system in which the model is the least durable layer. **Build the layers in this order and each one outlives the model news cycle above it.**

EXHIBIT 2 Five layers of an audit-ready agent stack

Oversight ACCOUNTABILITY LAYER	Model-risk integration (SR 11-7), audit trails, human sign-off chains, AI-use disclosure. Designed first, not bolted on for the exam.	BANK OWNS
Harness CONTROL LAYER	Mandates, versioned instructions, evaluation suites, exception queues. The agent's authority lives here, in writing, not in the prompt.	BANK OWNS
Model REPLACEABLE LAYER	Buy over build for nearly every bank. Treat open-versus-frontier as a procurement decision with exit clauses, not an identity.	VENDOR LED
Integration CONNECTIVE LAYER	API and event fabric between systems of record, risk tools and the agent. Governed at the API layer, with write-backs logged like any other transaction.	BANK OWNS
Data FOUNDATION LAYER	Quality-controlled, permissioned, lineage-tracked data. An agent that reads weak data produces confident nonsense at machine speed.	BANK OWNS

Reference stack: Clarium, aligned to delivery patterns across data platforms (Databricks, AWS), enterprise integration (MuleSoft-pattern) and cybersecurity monitoring. Ownership labels: Clarium analysis.

Notice what the ownership labels imply. The three layers that decide whether an agent is safe, the data it reads, the harness that bounds it, and the oversight that audits it, are all yours. The one layer everyone talks about, the model, is the one you should be most willing to swap.

This is also where model-risk management lands. Regulators' guidance on model risk was written in 2011 and confirmed as covering AI in 2021, and it asks questions any CIO can answer for a database but few can yet answer for an agent: what does it consume, who owns it, how is it tested, what happens when it fails. **A five-layer stack makes those questions answerable by construction.**

SO WHAT

If a vendor's proposal is mostly about the model, you are buying a demo. **Ask instead who owns the data layer, who writes the harness, and who signs the oversight ledger.** Those three answers predict deployment success better than any benchmark.

Scale and resilience are architecture decisions, *not pilot afterthoughts*

A pilot that works for 200 documents a day and a production system that works for 200,000 are different engineering objects. **The patterns below are the difference between an agent that demos and an agent that runs.**

EXHIBIT 3 Six patterns that turn a pilot into production

Deterministic fallback	**Every agentic flow has a non-AI path kept warm behind a switch.** The old workflow is not retired; it becomes the resilience plan. When the model degrades, the process degrades to yesterday, not to zero.
Queues and idempotency	Agent actions are enqueued and deduplicated. A retry can never double-pay an invoice or double-book a customer. **At scale, the queue is the control surface.**
Circuit breakers	Confidence, anomaly and volume thresholds automatically route flow back to humans. **Nobody tunes a prompt at 2 a.m. because the breaker already fired.**
Canary rollout	One percent of documents, one branch, one portfolio first. Service-level objectives measure quality and override rates, **not just uptime.**
Decision observability	Every decision logs inputs, model version and instruction version. Cost per decision is tracked like cost per transaction, because it is one.
Capacity honesty	Token, licence and review economics modeled at production volume. **A pilot priced at 200 documents is not a forecast; it is a rounding error.**

Patterns: Clarium engineering practice, informed by Anthropic's harness research on long-running agents (task lists, written state, mandatory verification) and standard distributed-systems practice. Source: Anthropic Engineering (Nov 2025); Clarium analysis.

Anthropic's research on long-running agents found that a frontier model fails without a harness: an explicit task list, a written record of state, and mandatory end-to-end testing before work is called complete. **What banks need is that harness at institutional scale.** The task list becomes a mandate register. The progress file becomes decision observability. The test becomes the evaluation suite and the exception queue.

Resilience is also what makes narrow deployments expandable. An agent with a warm deterministic fallback and a working circuit breaker can be trusted with slightly more authority each quarter, because every failure mode has a rehearsed answer. **Autonomy is not granted; it is earned through observed behavior under controlled failure.**

SO WHAT

Budget the harness before the model. If half your agentic spend is on fallbacks, queues, observability and evals rather than tokens, **you are building a bank system and not a science project.**

The 4Ds are your gate system: *fail one discipline, and the use case waits*

The AI Fluency Framework, developed by Rick Dakan and Joseph Feller with Anthropic, describes four disciplines for working with AI. Run each candidate through them as pass/fail gates **before any code is written.**

EXHIBIT 4 Four gates, eight outcomes: the deployment decision in one page

D Delegation

THE BOUNDARY GATE

PASS Decision rights are written down: an approved mandate register defines exactly what moves and what stays human.

FAIL "The team will figure out the boundaries as we go." If the boundary is vibes, the governance is fiction.

D Description

THE MANDATE GATE

PASS The mandate fits in versioned instructions with hard limits that a second team can audit without the author present.

FAIL The process cannot be described without a series of meetings. If it cannot be described, it cannot be delegated.

D Discernment

THE VERIFICATION GATE

PASS A test exists that catches a wrong answer before it reaches a customer or a ledger: reconciliation, evals, exception queues.

FAIL Verification means "we review the output". A confident wrong answer and a correct one look identical on a screen.

D Diligence

THE ACCOUNTABILITY GATE

PASS A named signatory, a data-sensitivity check, and an AI-use disclosure path exist before go-live, not after the incident.

FAIL Ownership is "shared", the data is "probably fine", and disclosure is "we'll deal with that later".

Gates derived from: Dakan & Feller, AI Fluency Framework (aifluencyframework.org), developed with Anthropic, CC BY-NC-SA. Pass/fail phrasing: Clarium analysis.

> *If you cannot describe the mandate, you are not delegating. You are abandoning.*

The gates are deliberately unglamorous, and that is the point. They convert an exciting vendor conversation into a boring underwriting decision, which is exactly what a bank's governance exists to do. Note also what the gates are not: they are not a model benchmark, not a security questionnaire, and not a compliance sign-off. Those happen later. **The gates decide whether the process deserves to become an agent use case at all.**

SO WHAT

Publish the gates internally, with the fail phrases included. When a business line cannot get past Delegation or Description, **that is not a blocker to work around; it is the filter working.** Most candidates should fail, and the ones that pass are your roadmap.

Where agents do not belong, yet: *a short list worth defending in writing*

Every bank needs an explicit negative space: use cases where agents are off the table for now, written down and owned. **The list is not pessimism; it is the fastest way to earn the regulator's trust for everything else.**

EXHIBIT 5 Two lists every board should be able to recite

Off the table

NO AGENT AUTONOMY

- **Payment and transfer execution:** autonomous money movement belongs to regulated rails, not model outputs
- **Client advice and suitability:** EU AI Act high-risk territory; a licensed human owns the recommendation
- **Credit decisions:** the EU AI Act designates credit scoring high-risk; agents may draft, never decide
- **Complaints adjudication:** discretion, empathy and regulatory duties that do not transfer to software
- **Anything touching market conduct rules:** if a wrong output is a reportable event, it is not autonomous

The right pattern

DRAFT, VERIFY, SIGN

- **Draft-then-sign:** agent produces, licensed human reviews and owns the outcome
- **Copilots at scale:** JPMorgan's LLM Suite reached 200,000 employees in eight months as a governed assistant, not an autonomous decision-maker
- **Exception queues:** agents triage and prepare; humans close files and absorb judgment
- **Monitoring and detection:** agents watch for anomalies and raise hands; humans investigate
- **High-volume verification:** reconciliation and extraction where every output can be checked against a source system

Left list: Clarium analysis informed by EU AI Act (Regulation (EU) 2024/1689) high-risk categories for credit scoring and insurance pricing, and by Federal Reserve/OCC model-risk guidance (SR 11-7, 2011; SR 21-8 and OCC Bulletin 2021-38, 2021). Right list: JPMorganChase on LLM Suite adoption (2025); Clarium delivery patterns.

The no-fly list ages well precisely because it is short. Payment execution, advice, credit decisions and adjudication stay off-limits until regulation, evidence and harnesses change together, which is a board-level decision and not an innovation-team one. Meanwhile the right-pattern column covers most of the value banks are actually finding: drafting, monitoring, verification and preparation work that makes humans faster without touching their authority.

There is a discipline angle too. The no-fly list is the Diligence gate applied at institution scale: a named accountability boundary that says the signatory, the data owner and the regulator will never be surprised by where an agent turned up.

SO WHAT

Write the list down and review it quarterly. **A bank that can state where agents do not go, in one page and in public, will deploy faster everywhere else** because its governance no longer has to improvise.

Worked example: *how a mid-size bank picks its first agent*

Five candidates from a real shortlist shape. Run each through the matrix and the gates, and the roadmap writes itself. **Notice how many common ambitions fail honestly.**

EXHIBIT 6 A shortlist, scored: volume, discretion, verdict

CANDIDATE PAIN POINT	VOLUME	DISCRETION	VERDICT
Invoice and PO reconciliation	HIGH	LOW	DEPLOY NOW. Agent plus rules, humans on exceptions. Every output checkable against source systems; the classic first move.
KYC periodic refresh and document chasing	HIGH	MEDIUM	AGENT WITH HARNESS. Agent drafts requests and tracks responses; the KYC officer closes the file. Expansion is earned by override rates.
Commercial credit memo drafting	HIGH	HIGH	AGENT DRAFTS, OFFICER SIGNS. The Vol. 01 control-chain pattern: bounded mandate, reconciliation, evidence retained.
Wealth and lending advice	LOW	HIGH	NO AGENT. Copilot at most, inside the EU AI Act high-risk perimeter. The licensed human owns the recommendation end to end.
Trading and payment execution	ANY	ANY	NOT AN AGENT USE CASE. Keep straight-through processing on regulated rails with existing controls; revisit only when law and harnesses change.

Scoring: Clarium analysis using the agent-fit matrix (Exhibit 1) and the four gates (Exhibit 4). Reference design; candidates are illustrative composites of common bank shortlists, not any single institution.

Two of five candidates fail, one of them loudly. That is a healthy scorecard: it means the filter is doing its job and the roadmap is honest. The three that pass share a property worth naming: **every one of them keeps a human between the agent's output and the customer.**

The sequencing matters as much as the picks. Reconciliation first, because it builds the data layer, the queues and the observability. KYC refresh second, because it reuses all three and adds an exception queue. Credit memos third, because by then the harness, the logs and the sign-off chains already exist. **Each deployment inherits the last one's plumbing, which is how a narrow first move becomes a durable capability.**

SO WHAT

Run this exact table with your own shortlist before the next vendor call. **If the exercise produces three honest failures, it has already paid for itself:** those are three pilots you will never have to quietly shut down.

SECTION 07 · THE CONVERSATION

Five questions before your next agent pilot, *and one of them is about stopping*

These are the questions we bring to the first working session with a bank, and the ones we ask of our own proposals. **None of them is about which model to buy.**

01 Which candidate survives all four gates, and which fails first?

Run the honest version. If every candidate passes, the gates are broken or the shortlist is theater. **A healthy shortlist kills at least one beloved idea.**

02 Is your model-risk inventory ready to classify an agent?

SR 11-7 was written in 2011 and confirmed as covering AI in 2021. An agent is a model, a process, and a vendor system at once. **Which column does it go in, and who at your bank signs that answer?**

03 Where is your harness, concretely?

Who signs the mandate register, which queue reviews exceptions, and which log answers the examiner's question about the decision made at 3 a.m.? **If those three answers take more than a minute each, the harness is not built yet.**

04 Which layers do you own, and which can your vendor switch off?

The model layer is replaceable; your data, harness and oversight layers should not be. **Scale without ownership is tenancy, and rent goes up.**

05 Where have you decided, in writing, that agents do not go?

The no-fly list is a governance asset. It earns regulator trust, protects customers, and speeds up everything outside its perimeter. **If it does not exist yet, writing it is the highest-value agentic AI work you can do this quarter.**

Which candidate fails first at your bank?

Clarium helps banks run this exact selection process: pain-point scoring, five-layer architecture, harness design and the no-fly list. Bring your shortlist; we will bring the gates.

Start the conversation · clarium.tech

Sources

[1] MIT Project NANDA, "The GenAI Divide: State of AI in Business 2025" (2025): 95% of organizations studied reported no measurable P&L impact from generative AI pilots; coverage in Fortune, August 18, 2025. Cited in the Executive Summary and Section 01.

[2] JPMorganChase, "LLM Suite named 2025 'Innovation of the Year' by American Banker" (2025): the platform went from zero to 200,000 onboarded users within eight months, operating as a governed assistant for employees. Cited in the Executive Summary and Section 05.

[3] Board of Governors of the Federal Reserve System and OCC, "Supervisory Guidance on Model Risk Management" (SR 11-7 / OCC 2011-12, 2011); Federal Reserve SR 21-8 and OCC Bulletin 2021-38 (2021), confirming the guidance's applicability to artificial intelligence and machine learning models. Cited in Sections 02 and 05.

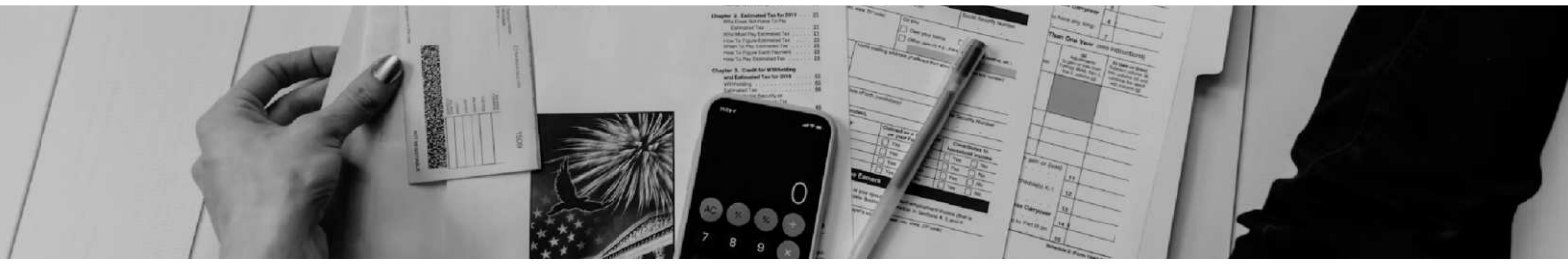
[4] European Union, AI Act, Regulation (EU) 2024/1689: high-risk classification of AI credit scoring and life/health insurance pricing; European Commission, "AI in finance" (June 2024); KPMG, "Setting the ground rules: the EU AI Act." Cited in Sections 04 and 05.

[5] Dakan, R. & Feller, J., "AI Fluency: Framework & Foundations," developed with Anthropic. aifluencyframework.org; Anthropic Academy. Released under CC BY-NC-SA 4.0. The four competencies, Delegation, Description, Discernment, Diligence, are adapted here as go/no-go gates in Section 04.

[6] Anthropic, "Effective harnesses for long-running agents," Anthropic Engineering (Nov 26, 2025): task lists, written state, and mandatory end-to-end verification for agents working across sessions. Adapted to institutional scale in Section 03.

[7] Clarium delivery patterns referenced as practical grounding: loan-origination platforms, PPP loan origination, middle-office mortgage portals, MuleSoft integration of Abrigo LOS with Fiserv Premier core, and AWS cloud migration for financial services. Reference design only; not a depiction of any single client.

About this report. Prepared by Clarium as part of the Clarium Insights series on responsible AI in financial services. Clarium delivers AI/ML, cloud, data analytics, cybersecurity, and enterprise integration solutions for financial institutions and regulated industries. Clarium partners with AWS, Microsoft Azure, Google Cloud, Databricks, Salesforce, MuleSoft, and leading security providers. **Disclaimer.** This document is for general information only and does not constitute legal, regulatory, or investment advice. Statements reflect publicly available sources as of publication and may be subject to change as regulation evolves. · clarium.tech



Selection is a ledger discipline. Photo: Unsplash