



AI for banking and financial services

Applying AI in high-control environments with strong governance, evidence, and human accountability.

WORKS WITH THE ASSISTANTS TEAMS ALREADY USE



EDITION 23 / EMBER & LILAC

EXECUTIVE WORKING EDITION - 08 OCT 2026

23

AT A GLANCE

Applying AI in high-control environments with strong governance, evidence, and human accountability.

Banks, lenders, insurers, wealth managers and payment firms operate under supervisory expectations for governance, model risk, fair treatment of customers, record keeping and operational resilience. AI can help with service support, document operations, financial-crime support and knowledge work, but only when it fits inside those controls. This playbook explains where AI fits first, how to classify risk, how to design human review and evidence trails, how to validate and monitor, and how to run a pilot that a risk committee can approve. It is general guidance, not legal, regulatory or financial advice.

WHO IT IS FOR

Heads of operations, risk and compliance, technology and data leaders, internal audit, and business line owners at banks, credit unions, lenders, insurers, wealth and payments firms.

TIME TO USE

75 minutes to read; one working session to apply

What you will have at the end

- A risk-tiered view of AI use cases that your second line can recognize
- Design patterns for human review and evidence that stand up to audit
- A validation and monitoring approach proportionate to each tier
- A pilot that starts in lower-risk work and builds credibility

Contents

01	AI in a control-heavy institution	03
02	A risk-tiered view of use cases	04
03	How the technology behaves, and why that matters to control owners	06
04	Governance, roles and the three lines	07
05	Controls, evidence and human review	09
06	Validation, monitoring and value measurement	10
07	A staged implementation approach	12
08	Common pitfalls and an illustrative scenario	13
09	Vendor due diligence questions	14
10	A first quarter that produces evidence	16
11	Terms used in this publication	17

EXECUTIVE BRIEF

AI in a control-heavy institution

01

K'

In financial services the question is not whether AI can do the task. It is whether you can show, afterwards, why it was safe to let it.

Financial institutions already manage risk systematically: three lines of defense, model risk management, change control, vendor due diligence, conduct and customer fairness frameworks, and record-keeping obligations. AI does not replace these. It enters them. A use case that cannot be explained to a risk committee, evidenced to an auditor and owned by an accountable executive should not go live, however good the demo.

The practical consequence is that the best first projects are those that improve productivity in knowledge-heavy work while keeping a person responsible for every consequential outcome. Examples include summarizing policies and procedures for service staff, extracting data from documents for operations teams to verify, preparing draft case narratives for financial-crime analysts, and helping relationship teams assemble client briefing material from approved sources. These cut effort without handing a credit, pricing or customer-eligibility decision to a model.

- Classify every proposed use by consequence for customers, the institution and regulators before choosing technology.
- Begin with internal, human-reviewed, lower-consequence use cases, and earn the right to go further.
- Design for evidence: what the tool saw, what it produced, who reviewed it and what they decided.
- Align with existing model risk, third-party risk, data protection and conduct frameworks rather than building a parallel process.
- Keep accountability with named individuals under your existing accountability regime.

II

NOT REGULATORY ADVICE

Supervisory expectations vary by jurisdiction, sector and institution type, and they change. Confirm current requirements and guidance with your compliance, legal and risk functions and your supervisors.

WORKING QUESTION

If a supervisor asked you tomorrow for a list of every AI use in the institution and its risk classification, could you produce it?

WHERE IT FITS

A risk-tiered view of use cases

02

Tiering lets you move quickly on safe work and slowly on consequential work, with reasons you can defend.

TIER	EXAMPLES	TYPICAL TREATMENT
Lower: internal productivity, no customer decision	Policy and procedure Q&A for staff, meeting notes, drafting internal emails, code assistance for engineers	Approved tool, usage rules, sampling-based review
Medium: customer-facing communication or operational document handling with review	Draft responses to customer queries, document data extraction for operations verification, contact-center agent assist	Named reviewer, tested prompts, quality sampling, complaints monitoring
Higher: supports risk or compliance judgments	Alert narrative drafting and triage support in financial-crime operations, regulatory change summarization, credit file summarization for underwriters	Formal validation, human decision, evidence trail, second-line review

TIER	EXAMPLES	TYPICAL TREATMENT
Highest: automated decisions affecting customers' access or terms	Credit decisions, claims decisions, pricing, fraud blocks with customer impact	Full model risk management, fairness testing, explainability, regulatory engagement; not a first project



Within each tier, ask about the human role. A person who reviews a summary and decides is different from a person who rubber-stamps a recommendation under volume pressure. The more a workflow pushes volume through a reviewer, the more you must check that review is real: time allowed, sampling by quality assurance and disagreement rates all matter.

Good first projects

- Service knowledge assistant answering staff questions from approved policy and product documents, with source links
- Document operations: extracting fields from applications, statements and correspondence into a work queue where staff verify
- Case preparation for financial-crime analysts: assembling facts and draft narrative from the case record, with the analyst deciding
- Knowledge work for relationship and advisory teams: summarizing approved research and client files, with compliance-approved templates



COMMON MISTAKE

Choosing a high-profile, customer-facing chatbot as the first project. It combines conduct risk, accuracy risk and reputational risk. Earn credibility internally first.



WORKING QUESTION

Which of your candidate use cases sit in the lower two tiers, and which of those has the clearest owner?



HOW IT WORKS

How the technology behaves, and why that matters to control owners

03

Control owners do not need to be technical, but they must understand the behaviors that create risk.

Generative models produce text one piece at a time based on patterns learned in training and on the material supplied in the request. The same question can produce different answers on different occasions. Output can be fluent and wrong. The model has no inherent awareness of your policies unless you supply them, and it cannot reliably tell when it lacks the information. Traditional scoring models in credit or fraud are different: they are statistical models with defined inputs and outputs, usually validated under your existing model risk framework.

01 Ground

Connect the tool to approved, current sources such as policy libraries and product documents, with access controls that match staff entitlements.

02 Constrain

Instruct the tool on role, scope and refusal behavior. Where possible, restrict output to structured fields or approved templates.

03 Log

Record the input, the sources retrieved, the output, the user and the reviewer's action.

04 Review

A competent person checks and decides. Disagreements and corrections feed monitoring.

Behaviors to design around

- Variability: the same input may not give the same output, which complicates reproducibility and evidence.
- Invention: confident statements with no basis in the sources.
- Stale content: answers based on superseded policy.
- Prompt injection: instructions hidden in documents or messages that try to redirect the tool.

- Sensitive data exposure: personal or confidential data passed to the wrong place or retained by a vendor.
- Bias: patterns in historical data or language that lead to uneven treatment.

DESIGN RESPONSE

Prefer retrieval from approved sources over free generation, keep a human decision on anything consequential, log everything needed for later review, and treat documents from outside the institution as untrusted input.

WORKING QUESTION

For your most promising use case, what would you need to log to reconstruct what happened six months later?

OPERATING MODEL

Governance, roles and the three lines

04

Fit AI into existing accountability structures, with clear owners at each line.

ROLE	RESPONSIBILITY	TYPICAL HOLDER
Accountable executive	Owns the use case, its risks and outcomes under the accountability regime	Business line head
Business owner and process owner	Defines the workflow, review standards, KPIs and training	Operations or product manager
Model risk and validation	Independent assessment of fitness for purpose, limits and monitoring	Second-line model risk function

ROLE	RESPONSIBILITY	TYPICAL HOLDER
Compliance and conduct	Customer fairness, communications standards, regulatory mapping	Compliance officer
Technology, data and security	Architecture, access, logging, vendor integration	CTO organization, CISO
Third-party risk and procurement	Vendor due diligence, contract terms, concentration and exit planning	Vendor management
Internal audit	Independent assurance over the framework and selected use cases	Third line

An AI governance forum, which can be a sub-committee of an existing risk body, should review each use case against a short template: purpose, tier, data, model or vendor, human role, controls, testing results, monitoring and exit plan. Keep the inventory of use cases current, including those embedded in vendor products. Many institutions find AI features arriving inside software they already use, which means procurement and architecture reviews must ask the question every time.

Define the human role carefully. Reviewers need the authority, information and time to disagree with the tool. Record their decisions. Where staff make decisions with AI support, update procedures and training so that accountability for the decision remains clear and does not slide to the tool.

Governance minimums

- Inventory of AI use cases with tier, owner and status
- Single intake template and approval path
- Defined role of model risk, compliance and security in approval
- Policy on staff use of external tools and what data may be entered
- Incident process that covers AI errors and complaints
- Annual review of the framework and each live use case

WORKING QUESTION

Who is the single accountable executive for your first AI use case, and have they agreed?

RISKS AND CONTROLS

Controls, evidence and human review

05

Evidence is the product. If you cannot show what happened and who decided, the control does not exist.

RISK	CONTROL	EVIDENCE
Inaccurate output reaches a customer or regulator	Named reviewer; approved templates; quality sampling	Review logs, sample results, error register
Unfair or inconsistent treatment of customers	Fairness testing where outcomes are affected; monitoring of complaints and outcomes	Test reports, complaint analysis
Confidential or personal data exposure	Approved environments, access entitlement, vendor terms on data use and retention	Architecture review, contract, access logs
Financial-crime process weakened	AI assists analysts; decision rationale remain human; no suppression of alerts without approval	Case records, QA results, tuning approvals
Third-party dependency and concentration	Due diligence, contingency and exit plans, change notification	Vendor file, tested fallback
Records cannot be reproduced	Retain inputs, outputs, prompts and versions per retention rules	Retention schedule, retrieval test

Record keeping deserves attention. Regulators and auditors may ask why a decision was made or what a customer was told. For AI-assisted work this means retaining the relevant inputs, the output, the version of the instructions and tool, and the human decision, for the period your retention policy and regulations require. Agree with legal and compliance which of these are records and how they are stored and retrieved.

Customer fairness and communication standards still apply. Anything said to a customer, whether drafted by a person or a tool, must be clear, fair and not misleading. Where customers are dealing with an automated system, consider transparency and an easy route to a person. In regulated advice or sales contexts, only approved content and licensed individuals should

shape what is said.



CHECK CURRENT REQUIREMENTS

Expectations for model risk management, outsourcing, data protection, conduct and operational resilience are set by your regulators and evolve. Have compliance map each use case to current rules and guidance.



WORKING QUESTION

For your highest-tier use case, which person would answer a supervisor's question about how a particular output was produced?



MEASURING VALUE

Validation, monitoring and value measurement

06

Test the tool against known answers, monitor it in production and measure business benefit against a baseline.

Validation should be proportionate to tier. For lower tiers, a documented test on representative cases and agreed acceptance thresholds may suffice. For higher tiers, involve independent validation: conceptual soundness, data, testing on challenging and adversarial cases, limitations, and ongoing monitoring. In every case, write down what the tool is for, what it must not do and how you will know if it drifts.



01 Build a test set

Assemble 50 to 100 real, representative cases with answers agreed by experienced staff, including difficult and edge cases and cases where the right answer is to decline.

02 Set thresholds

Agree in advance what accuracy and error types are acceptable, with zero tolerance for certain error classes such as wrong figures in customer communication.

03 Test adversarially

Try misleading inputs, conflicting documents and injected instructions. Record failures.

04 Monitor in production

Sample live outputs weekly, track reviewer corrections and complaints, and re-test when sources or tool versions change.



METRIC	MEASUREMENT	NOTE
Accuracy against test set	Share correct, partly correct, wrong or properly declined	Re-run on every change
Reviewer override rate	Share of outputs changed or rejected	Very low may signal rubber-stamping
Handling time	Minutes per case before and after including review	Baseline from sampled work
Rework and error rate	Cases returned or corrected downstream	Track by cause
Complaints and customer outcomes	Volume and themes linked to AI-assisted work	Where customers are affected
Analyst quality in financial-crime work	QA scores on case files, timelines	Do not trade quality for speed



ILLUSTRATIVE ARITHMETIC

If an operations team processes 400 documents a day at 5 minutes each and extraction support reduces that to 3.5 minutes with verification included, the saving is 10 hours a day. Treat these as placeholders: measure your own sample and include error correction time.



WORKING QUESTION

What error would be unacceptable even once, and does your test set contain cases that would reveal it?



IMPLEMENTATION

A staged implementation approach

07

Build trust with the second line by starting small, documenting thoroughly and expanding on evidence.

01 Inventory and classify

List current and proposed AI uses, including those inside vendor products, and assign tiers.

02 Pick a first use case

Choose a lower or medium tier workflow with a clear owner, available data and measurable effort, such as staff policy Q&A or document extraction.

03 Complete the intake pack

Purpose, data, vendor, human role, controls, test plan and monitoring plan, reviewed by risk, security and compliance.

04 Validate and pilot

Run the test set, fix issues, then pilot with a small team for four to eight weeks with logging and weekly review.

05 Go-live decision

The accountable executive, with risk input, approves limited rollout against thresholds, with a defined rollback.

06 Monitor and expand

Operate with sampling and reporting. Use the evidence to propose the next use case or tier.

Pilot gate

- Use case is tiered and has an accountable executive
- Data protection, security and vendor reviews are complete
- Test set and thresholds are agreed and results documented
- Human review role, time and training are defined

- Logging and retention meet policy
- Rollback and incident procedures are tested
- Customer-facing wording and disclosures are approved where relevant

Training matters as much as technology. Reviewers need to know what the tool does well, where it fails and what is expected of them. Provide worked examples of good and bad outputs. Reinforce that they are accountable for the decision, and that raising a concern about the tool is part of their role. Provide managers with guidance on workload so that review is not squeezed by volume targets.

WORKING QUESTION

What would your risk committee need to see in the intake pack to say yes to a limited pilot within a month?

PITFALLS

Common pitfalls and an illustrative scenario

08

Most problems arise when pace outruns control or when review becomes a formality.

- Building a parallel governance process that nobody integrates with model risk and third-party risk.
- Allowing shadow use of public tools by staff under productivity pressure.
- Treating a vendor demo as validation.
- Letting a human reviewer approve hundreds of items an hour and calling it oversight.
- Failing to log enough to reconstruct events later.
- Ignoring AI features that arrive inside existing vendor software.
- Starting with an automated customer decision.

ILLUSTRATIVE SCENARIO (HYPOTHETICAL)

A mid-sized lender pilots document extraction for loan applications. Early results look strong, but sampling shows income figures are sometimes read from the wrong column in one statement format. The pilot adds a mandatory verification of income fields and a monitoring check on that format. The lesson: sampling by error class caught what overall accuracy hid.

Another recurring issue is vague accountability. When a tool contributes to a decision, people sometimes say the system recommended it. Supervisors expect a named person or function to be responsible. Make sure procedures say who decides, what they must check and what they must record, so that accountability is real and not diffused.

Finally, do not over-index on caution to the point of paralysis. Institutions that block all use tend to find staff using unsanctioned tools. A well-governed, approved option with clear rules for lower-tier work is often safer than prohibition.

WORKING QUESTION

Where might your institution already have AI in use that nobody has classified?

BUYING AND VENDORS

Vendor due diligence questions

09

Treat AI vendors as material third parties with additional questions on data, models and change.

TOPIC	ASK	WHY IT MATTERS
Data	Where is data stored and processed? Is it used to train models? Retention and deletion terms?	Confidentiality and data protection
Model changes	How and when do underlying models change? Notice, testing windows and rollback?	Behavior can shift without your code changing
Transparency	What documentation, testing results and known limitations can you share?	Supports validation
Security and resilience	Independent assessments, incident notification, subprocessors, continuity plans?	Operational resilience
Audit and access	Can we audit, and can regulators obtain information?	Supervisory expectations
Exit	How do we export our data and configurations, and what are the transition terms?	Concentration and exit risk

Ask vendors to support your testing rather than rely on their metrics. Provide your test set, run it in a trial environment and have your validators review results. Pay attention to how the vendor responds when a test fails: a good partner explains the cause and fix; a poor one disputes the test.

Before you sign

- Third-party risk assessment is complete and approved
- Contract covers data use, change notification, audit, incident notice and exit
- Trial results on your test set are documented
- Integration respects entitlements and logging requirements
- Support and escalation contacts are agreed
- Regulatory notification requirements for outsourcing, if any, have been checked

WORKING QUESTION

If the vendor changed its underlying model next month, how would you find out and what would you re-test?

30 / 60 / 90

A first quarter that produces evidence

10

Use 90 days to build the inventory, approve a first low-risk use case and prove it with evidence.

First 30 days: ground it

- Build the AI use inventory, including AI embedded in vendor tools, and assign risk tiers
- Name the accountable executive and the governance forum for AI
- Publish interim staff rules on approved tools and data that may not be entered
- Select a first lower or medium tier use case and capture a baseline

Days 31 to 60: test it

- Complete the intake pack and the security, data protection and third-party reviews
- Build the test set with agreed thresholds and run adversarial tests
- Define logging, retention and the human review role, and train reviewers
- Draft the monitoring plan and rollback procedure

Days 61 to 90: decide

- Run a limited pilot with weekly sampling and review
- Present results to the risk committee against thresholds and baseline
- Decide to expand, adjust or stop, with documented reasons
- Propose the next use case and update the framework from lessons learned

GLOSSARY

Terms used in this publication

11

TERM	MEANING
Model risk management	The framework for developing, validating, approving and monitoring models used in decisions.
Three lines of defense	A structure of business ownership, independent risk oversight and independent audit assurance.
Second line	Independent risk and compliance functions that oversee the business.
Validation	Independent assessment that a model or tool is fit for its intended purpose.
Human in the loop	A named person reviews and decides, with the tool supporting rather than deciding.
Evidence trail	The record of inputs, outputs, versions and decisions needed to reconstruct what happened.
Prompt injection	Hidden instructions in content that try to redirect an AI tool.
Fair treatment	The expectation that customers receive consistent, non-misleading and equitable treatment.
Third-party risk	Risk arising from reliance on external vendors and service providers.
Alert narrative	The written explanation an analyst records for a financial-crime alert or case.
Retention schedule	The policy defining how long each type of record must be kept.
Override rate	The share of tool outputs that a reviewer changes or rejects.

A note on use

This playbook is general guidance. It is not legal, regulatory or financial advice. Check current supervisory expectations and requirements for your jurisdiction and institution type.



TAKE THE NEXT STEP

Take the free Halden assessment at halden.company/assessments, or talk to us at halden.company/contact about turning this into a plan for your organization.

