AI Orchestration Vendor Evaluation Scorecard

A bank-grade scorecard for comparing AI orchestration platforms. Eight categories, fifty-plus criteria, and a calibrated scoring scheme. Bring to vendor demos, third-party-risk reviews, and the build-vs.-buy conversation with the board.

From HitLai Institute — Module B3: "AI Strategy for Banking Leadership"

WHY THIS EXISTS

Most "AI platform" evaluation scorecards are written by people who have never sat across from an OCC examiner or filled out a NYDFS vendor questionnaire. They optimize for features. Banks need to optimize for survivability.

This scorecard inverts the priorities. Features matter, but a feature-rich platform that cannot pass third-party risk review is worse than useless. The categories below are weighted accordingly — security, audit, regulatory alignment, and deployment flexibility come first; productivity features come later.

Use this for any AI orchestration evaluation. Adapt the weights to your bank's specific risk appetite and regulator expectations.

HOW TO SCORE

For each criterion, score 0 / 3 / 5 / 8 / 10:

- **0** — Not present, vendor cannot do it
- **3** — On roadmap, not available today
- **5** — Present but limited or requires custom work
- **8** — Present, production-grade, with reasonable controls
- **10** — Present, production-grade, with strong evidence and references

Weight each category per the percentages below; total to a final score out of 100. Anything below 70 should not pass third-party risk review for material AI deployment in a bank.

CATEGORY 1 — DEPLOYMENT FLEXIBILITY (WEIGHT: 18%)

Most-failed category in our experience. Vendors love to say "we can self-host" until the contracting conversation.

Criterion	Score (0/3/5/8/10)	Notes
True self-hosted deployment inside the bank's network, no outbound dependencies		
Air-gapped deployment supported and documented		
Private-tenant cloud with documented tenant isolation		
Hybrid deployment — same platform, different deployment per workflow		
Support for open-weight local models (Llama, Qwen, Mistral) via Ollama or vLLM		
Documented sizing and architecture for 1-10K users		
Reference customer running self-hosted in regulated industry (banking, healthcare, defense)		

Pass bar: any reasonably-sized bank should be able to deploy self-hosted for KYC, AML, and investigation workloads without custom engineering.

CATEGORY 2 — MULTI-MODEL & VENDOR INDEPENDENCE (WEIGHT: 12%)

A platform that ties the bank to a single AI model creates the same kind of strategic risk as any other monoculture dependency.

Criterion	Score	Notes
Multiple frontier model providers supported in production		
Workflow-level model selection (different models per workflow)		
Switching between models without re-engineering workflows		
Local-model support equivalent to cloud-model support		
Open standards for model interoperability (e.g., model-agnostic prompt definitions)		
Documented model fallback / failover behavior		
No commercial coupling between platform vendor and any single model vendor		

CATEGORY 3 — GOVERNANCE & POLICY (WEIGHT: 15%)

Criterion	Score	Notes
Configurable trust settings per workflow (AI watches → suggests → acts-you-approve → handles routine)		
Trust-setting ceilings enforced at platform level (not policy hope)		
Role-based access control with bank-grade granularity		
Approval workflows with named verifiers per step		
Policy engine for data classification, PII handling, geographic constraints		
Per-workflow ownership, approval, and review cadence		
Integration with the bank's existing IAM / SSO		

CATEGORY 4 — AUDIT, LOGGING & EVIDENCE (WEIGHT: 15%)

The category examiners actually care about.

Criterion	Score	Notes
Every prompt, response, verification step, and approval logged with user attribution		
Tamper-evident audit log (hash chain, WORM, or equivalent)		
Examiner-grade query SLA: “what did AI do on matter X on date Y” answered in 60 seconds		
Export to bank’s SIEM / GRC platform		
Retention configurable to 5-7 years to meet bank records policy		
Documented controls mapped to SOC 2 TSC		
SOC 2 Type II report (recent, scope appropriate)		
Vendor-provided evidence for OCC/Fed/FDIC examiner walkthrough		

CATEGORY 5 — SECURITY (WEIGHT: 12%)

Criterion	Score	Notes
Encryption in transit (TLS 1.2+) and at rest (AES-256 or equivalent)		
Customer-managed encryption keys (CMK) supported		
MFA enforcement for all platform access		
Documented vulnerability management program		
Penetration testing — annual, with executive summary shareable under NDA		
Incident response process with notification SLA in contract		
Sub-processor list published and updated		
ISO 27001 / SOC 2 / equivalent certifications current		
No-training contract clause available without negotiation		
Data residency commitment in contract (US-only, EU-only, etc., as required)		

CATEGORY 6 — REGULATORY ALIGNMENT (WEIGHT: 10%)

Criterion	Score	Notes
Documented mapping to SR 11-7 model risk management		
Supports the bank's BSA/AML supervision requirements (no AI dispositioning or SAR filing)		
Supports fair-lending sampling and review on customer-facing AI		
Compatible with NYDFS Part 500 cyber requirements		
Compatible with FFIEC IT Examination Handbook expectations		
Vendor has banking customers and can name regulatory framework experience		
Vendor's contract language acceptable to bank's legal and third-party risk		
Audit-rights clauses meet FFIEC third-party risk guidance		

CATEGORY 7 — INTEGRATION & EXTENSIBILITY (WEIGHT: 10%)

Criterion	Score	Notes
Native connectors to bank-relevant systems (core banking, AML platform, KYC vendor, LOS, CRM, payments, DMS)		
Documented pattern for adding new connectors without custom engineering		
Self-extending agents that build missing connectors automatically (with IT review before live)		
Public agent protocols supported natively: Google A2A v0.2.5, MCP (client and server), OpenClaw Runtime Gateway		
Webhook and API extensibility		
Versioning of workflows and policies (with rollback)		
Test / sandbox environment separate from production		
Dry-run mode for new workflows or model changes		

CATEGORY 8 — VENDOR VIABILITY & RELATIONSHIP (WEIGHT: 8%)

Criterion	Score	Notes
Vendor company longevity and financial stability indicators		
Banking customer reference list shareable under NDA		
Named executive sponsor for the bank's account		
Documented escalation path for incidents and contractual disputes		
Roadmap shared under NDA with the bank's compliance and IT review		
Reasonable contract exit assistance (data export, knowledge transfer, transition period)		
Insurance: cyber, tech E&O, general liability at limits acceptable to bank		
Documented business continuity and disaster recovery for the platform		

SCORING TEMPLATE

Category	Weight	Score (avg of criteria × weight)	Weighted
1. Deployment flexibility	18%		
2. Multi-model & vendor independence	12%		
3. Governance & policy	15%		
4. Audit, logging & evidence	15%		
5. Security	12%		
6. Regulatory alignment	10%		
7. Integration & extensibility	10%		
8. Vendor viability & relationship	8%		
TOTAL	**100%**		**/100**

DECISION BANDS

Score	Interpretation
85-100	Strong candidate; proceed to contract; expect manageable third-party-risk review
70-84	Viable; address specific gaps before contract; longer third-party-risk review
50-69	Significant gaps; not recommended for material AI deployment without remediation commitments
Below 50	Not viable for banking deployment at this time

RED FLAGS THAT OVERRIDE THE SCORE

Some findings should block a deployment regardless of total score. If any of these are true, do not deploy in production for material workloads.

❌ No no-training clause available in writing ❌ No self-hosted option for KYC, AML, or any workflow touching investigation material ❌ Audit log cannot be searched at examiner SLA, or cannot be exported to bank archive ❌ Vendor cannot name even one banking reference customer ❌ Trust-setting ceilings cannot be enforced at platform level — only as policy guidance ❌ Vendor SOC 2 Type II report is more than 14 months old or scope excludes the platform you'd deploy ❌ No documented incident response plan with notification SLA ❌ Single-vendor model lock-in (no path to switch or run multiple models)

QUESTIONS TO ASK ON THE VENDOR DEMO

Ask all of these. The answers reveal the vendor's posture as much as the demo does.

1. Show me a real banking customer running self-hosted. How long did the deployment take?
2. Show me the audit log for a real workflow. Search for "what did this user do yesterday."
3. Show me how I disable a specific AI model and switch the entire bank to a different one — without rebuilding workflows.
4. Show me how a workflow's trust ceiling is enforced. Try to escalate beyond the ceiling — what stops it?
5. What happens to my data if I terminate the contract? Show me the export.
6. Walk me through your SOC 2 Type II findings and exceptions, not the cover page.
7. Show me your sub-processor list. When was it last updated?
8. If a regulator subpoenas your records for a workflow run by my bank, what's your protocol?
9. Show me a deployment for a bank under NYDFS Part 500. What's different from your standard pattern?
10. What's the longest a customer has been a reference for you? Why?

The vendor's comfort with these questions tells you more than the demo.

THE BUILD-VS.-BUY CONVERSATION

The honest framing for the board:

Path	Time to value	TCO over 3 years	Regulatory comfort	Vendor lock-in
Buy (governed AI orchestration platform)	1-2 quarters	Predictable	High if scorecard ≥ 80	Manageable with multi-model + exit clauses
Build (internal platform)	4-8 quarters	Higher; ongoing engineering	Earned through your own validation	None
Embedded vendor AI (no central platform)	Quick wins	Hidden cost in siloed governance	Low; each vendor's policy is different	High per-vendor
Do nothing	Immediate "no cost"	Compounding shadow-AI risk	Negative; eventual finding	N/A

For most regional and mid-market banks, **buy** is the right answer if a vendor scores ≥ 80. **Build** is rarely justifiable below tier-1 bank scale. **Embedded vendor AI without central platform** is the option that most often becomes a finding.

THE HITLAI / AICTRLNET ANSWER, IN ONE LINE

> AI tools, your team, your systems — running together, safely.

For a bank evaluating governed AI orchestration: a single platform that deploys self-hosted for sensitive workloads and private-tenant for the rest, with multi-model support including open-weight local models via Ollama and vLLM, with trust ceilings enforced at the platform level, with an examiner-grade audit log, and with three public agent protocols spoken natively (Google A2A v0.2.5, MCP client and server, OpenClaw Runtime Gateway) so the platform can talk to your existing systems and to whatever you adopt next. Mapped to OCC, FDIC, Fed, SEC, FinCEN, GLBA, NYDFS, and SOC 2. Reviewed and supervised at every step.

Want a working session on running this scorecard against a real shortlist? That's part of Module B3 — the 60-minute leadership-ready strategy session. → **hitlai.net/institute/banking**

This scorecard is not legal, regulatory, or procurement advice. It is operational guidance for vendor evaluation under the supervision of the bank's third-party risk, compliance, audit, and legal functions.