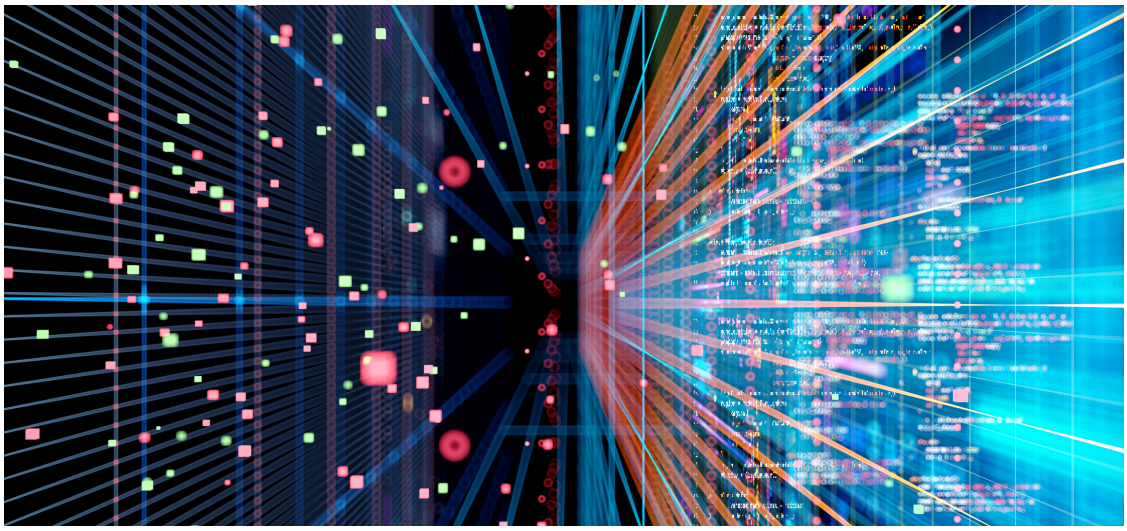
The Agentic AI Maturity Spectrum in Financial Crime

A Chartis Insight



Executive summary

The financial crime and compliance solutions market is shifting from siloed, transaction-level controls toward enterprise-wide, platform-led solutions that unify onboarding (e.g., Know Your Customer (KYC)/ Know Your Business (KYB), etc.), identity verification, payments, fraud risk management anti-money laundering (AML), and customer lifecycle risk. Artificial intelligence (AI), including both generative AI and agentic AI, is the primary force shaping buyers' priorities. The pace of vendors' claims with respect to the use of AI in their solutions has outstripped the market's ability to evaluate them. Vendors are claiming 'agentic AI' capabilities that do not, in our opinion, constitute agentic AI.

This research paper provides a tactical framework that buyers of financial crime technology can use to evaluate the maturity of agentic AI. It defines what agentic AI is and what it is not. It defines a maturity spectrum, from basic automation to fully agentic systems, and applies that spectrum to five core use cases: AML investigations, network analysis, fraud prevention, sanctions screening and onboarding. The objective of this paper is to give buyers a practical tool they can use to:

- Evaluate vendors' claims with respect to AI.
- Compare capabilities.
- Help them make informed procurement decisions.

Voice of the customer: AI risk and reward

AI is increasingly viewed as a strategic enabler in financial crime and fraud management: customers of financial crime management solutions cite its ability to materially reduce operational burdens, improve detection quality and assist in responding more effectively to evolving threats. Customers further state that AI, by automating high-volume review tasks and adapting controls in near real-time, can extend the capacity of lean compliance teams while supporting a more continuous and holistic approach to risk management.

At the same time, executives still have some reservations about:

- The effectiveness of new AI governance controls.
- The ability to scope an agent's authority.
- Escalation design.
- The potential for broader systemic risk as AI becomes more deeply embedded in financial crime control environments.

Customers also state that the paths to agentic financial crime solutions come with both benefits and risks. For example:

Benefits of AI

- **A shift from tools to a digital workforce**. AI agents execute end-to-end processes and reduce time and resource limitations. Humans supervise the higher-risk tasks. For example, AI can give a five-person compliance team the capability to operate at the capacity of a 50-person team.
- **A shift from static controls to adaptive systems**. AI enables controls to evolve dynamically with data and threats. Compliance becomes continuous and predictive, rather than sporadic and reactive. For example, AI agents can monitor transactions, detect a new fraud pattern and then update the detection logic automatically, tightening thresholds or adding a new condition when that pattern starts to form.

Risks of AI

- **A shift from model risk to agent risk**. The evaluation question shifts from 'is the model accurate?' to 'What actions is the AI agent authorized to take, what are the escalation triggers, and can it fail unpredictably?' Agentic AI reduces operational risk at the micro level while simultaneously introducing new systemic and tail risks at the macro level (i.e., coordinated agent failures, novel attack surfaces, concentration risk in shared models).

The agentic maturity spectrum

Rather than a binary 'agentic vs. not agentic' distinction, this paper proposes a four-level maturity spectrum for AI systems and provides examples of use cases that are typically satisfied at each level.

Level 1 – Rule-based automation

Deterministic workflows that are triggered by predefined conditions. These predefined conditions are not based on any AI reasoning or capability. Some examples of rule-based automation that do not involve AI include the following:

- Auto-routing alerts by risk score.
- Batch screening against watchlists.
- Threshold-based transaction blocking.

Level 2 – AI-assisted (copilot)

AI surfaces information, generates summaries or recommends actions in a chatbot style, but leaves all judgment and execution to the human operator. Some examples of AI-assisted activities include the following:

- ChatGPT-style interfaces over case investigations and workflows.
- Single-turn large language model (LLM) outputs that summarize alerts.
- Browser extensions with access to data that is on the open screen.
- Retrieval-augmented generation (RAG)-based document retrieval for investigators.

Level 3 – Agentic with human oversight

AI agent(s) execute multi-step workflows with tool calls and structured reasoning, but require human operator approval at decision points to continue moving through the workflow. The agent(s) perform investigation steps, generate evidence packages and draft narratives. The human operator reviews the narrative evidence and makes the decision. Many systems that are application programming interface (API)-first and do not offer capabilities for humans to conduct the review are limited to Level 3 as the maximum level of AI autonomy, because APIs are ultimately a data transfer mechanism. Some examples of this level of AI include the following:

- AI agents triaging alerts.
- Retrieval of transaction histories for human review.
- Checking of screening watchlists.
- Assembling evidence and documentation.
- Presenting a complete investigation for analyst review.

Level 4 – Autonomous agentic

AI systems that autonomously execute end-to-end workflows within governed constraints can self-evaluate against defined success criteria, and improve results relative to these criteria over time through feedback loops. Human oversight shifts from per-decision review to exception handling and periodic quality assurance (QA). Some examples of fully agentic AI include the following:

- Autonomous alert disposition with configurable escalation triggers.
- Self-improving detection logic based on outcome results and data.
- Cross-workflow coordination between detection and investigation agents.

Most vendors claiming 'agentic AI' in anti-financial crime use cases are operating at *Level 2: AI-assisted (copilot)* or the early stages of *Level 3: Agentic with human oversight*.

The practice of claiming to be using AI agents while actually using a combination of automation or chatbots and legacy processing is known as 'agent washing'. 'AI washing' is the practice of overstating or misrepresenting how or to what extent AI is considered or deployed, often to make it seem that solutions are more advanced or use-case appropriate than what is actually in practice.

Table 1 summarizes the appropriate responses to key questions across the agentic AI levels.

Table 1: Responses based on agentic AI level

Responses based on agentic AI level			
Characteristic	**What to ask**	**Level 2 response**	**Level 3+ response**
Goal decomposition	How are investigations broken into steps?	'AI summarizes the alert.'	'AI agent executes 5-8+ specialized signal agents per alert type, running in parallel with distinct context windows and toolsets per task.'
Context engineering	How do you scope data for each task?	'We send all available data to the model.'	'Each task has engineered context based on what analysts actually use.'
Code generation	How does the agent query data?	'The LLM reads the data directly.'	'AI agent generates SQL/Python to deterministically query specific fields.'
Evaluation sets	How do you validate accuracy?	'We test on sample data.'	'Evaluation datasets from N years of human-reviewed data, benchmarked against top analysts.'
Autonomy ladder	Can I configure automation levels?	'The human always decides.'	'Four levels configurable per queue, process, team or jurisdiction.'
Multi-model	How many models do you use?	'We use [single model name].'	'We orchestrate N models, dynamically selected per task, and updated continuously based on the latest benchmarks against specific task types.'
Feedback loop	How does the AI agent improve?	'We update our models periodically.'	'Analyst feedback aggregates into decision criteria improvements, and administrators have the ability to manually adjust or tune the agents in a self-service manner.'

Source: Chartis Research

Buyers should use the definitions below to understand vendor capabilities and establish realistic expectations.

Section 1: Defining agentic AI

'Agentic AI' refers to AI systems designed to pursue defined goals autonomously by planning, making contextual decisions and executing multi-step actions within governed constraints, often orchestrating multiple models, data sources and workflows. These systems frequently:

- Rely on an AI component – most commonly language-based or combinations of language and statistically oriented AI – to approximate and execute complex internal behavior.
- Exhibit proactive, goal-directed behavior: they initiate tasks, adapt to changing conditions and manage end-to-end processes with limited human intervention, while remaining aligned to policy, regulatory and explainability requirements.
- Operate within an agent-oriented architecture that decomposes large, complex monolithic processes into individualized agents, each with a scoped context and defined capabilities.
- Amplify human decision-making and automate complex, high-value workflows by combining autonomy, orchestration, continuous learning and transparent reasoning.

Section 2: Characteristics of agentic AI in financial crime solutions

Chartis specifies the attributes of agentic AI across seven dimensions. These characteristics should be used as evaluation criteria, and buyers should ask vendors to demonstrate each one against the specific use case being evaluated.

2.1 Design

Clear goal decomposition

Agentic tasks are broken into discrete, verifiable sub-goals that the agent can pursue either sequentially or in parallel. In financial crime solutions, this means decomposing an investigation into specific evidence-gathering steps, each with defined data sources and expected outputs.

Well-defined success criteria

Each step has explicit, testable outcomes so the AI agent can self-evaluate progress. This is where evaluation sets are increasingly important (please see Section 3).

Modular task architecture

Workflows are structured so individual components (tasks) can be delegated, retried or swapped without cascading failures. Mature, fully agentic AI platforms allow buyers to configure which tasks an agent performs per queue or workflow, rather than deploying monolithic agents.

An underappreciated dimension of modularity is who controls it. Systems in which investigation logic is configurable only by the vendor, through professional services engagements or support tickets, create an operational dependency that compounds over time. Institutions whose compliance and fraud teams can define, modify and extend investigation tasks themselves can iterate at the pace of their own risk environment. This distinction – vendor-controlled configuration versus institution-controlled configuration – is one of the clearest indicators of platform maturity.

Buyer evaluation questions include the following:

- Can I configure which tasks the AI agent performs per queue?
- Can I add or remove tasks without re-engineering the AI agent?
- Can my compliance or fraud team define new investigation tasks without an engineering dependency on the vendor?
- Can autonomy levels be set independently per queue or alert type, rather than applied globally?
- What investigation parameters can the institution modify without a vendor support engagement?

2.2 Information and context

Context engineering over prompt engineering

This is the single most important technical differentiator between production-grade agentic systems and demos. Context engineering is the discipline of optimizing what information you provide to the model for the exact decision you want it to make. Too little context and it lacks the information to decide correctly. Too much context and performance degrades. Models, just like people, get confused when given more information than they need to complete a task.

Production-grade agentic systems scope context per task. For example, a sanctions-check AI agent receives only the entity data, watchlist matches and relationship graph relevant to the specific sanctions name check. It does not receive the full transaction history, device data and behavioral analytics that would be relevant to a fraud prevention task.

Persistent state management

Workflows maintain a shared memory, or scratchpad, so AI agents don't lose progress across multi-step tasks.

Structured inputs and outputs

Data passed between steps is typed and schema-consistent, reducing data parsing errors and 'hallucinated' assumptions. The most reliable agentic AI systems generate deterministic code (SQL, Python) to query data, rather than having the LLM directly interpret raw data. This hybrid approach of AI reasoning combined with deterministic code execution is a hallmark of production-grade AI systems.

Buyer evaluation questions include the following:

- Does the AI agent generate deterministic code to query my data, or does the LLM directly interpret raw data?
- How do you scope context per task to avoid performance degradation?

2.3 Control and Oversight

Progressive autonomy

Mature agentic AI platforms offer a configurable autonomy ladder, not a binary on/off switch. Buyers should expect a minimum of four levels:

- **On-demand only**. The AI agent runs when a human analyst requests it.
- **Auto-review with human decision**. The AI agent prepares an evidence package and the human operator conducts the evaluation and makes the decision.

- **Auto-review with decision recommendation**. The AI agent recommends a specific disposition and the human operator either approves or disapproves.
- **Auto-close**. The AI agent makes and executes the decision within configured guardrails.
- **Backtesting**. Testing the process and results against historical data to confirm that desired outcomes are achieved.

Critically, these levels should be configurable per queue and use case, and distinct based on risk tier.

Graceful escalation paths

When the AI agent hits ambiguity or low confidence, there is a defined path to escalate rather than guess. The escalation criteria should be configurable by the buyer.

Audit trails

Every AI agent action is logged and includes reasoning, evidence citations and confidence indicators. This enables review and any necessary debugging, and is regulatorily defensible based on the audit trail.

Buyer evaluation questions include the following:

- Can I configure the level of autonomy per queue and per risk tier?
- What happens when the AI agent hits a low confidence result – does it escalate or guess?

2.4 Tool and resource fit

Multi-model orchestration

Production-grade agentic AI systems do not rely on a single LLM. Different LLMs excel at different tasks (document analysis, code generation, entity extraction, narrative drafting), and LLM benchmarks shift continuously. Mature platforms dynamically select the best-performing LLM for each task and can swap LLMs as benchmarks change, testing against evaluation sets prior to deployment in the production environment.

Right-sized tool availability

AI agents are given access to exactly the tools needed and no more, reducing the risk surface and decision overhead.

Latency-tolerant design

Workflows accept that agentic AI loops take time and, if necessary, adjust the downstream systems so they don't time out prematurely.

Buyer evaluation questions include the following:

- How many LLMs does your platform use?
- How do you select which LLM handles which task?
- What happens when an LLM's benchmark performance changes?

2.5 Feedback and iteration

Tight feedback loops

AI agents receive fast, actionable signals to self-correct in mid-task. Mature systems use an analyst's feedback with respect to the AI agent output, aggregating the analyst feedback and using it to improve decision criteria over time without retraining the underlying models.

Dual-direction quality assurance

The most mature agentic AI deployments involve AI checking the human operator's work *and* human operators checking the AI's work. This dual-direction QA is an indicator of production maturity.

Measurement

Workflow performance is measurable, enabling systematic, prompt and architecture refinement over time. Buyers should ask for specific accuracy metrics benchmarked against high-performing human analysts.

Buyer evaluation questions include the following:

- How does analyst feedback improve the AI agent over time?
- Can the AI agent conduct quality assurance/quality control (QA/QC) on the work produced by a human analyst, as well as the reverse?

2.6 Governance and security

Least-privilege execution

AI agents operate using a defined identity, with minimum permissions required for the task.

Scope constraints

Clear boundaries define what the AI agent can and cannot do.

Deterministic fallback behavior

When the AI agent is uncertain, the default action is safe and conservative, not creative (thus avoiding AI-driven 'hallucinations').

Regulatory compliance in production

Buyers should ask whether the vendor has achieved compliance with the strictest applicable AI regulation (e.g., the EU AI Act) in production, not just on a roadmap.

Buyer evaluation questions include the following:

- Is your institution in compliance with the European Union Artificial Intelligence (EU AI) Act based on what is in the production environment today?
- Can the appropriate report illustrating compliance with the EU AI Act be provided on demand?

2.7 Context engineering (new dimension)

This dimension is elevated to first-class status because it is the primary technical differentiator between production-grade agentic AI systems and marketing demos.

The core challenge

Every LLM has a context window, and even the best models degrade in quality as they approach the limit. A system that dumps all available data into the context window will produce worse results than one that carefully engineers the minimum sufficient context for each specific task.

What agentic AI is not

Buyers should be aware of the unique characteristics of AI systems when making decisions on how best to apply AI capabilities to different use cases. For example:

A chatbot or copilot is not agentic

If the system is prompted to summarize data and relies on the human operator to move the process forward, it is likely either AI-assisted (Level 2), or not agentic AI (Level 3). The defining characteristic of fully agentic AI (Level 4) is that it performs the work, rather than merely informs the work.

A single LLM call is not agentic

Agentic AI systems involve multi-step reasoning, tool use and state management across a workflow. A one-shot prompt that generates a summary of information and/or data is generative AI, not agentic AI.

RAG alone is not agentic AI

Retrieval-augmented generation improves the quality of AI responses by grounding them in relevant data. This is an AI framework, not an AI architecture. A fully agentic AI system architecture may use RAG, but RAG by itself does not make a system comparable to fully agentic AI.

Rule-based workflow automation is not agentic AI

If the system follows a predetermined sequence of steps without AI-driven reasoning or adaptation, it is automation (Level 1), not fully agentic AI (Level 4).

Red flags: how to spot non-agentic claims

When evaluating vendors, the following are indicators that the system may not be truly agentic:

- The vendor cannot explain how context is scoped per task.
- There are no evaluation datasets, or the evaluation datasets are synthetic, rather than derived from real human decisions.
- The 'AI agent' is a chatbot interface over a data lake with no structured workflow execution.
- There is no configurable autonomy ladder – it's all-or-nothing automation (attempting to fully automate an entire process, rather than in stages).
- The vendor cannot produce accuracy metrics benchmarked against human analysts.
- Audit trails show what the AI agent decided, but without the reasoning chain or evidence used to make the decision.
- The system cannot generate deterministic code to conduct data queries.
- There is no feedback loop from human analyst corrections to be used in improving AI agents.

How to evaluate

To evaluate a vendor's context engineering capabilities, ask the vendor to explain how they scope context for a specific task. A credible answer involves: (1) understanding which data fields are needed for the specific decision; (2) using deterministic code to extract only those fields; (3) testing context configurations against evaluation datasets; and (4) iterating as models and data evolve.

The historical data advantage

Vendors with years of human analyst behavior data have a structural advantage in context engineering for the following reasons:

- They know what data analysts actually used when making each type of decision.
- They can engineer AI agent context to learn those patterns.

Buyer evaluation questions include the following:

- Walk me through how you engineer context for [insert the name of the specific task].
  - What data does the agent see, and what data is excluded?
  - How did you arrive at that scoping?

Section 3: Reliability and trust – how to know if AI is real

3.1 The four pillars of AI reliability in financial crime solutions

Pillar 1: Temperature and determinism controls

Production AI systems set model temperature to zero or near-zero for compliance tasks, thus minimizing creative/random output ('hallucinations').

Pillar 2: Deterministic code generation

The AI agent should generate deterministic queries (SQL, Python) to retrieve and process data, rather than having the LLM directly interpret raw data. This hybrid approach dramatically reduces data accuracy errors.

Pillar 3: Evaluation datasets and golden-set benchmarking

Evaluation datasets are test datasets with known correct answers used to validate AI performance before deployment. Key questions for buyers include:

- Does the vendor have evaluation sets for each task type? How many examples?
- What is the source of the eval set? (Historical human review data is gold; synthetic data is weaker).
- Does the vendor benchmark against high-performing analysts specifically (the 'golden set'), or against average human performance?
- What accuracy threshold must be met before a task is deployed?
- Can the buyer see or backtest the evaluation dataset results for their specific configuration?

Pillar 4: Context engineering

(See Section 2.7 above.) The discipline of providing the appropriate and necessary information for each decision, without omitting required information and without including information that is superfluous for completing the decision.

3.2 The accountability framework

Humans are also non-deterministic decision-makers. Financial institutions already manage this through established processes: QA audits, random sampling, training programs and regulatory oversight. The same frameworks apply to agentic AI.

Production-grade agentic AI deployments already implement accountability through:

- **Random sampling of AI agent outputs** for defensibility review and at a level of granularity with the sample.
- **AI QA of human work** – the AI agent audits human analysts' decisions for consistency and standards of practice adherence.
- **Full audit trails** with reasoning chains, evidence citations and confidence scores.
- **Configurable escalation triggers** that pause the AI agent when conditions exceed defined thresholds.

The strategic question is not 'is the model accurate?' but rather 'what actions is the agent authorized to take, what are the escalation triggers, and can it fail unpredictably?' This shifts the evaluation from model risk to agent risk.

Section 4: Use case – AML investigation

What the workflow looks like today

Rule-based systems generate alerts when transactions match predefined patterns. Human analysts review, investigate and disposition alerts. False positive rates of 90%+ drive significant operational costs for AML investigations.

Level 1 – Rule-based automation

Batch-processed, threshold-based rules. Alerts are generated every 24 hours. Analysts manually pull transaction histories, check lists and write narratives.

Level 2 – AI-assisted

AI summarizes alert context and suggests which fields to review. Generative AI drafts narrative templates. But the analyst still performs every investigation step and makes every decision.

Level 3 – Agentic with human oversight (the production frontier in 2026)

AI agents autonomously execute investigation steps: pulling transaction histories, conducting internet research, checking watchlists, assembling evidence packages, running anomaly detection on counterparty behavior, and drafting investigation narratives. Handle time drops by 80%+. False positive volumes drop as the AI agent correctly identifies and auto-closes clear false positives.

Level 4 – Autonomous agentic (emerging)

Detection and investigation layers are connected. The investigation outcomes feed into rule recommendations, and the rules are designed to self-optimize. The system notices anomalies even before rules are developed to identify a potential money laundering pattern and generate an alert. Human oversight shifts from operational processing to exception handling and QA sampling.

Section 5: Use case – network analysis

Defining the scope

Network analysis spans two distinct capabilities:

- **Internal graph/link analysis**. Entity resolution, relationship mapping and network visualization within a single institution's data.
- **Cross-institutional network intelligence (consortium)**. Signals derived from investigation outcomes, fraud typology classifications and entity risk data are aggregated across multiple institutions.

The maturity spectrum applied

Level 1 – Rule-based automation

Static link analysis based on shared identifiers. Batch-processed.

Level 2 – AI-assisted

Machine learning (ML)-driven entity resolution and fuzzy matching. Graph visualization tools. AI agent suggests related entities.

Level 3 – Agentic with human oversight

Network analysis agents autonomously expand entity graphs, identify hidden relationships, and generate network risk assessments. Cross-institutional signals automatically flag entities associated with known bad actors.

Level 4 – Autonomous agentic

Network intelligence compounds automatically. Each institution's investigation outcomes improve network risk signals for all participants. The agent proactively identifies emerging fraud/money laundering networks before individual institutions detect them.

Key evaluation criteria

- Does the consortium data include investigation outcomes and fraud/money laundering typology classifications, or just binary good/bad flags?
- Does the network AI agent see cross-institutional signals, or only your institution's data?
- What is the diversity of the consortium? (Cross-vertical coverage vs. single-sector and cross-typology vs. single typology).
- Can the AI agent proactively flag entities based on cross-network intelligence before they transact on your platform?

Section 6: Use case – fraud prevention

What the workflow looks like today

Analysts review flagged transactions, gather evidence, determine fraud type, document findings and take action with respect to the disposition of the alert/investigation. Many institutions outsource L1 investigations to offshore teams.

The maturity spectrum applied

Level 1 – Rule-based automation

Manual investigation with checklist-based standard operating procedures (SOPs) for fraud identification.

Level 2 – AI-assisted

AI summarizes alert context or drafts narrative templates. The analyst still performs every evidence-gathering step in fraud prevention.

Level 3 – Agentic with human oversight

The investigation AI agent executes the full evidence-gathering workflow: pulling histories, analyzing device data, checking counterparty behavior, running external lookups, and drafting narratives in the institution's required format.

Level 4 – Autonomous agentic

Investigation AI agents coordinate with detection agents. The investigation outcomes automatically inform rule tuning for future identification of fraud. Self-service AI agent building allows fraud teams to create custom investigation tasks without engineering.

Key evaluation criteria

- Can the fraud team build custom investigation tasks without engineering?
- Does the AI agent produce three outputs: recommended action, regulator-ready narrative and transparent worklog?
- Can levels of autonomy be configured differently for different fraud types?
- Does the AI agent adapt and identify new fraud typologies?

Section 7: Use case – sanctions screening

What the workflow looks like today

Entities are checked against government and international watchlists. False match rates are extremely high due to name variations and transliteration differences.

The maturity spectrum applied

Level 1 – Rule-based automation

Batch screening with exact or fuzzy name matching. All potential matches are sent to human reviewers.

Level 2 – AI-assisted

AI agent improves match quality. AI summarizes match details for faster human review.

Level 3 – Agentic with human oversight

Sanctions agents autonomously evaluate the match quality, check corroborating evidence (date of birth, nationality, aliases, etc.), assess relationship proximity for politically exposed person (PEP) screening, and recommend disposition with confidence scores. Clear non-matches are auto-closed by the AI agent.

Level 4 – Autonomous agentic

The sanctions AI agent operates in real time at transaction speed. Adaptive learning incorporates new designations immediately. Cross-list correlation identifies entities sanctioned under different names across jurisdictions.

Key evaluation criteria

- Can the agent distinguish between sanctions screening (binary list check) and PEP screening (relationship analysis)?
- What is the auto-close rate for clear non-matches, and what is the accuracy rate?
- How quickly does the system incorporate new sanctions designations?
- Is the agent specialized for sanctions, or is it a general-purpose agent running sanctions alongside unrelated tasks?

Section 8: Use case – onboarding

What the workflow looks like today

Analysts collect information from the prospective customer/business prior to account opening, and review and confirm that all relevant information is valid and will complete the data collection required based on the account and customer type. Some of the information collected is often used to complete a customer risk rating, which may require the collection of additional data (in the case of enhanced due diligence). The risk rating will also commonly be used to determine the period review frequency.

The maturity spectrum applied

Level 1 – Manual data collection for onboarding

Manual data collection, often involving significant client contact, to collect all data required for customer onboarding, due diligence analysis and customer risk rating.

Level 2 – AI-assisted

AI agents collect certain portions of the data required for customer onboarding. The analyst completes the remaining required information for customer onboarding, and reviews and confirms the dataset for the purposes of completeness, accuracy and customer risk rating.

Level 3 – Agentic with human oversight

The onboarding agentic AI completes the full customer data collection and executes the onboarding case management workflow so that the customer onboarding record is complete, with all required information. An analyst reviews and approves the full dataset for onboarding purposes and runs the customer risk rating.

Level 4 – Autonomous agentic

Agentic AI collects all the data required for customer onboarding through the case management workflow. It then confirms that all information is comprehensive for the customer/account type, and includes the execution of the customer risk rating in the case management workflow.

Key evaluation criteria

- Can the AI agent consume and correctly interpret customer onboarding data from multiple sources?
- Does the AI agent integrate with the customer risk rating process, including determining the appropriate level of due diligence (e.g., customer due diligence [CDD] vs. enhanced due diligence [EDD]). Does it collect any incremental data?
- How does the AI agent address potentially conflicting data from two or more distinct sources?

Section 9: Pain points driving agentic AI adoption

These include large volumes of false positives, poor data quality, evolving fraud and money laundering tactics, cross-channel complexity, and the operational burden of manual review.

The evolutionary path that buyers envision

- **From tools to a digital workforce**. AI agents execute end-to-end processes. Humans supervise the higher-risk tasks. A five-person compliance team can operate at the capacity of a 50-person team.
- **From static controls to adaptive systems**. Controls evolve dynamically with data and threats. Compliance becomes continuous and predictive, rather than sporadic and reactive.
- **From model risk to agent risk**. The evaluation question shifts from 'is the model accurate?' to 'What actions is the AI agent authorized to take, what are the escalation triggers, and can it fail unpredictably?'
- **The tension: micro efficiency vs. macro risk**. Agentic AI reduces operational risk at the micro level while simultaneously introducing new systemic and tail risks at the macro level (such as coordinated agent failures, novel attack surfaces, concentration risk in shared models).

Section 10: Calls to action

For buyers

- Use the AI maturity spectrum to classify vendors' claims. Map capabilities to Levels 1-4.
- Demand transparency around evaluation datasets. Ask every vendor: what are your evaluation datasets, where do they come from, and what accuracy threshold must be met?
- Test the vendor's context engineering claims. Ask it to demonstrate how it scopes context for a specific task.
- Start with one workflow, not a platform overhaul. Begin with a single high-volume workflow. Validate the accuracy in parallel with human review. After completing this for one workflow, expand to an additional workflow. Repeat until all workflows have been addressed.
- Configure AI autonomy progressively. Start at Level 2 or 3 of the AI maturity spectrum to build trust, then move to a higher level as accuracy increases.

- Build accountability frameworks now. Implement random sampling, dual-direction QA and audit trail review to extend existing QA frameworks.
- Address the challenge with the appropriate level of urgency. The problems that buyers are debating today often have known solutions in production. Governance frameworks should enable speed, not prevent action.

For vendors

- Be candid about the level of AI maturity of the solution. Claiming Level 4 when delivering Level 2 damages buyer trust.
- Invest in evaluation datasets, not just models. The model is the engine, but the evaluation dataset is the QA.
- Build for self-service. If deploying a custom AI agent task requires months of engineering, then the system will not scale.

Section 11: What to watch

- **Agent-to-agent commerce**. AI agents making purchases on behalf of consumers break traditional fraud signals. Platforms must adapt their detection logic for agent-initiated transactions.
- **The accountability evolution**. Watch for Financial Crimes Enforcement Network (FinCEN) guidance on task-based AI agents with overlaying manager agents for QA/QC.
- **Use context engineering as a competitive moat**. As frontier models become increasingly commoditized, one of the primary competitive differentiators shifts to how well context is engineered for each task.
- **Self-service AI agent building**. Compliance teams want to build their own AI agents without engineering dependency. In our opinion, vendors that enable self-service AI building will capture significant market share.
- **Cross-institutional intelligence networks**. The value of agentic AI compounds when AI agents operate across a network of institutions. Keep an eye on potential consortium depth and diversity.
- **The interaction model shift**. Users increasingly operate through their own AI agents rather than logging into vendor user interfaces (UIs) directly. Agentic AI platforms will need 'receiving agents' – API and/or Model Context Protocol (MCP) layers that can accept instructions from the buyer's external agent.

About Unit21

Unit21 is a provider of AI risk infrastructure solutions, with over 200 customers across 90 countries. The platform unifies fraud and AML detection, investigation and regulatory filing with AI agents that execute investigations end-to-end, gathering evidence, drafting narratives and filing reports so teams can scale. Its AI agents have processed over 1.5m alerts, and its Fraud Consortium covers 100m+ U.S. consumers across a network of financial institutions, FinTechs and crypto platforms. Every AI decision is explainable, auditable and usable for regulatory compliance. Unit21 is backed by Tiger Global Management, Gradient Ventures (Google's AI-focused venture fund), ICONIQ Capital and others.

Learn more at **unit21.ai.**