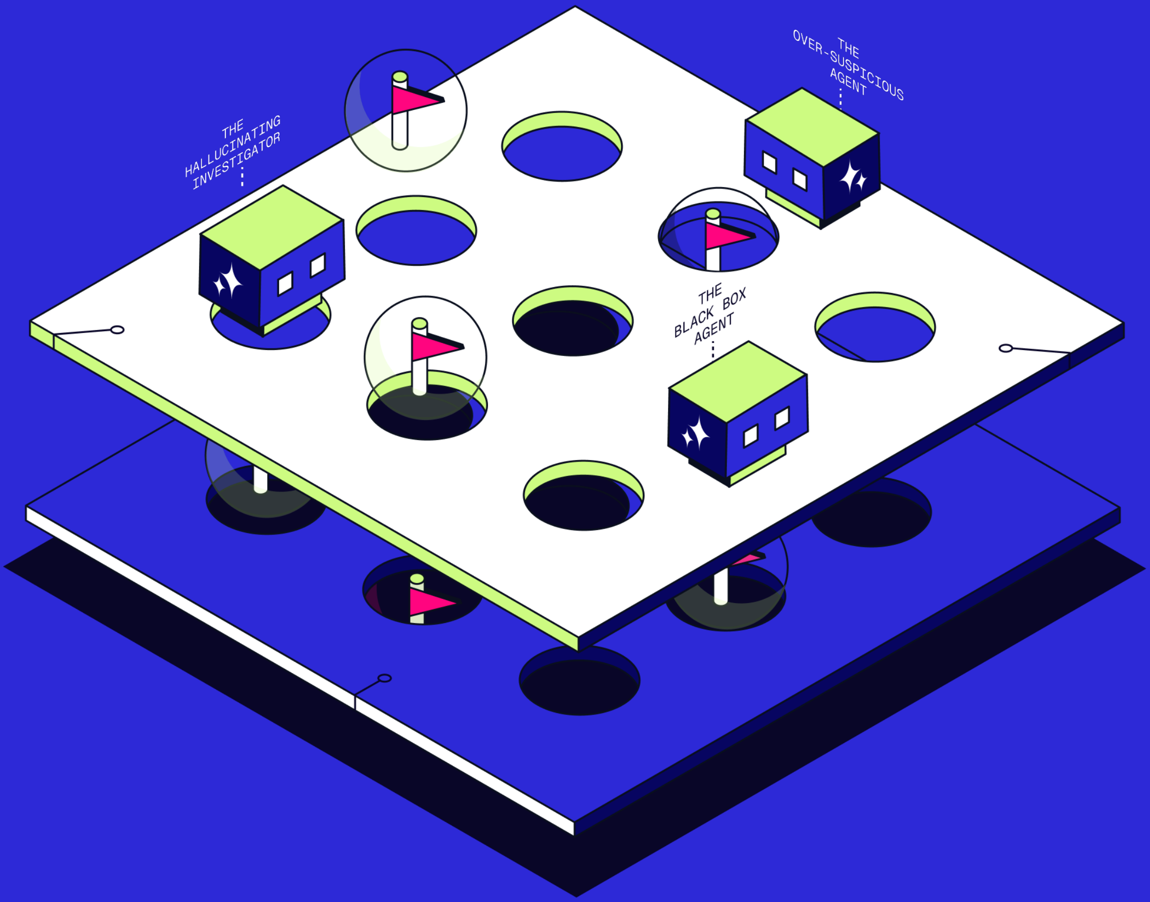
WHITEPAPER

How to deploy trustworthy AI agents for financial crime



Sardine

The era of the manual investigation is ending.

But the era of the autonomous investigator is not yet ready for the auditor.

For the last eighteen months, Agentic AI has been the primary focus of fintech innovation. Yet, within most financial crimes units, these systems remain siloed. They are used for drafting emails or summarizing notes while the core work like interpreting transactional risk and documenting evidence remains tethered to human research speeds.

In 2026, this gap will become a crisis.

As criminals leverage AI to shrink attack cycles from days to seconds, the traditional compliance model of "throwing more bodies" at the alert queue has reached a breaking point. But the rush to automate carries a silent risk: deploying agents that prioritize speed over defensibility. When AI produces a decision that an investigator cannot explain or an auditor cannot trace, the institution has simultaneously automated a process and a liability.

The current industry approach to AI agents is filled with optimistic promises but lacks a rigorous operational framework.

Designed for COOs, heads of fraud, and risk architects who need to move beyond the hype, this paper:

- Exposes the structural flaws in "super-agent" deployments that lead to hallucinations
- Defines the shift from black-box narratives to "data-lineage" as the new regulatory standard
- Introduces the **Atomic Agent** framework for modular, scalable investigations
- Provides a roadmap for shifting human investigators from "data gatherers" to "strategic decision-makers"

As financial crime scales with the deployment of AI, the window for manual intervention is closing. This paper provides the architecture to ensure your operational defense keeps pace without sacrificing defensibility.

Executive summary

Agentic AI in financial crime introduces a fundamental shift in investigative logic: the transition from human led research to autonomous evidentiary construction. As financial crime grows in both velocity and sophistication, traditional operations, which are reliant on manual queues and human contextual review, have reached a point of structural exhaustion. While AI agents offer a scalable alternative for gathering evidence and interpreting risk, their deployment introduces a new set of failure modes that challenge existing regulatory and operational trust assumptions.

In conventional financial crime operations, the integrity of a decision relies on the investigator's ability to document a clear and reproducible chain of reasoning. Agentic flows often fragment this process. When improperly deployed, agents produce narratives that are either hallucinated through excessive context, over sensitized to patterns without domain grounding, or presented as black boxes devoid of audit ready data lineage. In these cases, the AI does not reduce risk; it obscures it, creating a new layer of liability for the institution.

The transition to agentic operations, therefore, requires a shift in how investigative work is decomposed. Rather than attempting to build singular, omniscient models, effective defense relies on the use of **Atomic Agents.** These are specialized units designed to execute narrow investigative primitives such as Data Analysis, OSINT, and KYB. By chaining these specialized agents into structured workflows, institutions can maintain the documentation standards required by regulators while achieving the speed necessary to counter AI driven fraud.

This paper outlines the three primary failure modes of AI agents in financial crime, analyzes where traditional investigative controls degrade, and proposes a layered architecture for deploying agents that are not only efficient but fundamentally defensible.

01

The current landscape

We've hit a breaking point

Fraud isn’t just increasing, it’s accelerating faster than any human team can respond

A typical financial services company has an enormous percentage of its workforce tied up in investigation-heavy workflows: fraud operations, transaction monitoring, sanctions review, onboarding compliance, disputes, and customer support.

Fraud isn’t just increasing, it’s accelerating faster than any human team can respond. Scammers are using AI. Deepfakes are getting cheaper and more convincing. Attack cycles are shrinking. And yet inside most institutions, the day-to-day reality looks the same: alert queues dominated by false positives, customers locked out for no clear reason, legitimate businesses stuck in review limbo.

Just two years ago, the only lever banks and fintechs could pull to address it was to throw even more warm bodies at it. But it’s 2026, and the industry is doing what it always does when manual work becomes unscalable: it’s turning to automation. Only this time, it actually has the means to do so with Agentic AI.

Specifically, AI systems that can gather evidence, interpret context, produce investigation-ready conclusions, and in some cases, even take action autonomously. And yet, despite all the excitement, most early deployments of AI agents in financial crime fail. Not because the models are weak, but because how they are deployed is wrong.

The risk of incorrect deployment

How so?

In regulated environments, an agent that produces confident narratives without guardrails doesn't reduce risk, it creates a new one. And risk teams, surprisingly, don't like risk. These teams quickly realize they end up spending their time on validating, correcting, and hand-holding AI agents. They aren't more efficient. They aren't spending their time on more important tasks. Frustration builds, and the project gets bogged down, or in the worst case scenario, it gets completely sidelined.

The evidence of success

But if you deploy it correctly? When we A/B tested on Sardine's own on-ramp platform, we found that flows supported by AI agents didn't just reduce labor. **They also showed 49% faster time-to-revenue.**[1] That's a quantifiable impact on business growth, not just efficiency savings.

However, to reach these metrics, teams must first navigate the structural traps that sink most AI projects. After building and deploying AI agents in real-world fraud and AML workflows, we've seen three predictable failure modes show up again and again. **Understanding these is the difference between building a force multiplier and building a liability.**

[1] Sardine: The Agentic AI Oversight Framework
https://www.sardine.ai/whitepapers/the-agentic-ai-oversight-framework

02

The 3 failure modes of AI Agents

Failure mode #1: The Hallucinating Investigator

Failure mode #2: The Over-Suspicious Agent

Failure mode #3: The Black Box Agent

The 3 failure modes of AI Agents

MODE	WHAT IT LOOKS LIKE	ROOT CAUSE	BUSINESS IMPACT	HOW TO AVOID IT
The Hallucinating Investigator	Confident, well-written narratives that include subtle inaccuracies or unsupported assumptions	Agents given too much context, too broad a mandate, and vague prompts	Hidden errors, regulatory risk, loss of trust in AI outputs	Use **atomic agents** with narrow scopes and deterministic data retrieval
The Over-suspicious Agent	Flags normal business behavior as suspicious; escalates too many alerts	Pattern recognition without domain grounding; no structured reasoning framework	Increased false positives, analyst fatigue, new bottlenecks	Force agents to reason within structured investigative questions and domain constraints
The Black Box Agent	Produces conclusions without clear evidence trails or reproducible logic	Narrative generation without documentation architecture	Decisions that cannot be defended to auditors or regulators	Design agents to behave like structured investigators, such as deterministic pulls, clear sourcing, explicit uncertainty

Failure mode #1: The Hallucinating Investigator

The most common mistake teams make is deploying AI agents with too much context and capabilities. The thinking is understandable: ***“We have a lot of data. Let’s feed the model everything. It will figure it out.”*** In practice, this is the fastest path to hallucination.

The problem isn’t that modern models are incapable. The problem is that fincrime investigations are fundamentally adversarial and evidence-driven, and large, open-ended prompts increase the chance that the model fills gaps with assumptions. If you ask an agent to “review this customer’s entire profile and determine if this is suspicious,” you’ve given it a task that is too vague, too broad, and too interpretive. And the model will do what models do best: it will generate a plausible narrative. Even if that narrative is wrong.

This gets worse by the fact that hallucinations in fraud and AML are often subtle. The agent might infer a business category incorrectly, assume a relationship between two counterparties, or claim that a transaction is unusual when it is completely normal for that industry.

Less context often produces more reliable conclusions. Avoid deploying a “super-agent” that knows everything. Instead, implement a set of narrow agents that each know how to do one thing extremely well. This is one of the reasons why Sardine is focused on developing what we call “atomic agents”, or AI agents designed around specific investigation primitives. For example:

- A **Data Analyst** agent that can interpret transactional data
- An **OSINT agent** that can perform external research and summarize findings
- A **KYB agent** that can validate business identity and ownership relationships
- A **Graph Analyst agent** that can interpret entity networks and shared identifiers

The fundamental design philosophy is that agents do not replace humans, but instead replace a specific skill. Since today’s fraud investigations require a variety of skills, and each step in the process demands a dedicated AI agent. This approach reduces hallucination risk dramatically because each agent operates inside a smaller decision boundary. Instead of “guessing” what matters, it is asked to retrieve, interpret, and summarize specific data in a specific context.

Failure mode #2: The Over-Suspicious Agent

The second failure mode is more operational than technical. Even when AI agents don’t hallucinate, they often suffer from a different flaw: they are too quick to assume something is nefarious.

This isn’t because models are “paranoid.” Fraud detection is inherently pattern-driven, and models are trained to recognize signals. But in financial crime, signal recognition without contextual grounding produces a predictable outcome: over-escalation.

The agent sees a high-value payment to a foreign counterparty and flags it. The agent sees multiple entities connected through shared addresses and assumes it’s a shell network. The agent sees circular flows and suggests layering.

To a human reviewer, the output sounds intelligent. It uses the right vocabulary and "sounds" like an experienced investigator. But it often misses the most important reality of fincrime operations: most alerts are false positives. This is where many AI deployments accidentally create a new bottleneck by increasing workloads, not reducing them. Critically, the goal of an AI Agent isn't just to find "more" fraud; it's to identify **legitimate behavior** at scale so investigators can focus on the truly anomalous.

This failure mode shows up most clearly in transaction monitoring, where a huge percentage of flagged behavior is not criminal at all. A good example is **self-dealing**, one of the most common patterns that triggers alerts, especially for high-volume businesses. It can look like suspicious circular movement, but in reality it is often just money moving across related accounts, subsidiaries, or internal entities.

To an AI agent trained on "fraud signals," these can look like suspicious anomalies. To a real investigator, they are normal business operations.

Why does this happen? True domain expertise is the insights we have on that data. And these insights are by default outside of the dataset. Tinkering with the dataset or with prompts can only get you so far. To inject real-life domain expertise we need to design agents that are explicitly trained and constrained to ask the right questions:

- Does this counterparty category make sense for this business?
- Is this transaction self-dealing or third party?
- Are the entities connected through ownership, shared identifiers, or operational structure?
- Is this pattern consistent with the industry's normal flows?

When agents are forced to reason in these frameworks, they become far less likely to "jump to fraud" as the default conclusion.

Failure mode #3: The Black Box Agent

Even if your AI agent avoids hallucination and over-suspicion, it can still fail in the most important way: **It can produce conclusions that are not defensible.**

In financial crime, accuracy is not enough. If an agent flags a business as suspicious, a compliance team still has to answer: What evidence did the agent use? What data sources did it rely on? Can the decision be reproduced? Would an auditor accept this reasoning? Would a regulator accept it? Within the 2026 compliance environment, an AI's "reasoning" is only as good as its data lineage; if the specific source is not cited and reproducible, the conclusion effectively does not exist for audit purposes.

This is where many AI tools break down. They output a narrative, but they don't show the chain of evidence. They cite external information vaguely without clear sourcing. They provide a recommendation without documenting why. But in regulated environments, **AI cannot be a black box.**[2] It has to be an evidence machine.

A good AI agent should behave less like a chatbot and more like a structured investigator:

- It should pull data deterministically (e.g., from defined queries)
- It should summarize findings in consistent formats
- It should explain why a counterparty is relevant
- It should highlight uncertainty instead of masking it

This changes our definition of what "black box AI" means. **It's no longer about interpreting a model's internal weights; rather, it's about auditing its external evidence.** When agents operate inside investigation workflows, such as alert queues, case management, analyst escalation, they need to produce outputs that humans can approve, defend, and audit. In agentic workflows, documentation is as vital as reasoning.

[2] Sardine Blog: Agentic AI in Banking
https://www.sardine.ai/blog/agentic-ai-in-banking

03

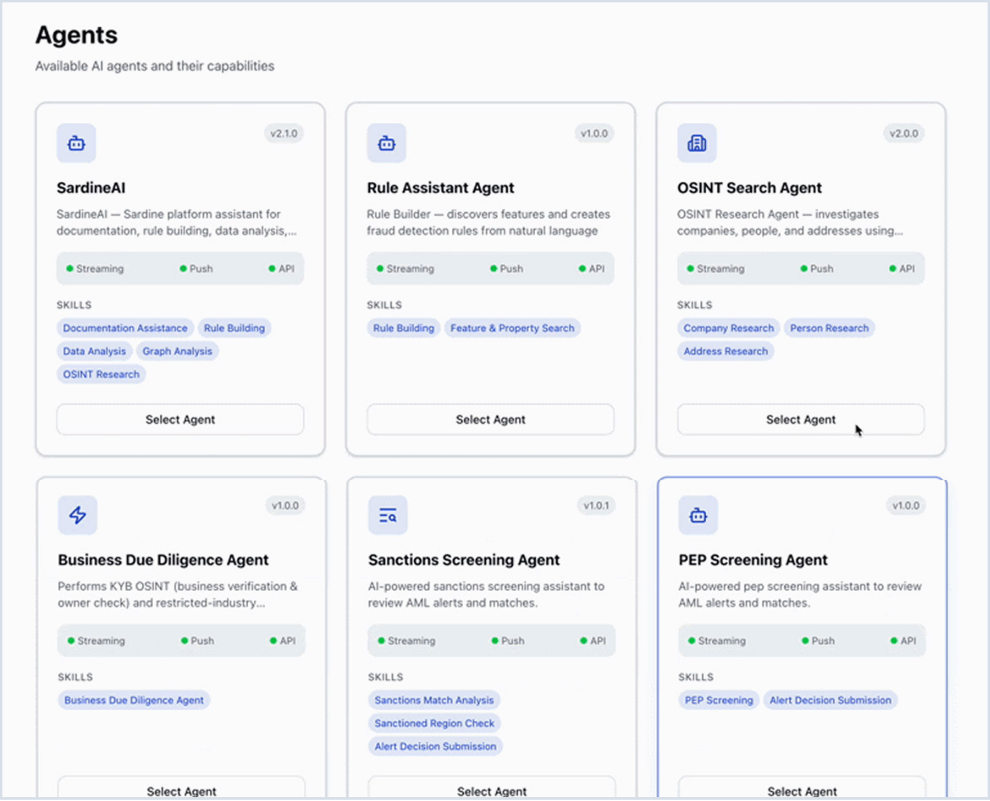
How to deploy AI agents for optimized ROI

Sardine's AI Agent Garden

An AI Agent garden of investigation skills

If hallucination, over-suspicion, and black-box outputs are the predictable failure modes, then the solution isn't incremental tuning. It's structural. You don't fix these problems by asking the model to "be more careful." You fix them by redesigning how investigation work is decomposed, executed, and documented.

That's the foundation behind how we built **Sardine's AI Agent Garden**, a set of atomic agents, each one replacing a specific investigation skill, not a whole investigator.



Agents

Available AI agents and their capabilities

SardineAI v2.1.0

SardineAI — Sardine platform assistant for documentation, rule building, data analysis,...

● Streaming ● Push ● API

SKILLS

Documentation Assistance · Rule Building · Data Analysis · Graph Analysis · OSINT Research

Select Agent

Rule Assistant Agent v1.0.0

Rule Builder — discovers features and creates fraud detection rules from natural language

● Streaming ● Push ● API

SKILLS

Rule Building · Feature & Property Search

Select Agent

OSINT Search Agent v2.0.0

OSINT Research Agent — investigates companies, people, and addresses using...

● Streaming ● Push ● API

SKILLS

Company Research · Person Research · Address Research

Select Agent

Business Due Diligence Agent v1.0.0

Performs KYB OSINT (business verification & owner check) and restricted-industry...

● Streaming ● Push ● API

SKILLS

Business Due Diligence Agent

Select Agent

Sanctions Screening Agent v1.0.1

AI-powered sanctions screening assistant to review AML alerts and matches.

● Streaming ● Push ● API

SKILLS

Sanctions Match Analysis · Sanctioned Region Check · Alert Decision Submission

Select Agent

PEP Screening Agent v1.0.0

AI-powered pep screening assistant to review AML alerts and matches.

● Streaming ● Push ● API

SKILLS

PEP Screening · Alert Decision Submission

Select Agent

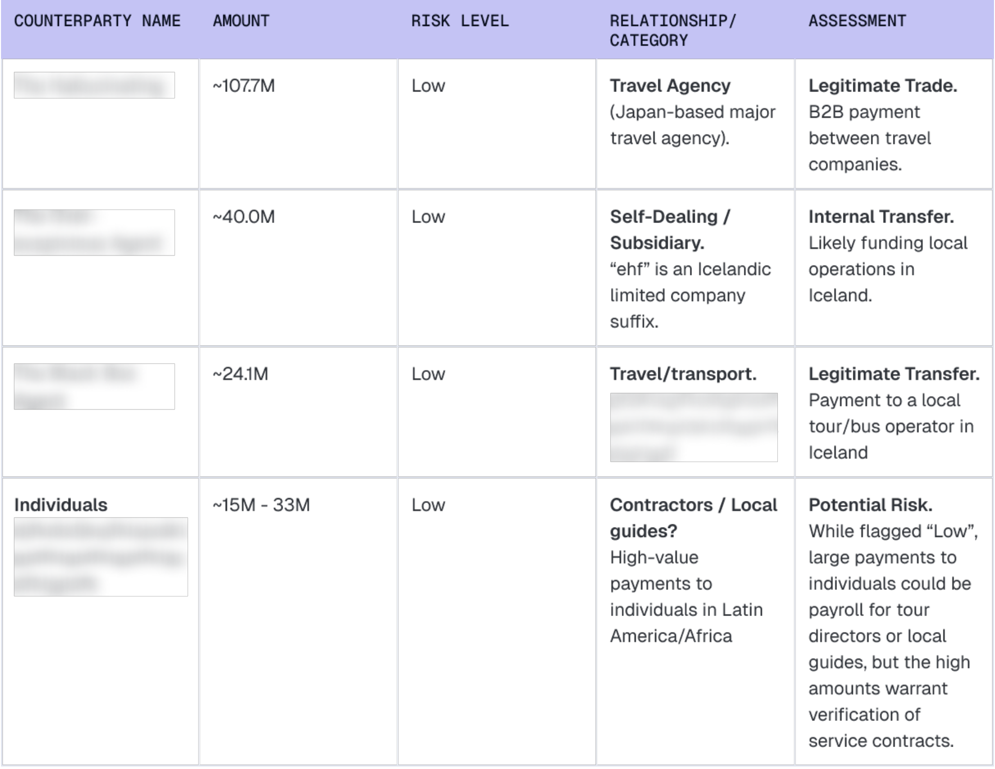
Our goal isn't to create a single omniscient "fraud brain" that can do everything. Instead, we're creating a system that behaves like a seasoned investigation team: specialists that do their piece of the work, leave a clean trail, and hand off to the next step.

The Data Analyst Agent: “What happened?”

Most investigations start with the same problem: the raw data is unreadable at human speed. Transaction monitoring is the clearest example. You’re staring at hundreds or thousands of transfers and trying to answer questions that are simple to ask but painful to answer: ***Who are the top counterparties? What’s the distribution of risk? What’s outward versus inward flow? Is the “weird” behavior a pattern or a one-off?***

The Data Analyst Agent is designed to do that first pass, fast. It produces a structured view of activity, including:

- Sessions and transactions in the recent time window.
- Ranked counterparties.
- High-value and high-risk items surfaced first
- Context like directionality and timing.



COUNTERPARTY NAME	AMOUNT	RISK LEVEL	RELATIONSHIP/ CATEGORY	ASSESSMENT
[redacted]	~107.7M	Low	**Travel Agency** (Japan-based major travel agency).	**Legitimate Trade.** B2B payment between travel companies.
[redacted]	~40.0M	Low	**Self-Dealing / Subsidiary.** “ehf” is an Icelandic limited company suffix.	**Internal Transfer.** Likely funding local operations in Iceland.
[redacted]	~24.1M	Low	**Travel/transport.** [redacted]	**Legitimate Transfer.** Payment to a local tour/bus operator in Iceland
Individuals [redacted]	~15M - 33M	Low	**Contractors / Local guides?** High-value payments to individuals in Latin America/Africa	**Potential Risk.** While flagged “Low”, large payments to individuals could be payroll for tour directors or local guides, but the high amounts warrant verification of service contracts.

In our Transaction Monitoring setup, that starts from deterministic retrieval, or pulling a defined set of transactions within a defined time range so the agent isn’t improvising what it analyzes. This compresses data querying and analysis into a single brief that a human can scan and immediately know where to look next.

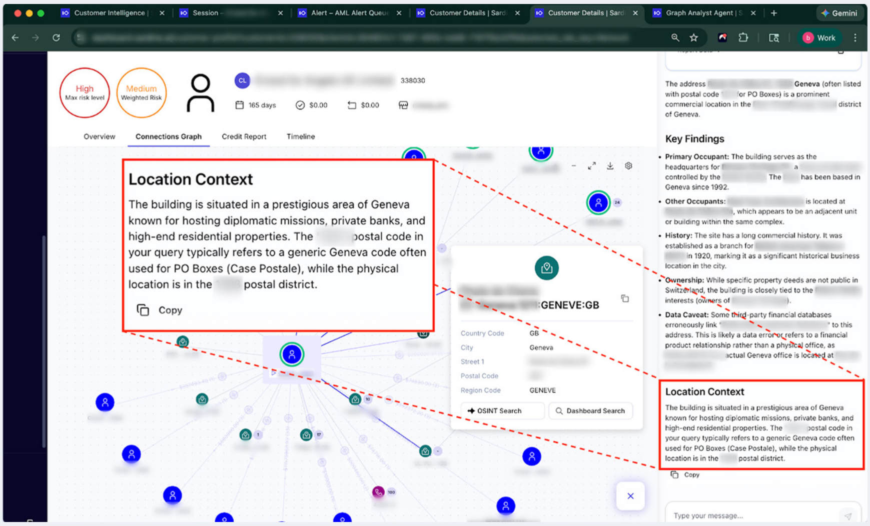
The OSINT Agent: “What’s the context?”

Once you can see what happened, the next bottleneck is interpretation. Most compliance teams aren’t short on data, they’re short on context. A counterparty name by itself rarely tells you whether a transaction makes sense.

The OSINT Agent is built for the simplest but most time-consuming task: turning an entity string into an intelligible business category and a credibility check: Is this a real business? What do they do? Are they plausibly connected to the customer’s line of business? Do they have signals that indicate mismatch, shell behavior, or outright fabrication?

Crucially, this isn’t done for every counterparty. It’s done for the ones that matter: the highest-risk and highest-value counterparties surfaced earlier. That’s how you keep OSINT from becoming an endless rabbit hole (and token sink) and turn it into a repeatable investigative step.

A good example is when a human investigator is looking at a business connection graph and sees a high-risk entity tied to an address in Geneva. On paper, the address looks “legit,” but it could just as easily be a virtual office, a PO box, or a shared services location used by dozens of unrelated businesses.



Location Context

The building is situated in a prestigious area of Geneva known for hosting diplomatic missions, private banks, and high-end residential properties. The [redacted] postal code in your query typically refers to a generic Geneva code often used for PO Boxes (Case Postale), while the physical location is in the [redacted] postal district.

[redacted]:GENEVE:GB

Country Code	GB
City	Geneva
Street 1	[redacted]
Postal Code	[redacted]
Region Code	GENEVE

The address [redacted] Geneva (often listed with postal code [redacted] for PO Boxes) is a prominent commercial location in the [redacted] district of Geneva.

Key Findings

- **Primary Occupant:** The building serves as the headquarters for [redacted] a [redacted] controlled by the [redacted] The [redacted] has been based in Geneva since 1992.
- **Other Occupants:** [redacted] is located at [redacted], which appears to be an adjacent unit or building within the same complex.
- **History:** The site has a long commercial history. It was established as a branch for [redacted] in 1920, marking it as a significant historical business location in the city.
- **Ownership:** While specific property deeds are not public in Switzerland, the building is closely tied to the [redacted] interests (owners of [redacted]).
- **Data Caveat:** Some third-party financial databases erroneously link "[redacted]" to this address. This is likely a data error or refers to a financial product relationship rather than a physical office, as [redacted] actual Geneva office is located at [redacted]

Location Context

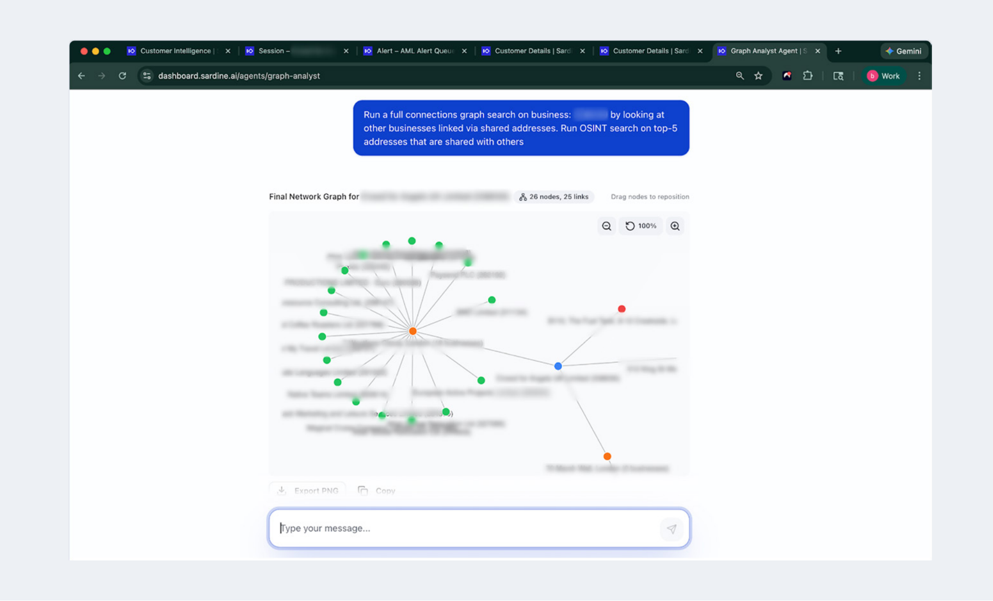
The building is situated in a prestigious area of Geneva known for hosting diplomatic missions, private banks, and high-end residential properties. The [redacted] postal code in your query typically refers to a generic Geneva code often used for PO Boxes (Case Postale), while the physical location is in the [redacted] postal district.

Instead of forcing an analyst to manually research the address across multiple sources, the OSINT Agent runs directly from the address node in the graph and produces a structured summary of what that location actually represents. In this case, the agent finds plenty of data that positions the address as an exclusive office building located in a prestigious area of Geneva. But it also highlights the fact that it's likely being used as a PO Box hub that can potentially harbor illicit activity. In fincrime, investigators don't just need more data; they need to know which data is misleading.

The Graph Analyst Agent: "What's connected that shouldn't be?"

If the OSINT Agent tells you what an address is, then the Graph Analyst Agent tells you who else is using it. In the example above, once the OSINT Agent clarified the Genevan address context and highlighted the possibility of shared usage and data inconsistencies, the next logical question wasn't "is this address prestigious?" It was: *how many other businesses are actually tied to it?*

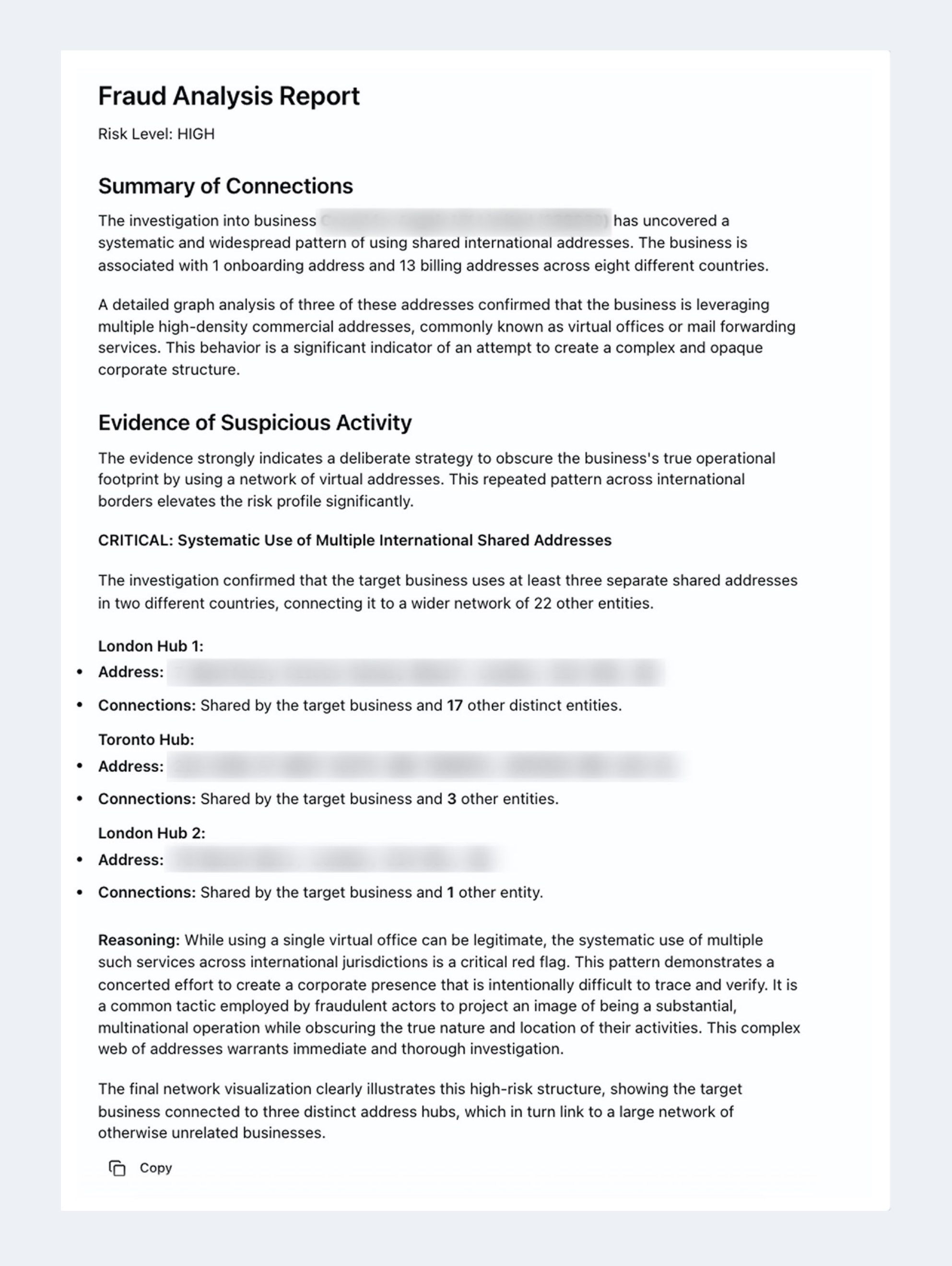
Instead of forcing an investigator to manually investigate all related entities one by one, the agent runs a full Connections Graph search on the business and surfaces every linked entity through shared identifiers. In the case shown, the graph immediately surfaced multiple businesses connected through shared addresses, including one address that was shared across a large cluster of entities.



Run a full connections graph search on business: [redacted] by looking at other businesses linked via shared addresses. Run OSINT search on top-5 addresses that are shared with others

Final Network Graph for [redacted] 26 nodes, 25 links

What matters here is not the visualization itself. It's the pattern. The Graph Analyst Agent summarizes the network structure: How many nodes and links are present? Which identifiers are shared most frequently? Where do the densest clusters sit? It then translates that structure into an explicit risk assessment, allowing investigators to quickly see both what the agent concluded as well as why it concluded it.



Fraud Analysis Report

Risk Level: HIGH

Summary of Connections

The investigation into business [redacted] has uncovered a systematic and widespread pattern of using shared international addresses. The business is associated with 1 onboarding address and 13 billing addresses across eight different countries.

A detailed graph analysis of three of these addresses confirmed that the business is leveraging multiple high-density commercial addresses, commonly known as virtual offices or mail forwarding services. This behavior is a significant indicator of an attempt to create a complex and opaque corporate structure.

Evidence of Suspicious Activity

The evidence strongly indicates a deliberate strategy to obscure the business's true operational footprint by using a network of virtual addresses. This repeated pattern across international borders elevates the risk profile significantly.

CRITICAL: Systematic Use of Multiple International Shared Addresses

The investigation confirmed that the target business uses at least three separate shared addresses in two different countries, connecting it to a wider network of 22 other entities.

London Hub 1:

- **Address:** [redacted]
- **Connections:** Shared by the target business and **17** other distinct entities.

Toronto Hub:

- **Address:** [redacted]
- **Connections:** Shared by the target business and **3** other entities.

London Hub 2:

- **Address:** [redacted]
- **Connections:** Shared by the target business and **1** other entity.

Reasoning: While using a single virtual office can be legitimate, the systematic use of multiple such services across international jurisdictions is a critical red flag. This pattern demonstrates a concerted effort to create a corporate presence that is intentionally difficult to trace and verify. It is a common tactic employed by fraudulent actors to project an image of being a substantial, multinational operation while obscuring the true nature and location of their activities. This complex web of addresses warrants immediate and thorough investigation.

The final network visualization clearly illustrates this high-risk structure, showing the target business connected to three distinct address hubs, which in turn link to a large network of otherwise unrelated businesses.

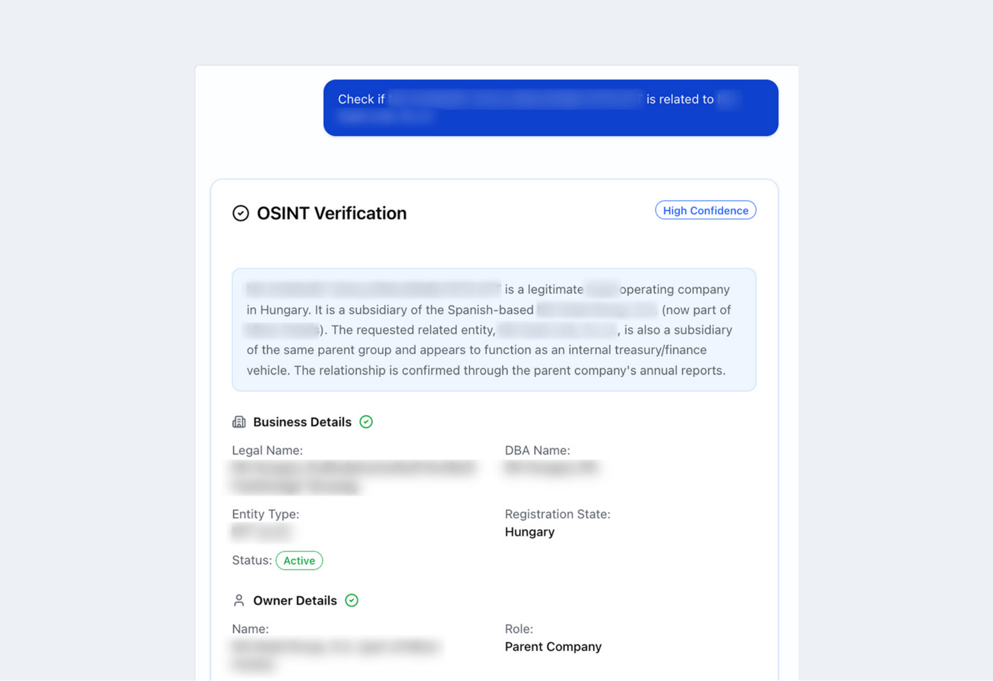
This interplay is deliberate. The Graph Agent identifies structural patterns like shared addresses, recurring nodes, and clusters, while the OSINT Agent helps contextualize those high-signal nodes. Together, they transform what would normally be dozens of manual searches into a single coherent view of how a business sits inside its broader network.

The KYB Agent: “Are these businesses actually related?”

A surprising amount of alert fatigue is caused by one simple limitation: most monitoring systems don’t understand corporate structures. When a group operates with subsidiaries, internal financing vehicles, treasury entities, shared directors, or brand-adjacent names, the system treats those transactions like third-party risk.

Our KYB Agent is designed to answer a question investigators ask constantly and painfully: *are these two businesses actually connected?*

For example, an analyst looks at a cross-border transaction between Hungarian and Spanish entities. The KYB agent immediately exposes that the two are subsidiaries of the same parent company, with one operating as the group’s treasury function. This provides a perfectly legitimate explanation for a transaction that would otherwise be flagged as high-risk.



Check if [redacted] is related to [redacted]

[redacted] is a legitimate [redacted] operating company in Hungary. It is a subsidiary of the Spanish-based [redacted] (now part of [redacted]). The requested related entity, [redacted], is also a subsidiary of the same parent group and appears to function as an internal treasury/finance vehicle. The relationship is confirmed through the parent company's annual reports.

Business Details ✓

Legal Name: [redacted]

DBA Name: [redacted]

Entity Type: [redacted]

Registration State: Hungary

Status: Active

Owner Details ✓

Name: [redacted]

Role: Parent Company

04

Chaining and hybrid AI agents

How we deploy as holistic workflows

Chaining and hybrid AI agents

If you stop at individual agents, you end up with an AI toolbelt. Helpful, but not transformational. The real leverage comes from how we deploy these agents as holistic workflows.

Chaining: The sequential workflow

Chaining allows different agents to work sequentially within the same case investigation:

- The **Data Analyst** agent produces the short list.
- The **OSINT agent** enriches the list with external data.
- The **Graph Analysis agent** adds network risk.
- The **KYB agent** confirms relationships.

This keeps each step narrow and testable. It prevents the agent from jumping to conclusions early, because the system is designed to gather evidence before it concludes. Most importantly, it produces documentation as a byproduct, which is the only kind of AI output that survives a regulated reality. This architecture transforms the AI from an opaque "Magic 8-Ball" into a standardized, transparent assembly line for evidentiary construction.

Hybrid deployment: Co-pilot vs. autonomous

Every business has different comfort levels around automation. This is why we made the AI Garden usable in multiple modes:

- **Co-pilot:** The agent assists a human-in-the-loop for complex investigations.
- **Fully Autonomous:** Resolving high-confidence “green” checks or low-risk alerts without human intervention.

05

The real promise of agentic AI

Agentic AI is the beginning of a new operating model where evidence is gathered automatically and false positives don’t consume entire teams.

Redesigning investigations, one step at a time

But that future happens when agents are deployed thoughtfully with guardrails and modularity. Ultimately, Agentic AI does not seek to replace the investigator, but rather to eliminate the manual drudgery of research, facilitating a shift where the human moves from a "searcher" of data to a "decision-maker" of cases.

Our implementation advice: Don’t start with a grand vision of an autonomous fraud brain. Start smaller. Pick one workflow that is high volume and repetitive. Decompose it into skills. Replace one skill, not the whole investigator. Measure the error rate, time saved, and false positives.

If you can’t audit it, don’t automate it. If you can’t explain it, don’t deploy it.

AI agents in fincrime shouldn’t feel like a leap of faith. They should feel like hiring a very fast, very literal junior analyst, with guardrails.

In practice, this means you don’t “roll out AI.” You redesign investigations, one step at a time.