

AI Workflows versus Swarm AI



Comparing investigation quality through real security cases. What two AI approaches found, what they missed, and how the evidence changed the decision.

Fluency Security | October 2026

Abstract

This paper compares AI workflow and swarm investigation through Fluency's reported 253-incident evaluation and five complete case examples. Both approaches operated after the same ML and UEBA detection front end identified cases. In the documented examples, swarm investigation identified an unsafe organization-wide OAuth grant that workflow triage would have closed, categorized security issues more precisely, and produced clearer, incident-specific response recommendations. Raw logs, historical context, and tenant-wide comparisons helped distinguish failed credential abuse, cross-host malware activity, and administrative changes requiring human verification.

The study reports 238 false-alarm closures and 15 escalations or confirmations for human review. These are dispositions, not independently measured accuracy. The available source does not include the full case dataset or execution traces, so the results demonstrate investigative gains in the documented examples without isolating parallelism as their cause. The central finding is that AI investigation quality should be judged by evidence, verdicts, scope, and useful next actions.

Contents

Section	Page
01 What are Workflow AI and Swarm AI	2
02 Comparing investigation quality through real case results	3
03 Where the investigation changed the decision	3
04 Where investigation expanded incident scope	4
05 How the AI processes differ	5
06 Comparing the investigation responses	6
07 Measure quality before claiming superiority	7
08 What this means for buyers and research	8

What are Workflow AI and Swarm AI

Workflow AI evolves the SOAR playbook

The workflow used in this comparison collected evidence through an AI-executed playbook, as described by Fluency. It looked up the user’s configuration and recent activity, reviewed the latest logins when login data was available, including GeoIP context, and examined notable events in the case time window. It then used that evidence to assess the case and recommend closure or escalation.

The cluster’s apparent attack type selected the playbook, which shaped the data collection and subsequent analysis. This is the migration of Security Orchestration, Automation and Response (SOAR) into AI execution: AI implements the investigative steps, while the playbook defines their order and branches. The distinction is the prescribed path through evidence, not an absence of evidence collection.

Swarm AI searches across schemas in parallel

In Fluency’s approach, events reside across multiple data lakes and data schemes. Swarm tasks search for the user entity according to the schema of the table being queried. Each search is tailored to that data type, including how the entity is represented, which fields hold useful context, and how the returned events should be interpreted.

This is parallel analysis across different event schemas, rather than a division into generic identity, permissions, or endpoint roles. Schema-specific searches are more targeted and efficient, as described by Fluency, and better suited to the structure of each source. The study does not quantify the efficiency gain.

A map and reduce analogy with diverse tasks

The map stage distributes the entity investigation across data lakes and tables. Each task uses the relevant schema to retrieve and analyze events. These tasks are diverse because their data types, entity fields, and analytical methods differ. The reduce stage synthesizes those findings into incident scope, a verdict, and a proposed response.

The parallel tasks need not wait on one another’s preliminary conclusions. This removes the serial analytical gating of a playbook while retaining genuine data prerequisites and a synthesis stage that reconciles evidence and uncertainty.

Dimension	Workflow AI	Swarm AI
Analysis structure	AI executes attack-specific playbook steps	Parallel entity searches tailored to event schemas
Data collection	Playbook shapes which context is collected	Each task searches the relevant lake and table
Decision	Reached through the workflow	Synthesized from schema-specific findings

How the study labels its outputs

Both approaches received cases from the same ML and UEBA detection front end in the video account. The PDF compares AI Assist triage with AI Investigator, which pulls raw logs, checks history, and compares activity across the tenant. The video frames these as workflow and swarm. [1, p. 1; 2]

Comparing investigation quality through real case results

Not all AI SOC's, AI SIEMs, or AI investigations produce the same quality of analysis. In the documented Fluency comparisons, swarm investigation improved recognition of material risk, categorization after detection, and the clarity of recommended responses. [1, pp. 1-4; 2]

1 Swarm identified material risk the workflow missed

The workflow recommended closing an OAuth consent case. The swarm identified a personally registered application with organization-wide mailbox permissions and an Entra risky-application signal, then escalated it. This was a missed unsafe grant requiring human verification; malicious intent or a completed attack was not established. [1, pp. 2-3]

2 Swarm produced more precise issue categorization

ML and UEBA identified behavior worth investigating. Swarm investigation then explained its security significance: an unsuccessful targeted password attack, a dropper affecting two hosts, an unsafe application grant, and an external inbox forward requiring authorization checks. Detection found the cases; investigation classified what the evidence supported. [1, pp. 1-4; 2]

3 Swarm responses were clearer and more specific

The investigator moved beyond escalation labels to incident-specific next actions: isolate both affected hosts, confirm and narrow or revoke the application grant, verify the external forward through a trusted channel, or remove the phishing file and confirm removal. These were recommendations, not records of actions executed. [1, pp. 1-4]

The reported result

253 incidents	238 closures	15 escalations
Investigated end to end	False alarms reported closed with evidence	Escalated or confirmed for human review

These totals correspond to 94.1% reported closures and 5.9% escalations. They describe disposition, not independently measured accuracy, recall, or safety. The five complete examples in the available PDF comprise two confirmed threats and three cases escalated to a human. [1, pp. 1-4]

The findings support better investigative depth in these examples. They do not establish that every swarm outperforms every workflow or that parallel execution alone explains the gains.

Where the investigation changed the decision

Organization-wide OAuth consent

Initial recommendation: Close ticket. Triage recognized successful admin consent for all principals, including Mail.ReadWrite and Mail.Send, and treated correlated activity as consistent. Investigator disposition: Escalated to a human. [1, pp. 2-3]

The investigation connected the grant to a Global Administrator who personally registered the application with a local developer-tool redirect URI. It also found Entra audit evidence of a risky application at consent time. The apparently

new ISP was the tenant's most common ISP, weakening a takeover interpretation without making the grant safe.

Why it matters: A successful and internally consistent administrative action can still create unsafe exposure. The investigator separated account-takeover evidence from permission risk. Its recommended next action was to confirm the grant with the administrator, then revoke or narrow it. The record supports an unsafe grant requiring human ownership; it does not establish malicious intent or completed mailbox exfiltration.

An external inbox forward

Initial recommendation: Escalate to Tier 2. Triage saw rule creation and disablement from corporate egress, but no clear external forwarding behavior. Investigator disposition: Escalated to a human. [1, pp. 3-4]

The investigator tested eight grounds against raw audit and sign-in logs and cleared seven: the activity belonged to one MFA-satisfied web session on the user's registered laptop. The remaining concern was material: a rebuilt field-service alert rule forwarded outside the company and was the tenant's only external forward created in 30 days.

Why it matters: The final disposition stayed escalated, but its rationale became more precise. The investigator reduced unsupported suspicion while retaining a concrete exception for verification. It considered a sanctioned second mailbox plausible and recommended confirming with the user through a trusted channel. This is unresolved authorization risk, not proof of business email compromise.

What these two examples demonstrate

The OAuth case changed the decision from closure to escalation. The inbox case improved the explanation behind an unchanged escalation. A quality comparison should measure both decision correctness and the usefulness of the reasoning, rather than count different verdicts alone.

04

Where investigation expanded incident scope

A dropper on two hosts

Triage identified repeated SentinelOne detections of Fee Calculator.exe and related artifacts within a nine-minute window on one host, with no recorded containment. The investigator used the same SHA-256 to connect activity across two hosts and multiple dates, finding 23 detections with no mitigation. [1, pp. 1-2]

A control search found 155 kills and 47 quarantines on 14 other hosts in the same window. That comparison supported the interpretation that containment operated elsewhere but had not addressed these detections. Outlook cache artifacts suggested email delivery. The investigator recommended isolating both hosts, remediating by hash, and examining both users' mailboxes.

Quality gain: Broader affected-asset scope and a control comparison that made the absence of mitigation meaningful. The PDF's headline and narrative describe the recurrence timing differently; the cross-host finding is clear, but the exact timeline should be reconciled against the raw events before publication as a chronology.

A targeted failed password attack

Triage recognized an Okta lockout but lacked baseline and history. The investigator pulled the raw Okta System Log and found 19 failed logins from nine residential IP addresses in the United States, Türkiye, India, and Morocco over 6 hours 41 minutes. No other account among roughly 150 with failed logins showed that pattern, and the source IPs did not target other accounts. [1, p. 2]

The source describes a targeted credential-stuffing run against a Global Administrator account. All attempts failed on the password; MFA was never reached. The evidence supports attempted credential abuse, not successful access. The recommendations included phishing-resistant MFA, just-in-time privileged access, and blocking the source

addresses.

Quality gain: Account-specific scope, tenant comparison, and an explicit boundary between attack and compromise. The PDF reports 19 failed logins; the supplied video description reports 18 failed passwords. These may count different event categories, but the PDF does not explain the difference.

A phishing file in an endpoint blind spot

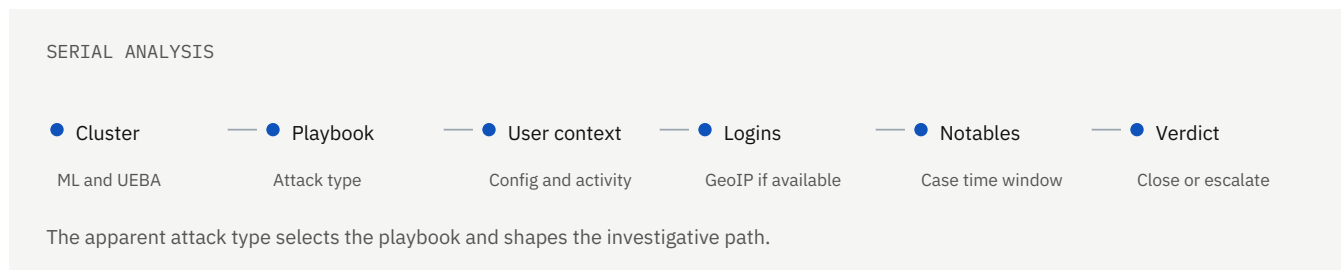
Triage escalated an agentless Defender detection on a shared VDI host. The investigator recognized three of five triage grounds as placeholders common to the tenant's agentless alerts. Two concerns remained: real-time EDR had not flagged the branded SVG lure, and removal was unconfirmed. Prior blocked phishing attachments on the same host supplied historical context. The case remained escalated, with removal verification and analysis of the detection gap recommended. [1, p. 4]

05

How the AI processes differ

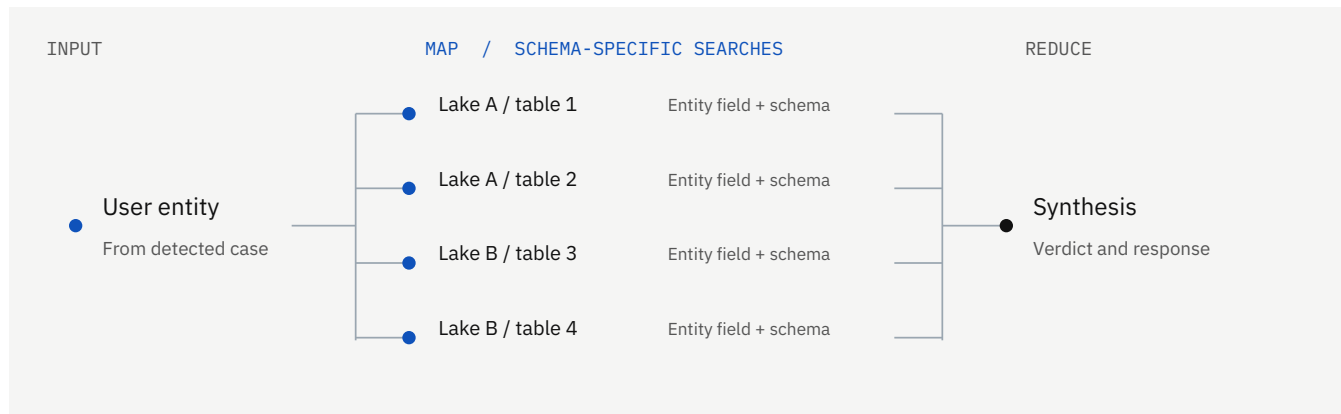
Both approaches start with a case identified by ML and UEBA. Workflow AI executes a prescribed investigative path. Swarm AI searches the user entity across different event schemas in multiple data lakes, then synthesizes the findings. These diagrams illustrate the process distinction; they are not execution traces from the case study.

Figure 1 Workflow AI executes a dependent sequence



The workflow collected user configuration, recent activity, available login and GeoIP data, and notable events. Its attack-specific playbook organized that collection and analysis.

Figure 2 Swarm AI searches across data lakes and schemas



Each task searches the user entity using its table's schema, then interprets that event type. Synthesis combines the findings. Lake and table identifiers illustrate the structure, not a specific deployment.

The dependency that changes

Process property	Workflow AI	Swarm AI
Task selection	Attack type selects the playbook	Event schema defines each entity search
Analytical dependency	Conclusions can gate subsequent analysis	No required conclusion-to-conclusion chain
Final decision	Produced through the prescribed path	Depends on synthesis of collected findings

Removing serial gating lets diverse analyses contribute before a decision is formed. Data prerequisites and evidence reconciliation still apply.

06

Comparing the investigation responses

Table 3 From triage recommendation to a specific response

Case	Workflow triage	Swarm finding and next action
OAuth grant	Close ticket	Unsafe mailbox grant. Confirm with admin; revoke or narrow permissions.
Dropper	Escalate to Tier 2	Two hosts, 23 unmitigated detections. Isolate both; remediate by hash.
Password attack	Escalate to Tier 2	Targeted failed attack. Strengthen MFA and privileged access.
Inbox rule	Escalate to Tier 2	Only external forward in 30 days. Confirm authorization with user.
Phishing file	Escalate to Tier 2	EDR blind spot and unconfirmed removal. Remove, verify, investigate gap.

Workflow triage and swarm findings correspond to AI Assist and AI Investigator in the PDF. The response column is a concise paraphrase; recommendations are not completed actions. [1, pp. 1-4]

What changes in the response

The workflow collected user and event context before explaining its closure or escalation recommendation. The investigator’s published outputs added more precise incident categorization, scope, and specific next steps. The OAuth case changes the decision itself; the other examples make an escalation more useful by explaining the affected scope, unresolved risk, and verification needed.

Why task diversity matters

Different event schemas contribute different evidence. Each parallel task searches the user entity in the appropriate table and interprets its events using that data type’s fields and semantics. Synthesis connects these findings across lakes into a response written for the incident. The case narratives show the resulting scope and interpretation; they do not document the individual search traces.

What the comparison does not establish

Clarity and specificity here are qualitative judgments based on the published narratives. The study does not provide blinded response ratings or evidence that the recommendations were executed. The response comparison should therefore be read alongside the original case facts and the evaluation limits.

Measure quality before claiming superiority

The visible cases demonstrate added evidence and more specific conclusions. They do not provide a complete confusion matrix or an independently adjudicated benchmark. Reporting a 94.1% closure share as 94.1% accuracy would erase the distinction between how many cases were closed and whether those closures were correct.

Evaluation question	What this source supports	What a stronger study needs
Did the system close safely?	238 false-alarm closures reported in aggregate	Review of closed cases and missed-threat rate
Did it catch material risk?	OAuth closure recommendation corrected to escalation	Independent case labels and false-negative counts
Did it scope the incident?	Two hosts linked by hash; tenant-wide comparisons	Asset and identity scope completeness against ground truth
Did parallelism help?	Video frames the techniques as workflow and swarm	Matched models, tools, budgets and execution traces
Was it efficient?	No latency or cost measurements supplied	Wall time, query count, tokens and cost per case
Did response succeed?	Recommendations recorded	Approval, execution and verified remediation outcomes

A defensible next comparison

Run both designs on a defined, identical case set with equivalent telemetry access and a fixed evaluation budget. Preserve the evidence window, model version, prompts, tools, failures, and decision trace. Have reviewers adjudicate outcomes without knowing which design produced them, and track disagreements explicitly.

Score unsafe closures separately from unnecessary escalations. Measure evidence completeness, affected-entity scope, calibration of uncertainty, specificity of recommendations, and cost per correctly resolved case. Repeat runs to identify variability, and test improvements on cases outside the examples used to develop them.

Risk score is not a verdict

The PDF includes a Linux workstation case with a score of 72,800 but stops before its triage and closure. No benign or malicious conclusion can be drawn from that fragment. The video’s claim that the three highest-scoring cases closed benign cannot be checked from the complete cases in this PDF. A score prioritizes review; evidence determines the disposition. [1, pp. 4-5; 2]

What this means for buyers and research

In the documented comparisons, the swarm approach produced better results than workflow-style triage in recognizing material security issues and proposing responses. The detection front end was shared; the gains appeared in how the AI investigated and interpreted the detected cases. The swarm identified an unsafe grant the workflow would have closed, refined issue categorization, and proposed clearer incident-specific actions.

Swarm analysis removes the serial logic gates that can make one preliminary conclusion determine which analyses happen next. Schema-specific tasks can search and analyze the entity across data lakes without waiting for that conclusion. This appears to improve investigation completeness and the correctness of the resulting decisions in these examples. It is a plausible explanation for the observed gains, rather than a mechanism isolated by the study.

Questions to ask during an evaluation

Can the AI inspect raw events and history, challenge its first interpretation, and follow affected entities beyond the starting alert? Can it explain why closure is safe or why escalation is necessary? Does the provider demonstrate improvement with new evaluations and regression results?

Fluency's research position

Fluency's objective is to deliver the best analysis it can today and keep improving it through results-based research on real incidents. The documented comparisons show why raw evidence, historical context, and tenant-wide checks matter. Continued improvement should be demonstrated through reproducible evaluations and clear reporting of both gains and remaining failures.

For buyers, the practical question is whether the system's logic gates restrict the investigation, and whether its conclusions and responses hold up against the evidence. Ask providers to demonstrate completeness, missed issues, and response quality on real cases. For research, test whether removing analytical gating accounts for the gains under matched models, evidence access, and budgets.

Evaluate Fluency on your own data

Start a free 30-day trial, connect an appropriate source such as Microsoft 365, and review available case evidence and investigation outputs with your team. Judge the conclusions and recommended actions against your environment.

[Start your Fluency evaluation.](#)

References and evidence boundaries

[1] Fluency, AI-Powered-SOC, user-supplied PDF, 5 pages. Aggregate results on p. 1; five complete case comparisons on pp. 1-4; incomplete Linux case on pp. 4-5. The document introduces ten cases, but this file does not contain ten complete comparisons. It is a narrative case summary, not the underlying 253-case event dataset.

[2] Swarm AI vs AI workflow: what 253 cases show, YouTube video NsGC6C-yjJ8 and description supplied with the project. Used for the workflow-versus-swarm framing and stated common front end; claims absent from [1] are identified separately. <https://youtu.be/NsGC6C-yjJ8>

Percentages are calculated from the reported totals. Architectural benefits and evaluation recommendations are analysis, not measured outcomes. This study does not establish an industry-wide best-analysis ranking. Prepared October 13, 2026.