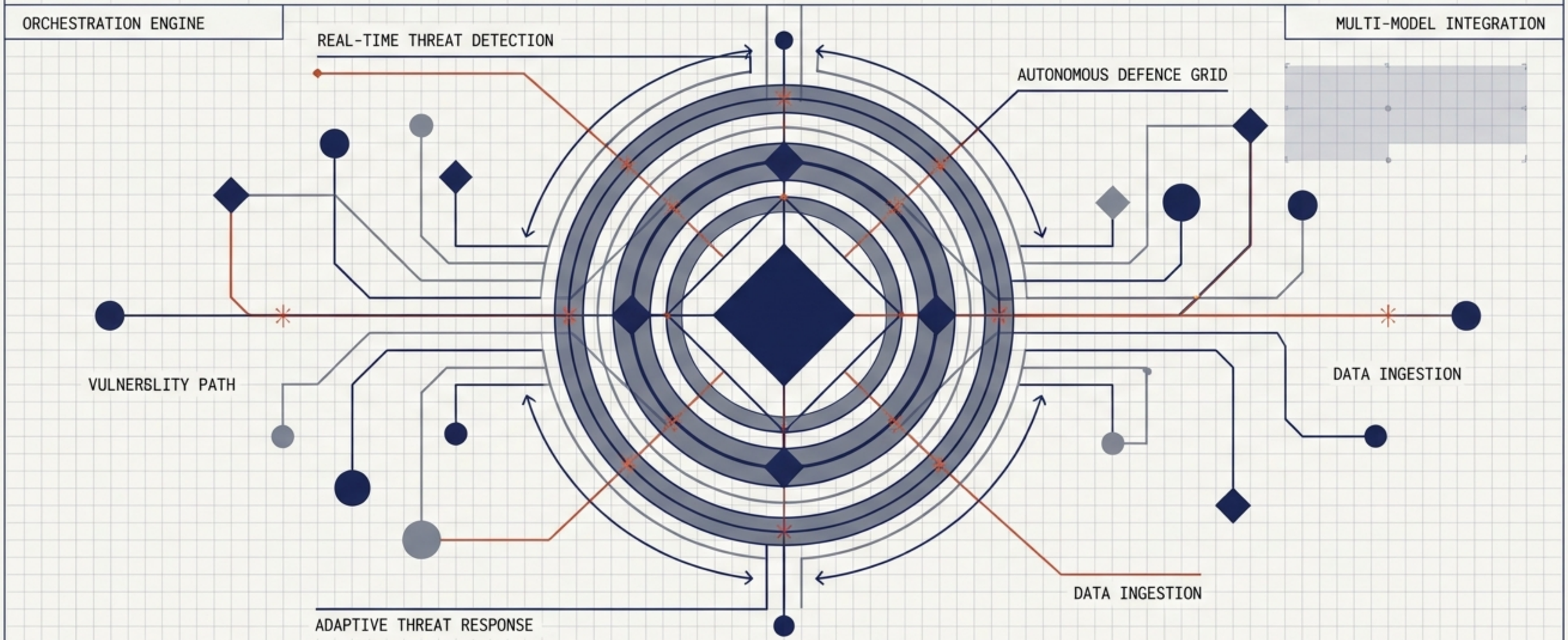
Defence at AI Speed

The Autonomous System Outperforming Single-Model Security



From Research to Production: The 5.12.2026 Cohort

AI vulnerability discovery is actively defending enterprise-scale, high-value targets.

16

Zero-Day Vulnerabilities Discovered & Patched

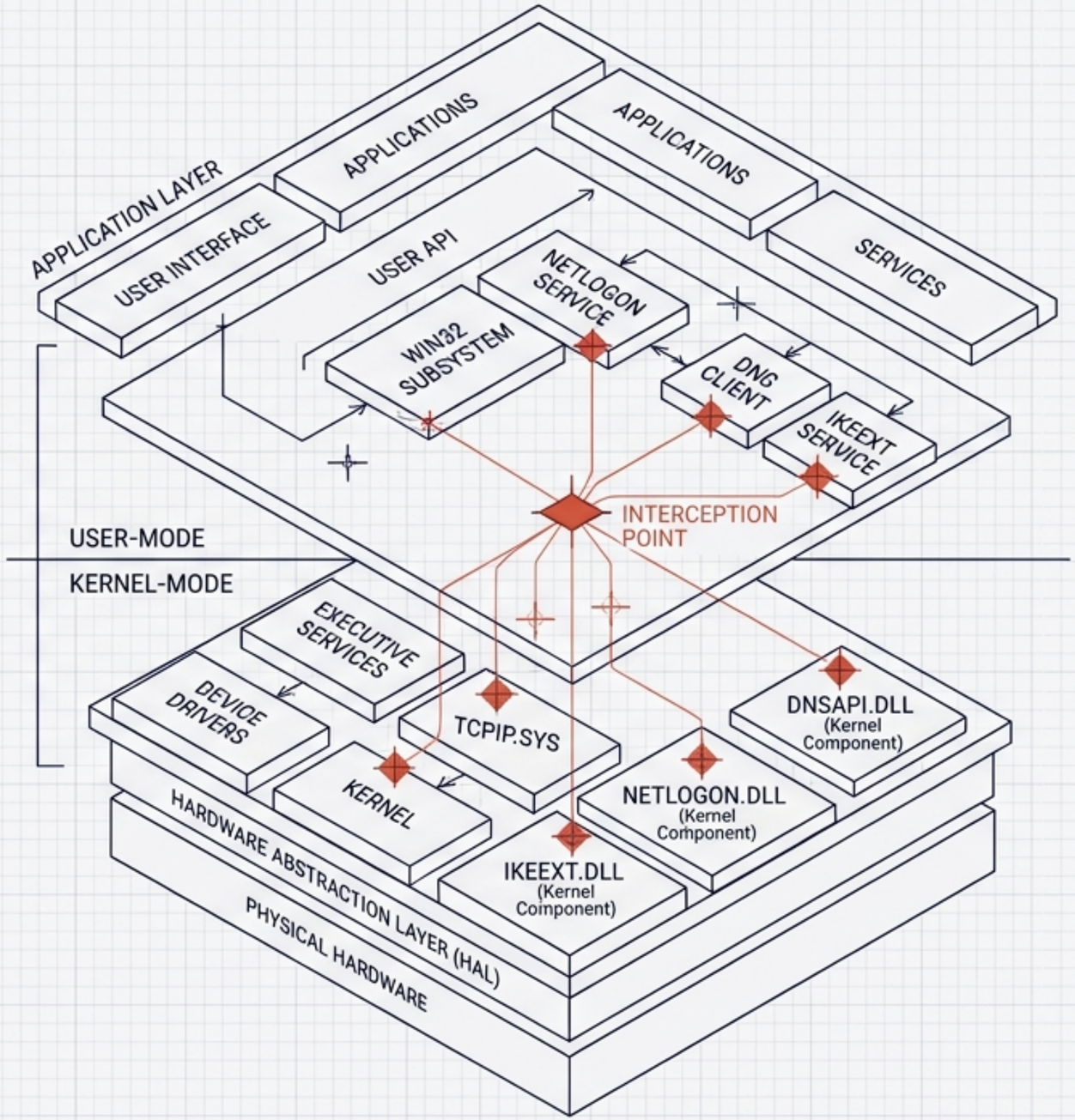
4 Critical RCEs

(Remote Code Execution flaws intercepted, highlighted with vermilion accents).

10 Kernel-Mode / 6 User-Mode

tcpip.sys, ikeext.dll, netlogon.dll, dnsapi.dll

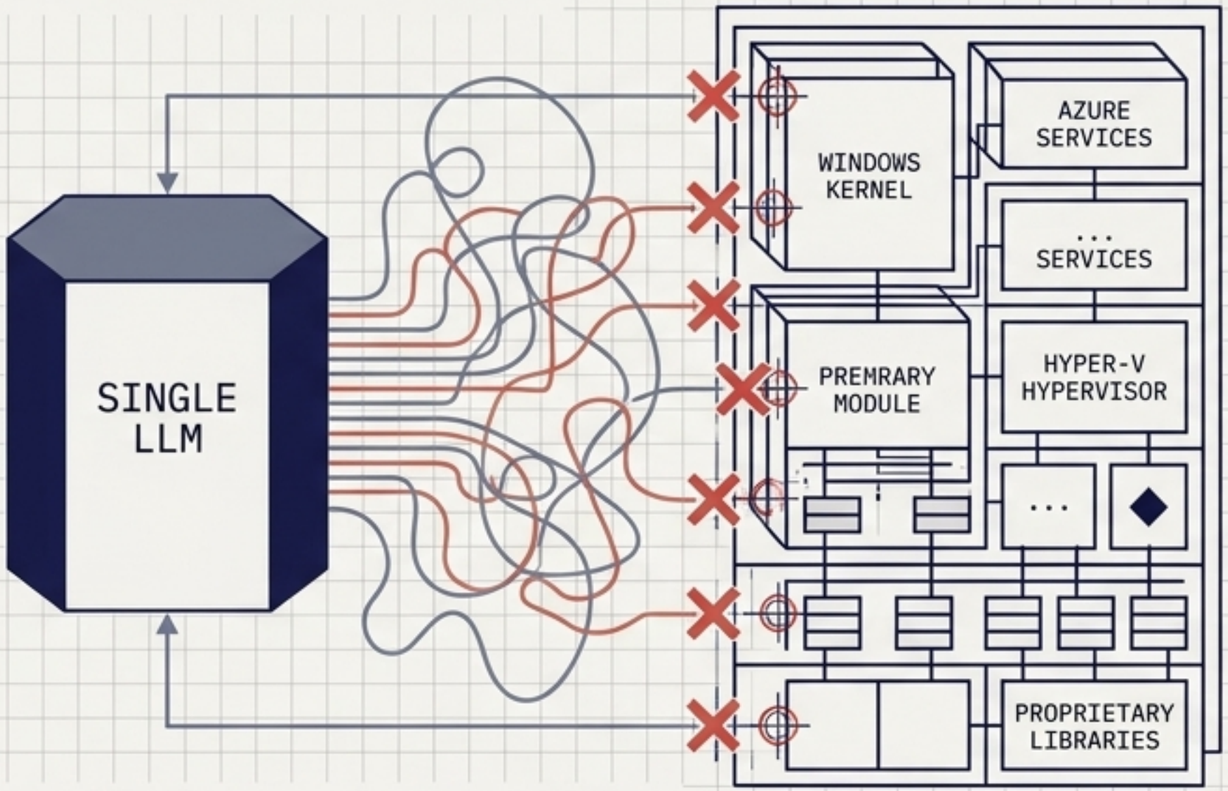
Pre-Authentication: Majority reachable from a network position with no credentials.



The Limits of Single-Model Analysis

Massive proprietary surface areas defeat simple pattern matching. Models must reason, not just memorise.

The Single-Model Failure



Chaotic, pattern-matching attempts fail to navigate complex, unpublished system architectures, resulting in blind spots.

The Three Barriers

Complexity & Scale

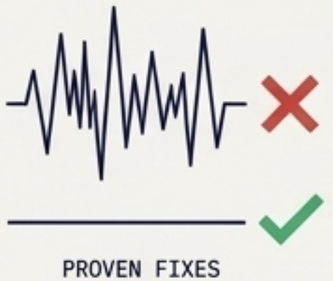
Proprietary Windows, Hyper-V, and Azure codebases are absent from commodity LLM training corpora.

Deep System Invariants

Kernel calling conventions, IRP rules, and IPC trust boundaries demand deep, multi-step reasoning.

The Noise Problem

In DevSecOps at scale, false positives are prohibitively expensive. Single models generate speculative noise; enterprise teams require proven fixes.



Introducing MDASH

Microsoft's Multi-Model Agentic Scanning Harness

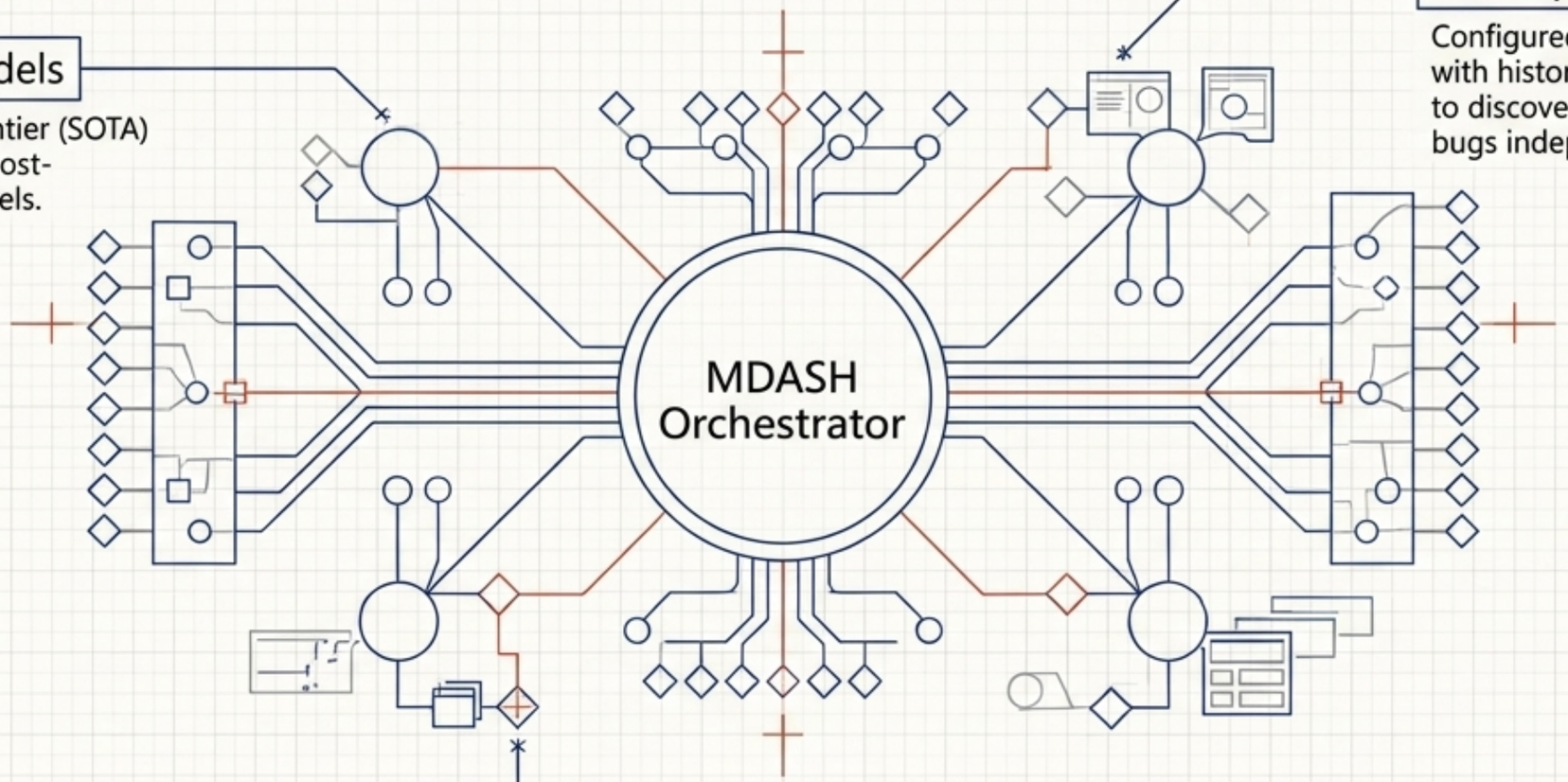
`Core Philosophy: The model is one input. The system is the product.`

Ensemble of Models

Orchestrates both frontier (SOTA) heavy reasoners and cost-effective distilled models.

>100 Specialised Agents

Configured via deep research with historical CVEs and patches to discover, debate, and prove bugs independently.



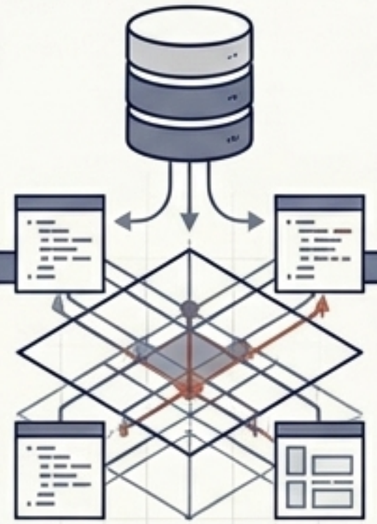
End-to-End Orchestration

Moves beyond isolated scanning to deliver a fully autonomous discovery and remediation pipeline.

The Autonomous Blueprint: A 5-Stage Pipeline

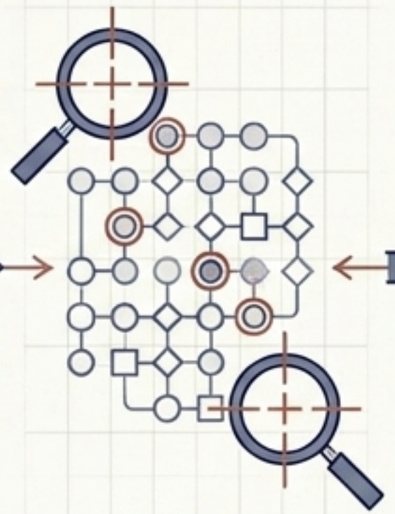
1. Prepare

Ingests source code, builds language-aware indices, maps attack surface, and models threats via commit history.



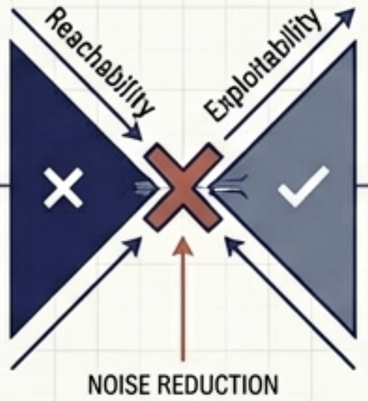
2. Scan

Specialised Auditor Agents target candidate paths, generating hypotheses and initial evidence.



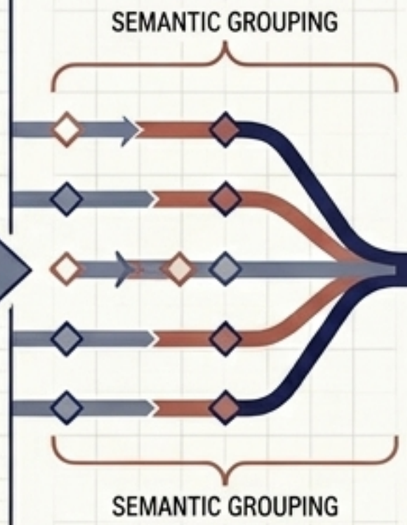
3. Validate

Debater Agents argue for and against reachability and exploitability.
(Noise reduction).



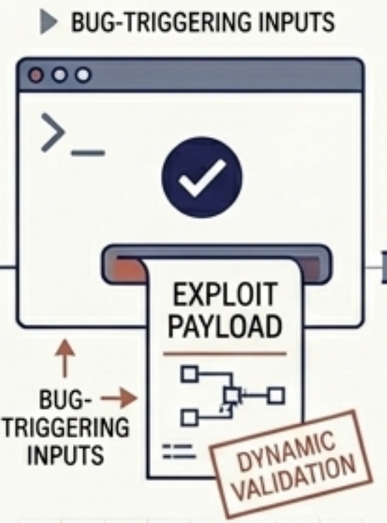
4. Dedup

Collapses semantically equivalent findings and groups by required patch.



5. Prove

Formulates bug-triggering inputs to validate pre-conditions dynamically.



1. Prepare

Ingests source code, builds language-aware indices, maps attack surface, and threats via commit history.

2. Scan

Specialised Auditor Agents target candidate paths, generating hypotheses and initial evidence.

3. Validate

Debater Agents argue for and against reachability and exploitability. (Noise reduction).

4. Dedup

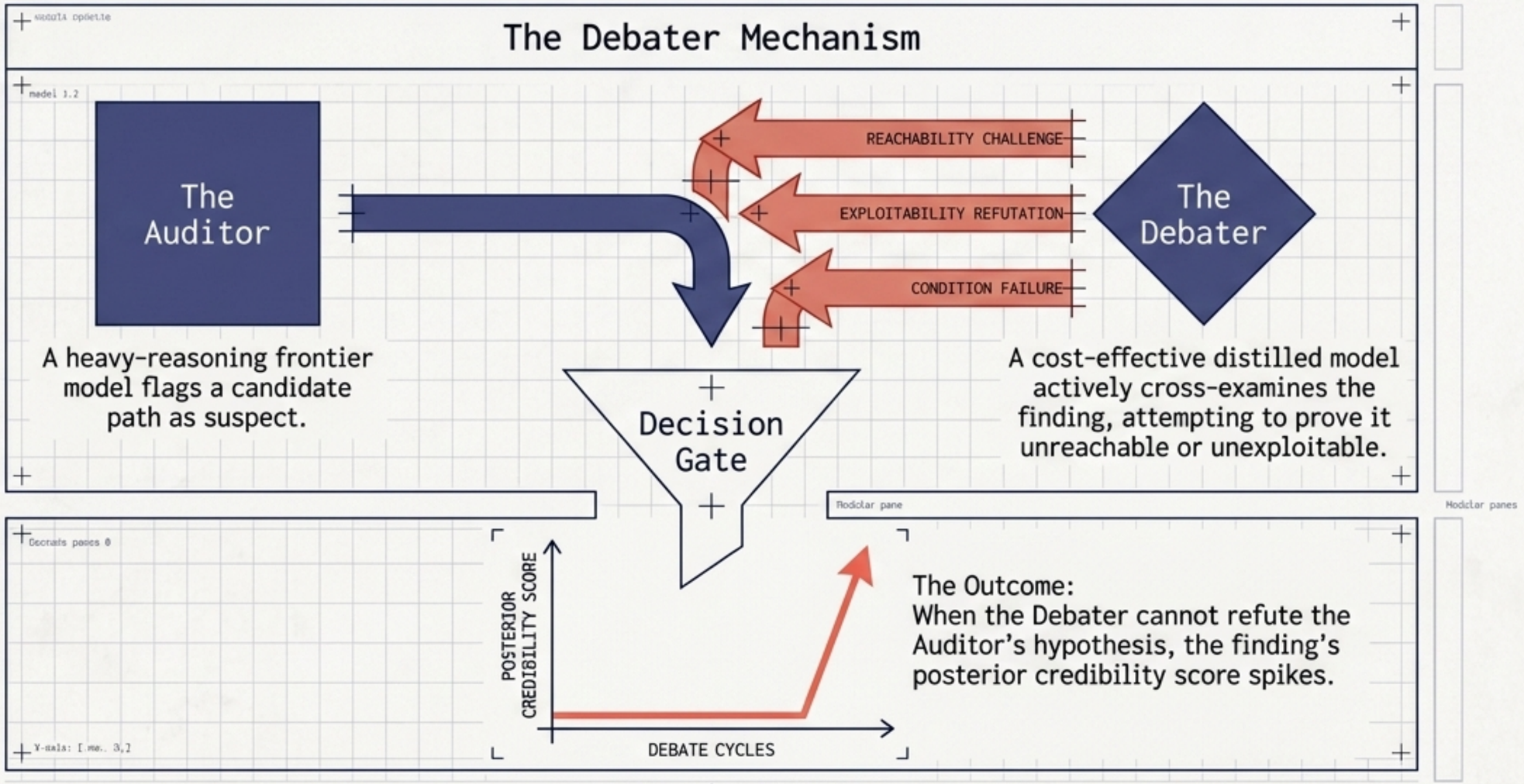
Collapses semantically equivalent findings and groups by required patch.

5. Prove

Formulates bug-triggering inputs to validate pre-conditions dynamically.

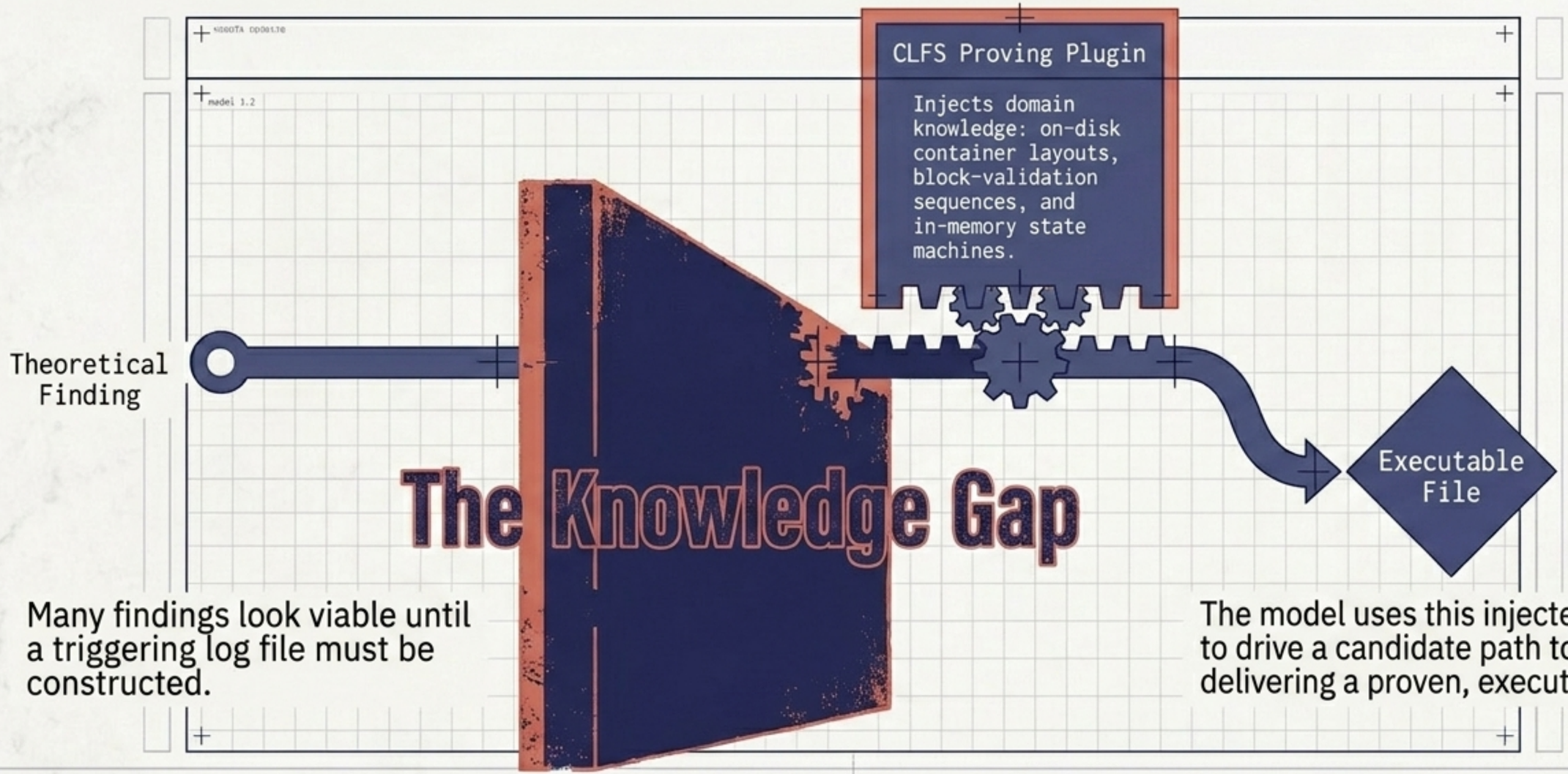
Validation Through Adversarial Debate

Disagreement between models is a high-fidelity signal.
Transforms a raw triage backlog into high-confidence, actionable intelligence.



The Prove Stage: Bridging Theory and Execution

Foundation models cannot internalise private filesystem invariants.
Plugins inject the missing context.



Many findings look viable until a triggering log file must be constructed.

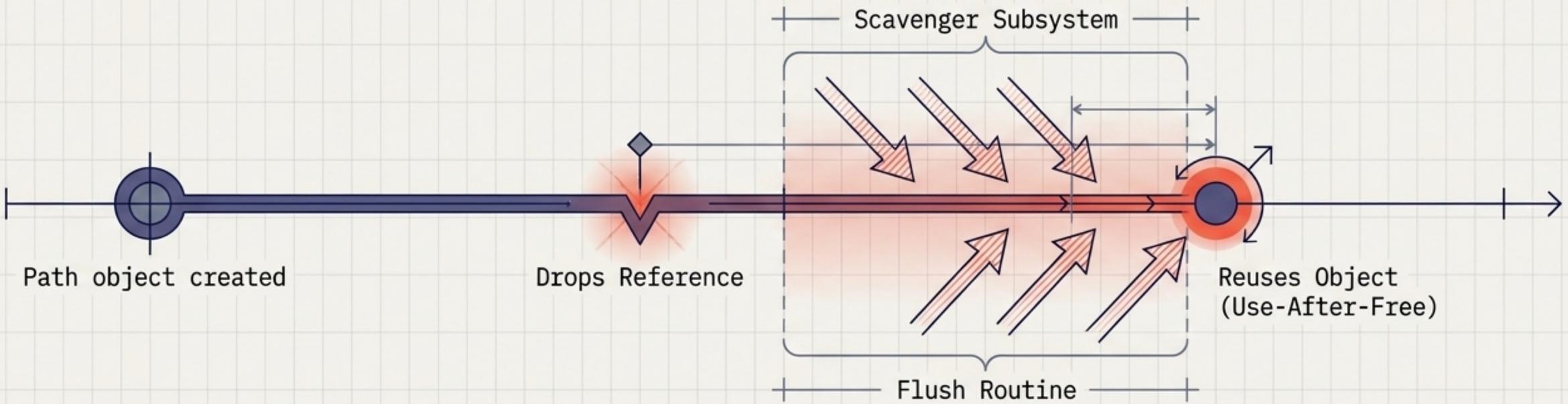
The model uses this injected logic to drive a candidate path to its sink, delivering a proven, executable file.

Diagnostic Matrix: Model vs. System

	Single-Model AI	MDASH Agentic Harness
Reasoning Scope	Isolated functions; loses context.	**Cross-file, multi-step reachability analysis.**
Validation Method	Assumed or hallucinated exploitability.	**Adversarially debated and dynamically proven.**
Architecture	One monolithic, fragile prompt.	**>100 specialised agents (Auditors, Debaters, Provers).**
Lifespan & Portability	Obsolete when the next model generation drops.	**Model-agnostic; retains context and plugins across generations.**

The Concurrency Blind Spot

Case Study: CVE-2026-33827 | Remote unauthenticated UAF in tcpip.sys via SSRR (Critical)



The Vulnerability

A reference-counted `Path` object drops its reference but is reused during `Strict Source and Record Route` processing.

Concurrently, independent subsystems can reclaim the freed object on SMP systems.

Why Single Models Miss It

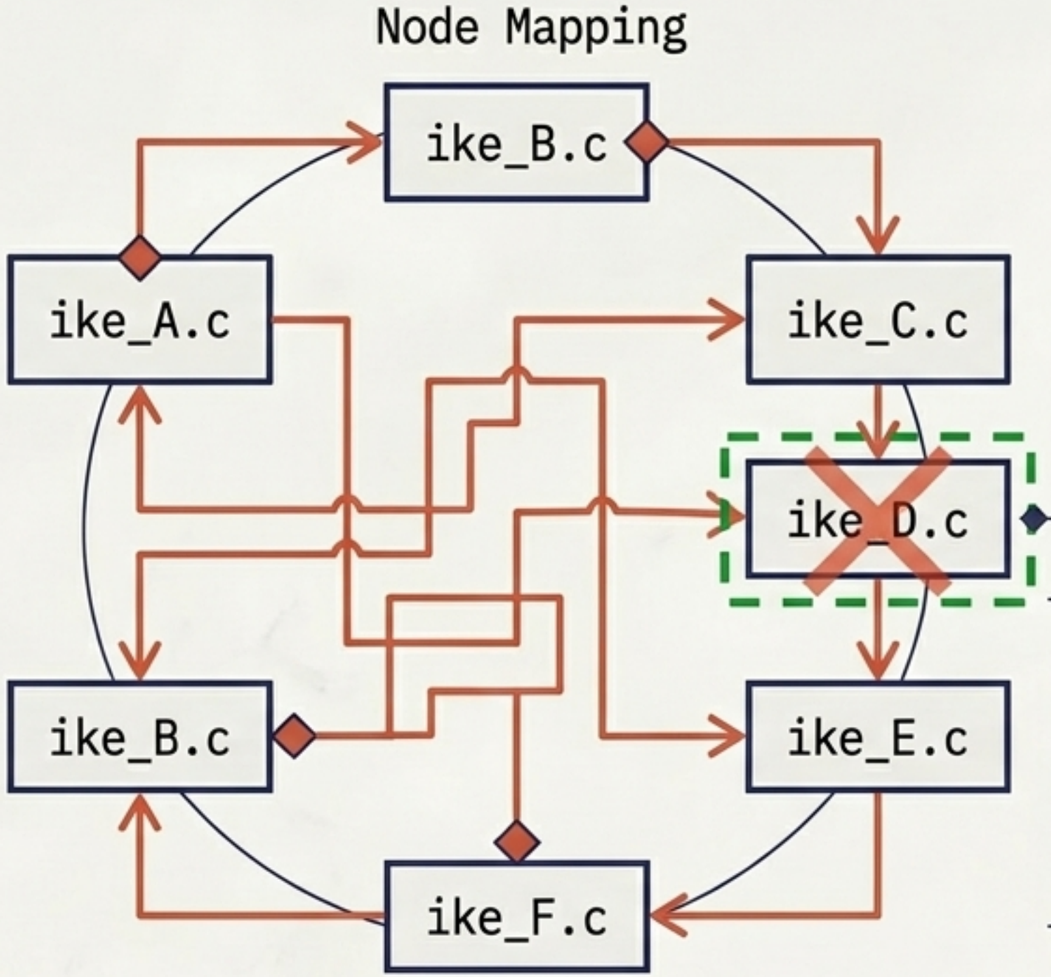
Temporal Dependency: The release and reuse are separated by non-trivial control flow.

Lack of Local Visibility: The single model sees two independent operations.

MDASH connects the ownership violation to the concurrency model.

Anatomy of a Cross-File Blind Spot

Case Study: CVE-2026-33824 | Unauthenticated IKEv2 double-free → LocalSystem RCE



The Correct Pattern: Used as a comparative baseline by MDASH during debate.

The Vulnerability

An **aliasing lifecycle bug.**

A flat memcpy clones a struct's bytes but not its heap allocations, causing the live Main Mode SA and the queued context to both free the same pointer.

Why Single Models Miss It

No single-file analysis can see the bug.

The crucial evidence is the contrast: identifying the correct pattern (ike_D.c) and noticing the deviation elsewhere.

MDASH's debate stage forces cross-file examination.

Ground Truth Telemetry: Validating the Engine

MDASH is not finding theoretical weaknesses; it is finding the bugs that required a Patch Tuesday.

Performance Telemetry
StorageDrive Private Driver
(0-Day Simulation)

ground-truth vulnerabilities found.

Precision: 0 False Positives.
Demonstrates professional offensive researcher capability on unseen code.

Performance Telemetry
MSRC clfs.sys
(5-Year Retrospective)

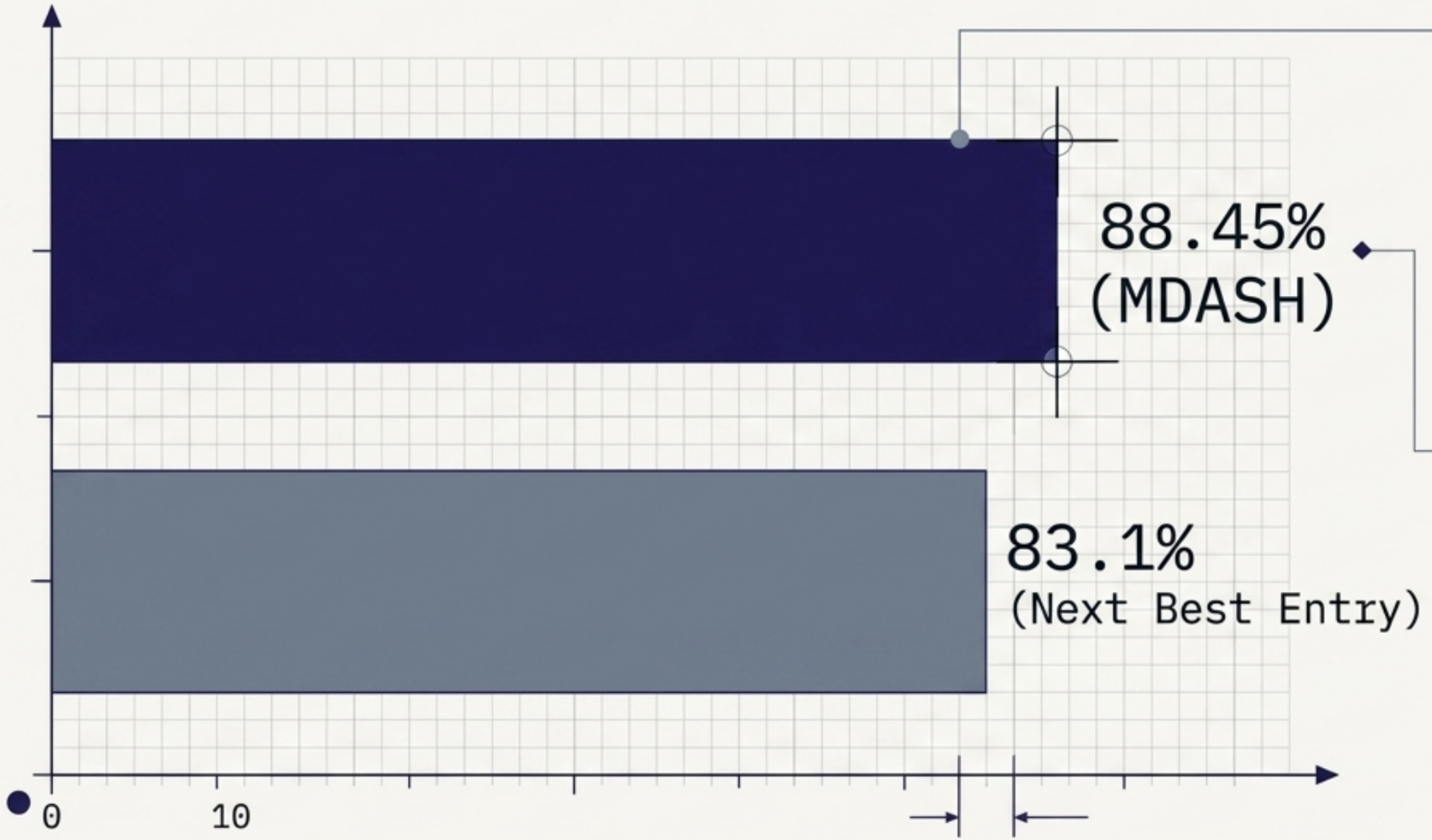
Result: 96% Recall on 28 historically confirmed MSRC cases.

Performance Telemetry
MSRC tcpip.sys
(5-Year Retrospective)

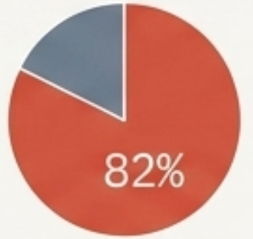
Result: 100% Recall on 7 historically confirmed MSRC cases.

CyberGym: Setting the Industry Benchmark

The Test: 1,507 real-world vulnerability reproduction tasks across 188 OSS-Fuzz projects.



Failure Analysis (The Final 11.55%)



- 82% of targeting failures resulted from vague task descriptions lacking file/function identifiers.

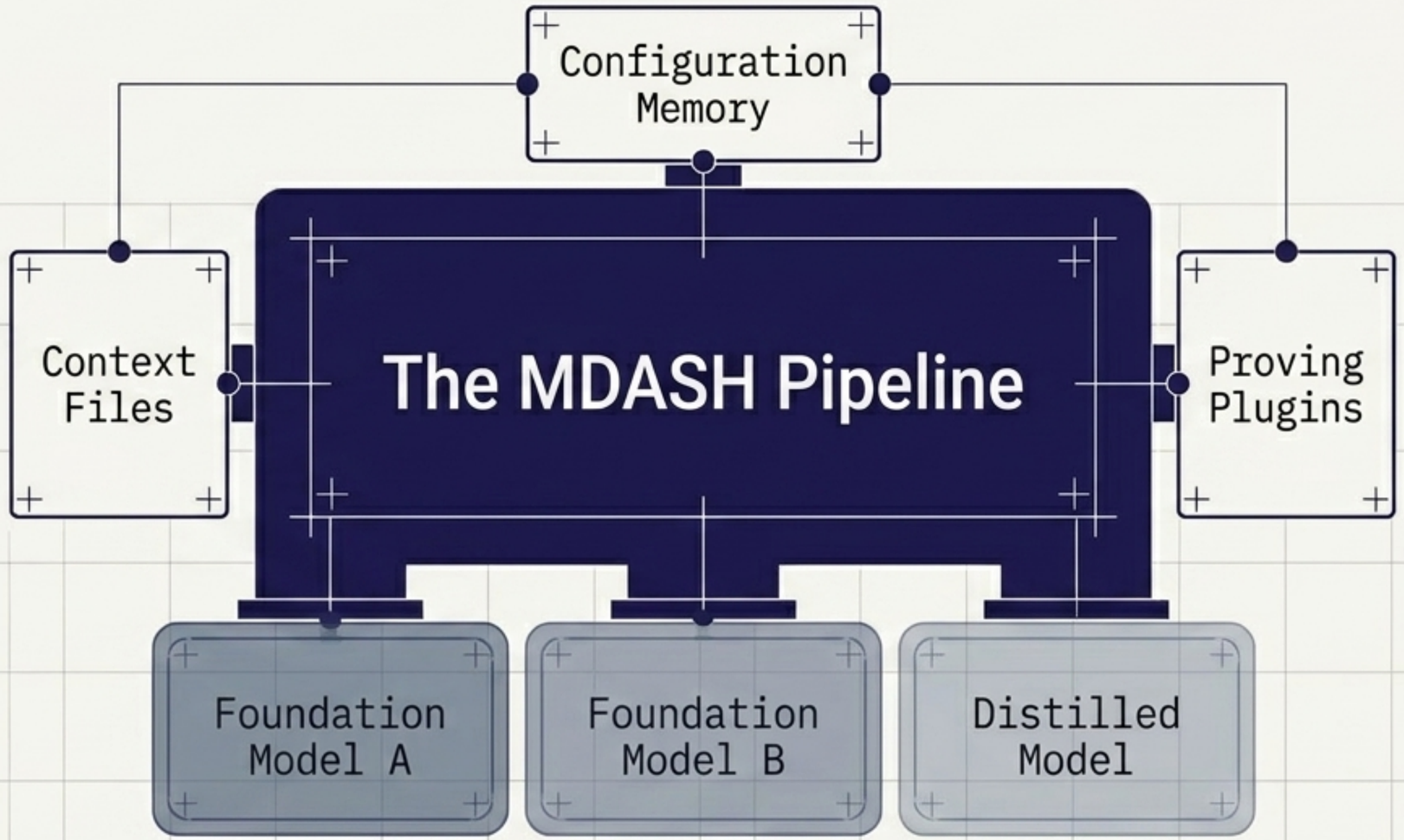
Formatting mismatches (e.g., generating libFuzzer inputs when honggfuzz was required) masked otherwise sound reproductions.

Escaping the Model Lottery

The Paradigm Shift: The model is merely an input. The system is the product.

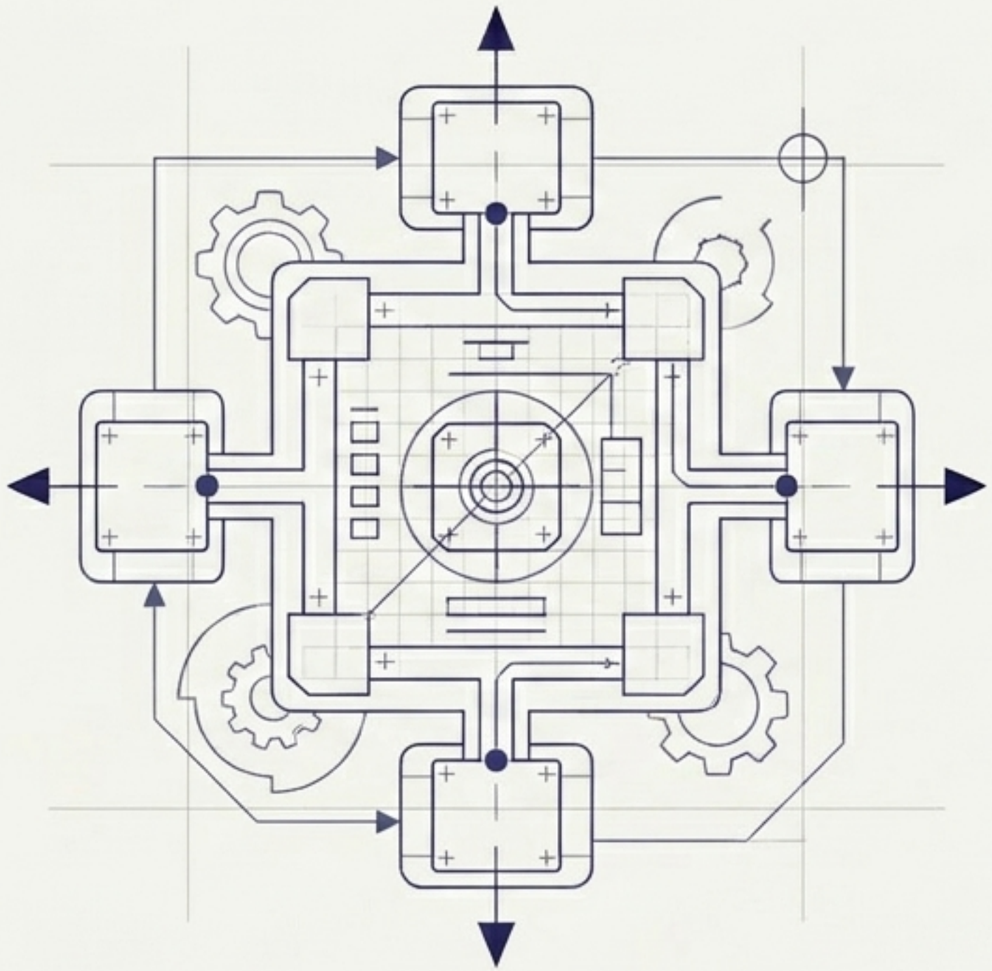
Future-Proof Defence

1. Systems gated on a single model must be rebuilt every six months.
2. MDASH absorbs model improvements instantly. Changing a model is just a configuration flip.
3. The enterprise investment—scope files, proving plugins, configurations, and organisational context—permanently resides in the harness.



The Durable Advantage in the AI Era

AI vulnerability discovery has crossed from a research curiosity to an engineering discipline.



The Mandate for Defenders

Discovery requires multi-step composition that no single prompt can achieve.

Validation through debate is the difference between an actionable fix and a noisy triage backlog.

The Final Question

The right question to ask of an AI security tool is not “Which model does it use?” but “What does it do with the model, and what survives when the next model arrives?”