Agentic AppSec Unleashed '26
by Checkmarx

Getting to High Fidelity

Why Completeness and Noise are Your Developers' Biggest Concern

Ori Bendet
VP of Product Management, Checkmarx

Frank Emery
Senior Director of Product Management, Checkmarx

Agentic AppSec Unleashed '26
by Checkmarx

Agenda

Frontier Models Recap

The Impact on Application Security

From Noise to Signal

How to Move Forward

ICYMI – Mythos and Glasswing

Claude Mythos Preview, a new LLM from Anthropic - its most capable frontier model to date showing a striking leap in scores on many evaluation benchmarks compared to Claude Opus 4.6.

Mythos Preview has found what it estimates are 6,202 high- or critical-severity vulnerabilities in these projects (out of 23,019 in total, including those it estimates as medium- or low-severity).

Finding include Linux, Firefox, OpenBSD, and FFmpeg, significantly shrinking the industry's traditional weeks-to-months patch window.

Source: anthropic.com/research/glasswing-initial-update

What Does it Mean for Application Security

Observations and Implications

Mythos is a quantum leap compared to most models

Mythos dramatically improves vulnerability discovery (new Zero Day)

Anthropic forcing a reckoning with known vulnerabilities with AppSec

Challenges

Exposes new zero-day vulnerabilities and 5x increase in risk

Democratizes exploitability of known vulnerabilities

It won’t address:

Generating secure code … it's still insecure

Proven vulnerabilities LLMs fail to detect

Consistency of results

Performance and scale

Cost effectiveness

Mythos is Only the Beginning

Capture-the-Flag (CTF)

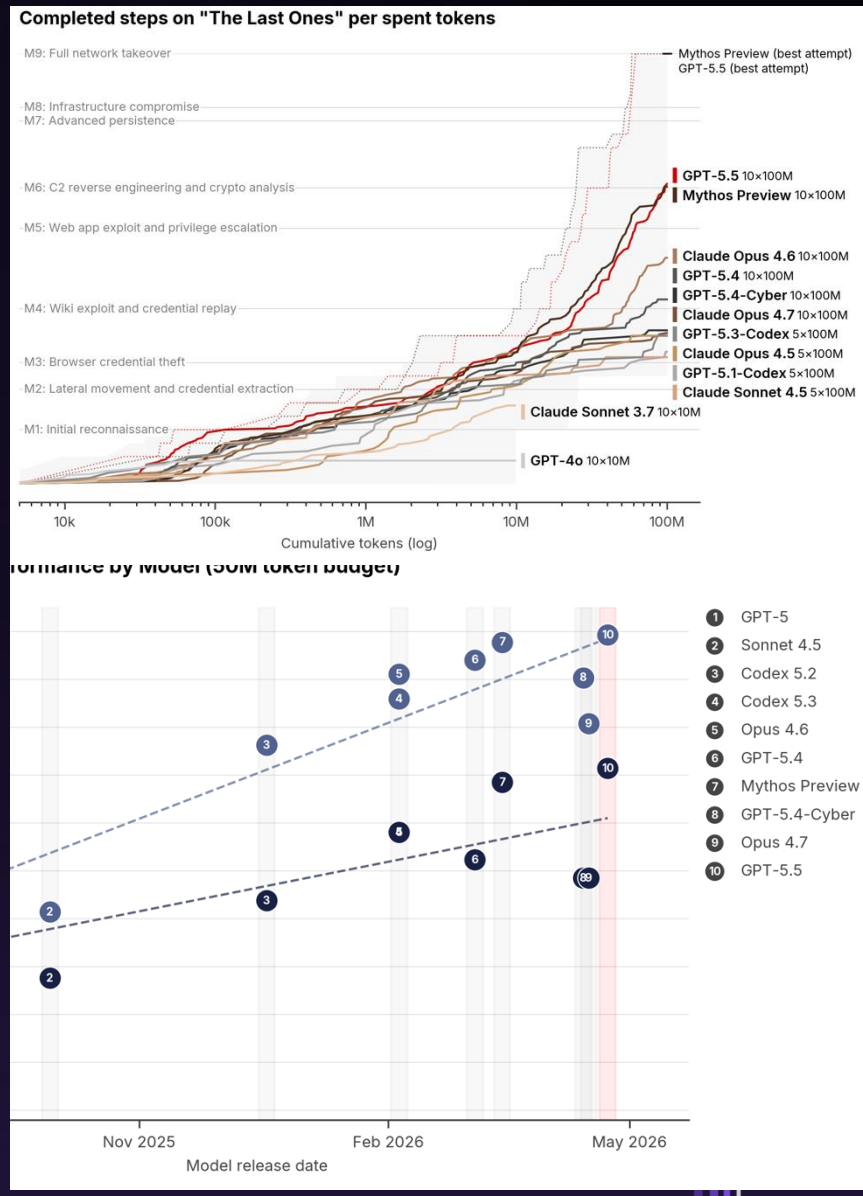
GPT-5.5 scored 71.4%

Mythos Preview scored 68.6%

End-to-End Attack Simulation

Mythos completed it in 3/10 of the times

GPT-5.5: 2/10



Mythos is a Point in Time.

You Need to be Ready for Whatever is Next

AI-Gen Code is Still Insecure

45%+ of the solutions generated by even the best models are **either incorrect or contain a security vulnerability**

On average, around **50% of the correct solutions are insecure**

Security requirements add coding complexity and lead to correctness tradeoffs in the models

Source: baxbench.com/

No Security Reminder | Generic Security Reminder | Oracle Security Reminder

Rank	Model	Correct & Secure ↓	Correct	% Insecure of Correct
1	GPT-5	53.8%	67.1%	19.8%
2	OpenAI o3	47.7%	65.1%	26.7%
3	Claude 4 Sonnet Thinking	46.9%	72.2%	35.0%
4	GPT-4.1	41.1%	55.1%	25.3%
5	Claude 3.7 Sonnet Thinking	39.0%	61.7%	36.8%
6	OpenAI o3-mini	37.0%	61.2%	39.6%
7	DeepSeek R1	34.9%	55.6%	37.2%
8	Claude 3.5 Sonnet	34.1%	56.2%	39.3%
9	Grok 4	33.9%	55.9%	39.3%
10	Gemini 2.5 Pro	33.8%	49.7%	32.1%
11	OpenAI o1	31.1%	62.2%	50.0%
12	Qwen3 Coder	30.8%	52.1%	40.9%
13	GPT-4.1 Mini	28.9%	46.4%	37.6%

Where LLMs Fail as a Static Scanner

Empirically Proven Vulnerability Classes That LLMs Consistently Fail to Detect — Regardless of Model Size or Prompt Engineering

1

Memory Safety and C/C++ Vulnerability Detection

CWE-119, CWE-125, CWE-787

Best model (Claude 3.7 Sonnet) achieved only 23.83% F1 on real-world C/C++ vulnerable statement detection with correct reasoning — GPT-4.1 performed similarly.

Source: SECVulEval Benchmark, 2025

2

Business Logic and Authentication Flow Flaws

CWE-840, CWE-287

GPT-4 exploited 87% of known vulnerabilities when given a CVE description, but only 7% when discovering them independently — confirming LLMs match patterns, not reason about intent.

Source: UIUC / arXiv LLM Exploit Study, 2024

3

OS Command Injection via Gadget Chains

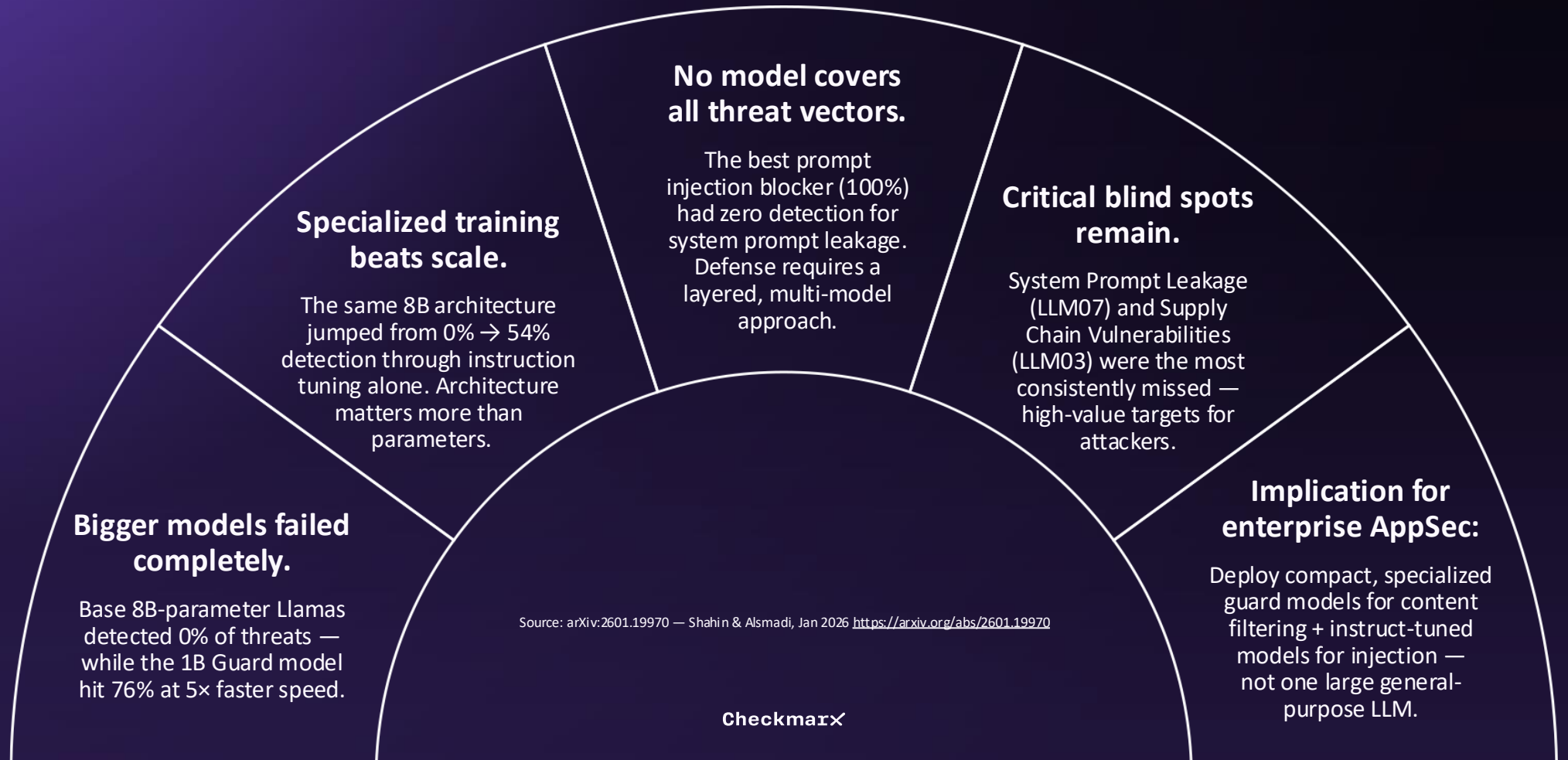
CWE-78

GPT-4-backed IRIS found only 3/13 CWE-78 cases — the lowest detection rate across all tested categories. Authors cite external side-effects and chained execution as root causes.

Source: IRIS / CWE-Bench-Java, 2024

Additional LLM Security Gaps

OWASP Top 10 for LLM Applications — Llama Benchmark (Texas A&M, Jan 2026)



Consistency of Results: The Hidden Story

Our results reveal high degrees of non-determinism: the ratio of coding tasks with zero equal test output across different requests is 75.76%

…. setting the temperature to 0 does not guarantee determinism in code generation, although it indeed brings less non-determinism than the default configuration

---

AI Security Review Results

The results showed major inconsistencies. In addition to the 2 “planted” vulnerabilities the Opus generated code generated code with many vulnerabilities.

32%
of vulnerabilities were consistent across all 5 scans

60%
of findings were false positives in dead and non exploitable code

60%
of scans missed the critical vulnerabilities we planted in the code

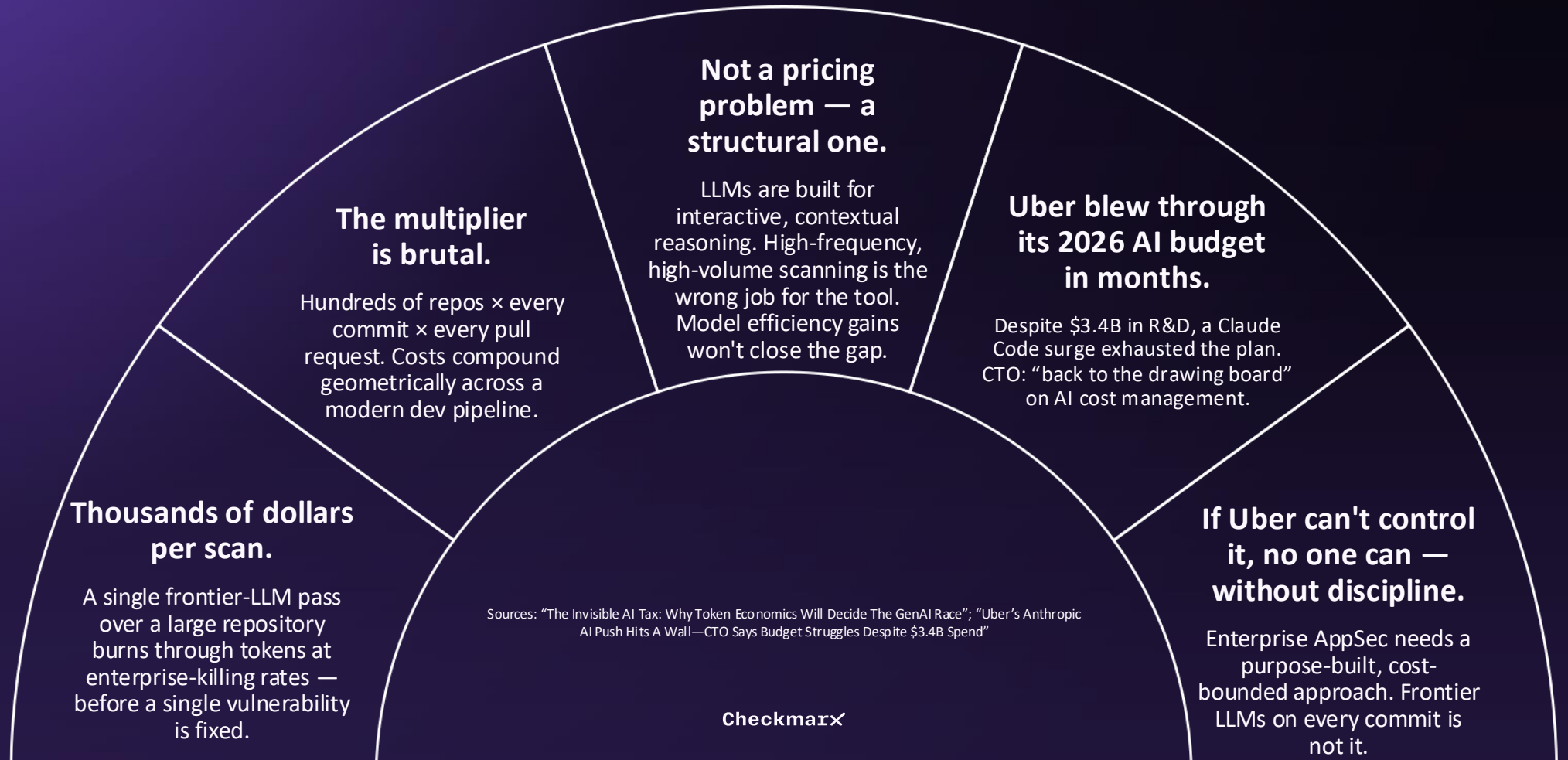
100%
of scans flagged a critical finding in dead code that is not exploitable

Multiple vulnerabilities were detected only once or twice

Sources: medium.com/@gadi_50049/ai-security-review-fails-in-practice-an-analysis-of-opus-4-6-87af7486efa3 | arxiv.org/abs/2308.02828

LLM Cost Effectiveness: The Token Tax

Why Frontier-LLM Scanning is Economically Problematic at Enterprise Scale



LLM Cost Inefficiencies: Token Tax Details

Scanning 1M LOC with 1K Vulnerabilities Found

Model	Input /1M	Output /1M	Input Cost (50M)	Output Cost (0.5M)	Total Cost
GPT-5.4 Pro >272k	$60	$270	$3,000	$135	$3,135
GPT-5.5 Pro ≤272k	$30	$180	$1,500	$90	$1,590
GPT-5 Pro	$15	$120	$750	$60	$810
Claude Opus 4	$15	$75	$750	$37.50	$787.50
Gemini 3.1 Pro >200k	$4	$18	$200	$9	$209
GPT-5.1 Codex	$1.25	$10	$62.50	$5	$67.50

46x

more expensive: GPT-5.4 Pro vs GPT-5.1 Codex

Claude Opus 4 is

11.7x

more expensive than Codex on this workload

Source: https://www.llm-prices.com/

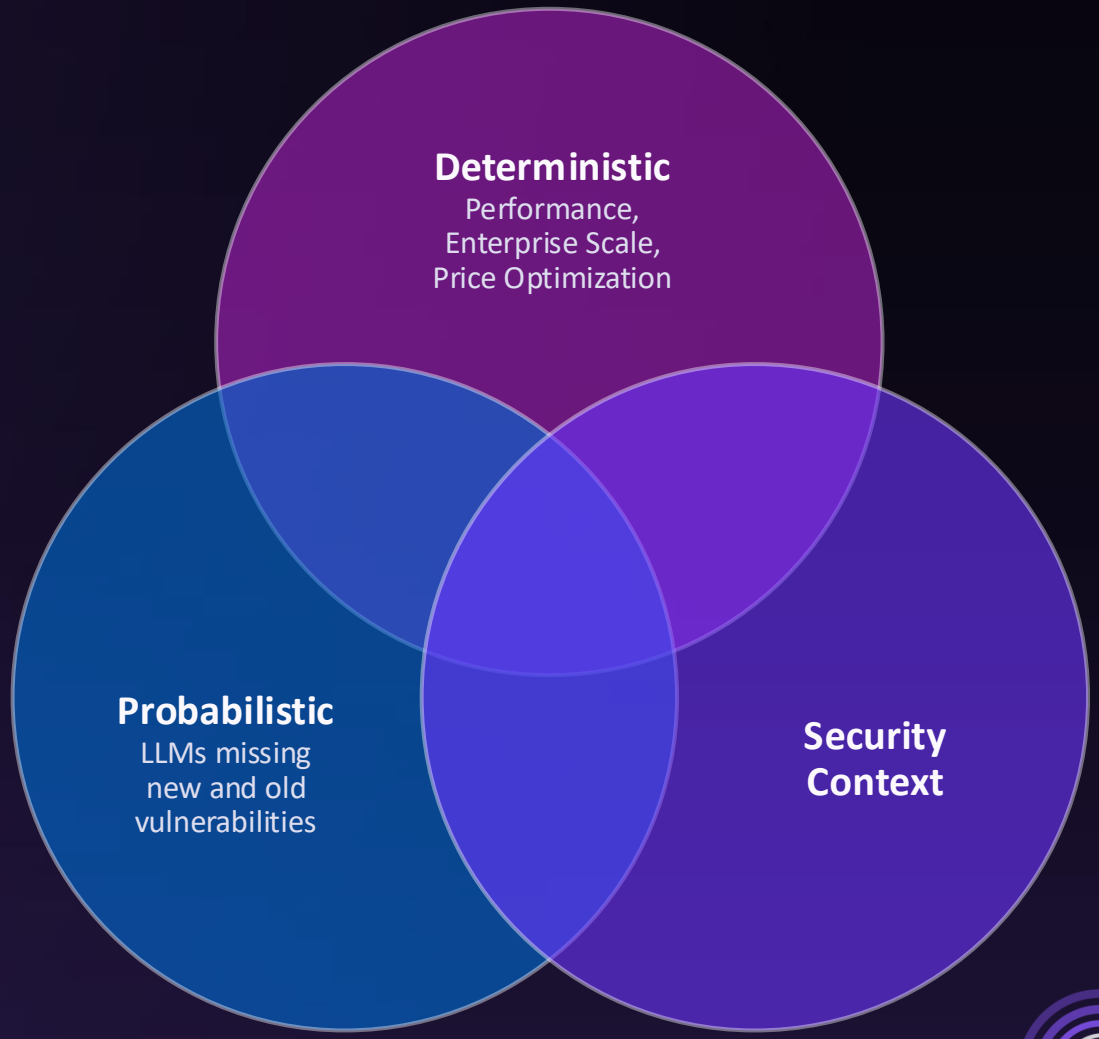
The Bottom Line: What's Needed

The highest possible fidelity

Remediation workflow approach (guided and autonomous)

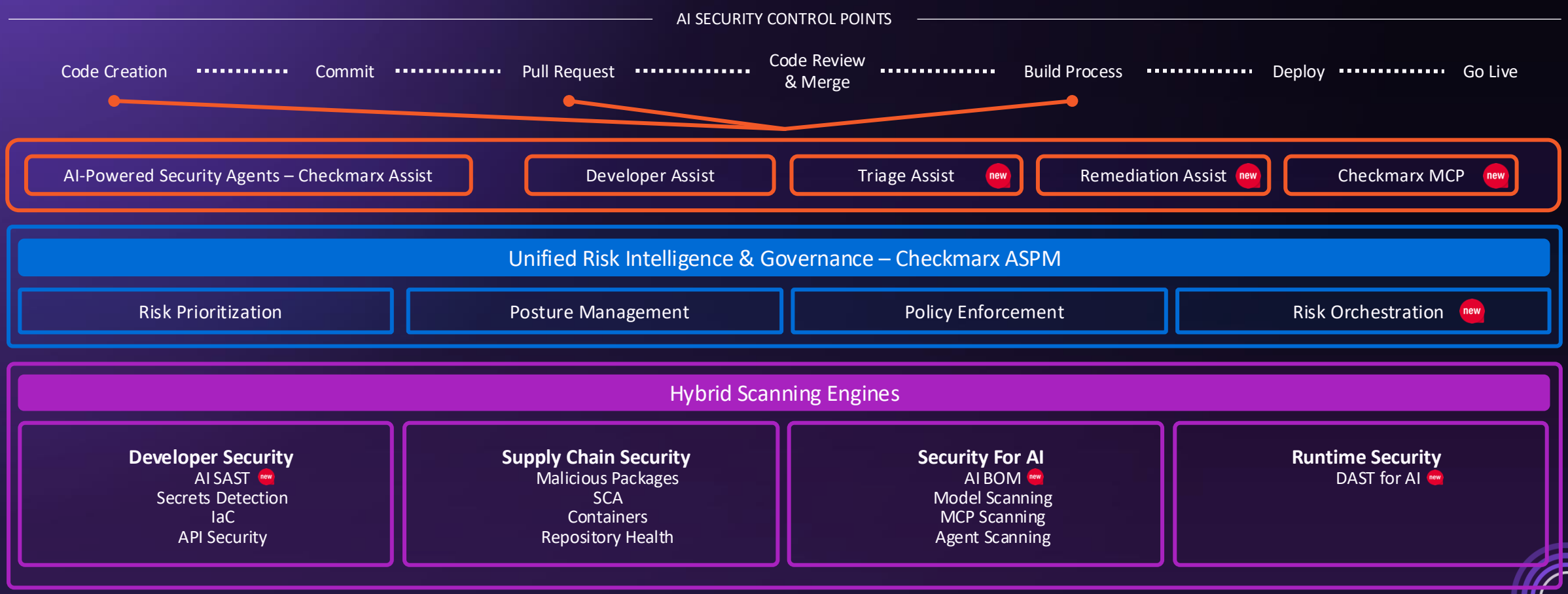
Governance, Compliance, and Policy Management

Consistency Across Scanning Results



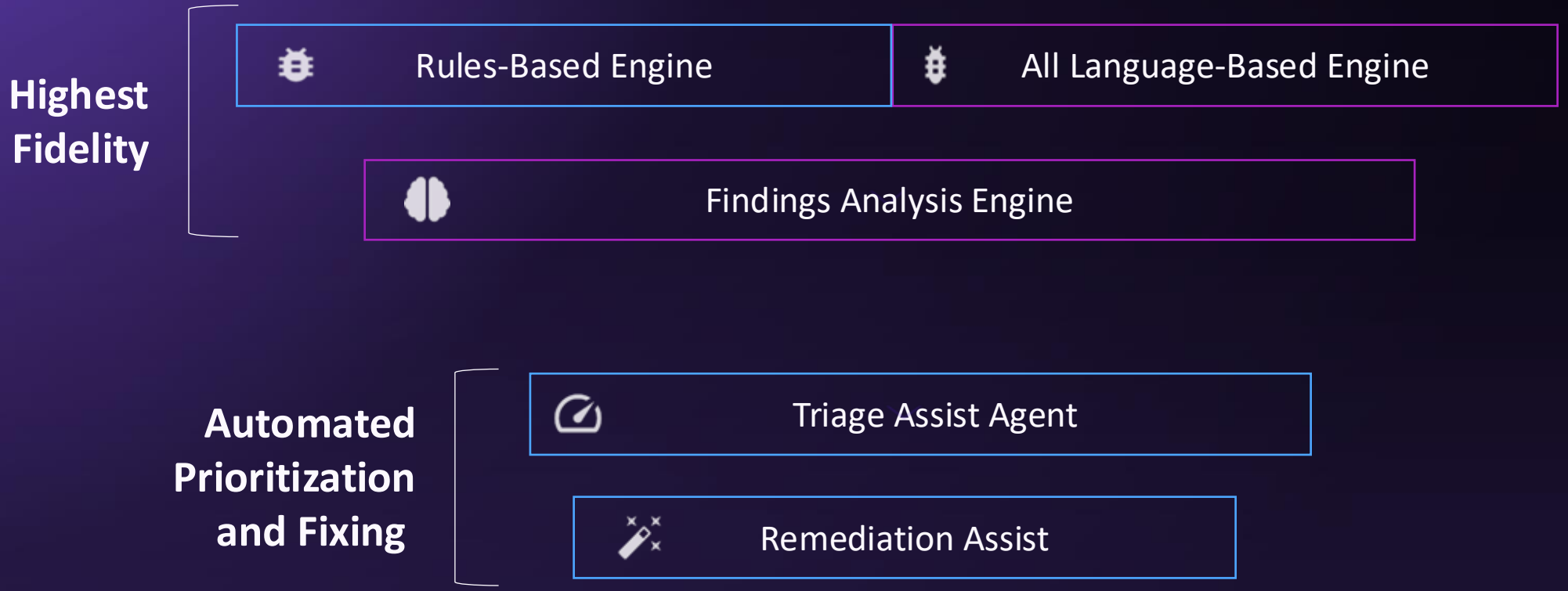
So How Do You Get to the Highest Fidelity Possible?

Checkmarx One:
Unifying Application Security for the AI Era



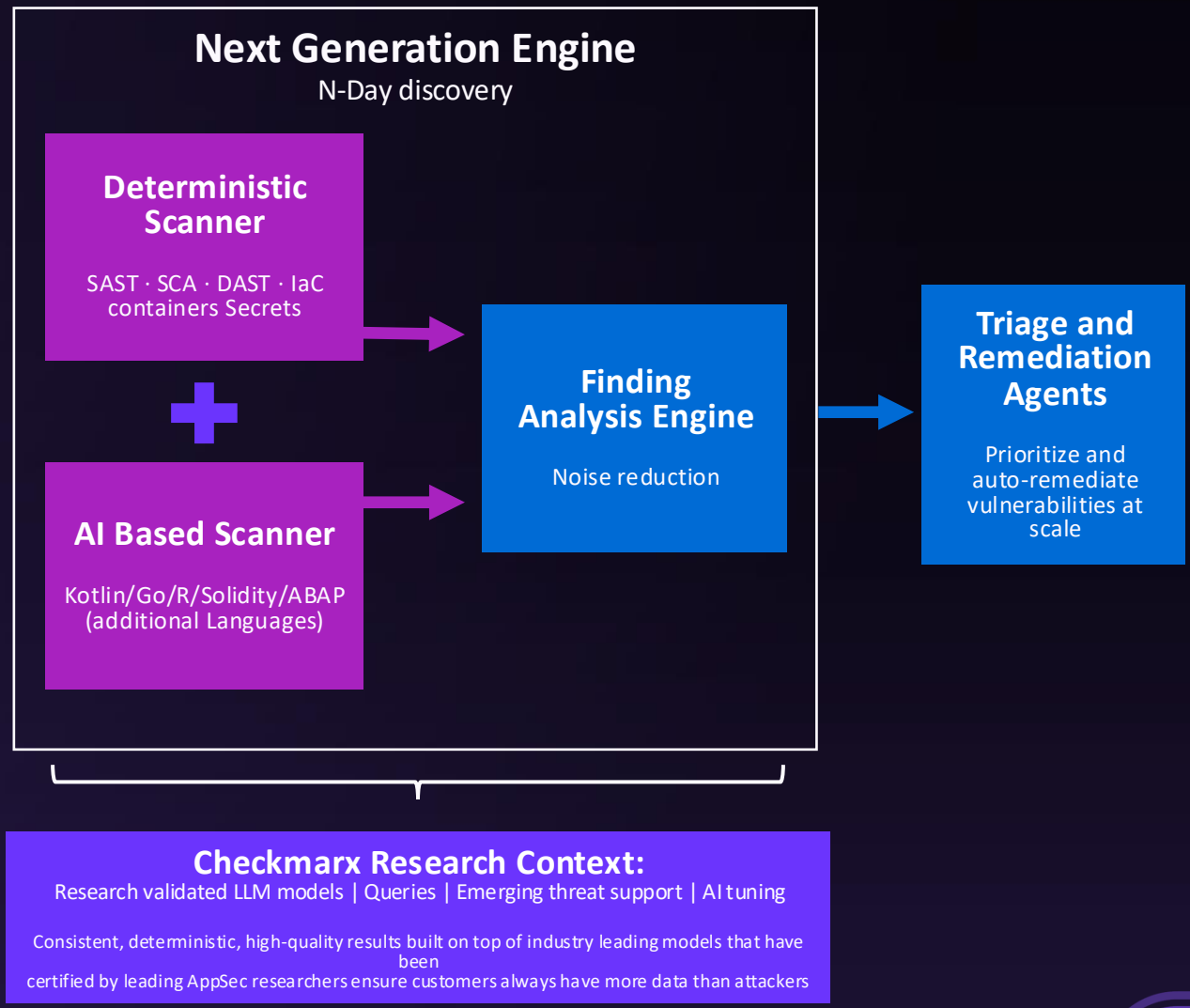
Next Generation Scanning Engine in Checkmarx One

60% Reduction in False Positives, Unlimited Language Support, 70% Improved Fidelity Over Traditional SAST



Next Generation Scanning Engine

Available Now in Checkmarx One - Creating the Most Accurate with Best Fidelity Results in the Market

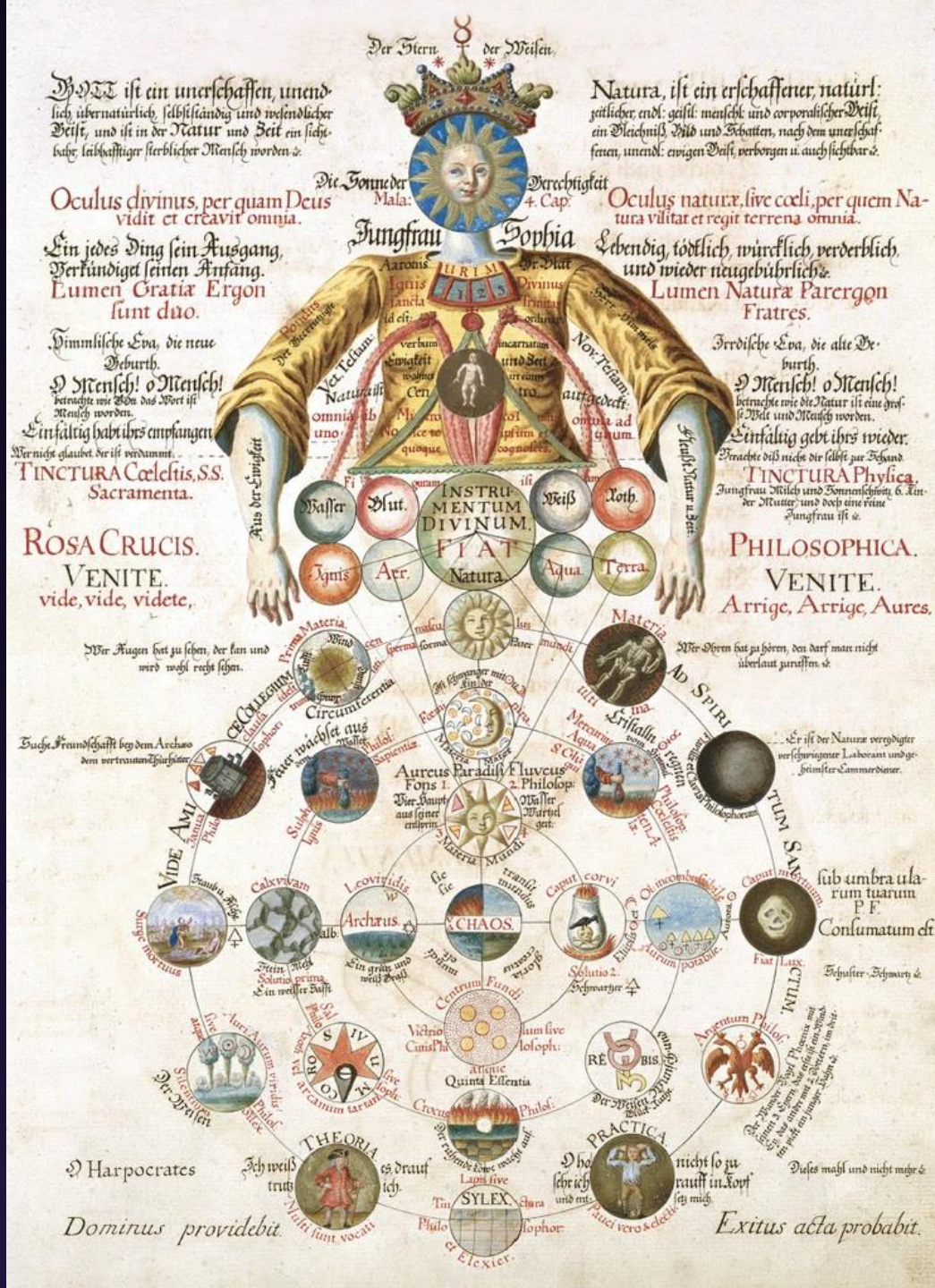


What is Pansophia?

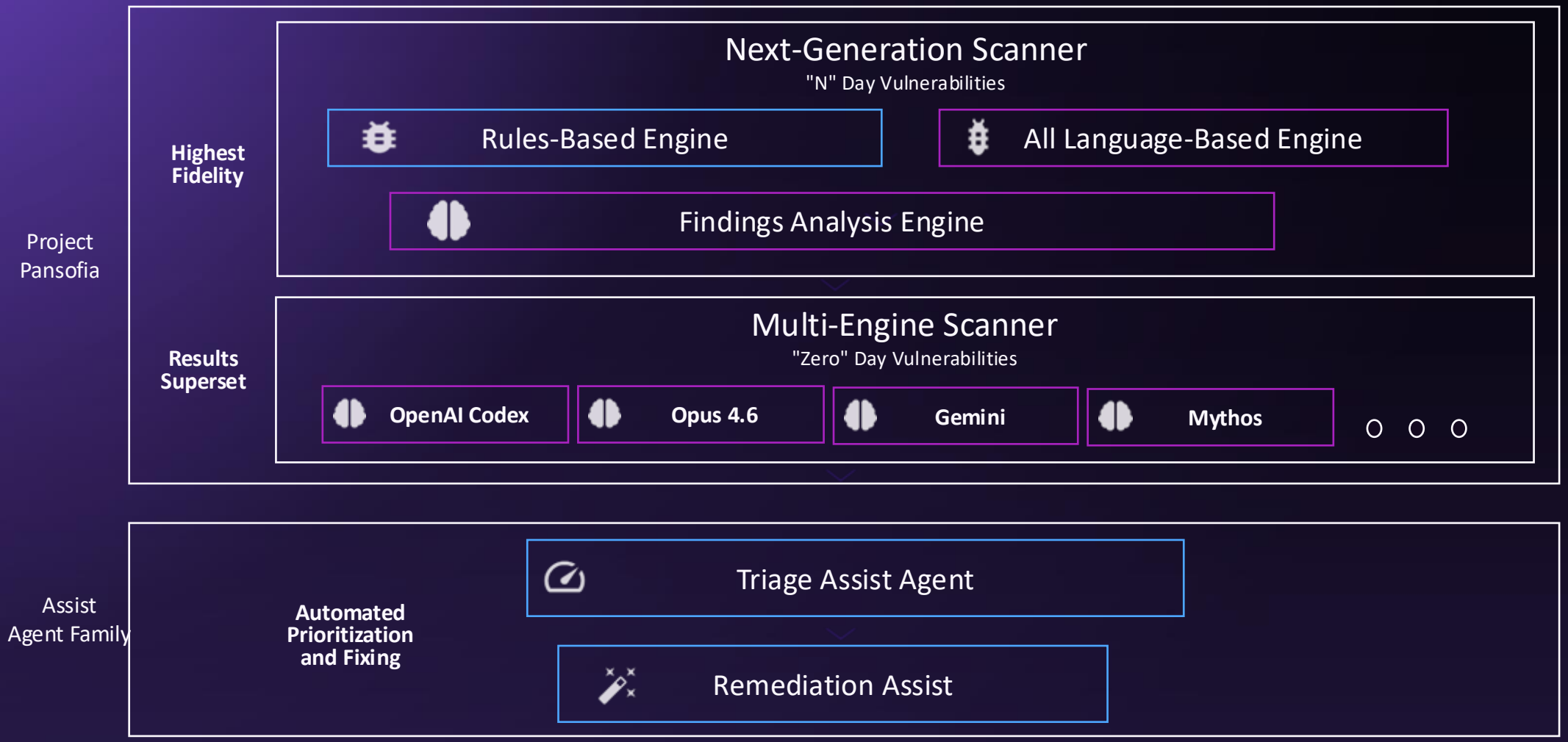
"All-Knowledge“: It suggests a universal wisdom that isn't just a list of facts, but a deep understanding of how everything (past, present, and future) fits together.

In later Renaissance philosophy, which heavily borrowed from Greek, the goal was:

A single system that contained everything known to humans.

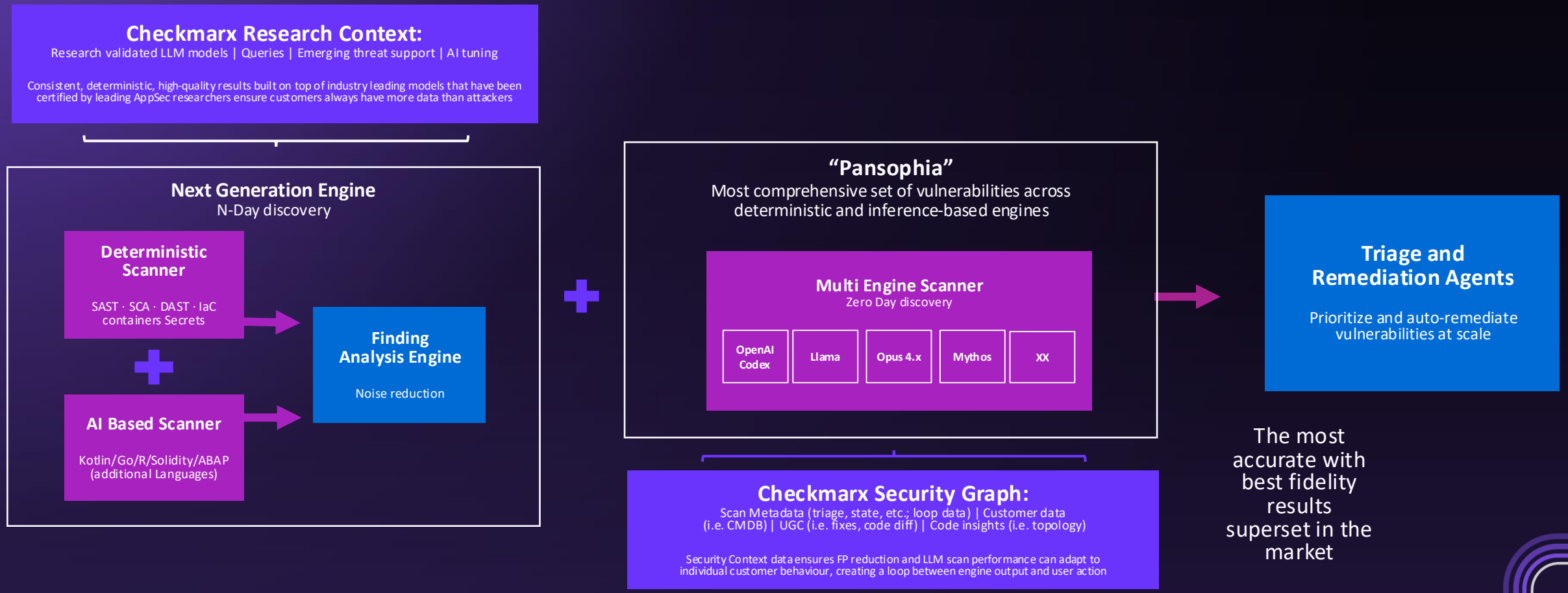


Project “Pansophia” Builds Upon Next Gen SAST



Pansophia – Advanced Scanning Engine Technology

Coming Soon - Creating the Most Accurate with Best Fidelity Results Superset in the Market



The most accurate with best fidelity results superset in the market

Agentic AppSec Unleashed '26

by Checkmarx