



GenAI Runtime Security

Where AI risk actually lives, and what controls it

ISC2 East Bay Chapter | October 8, 2026

Amit Kharat, Co-founder, **Mavs AI**



Amit Kharat



Co-Founder, Mavs AI

15+ years in cybersecurity | Ex-Seclore

CURRENT FOCUS

AI Security

AI Governance

BACKGROUND

Data Classification

DLP

DSPM

EDRM

Encryption

Data Security



Block or Bleed: The Choice Security Teams Are Handed

BLOCK

Ban or restrict AI tools.

Use moves to personal accounts and unmanaged devices

Visibility drops to zero

Security becomes the team that said no

BLEED

Allow AI without controls.

Business-sensitive data leaves in prompts

No record of what was shared, or with which model

Security owns the incident

**Both fail for the same reason:
no control operates on the conversation (language layer) itself.**



The most abused term in this room isn't "AI."

It's "AI Security and Governance."

Everyone says it. Almost no one means the same thing.



Four Different Things, that get Confused

AI SECURITY

Protecting the AI System itself: the model, the data, the runtime use of AI.

AI GOVERNANCE

Policy and Oversight: Who's allowed to use what, and why?

RESPONSIBLE AI

Fairness, Bias and Explainability. It's a different discipline.

SHADOW AI

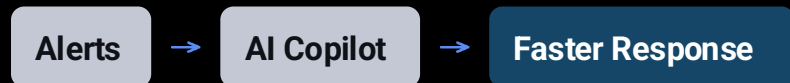
AI Apps and services used without IT or Infosec approval.



Two Pairs Worth Getting Right

AI-for-Security vs. Security-for-AI

AI-FOR-SECURITY



Using AI to strengthen your existing DEFENSE.
E.g., SOC, Anomaly detection in SIEM / XDR.

SECURITY-FOR-AI



Defending the AI System itself.
E.g., Runtime controls for LLMs, Agent action /
tool call gating.

Generative AI vs. Agentic AI

GENERATIVE AI



Produce content in response to a prompt.
E.g., ChatGPT/Claude, Document drafting

AGENTIC AI



Autonomous actions. Plans a sequence.
E.g., Coding agents, An agent handling loan
approval.

WHERE AI RUNS

WHERE AI RUNS: FOUR SURFACES



EMPLOYEE AI USAGE IN BROWSER

Employees using ChatGPT,
Gemini, etc in browser



AI AGENTS RUNNING ON DESKTOP

Employees using Claude co-
work, Open-Claw, Embedded AI



HOMEGROWN AI APPS & AGENTS

Customer/Agent chatbots,
Agentic workflows



AI ENABLED SAAS PRODUCTS

Products leveraging AI within
their offering



Which of These Four Is Your Biggest Blind Spot Today?

- A** Employees using AI in the browser – ChatGPT, Claude, CoPilot
- B** AI agents on the desktop – Claude Desktop, Cursor, Copilot
- C** Homegrown AI apps and agents
- D** AI inside SaaS products

THREAT MODEL



9:14 AM

An analyst at a regional bank pastes a client's loan file into a chatbot to summarize it faster.

9:15 AM

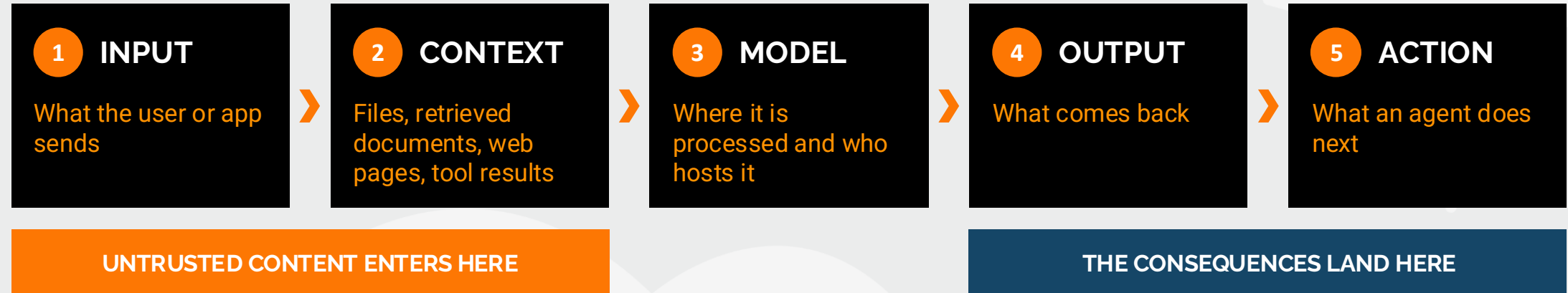
That data is now sitting on a server outside your jurisdiction, and outside your control.

That's Risk A, and only half the picture.





Structure of an AI Interaction

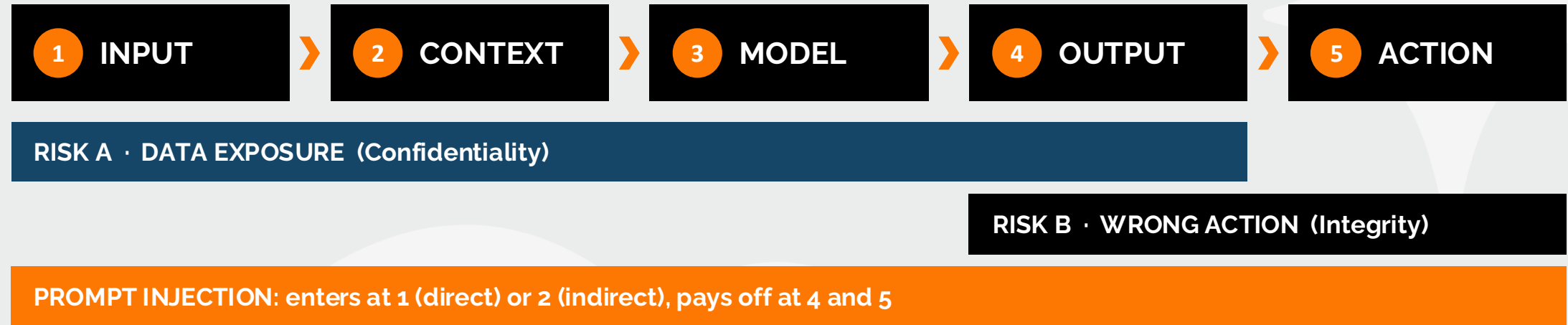


Content is now *instruction*.

Anything the model reads at points 1 and 2 can change what happens at points 4 and 5.



Two Failure Modes, One Mechanism



Risk A: Data exposure

Sensitive data is pasted into a prompt, pulled in by over-broad retrieval, or returned to the wrong person.

Samsung staff pasted source code and meeting notes into ChatGPT. The company then restricted generative AI tools.

Risk B: Wrong action

The agent does something the user never asked for, using the user's own permissions.

CVE-2025-32711: A crafted email made Microsoft 365 Copilot send out internal data. Zero clicks. Patched by Microsoft.



OWASP: The Risks Behind Our Two Failure Modes

OWASP Top 10 for LLM Applications (2025)

LLM01 Prompt Injection	Input or content that overrides the intended instructions
LLM02 Sensitive Information Disclosure	Private data exposed in prompts or responses
LLM05 Improper Output Handling	Output passed downstream without validation
LLM06 Excessive Agency	More permissions or autonomy than the task needs

ALSO ON THE LIST

LLM03 Supply Chain	LLM04 Data and Model Poisoning
LLM07 System Prompt Leakage	LLM08 Vector and Embedding Weaknesses
LLM09 Misinformation	LLM10 Unbounded Consumption

OWASP Top 10 for Agentic Applications (2026)

ASI01 Agent Goal Hijack	Injected content redirects what the agent is trying to do
ASI02 Tool Misuse and Exploitation	Legitimate tools used in unsafe ways
ASI03 Identity and Privilege Abuse	Agent acts with inherited or excess privileges
ASI10 Rogue Agents	Agent drifts from its intended purpose while still looking legitimate

ALSO ON THE LIST

ASI04 Agentic Supply Chain	ASI06 Memory and Context Poisoning
ASI05 Unexpected Code Execution	ASI07 Insecure Inter-Agent Communication
ASI08 Cascading Failures	ASI09 Human-Agent Trust Exploitation



What Do You Rely On Today to See Inside an AI Prompt?

- A DLP
- B CASB / SSE / Proxy / Firewall
- C Endpoint or browser controls
- D An AI-specific security tool
- E Nothing yet OR Evaluating

THE GAP



What the Existing Stack Sees in an AI Interaction

Typical capability. Varies by product and configuration.

	INPUT	CONTEXT	OUTPUT	AGENT ACTIONS	UNDERSTANDS MEANING
DLP endpoint and network	Partial Paste and upload in managed browsers; not native apps or APIs	No Assembled inside the AI service, past any DLP checkpoint	No Built to inspect data leaving, not answers coming back	No Tool and API calls run outside DLP channels	Partial Patterns, fingerprints, classifiers; misses paraphrase and intent
CASB / SSE cloud app control	Partial Inline for known AI apps on managed, decrypted traffic	No Retrieval happens inside the AI service	Partial Possible inline; streamed answers are rarely parsed	Partial Logs SaaS actions by connected apps, after the fact	Partial Same DLP engine, same pattern limits
Web proxy / SWG secure web gateway	Partial Request body with TLS inspection; otherwise only the domain	No Retrieval happens inside the AI service	Partial Scans responses for malware, not for leaked data	No Sees a web request, not that an agent made it	No Judges URL, category and reputation, not content
EDR / XDR endpoint detection	No Watches processes, not text typed into a page	No No view inside the AI service	No Does not inspect application content	Partial Process, file and network activity of local agents only	No Detects malicious behavior, not sensitive meaning

Each tool sees a fragment. None follows one interaction from **prompt to action**.



Why the Gap Is Structural, Not a Missing Feature

What gets inspected

BUILT FOR

Files and network flows

AI NEEDS

The conversation turn, with the context that came with it

How risk is detected

BUILT FOR

Patterns, labels and signatures

AI NEEDS

Meaning. A paraphrased M&A plan matches no pattern

What the verdict is

BUILT FOR

Allow, Block or Mask

AI NEEDS

Options in between with context preserved, so the work can continue safely

FOUR LAYERS OF AI SECURITY

Model security

weights, training, supply chain

Data security

data at rest and in motion

Application security

the code around the model

Runtime security

the live interaction

The first three inspect something standing still. Only the fourth watches the moment of use.

RUNTIME CONTROLS

CONTROL POINTS: ONE PER SURFACE



EMPLOYEE AI USAGE IN BROWSER

Employees using ChatGPT,
Gemini, etc in browser



Browser extension
or Secure chat app



AI AGENTS RUNNING ON DESKTOP

Employees using Claude co-
work, Open-Claw, Embedded AI



Endpoint agent



HOMEGROWN AI APPS & AGENTS

Customer/Agent chatbots,
Agentic workflows



API / LLM gateway



AI ENABLED SAAS PRODUCTS

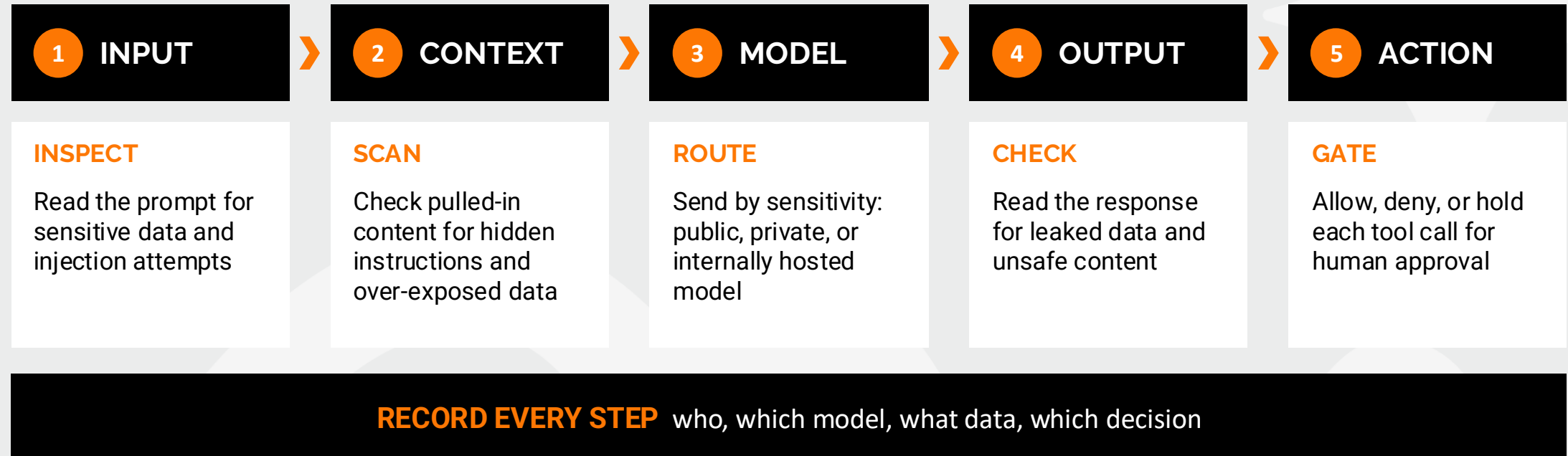
Products leveraging AI within
their offering



API / LLM gateway



What a Runtime Control Does at Each Point



Inspect. Detect. Enforce. Prove.



Between Allow and Block: The Dilemma returns

ALLOW	BLOCK
WHAT HAPPENS Passes untouched	WHAT HAPPENS Stopped
WHAT IT COSTS Nothing, if the risk call was right	WHAT IT COSTS Lost productivity, and use moves to shadow AI

A control that can only allow or block rebuilds the **Block or Bleed dilemma.**



Between Allow and Block: The Decision Spectrum

ALLOW	MONITOR	RE-PROMPT	ROUTE	JUSTIFY	BLOCK
WHAT HAPPENS Passes untouched	WHAT HAPPENS Passes, logged and alerted	WHAT HAPPENS Sensitive values replaced with synthetic stand-ins first	WHAT HAPPENS Sent to a private or internal model	WHAT HAPPENS Held for justification or human sign-off	WHAT HAPPENS Stopped
WHAT IT COSTS Nothing, if the risk call was right	WHAT IT COSTS You find out after the fact	WHAT IT COSTS Plain redaction can strip the context the model needs	WHAT IT COSTS Needs an approved alternative model	WHAT IT COSTS Adds delay for the user	WHAT IT COSTS Lost productivity, and use moves to shadow AI



What Changes With Agents

Two questions a human-driven control never had to answer:

Is it still doing what it was asked? What did it actually do?

IDENTITY

Whose permissions is it using?

An agent inherits the access of the user or service behind it, and acts with nobody watching. Least privilege has to be set per agent and per task.

TOOL CALLS

What is it allowed to do?

Every tool call is an action: read, write, send, execute. Gate them at the point where the agent reaches its tools, such as an MCP gateway.

BEHAVIOR

Is this normal for this agent?

Keep a full record of every step and a baseline of normal behavior, so drift and goal hijack are caught while they happen.

With agents, the control point moves from the prompt to the **action.**

GOVERNANCE



What an Auditor Will Ask About AI

"We have an AI policy"

THE GOVERNANCE GAP



"We can enforce it and prove it"

THE QUESTION

Who used AI last quarter, and with which tools?

Which models did our data go to, and who hosts them?

Did regulated data go into a prompt?

What did you stop, and what did you let through?

What did your agents do on a user's behalf?

THE EVIDENCE THAT ANSWERS IT

Usage record by user and application

Model and destination for every interaction

Inspection result for every prompt

Decision log: allowed, transformed, routed, blocked

Action record for every tool call

RUNTIME CONTROLS CHECKLIST



Runtime Controls Checklist: Nine Questions

SEE	CONTROL	PROVE
☐ 1 Can you list the AI in use across all four surfaces?	☐ 4 Can you detect sensitive data by meaning, including business data?	☐ 7 Is every interaction recorded: who, which model, what data, which decision?
☐ 2 Can you see AI conversation content, not just the destination?	☐ 5 When a prompt is risky, can the user still get the work done safely?	☐ 8 Can you answer an auditor's question in minutes, not weeks?
☐ 3 Can you see what retrieval and tools pull into the context?	☐ 6 Are agent tool calls gated, with least privilege per agent?	☐ 9 Can you show what was stopped, not only what happened?

INTRODUCING **MAVS AI**

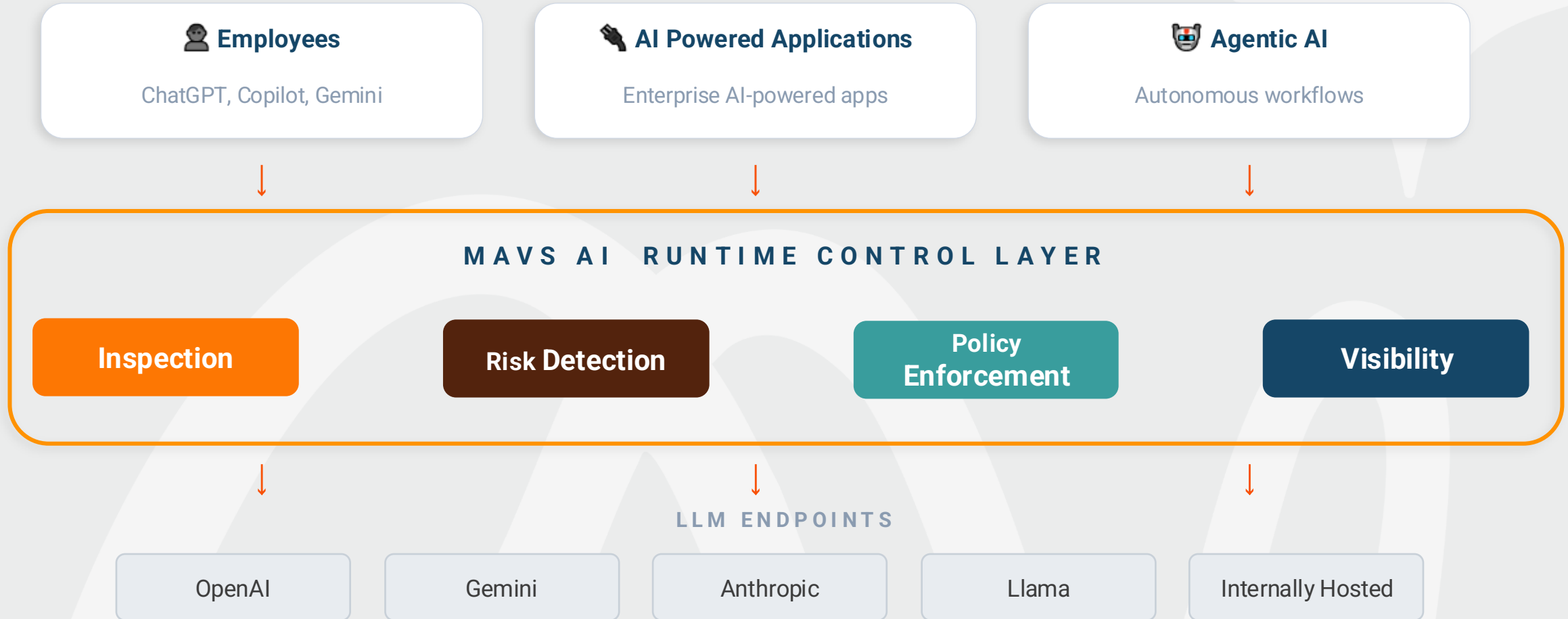
**CONTEXT PRESERVING CONTROL PLANE
THAT ENABLES AND GOVERNS AI. SECURELY.**

**Not block. Not restrict.
Just **ENABLE**.**

HOW IT WORKS



One Layer. Every AI Interaction.





Reagent Solutions

INVOICE

BNM#45360410D

1180 Harbor View Drive, Suite 210
Portland, OR 97035, United States
+1 (416) 987-4665
info@reagents.com

Due Date April 28, 2023

BETTER PHARMA
Austin, TX 78701, United States
info@betterpharma.com

Total Due: \$56,700.00

Item Description	Unit Price	Unit Price	Quantity	Total
[Drug] Novaxel SR	per pack (30 tablets)	\$120.00	150	\$18,000.00
[Compound] KP-338	per gram (R&D)	\$2,400.00	5	\$12,000.00
[Active ingredient] Dapagliflozin base	per kg (API)	\$3,000.00	8	\$24,000.00

Payment Method: Paypal, Visa, Mastercard

Accounts Manager

MARCUS DELROY

Thank you for your business!

SUB TOTAL	\$54,000.00
TAX VAT 10%	\$5,400.00
DISCOUNT 5%	\$2,700.00
TOTAL DUE	\$56,700.00

After

THANK YOU



Three Things to Remember

1

AI risk lives in the interaction

Content is now instruction.

2

Your stack sees fragments

No existing tool follows one interaction from prompt to action.

3

Runtime control layer

Inspect, detect, enforce, prove. With real options between allow and block.



Unleash AI. Securely.

DEMO

info@mavsai.ai | mavsai.ai