Beyond "Digital Assistants"

Securing Agentic AI Systems and Multi-Agent Workflows in High-Stakes Environments



Andra Lezza

Sage

Jeremiah Edwards

Who are we?



Andra Lezza

Principal AppSec Specialist
OWASP London Lead



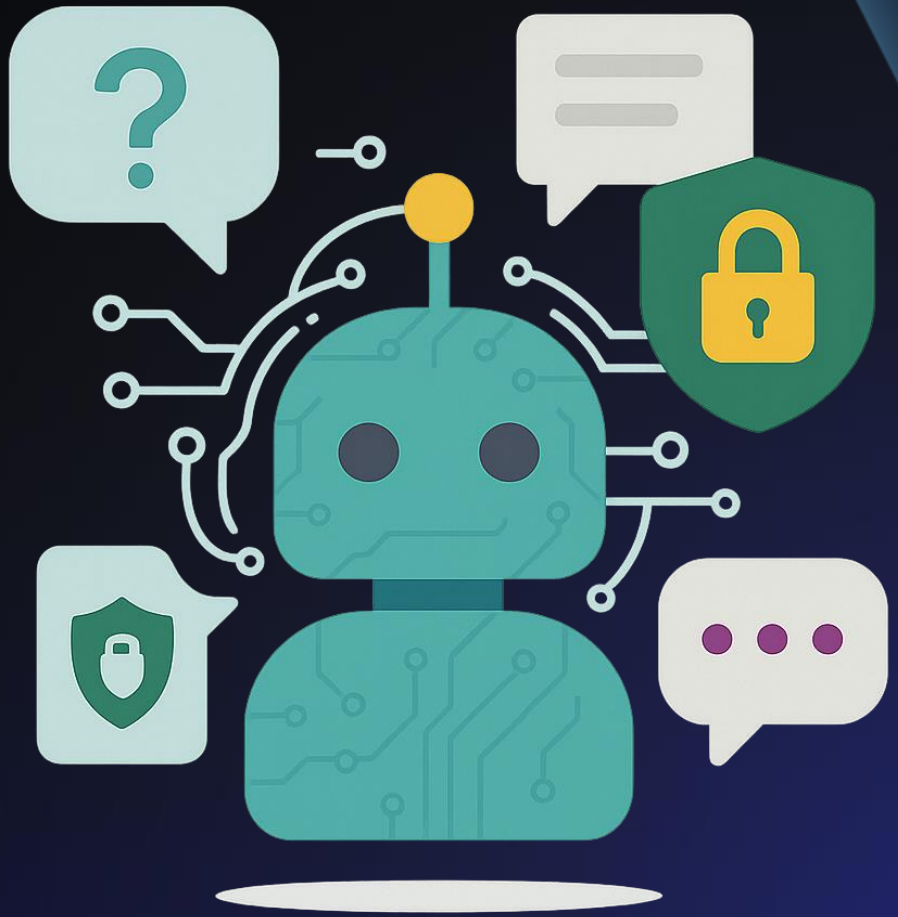
Jeremiah Edwards

Head of Sage AI

*Disclaimer:

The insights in this presentation are solely based on our personal experience and observations and do not represent Sage’s practices or official stance on AI.

How it started



How it’s going



Table of contents

01

⚠️ Real world risks & emerging standards 📝

02

✨ AI Top 10s

03

🤖 AI Agents & MCP

04

💣 Threat modelling 🖼️

05

🔐 (WIP) Security patterns & best practices

06

🏁 Key Takeaways

⚠️ Real world risks & emerging standards 📝

Why it matters



Things are changing

Google's AI agent 'Big Sleep' foils cyberattack in groundbreaking first, says Sundar Pichai

Google CEO Sundar Pichai revealed that its AI agent Big Sleep successfully stopped a cyber exploit before it was executed—marking the first time an AI has proactively prevented a cyberattack. The breakthrough signals a shift in cybersecurity toward AI-powered preemptive defense.

Written By **Ravi Hari**

Published • 15 Jul 2025, 11:32 PM IST

State of Agentic AI

Table of contents

Levels of Agents

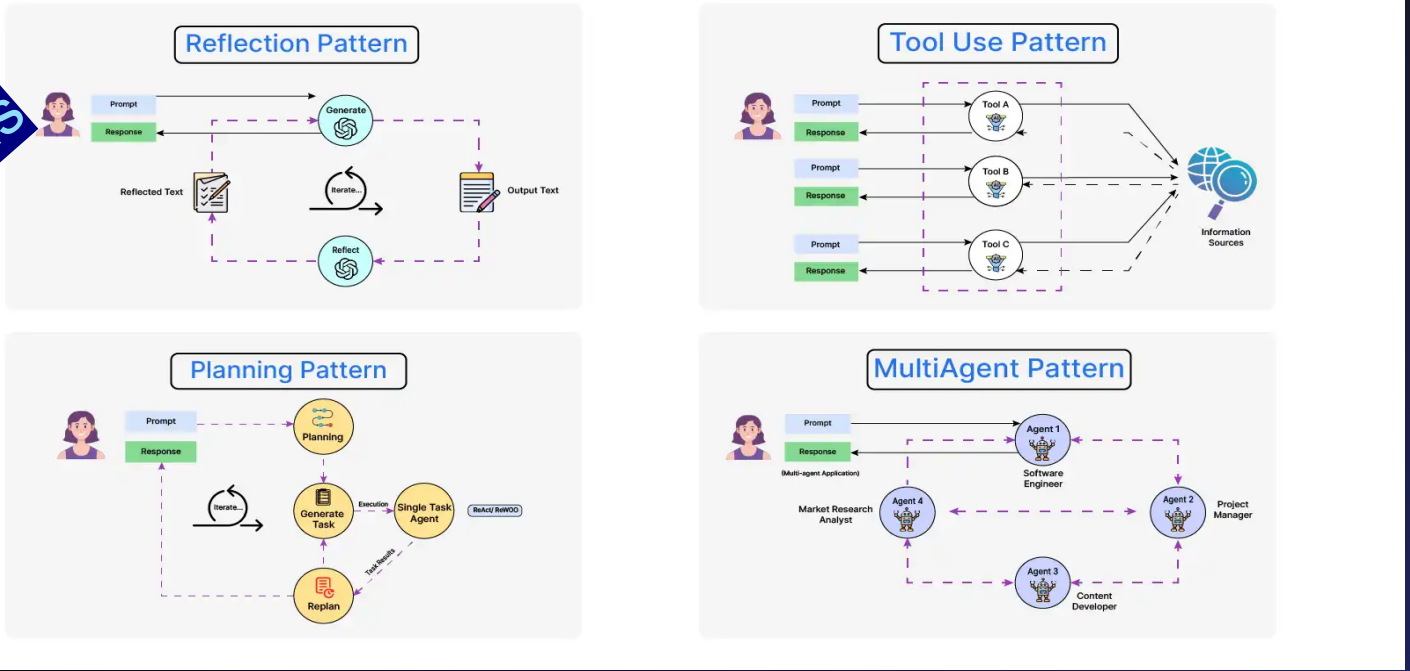
What Are AI Agents?

Your Field Guide to Agents

- Level -1: Fixed Automation – The Digital Assembly Line
- Level 0: LLM-Enhanced – Smarter, but Not Exactly Einstein
- Level 1: ReAct – Reasoning Meets Action
- Level 2: ReAct + RAG – Grounded Intelligence
- Level 3: Tool-Enhanced – The Multi-Taskers
- Level 4: Self-Reflecting – The Philosophers
- Level 5: Memory-Enhanced – The Personalized Powerhouses
- Level 6: Environment Controllers – The World Shapers
- Level 7: Self-Learning – The Evolutionaries

State of Agentic AI

Design Patterns

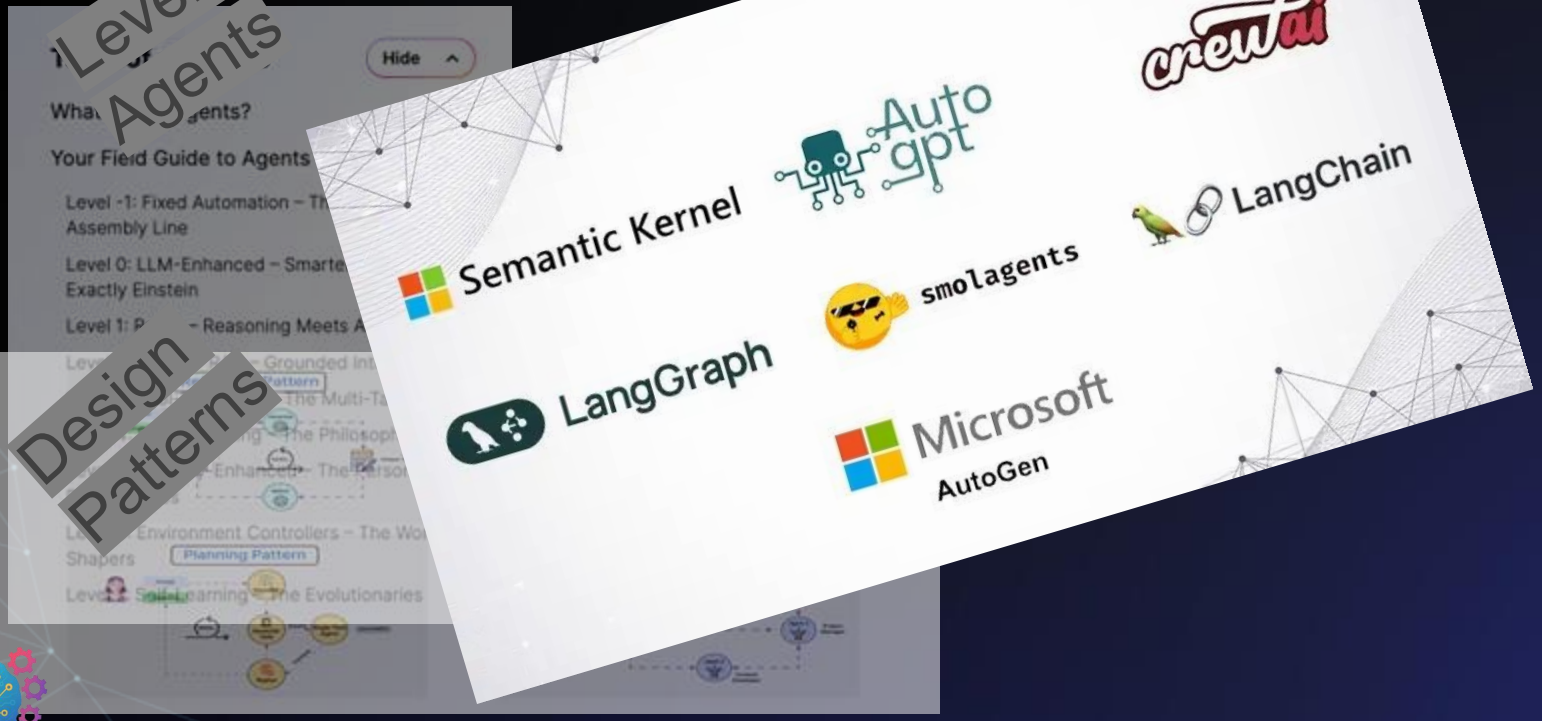


State of Agentic AI

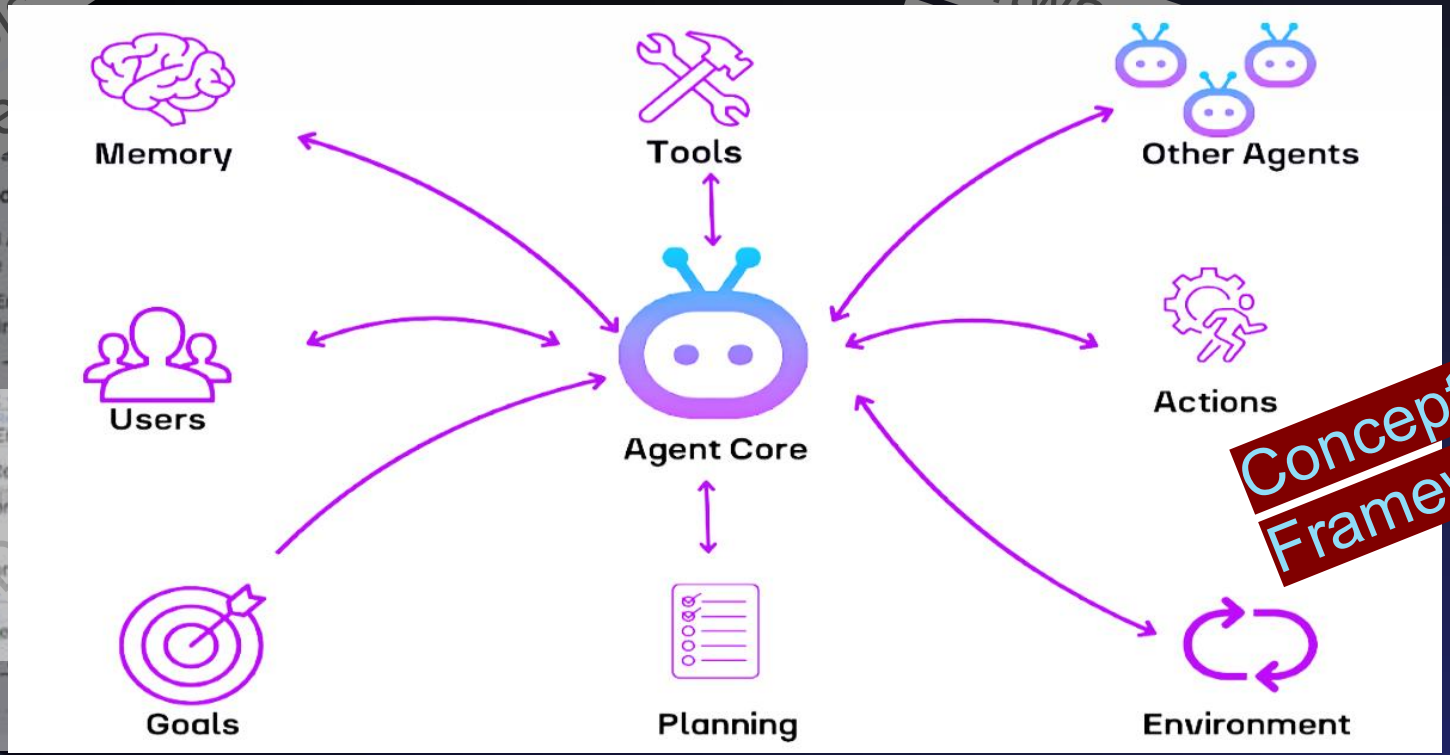
Levels of Agents

Design Patterns

Software Frameworks



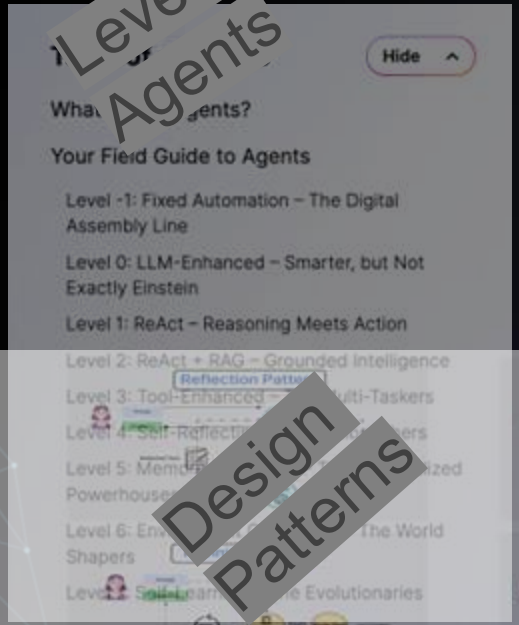
State of Agentic AI



Conceptual Frameworks

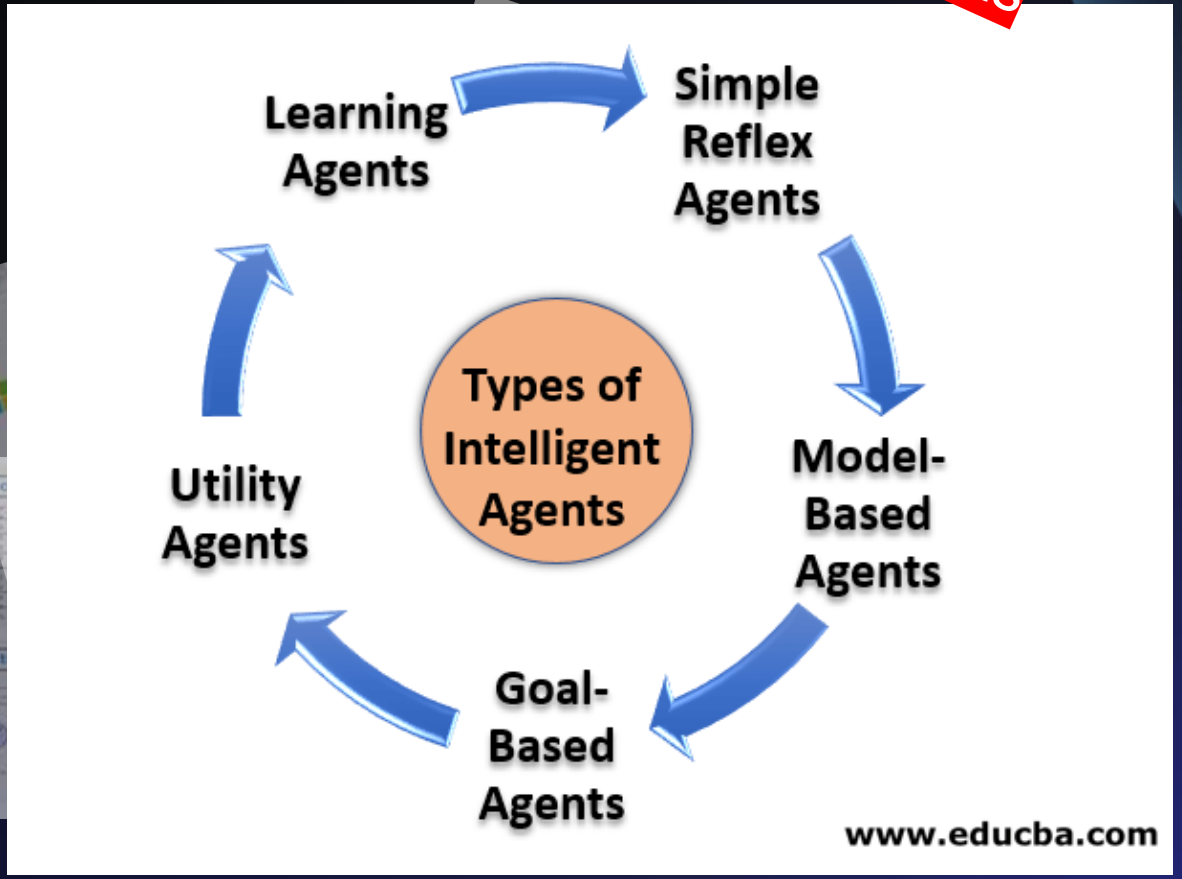
State of Agentic AI

Levels of Agents

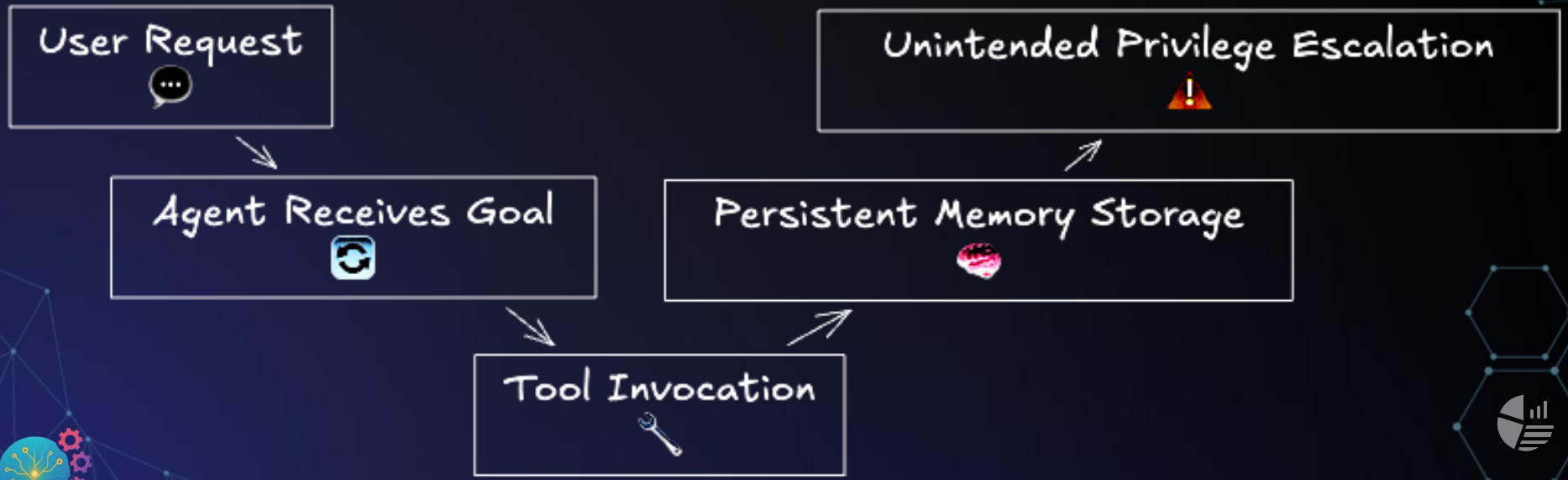


Design Patterns

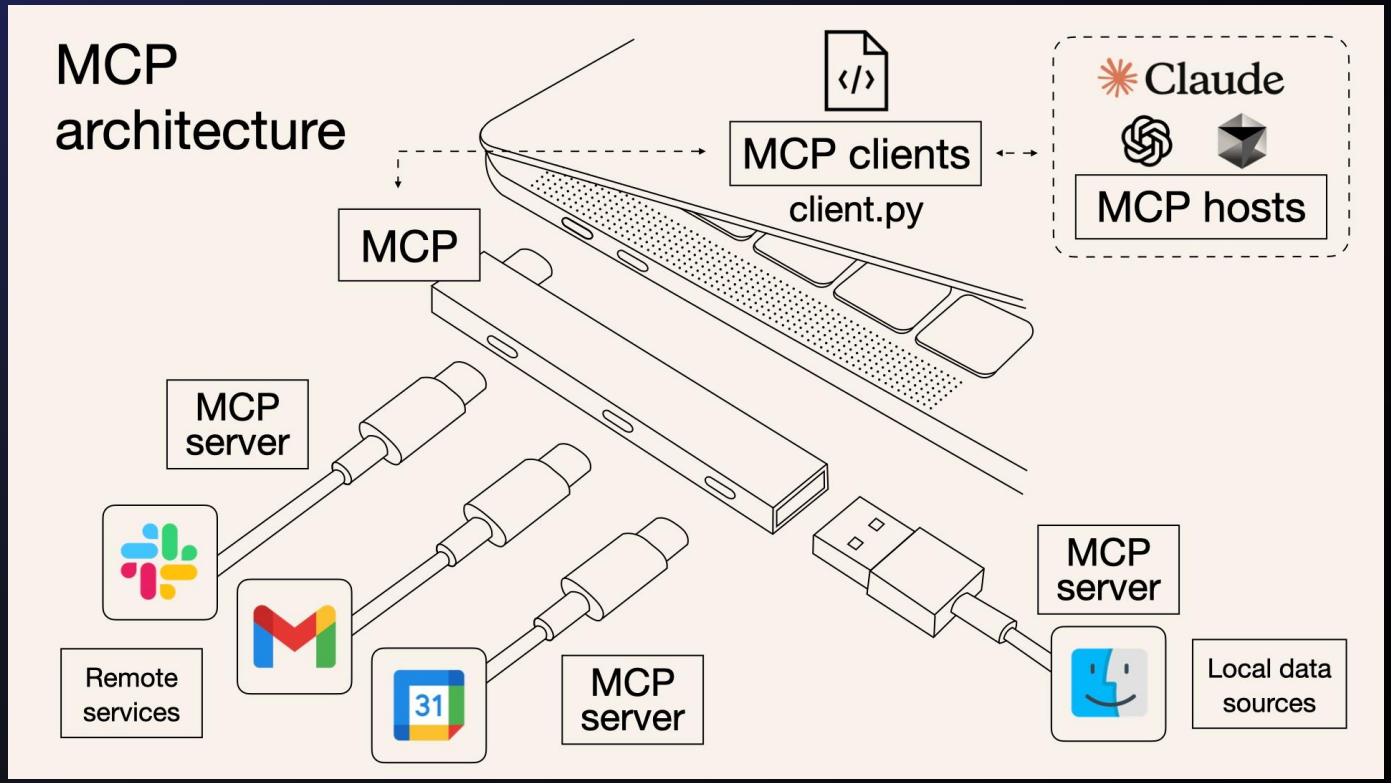
Types of Agents



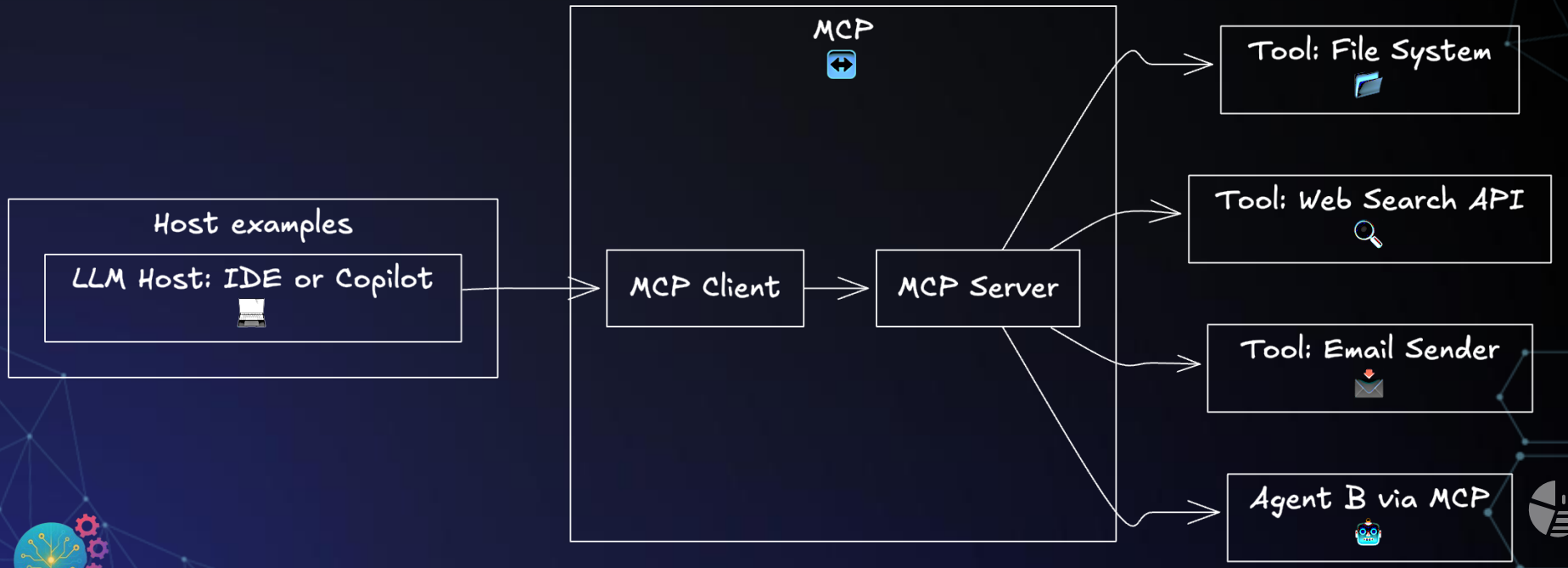
When Autonomy Backfires: Real-World Escalation in Agentic Systems



Model Context Protocol (MCP): what is it?



MCP: Enabler or Risk Multiplier?



High-stakes and trusted AI

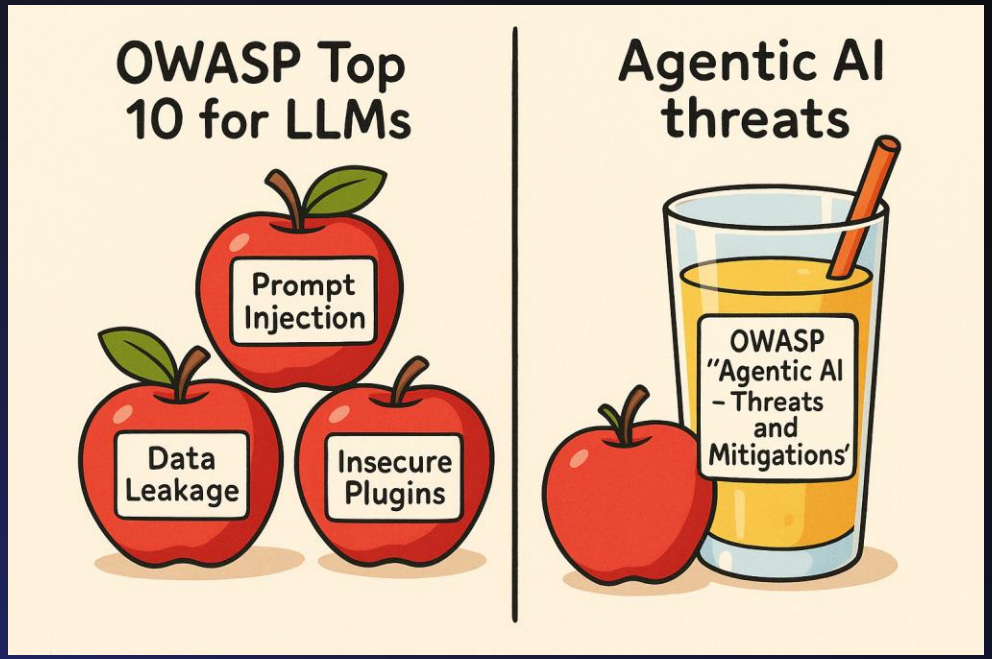


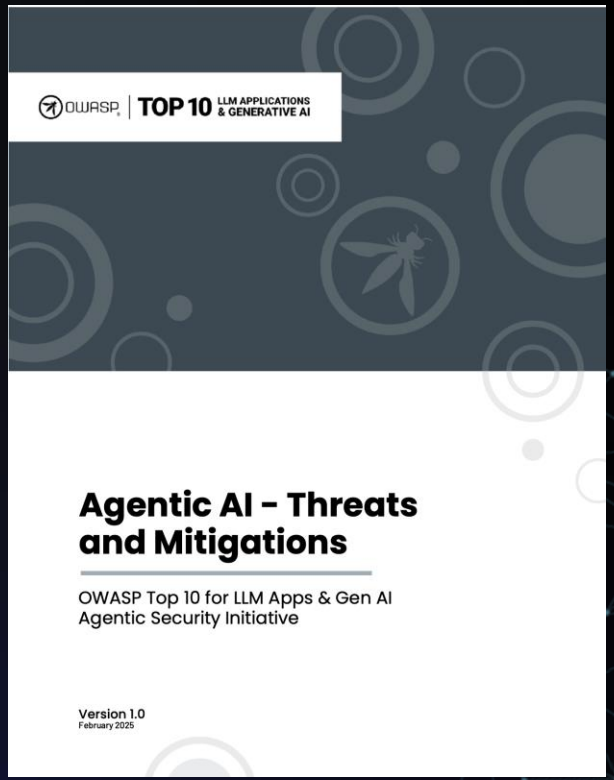
What are we collectively looking to achieve today?

Build trusted and secure AI systems

BY DESIGN AND PROCESS

✨ AI Top 10s


OWASP Top 10 for LLMs
Prompt Injection
Data Leakage
Insecure Plugins
Agentic AI threats
OWASP "Agentic AI – Threats and Mitigations'


OWASP | TOP 10 LLM APPLICATIONS & GENERATIVE AI
Agentic AI - Threats and Mitigations
OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative
Version 1.0
February 2025

OWASP Top 10

Web Apps	LLMs
Broken Access Control	Prompt Injection
Cryptographic Failures	Sensitive Information Disclosure
Injection	Supply Chain
Insecure Design	Data and Model Poisoning
Security Misconfiguration	Improper Output Handling
Vulnerable and Outdated Components	Excessive Agency
Identification/Authentication Failures	New **System Prompt Leakage**
Software and Data Integrity Failures	New **Vector & Embedding Weaknesses**
Security Logging/Monitoring Failures	New **Misinformation**
Server-side Request Forgery	Unbounded Consumption

OWASP Top 10

LLM	Agentic AI Threats
LLM01:2025 Prompt Injection	T6: Intent Breaking & Goal Manipulation T2: Tool Misuse
LLM02:2025 Sensitive Information Disclosure	T1: Memory Poisoning T12: Agent Communication Poisoning
LLM03:2025 Supply Chain	T3: Privilege Compromise T13: Rogue Agents T21: Service Account Exposure
LLM04:2025 Model Poisoning	T1: Memory Poisoning T17: Semantic Drift in Embeddings
LLM05:2025 Improper Output Handling	T2: Tool Misuse T11: Unexpected RCE and Code Attacks



OWASP AgenticAI – Threats and Mitigations, Version 1.0, Feb 2025

OWASP Top 10

LLM	Agentic AI Threats
LLM06:2025 Excessive Agency	T3: Privilege Compromise T2: Tool Misuse T13: Rogue Agents
LLM07:2025 System Prompt Leakage	**T8: Repudiation & Untraceability** **T9: Identity Spoofing**
LLM08:2025 Vector & Embedding Weaknesses	**T17: Semantic Drift** **T18: RAG Input Manipulation** **T28: RAG Data Exfiltration**
LLM09:2025 Misinformation	**T5: Cascading Hallucination Attacks** **T7: Misaligned/Deceptive Behaviours**
LLM10:2025 Unbounded Consumption	T4: Resource Overload T10: Overwhelming HITL

OWASP AgenticAI – Threats and Mitigations, Version 1.0, Feb 2025

OWASP Top 10

LLM	Agentic AI Threats
LLM06:2025 Excessive Agency	T3: Privilege Compromise T2: Tool Misuse T13: Rogue Agents
LLM07:2025 System Prompt Leakage	**T8: Repudiation & Untraceability** **T9: Identity Spoofing**
LLM08:2025 Vector & Embedding Weaknesses	**T17: Semantic Drift** **T18: RAG Input Manipulation** **T28: RAG Data Exfiltration**
LLM09:2025 Misinformation	**T5: Cascading Hallucination Attacks** **T7: Misaligned/Deceptive Behaviours**
LLM10:2025 Unbounded Consumption	T4: Resource Overload T10: Overwhelming HITL

OWASP AgenticAI – Threats and Mitigations, Version 1.0, Feb 2025

OWASP Top 10

LLM	Agentic AI Threats
LLM06:2025 Excessive Agency	T3: Privilege Compromise T2: Tool Misuse T13: Rogue Agents
LLM07:2025 System Prompt Leakage	**T8: Repudiation & Untraceability** **T9: Identity Spoofing**
LLM08:2025 Vector & Embedding Weaknesses	**T17: Semantic Drift** **T18: RAG Input Manipulation** **T28: RAG Data Exfiltration**
LLM09:2025 Misinformation	**T5: Cascading Hallucination Attacks** **T7: Misaligned/Deceptive Behaviours**
LLM10:2025 Unbounded Consumption	T4: Resource Overload T10: Overwhelming HITL

OWASP AgenticAI – Threats and Mitigations, Version 1.0, Feb 2025

OWASP Top 10

LLM	Agentic AI Threats
LLM06:2025 Excessive Agency	T3: Privilege Compromise T2: Tool Misuse T13: Rogue Agents
LLM07:2025 System Prompt Leakage	**T8: Repudiation & Untraceability** **T9: Identity Spoofing**
LLM08:2025 Vector & Embedding Weaknesses	**T17: Semantic Drift** **T18: RAG Input Manipulation** **T28: RAG Data Exfiltration**
LLM09:2025 Misinformation	**T5: Cascading Hallucination Attacks** **T7: Misaligned/Deceptive Behaviours**
LLM10:2025 Unbounded Consumption	T4: Resource Overload T10: Overwhelming HITL

Top 10 LLMs vs. Agentic AI Threats...

→ Top 10 for Agentic Applications



GenAI SECURITY PROJECT
genai.owasp.org

2025 Project Asset Roadmap, Q2 Update

INITIATIVES	Q1'25 (Jan-Mar)	Q2'25 (Apr-Jun)	Q3'25 (Jul-Sept)	Q4'25 (Oct-Dec)	Q1'26 (Jan-Feb)
Top 10 List for LLMs & Gen AI Applications				Nov announcement - T10 for LLM 2026	Aimed publish release of T10 for LLM 2026
Solutions Landscape Security Solutions and Trends for LLM and Gen AI apps	Mar - Landscape Doc Online Solution Dir - Monthly	Jun - Landscape Doc Online Solution Dir - Monthly Jun Incident Round-up Blog	Sept - Landscape Doc Online Solution Dir - Monthly Aug - Incident Round-up Blog	Dec - Landscape Doc Online Solution Dir - Monthly Nov - Incident Round-up Blog	Dec- Landscape Doc Online Solution Dir - Monthly
AI Threat Intelligence *Ongoing research & guidance for Gen AI incident response*	Jan - Leveraging LLMs for Exploitation		Jul - AI Incident Response Guide	Dec - DeepFake Guide Nov - "Three Years of GenAI" Threat Intel Roundup October - Second Exploit Generation Report	
			Jul - Data Call (Community Ask - data share) (July-Aug) Aug - Datasets update (Exploit/ Risk Management/Vulnerabilities)	Nov - Start Data Mappings T10 LLM 2026 (aligned with T10) Dec - LLM Data Security Best Practices v2 - (comment)	Jan - Data Mappings T10 LLM 2026 (CIS, etc) Jan - LLM Data Security BP v2
			Jul - OWASP AI Red Team Compass v1.5 Jul/Aug - GenAI Red Team Testing Handbook (Public Draft) Aug - A CSO's Guide to Gen AI Red Teaming (An OWASP Publication)	Nov - OWASP AI Red Team Compass v2.0 Nov - Gen Ai Red Teaming Guide 1.1	
			Jul - Cklist COMPASS v1.0 2025 Jul 4-10 minute training videos Sept - update Governance Checklist	Nov - AI Security CoE 2.0	Mar - AI Security CoE
			July - Securing Agentic Apps July - State of Agentic AI Security and Governance Aug - Agentic AI Threat Model and Vulns 1.1	Nov: Top 10 for Agentic Applications	February: Reference example Agentic Applications
Localization/Translations	Top 10: • Spanish • German • Simplified Chinese • Traditional Chinese • Portuguese • Russian	Top 10: • Hindi • Greek • Persian • Korean • Japanese	Top 10: • Dutch • Bahasa Indonesia • Italian • Khmer • Vietnamese • Arabic • French	CISO Governance Checklist + RT CSO RT Guide + RT Compass • Portuguese • Simplified Chinese • Traditional Chinese • German • Hindi • + TBD	Top 10 2026: • Portuguese • Simplified Chinese • Traditional Chinese • German • Hindi • + TBD



https://genai.owasp.org/roadmap/



AI Agents & MCP



What is an Agent anyway?

OpenAI

AI systems that can pursue complex goals with limited direct supervision.

Degree of Agenticness the degree to which a system can adaptably achieve complex goals in complex environments with limited direct supervision

Anthropic

Workflows - systems where LLMs and tools are orchestrated through predefined code paths.

Agents - systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish task

NVIDIA

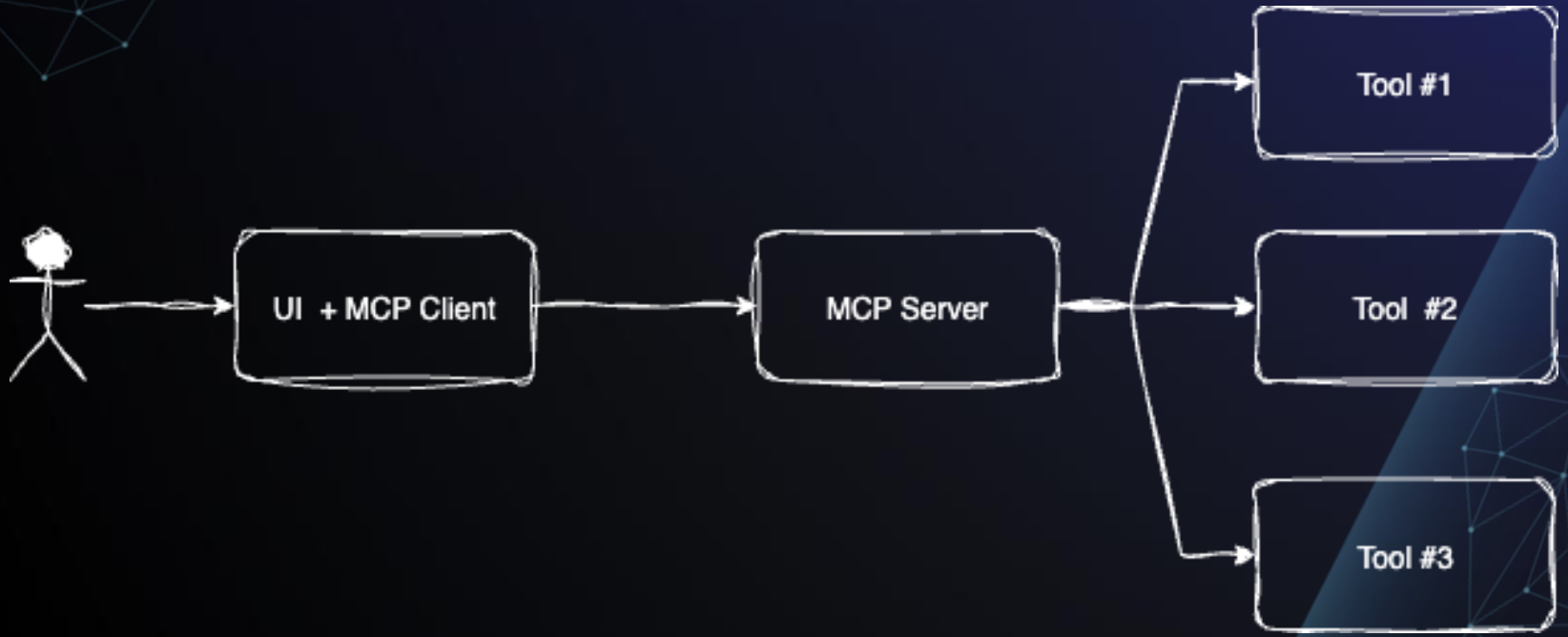
Agentic AI uses sophisticated reasoning and iterative planning to autonomously solve complex, multi-step problems.

Degree of perceived intelligence and capability

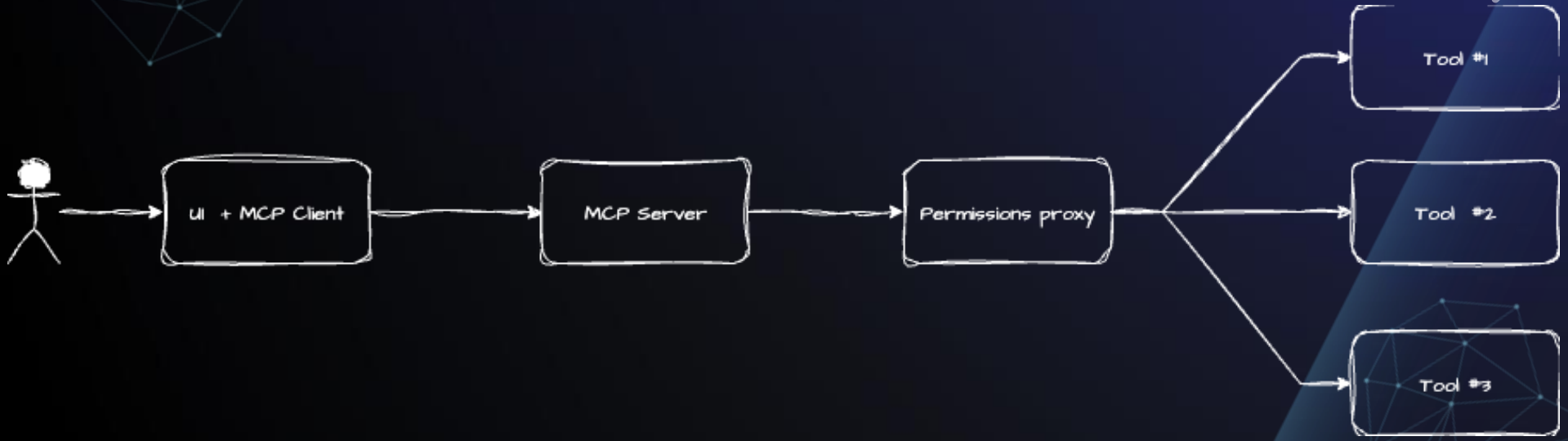


Russell & Norvig (2003). Artificial Intelligence: A Modern Approach

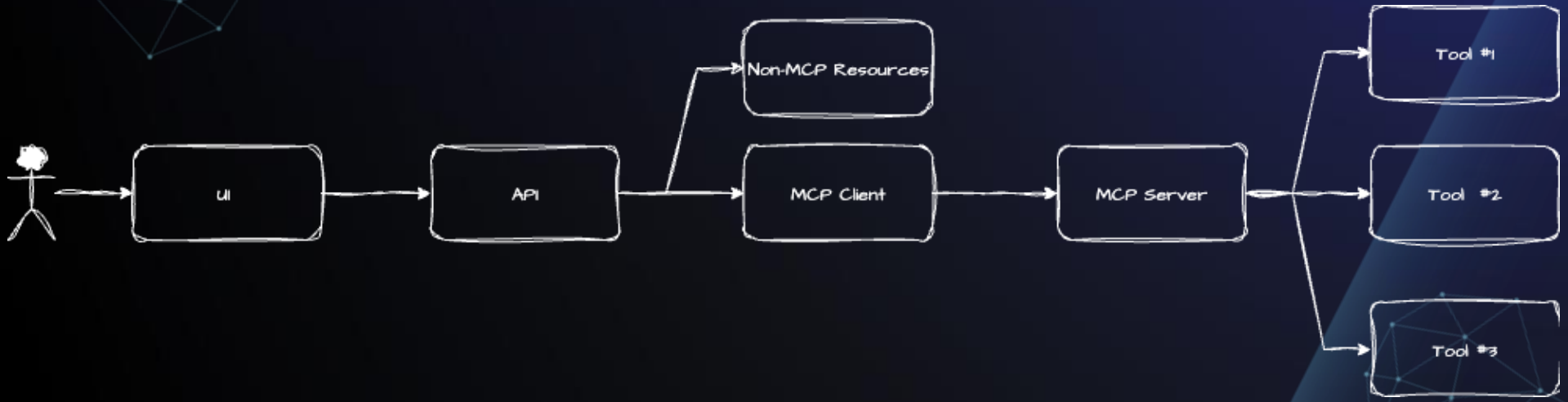
MCP: More clients, more servers, more problems



MCP: More clients, more servers, more problems



MCP: More clients, more servers, more problems



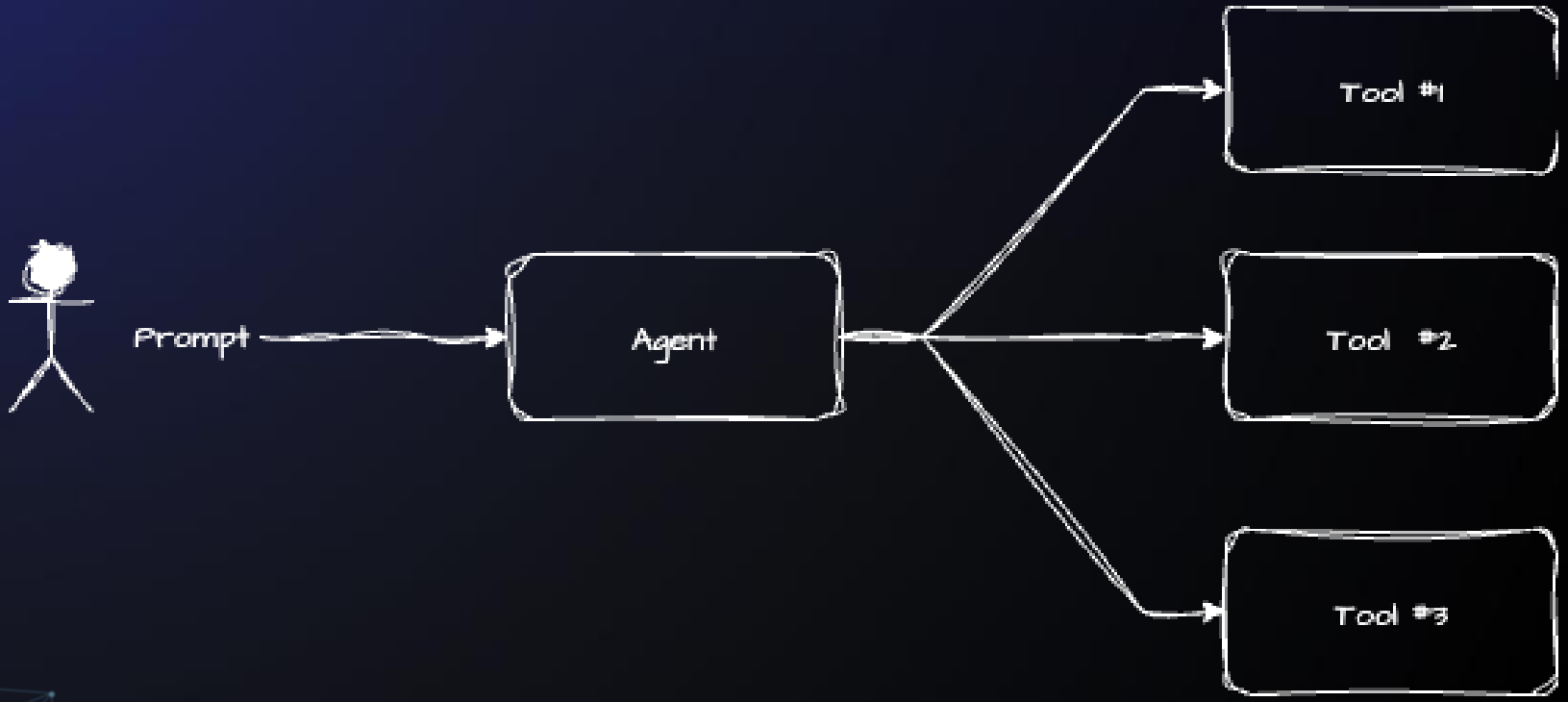


Threat modelling

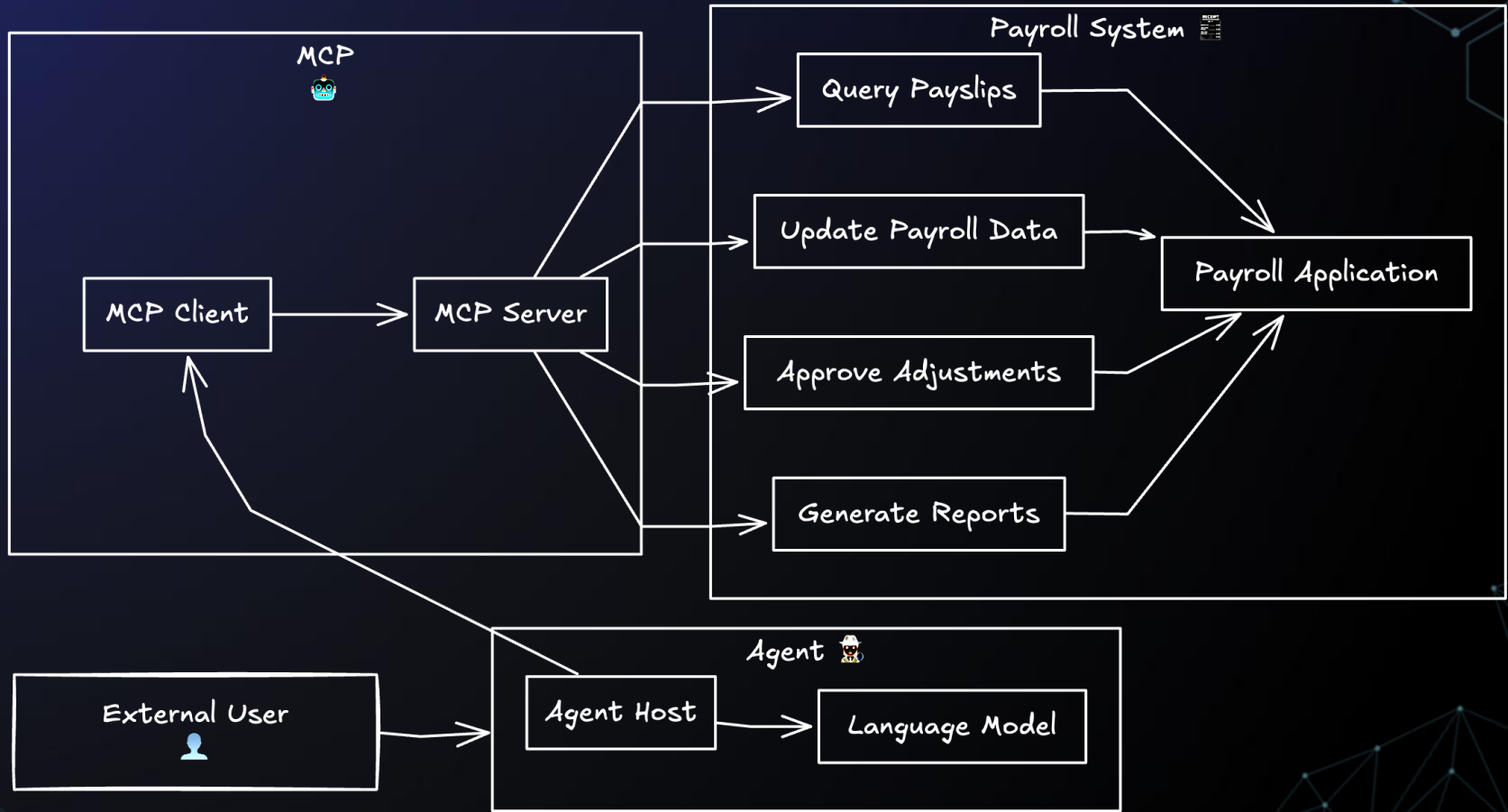




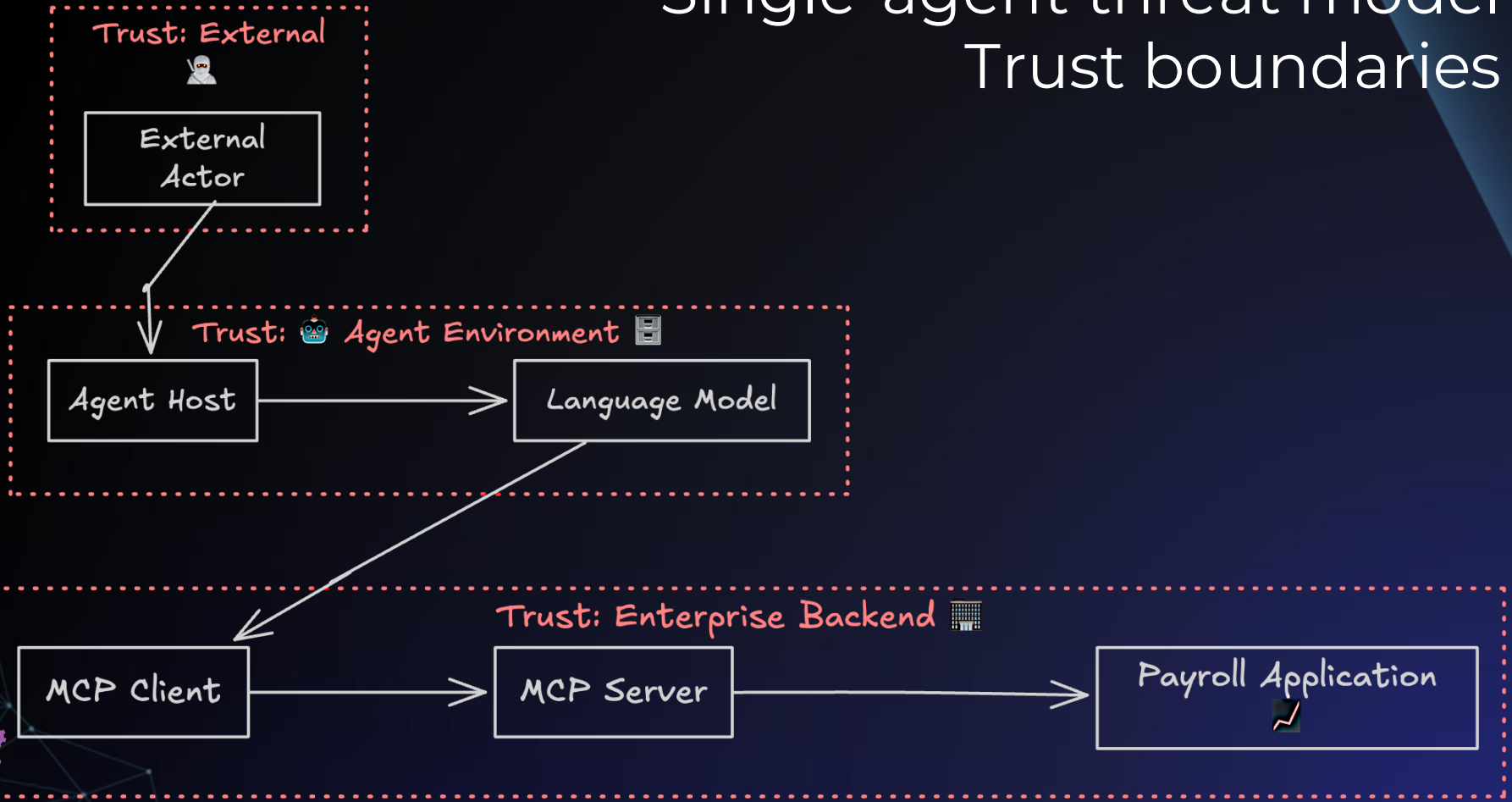
Single Agent, many tools



Single-agent architecture example



Single-agent threat model
Trust boundaries



Threat scenarios



Prompt Injection



Data Leak



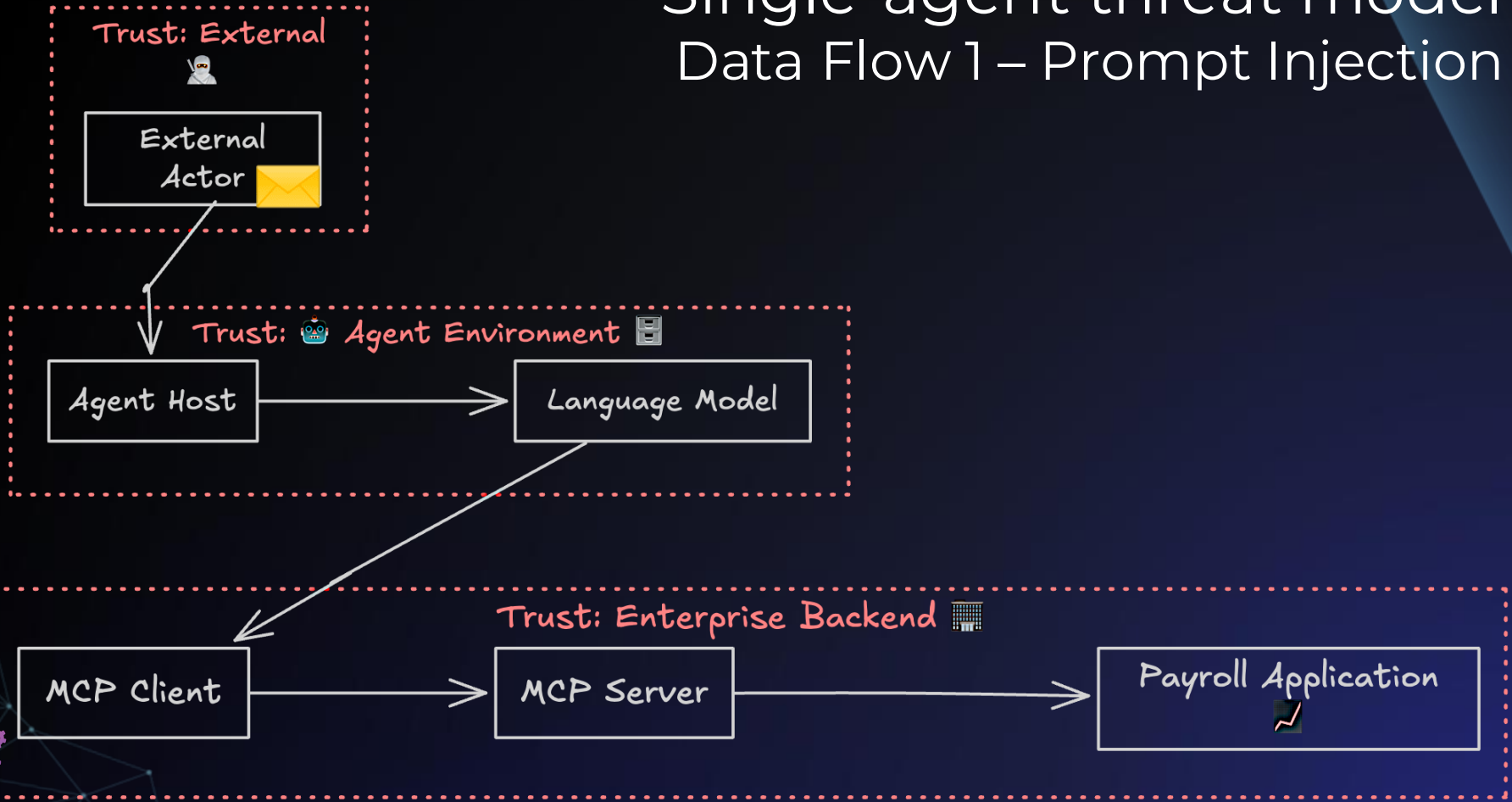
Business Logic



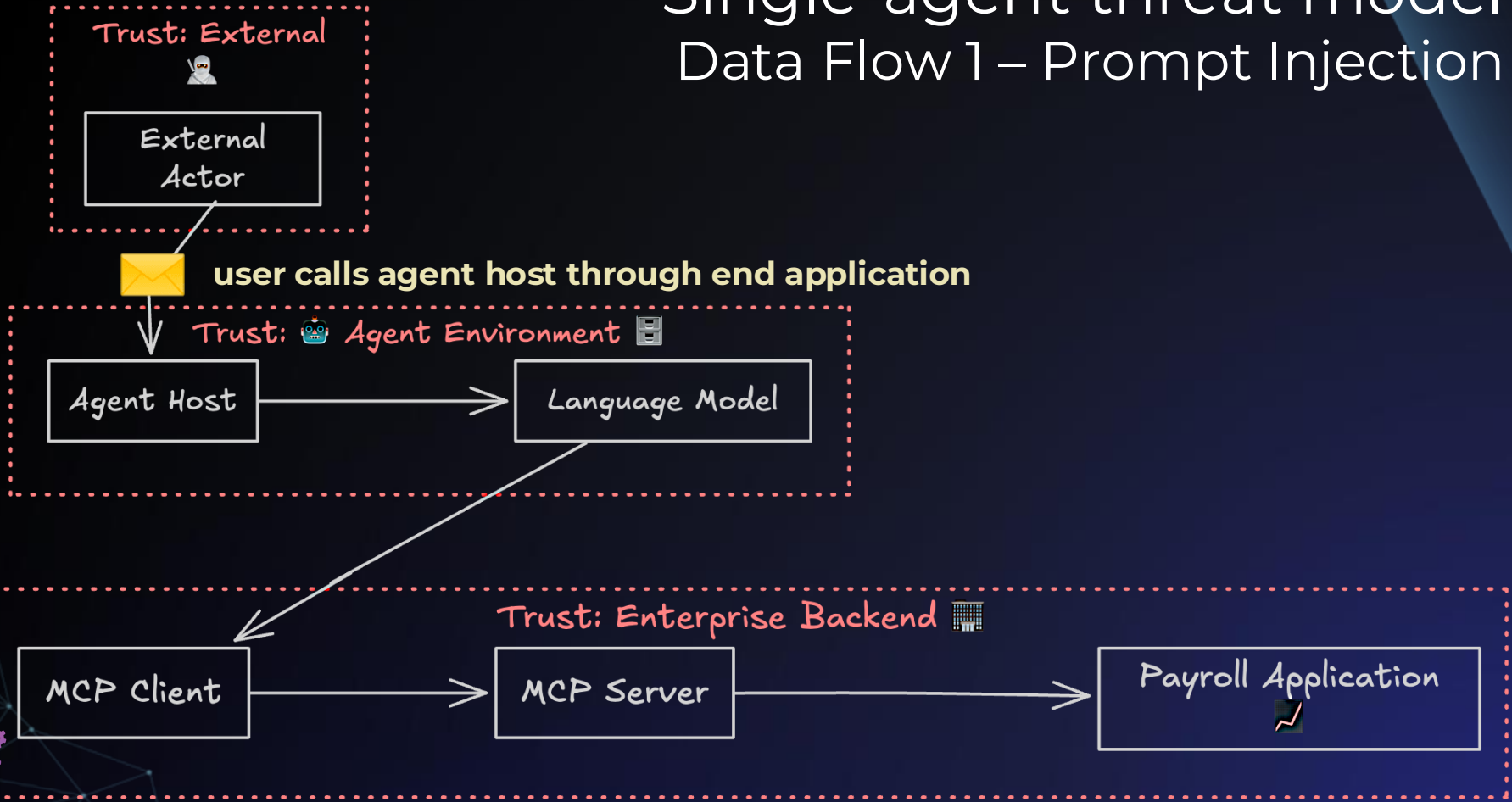
Prompt Injection



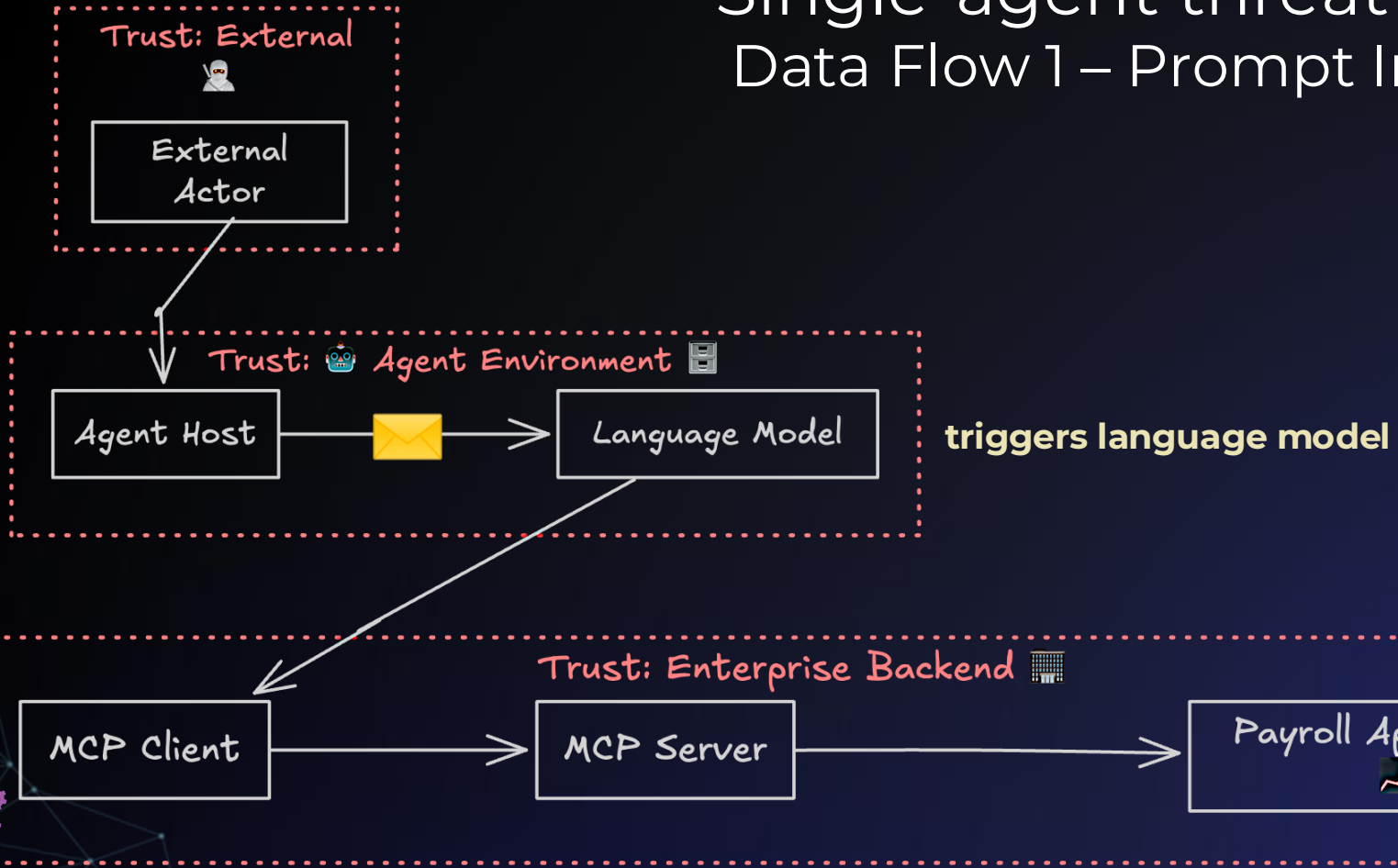
Single-agent threat model
Data Flow 1 – Prompt Injection



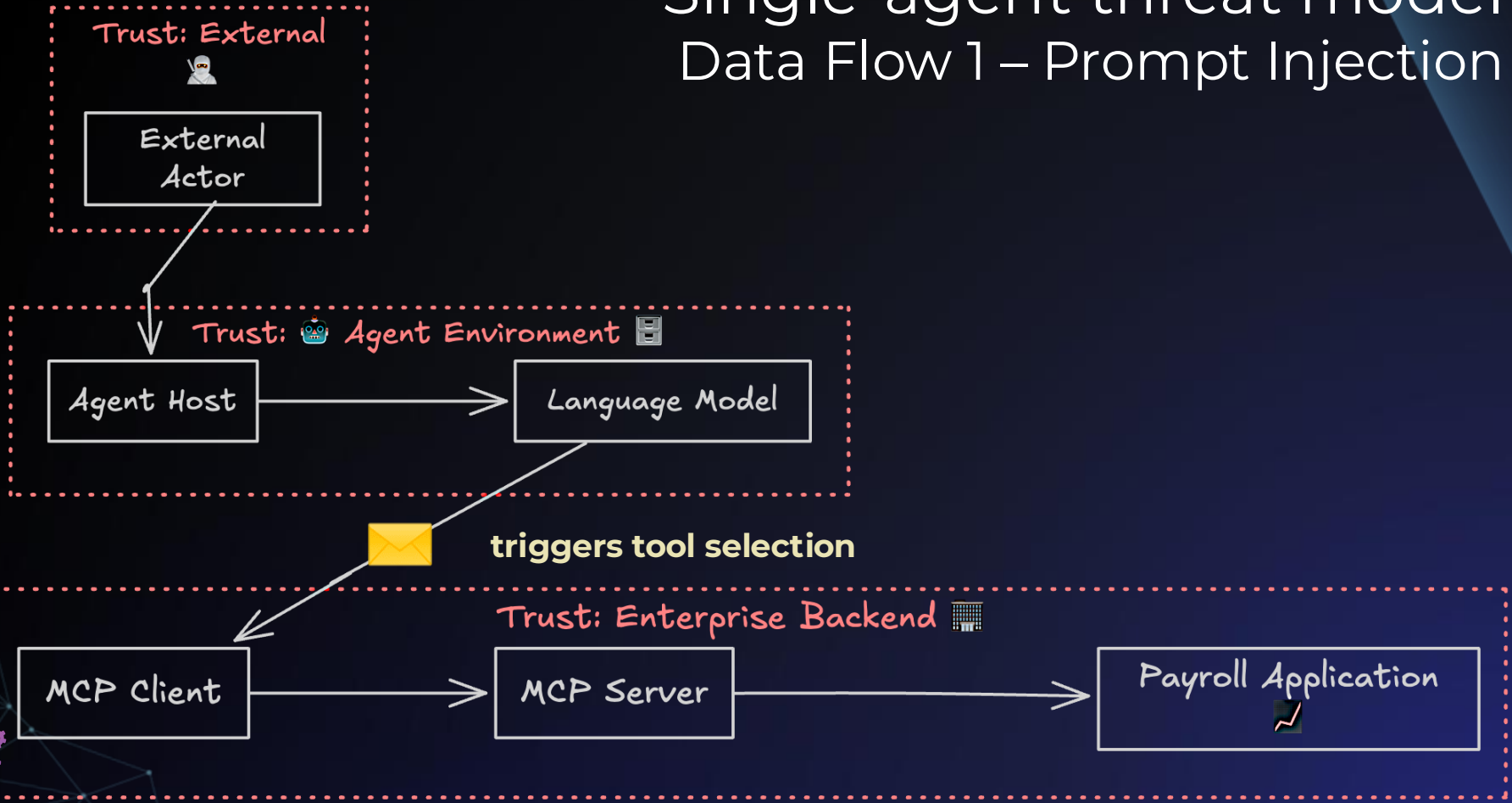
Single-agent threat model
Data Flow 1 – Prompt Injection



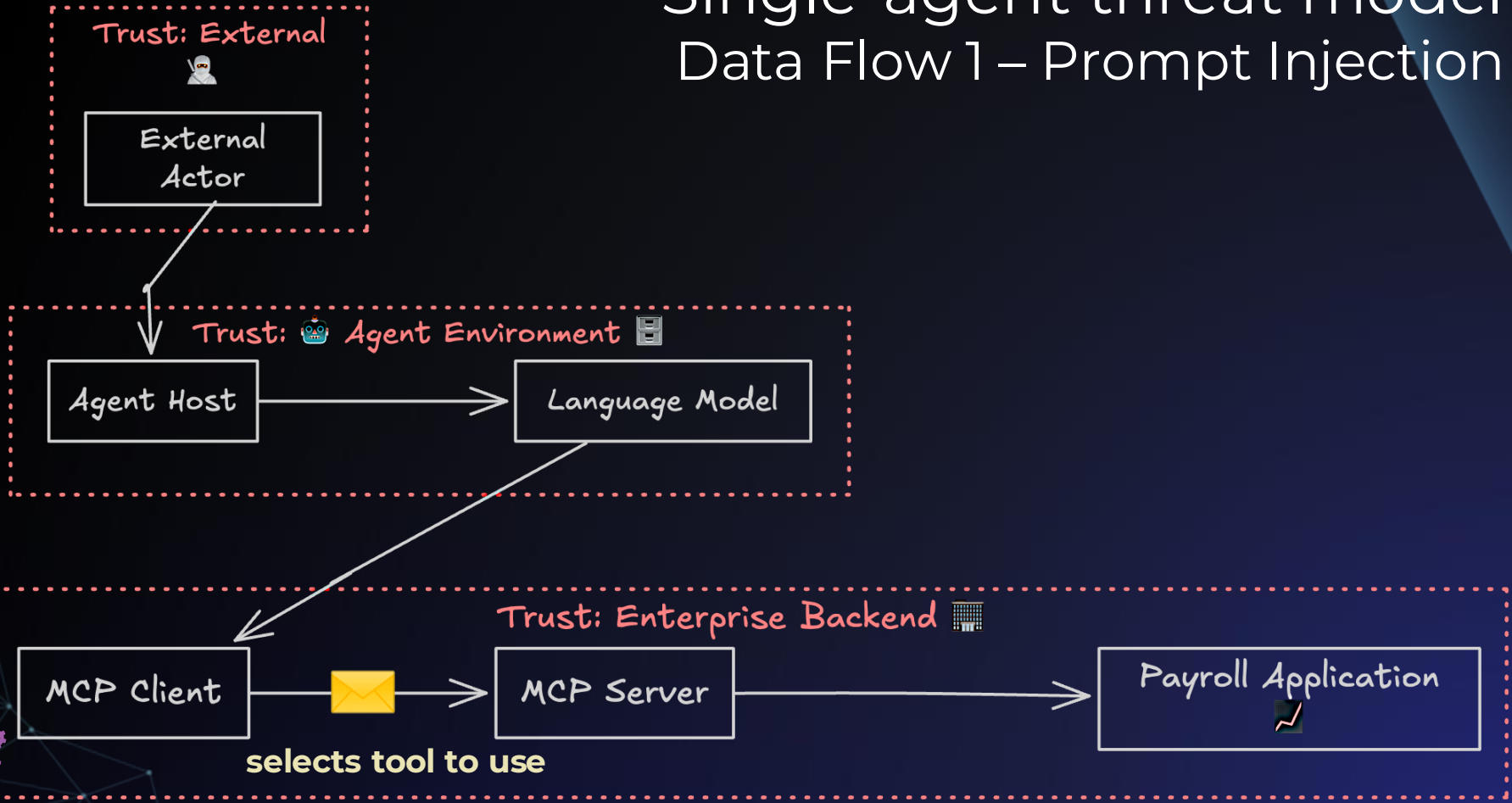
Single-agent threat model
Data Flow 1 – Prompt Injection



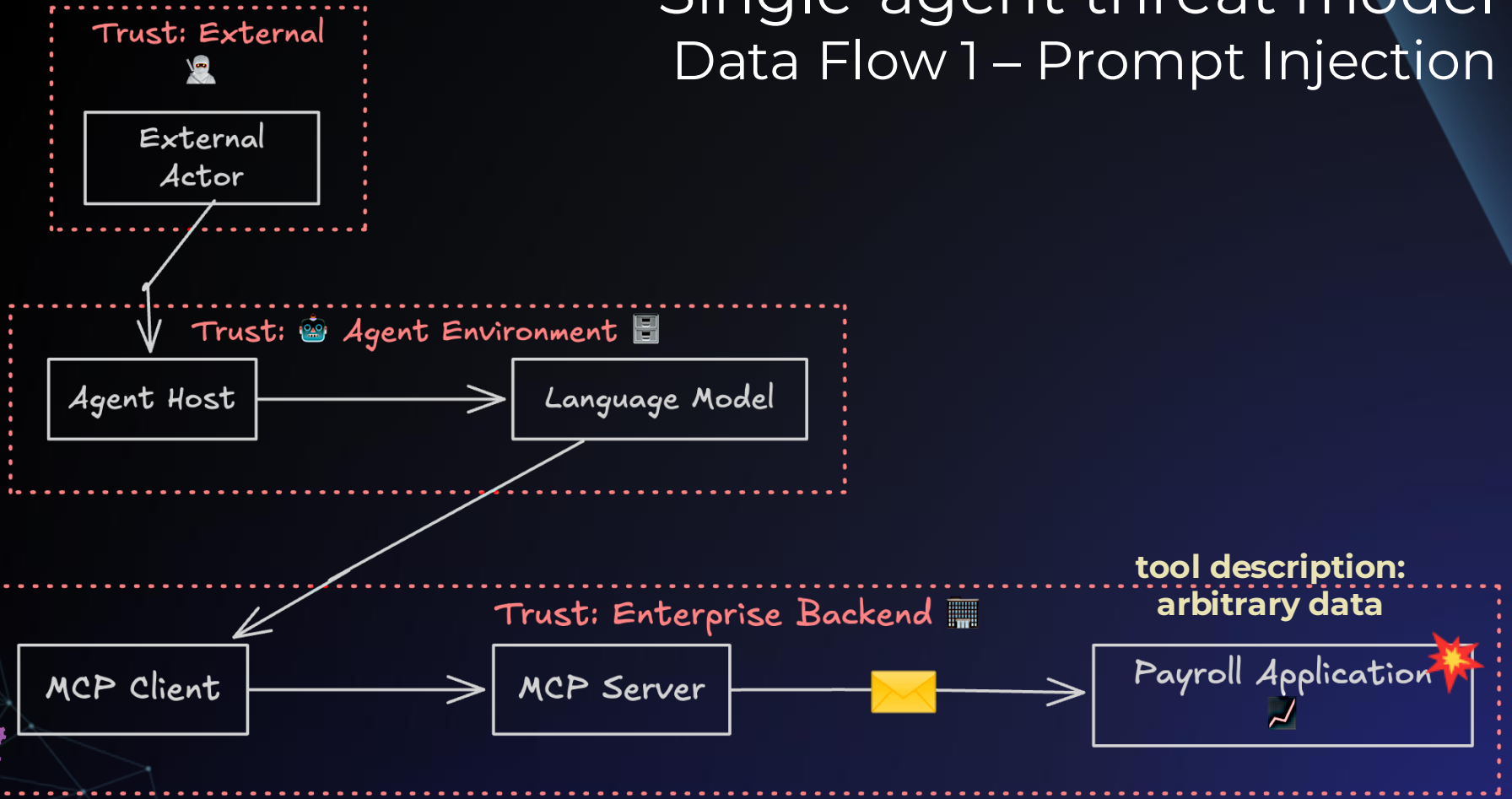
Single-agent threat model
Data Flow 1 – Prompt Injection



Single-agent threat model
Data Flow 1 – Prompt Injection



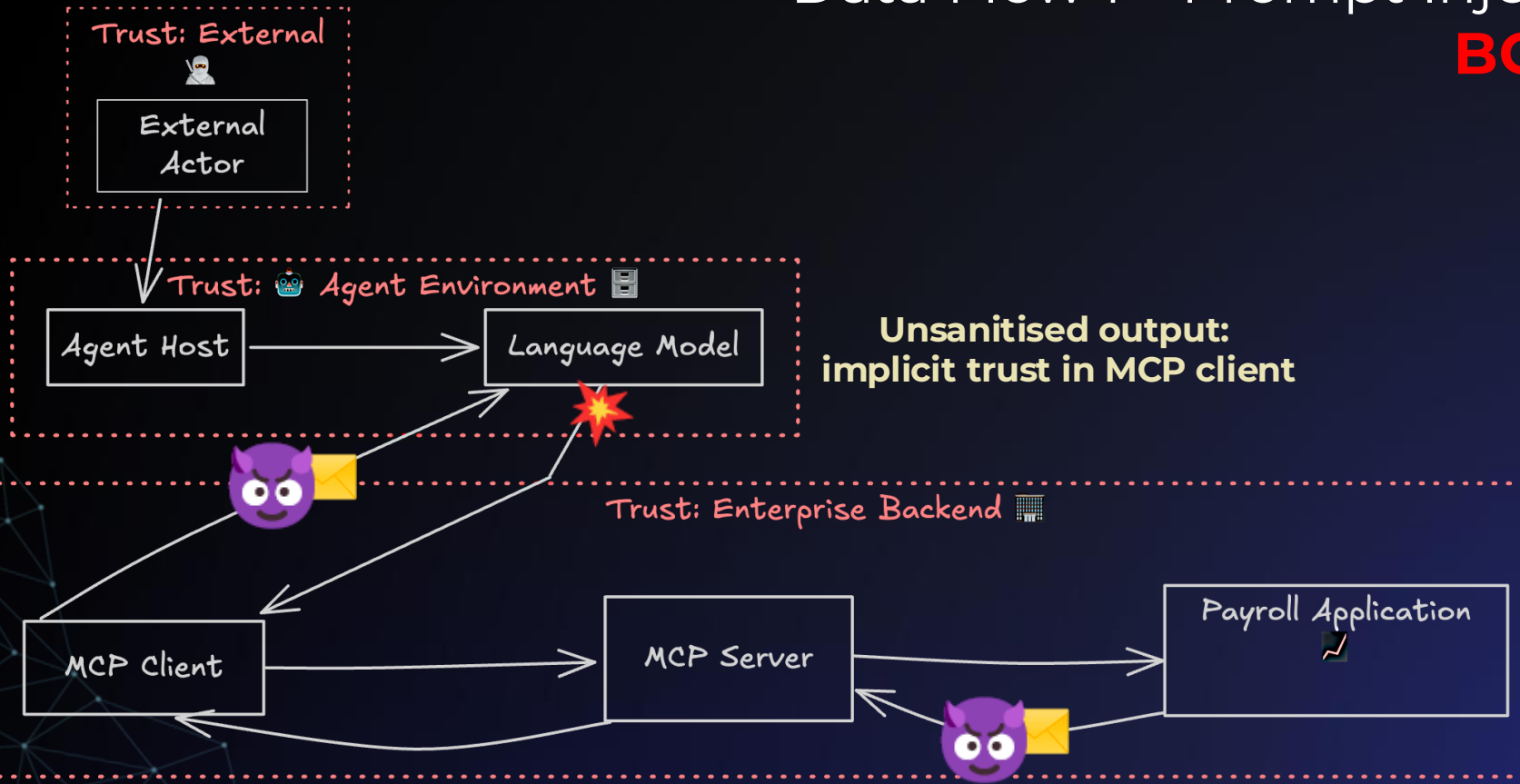
Single-agent threat model
Data Flow 1 – Prompt Injection



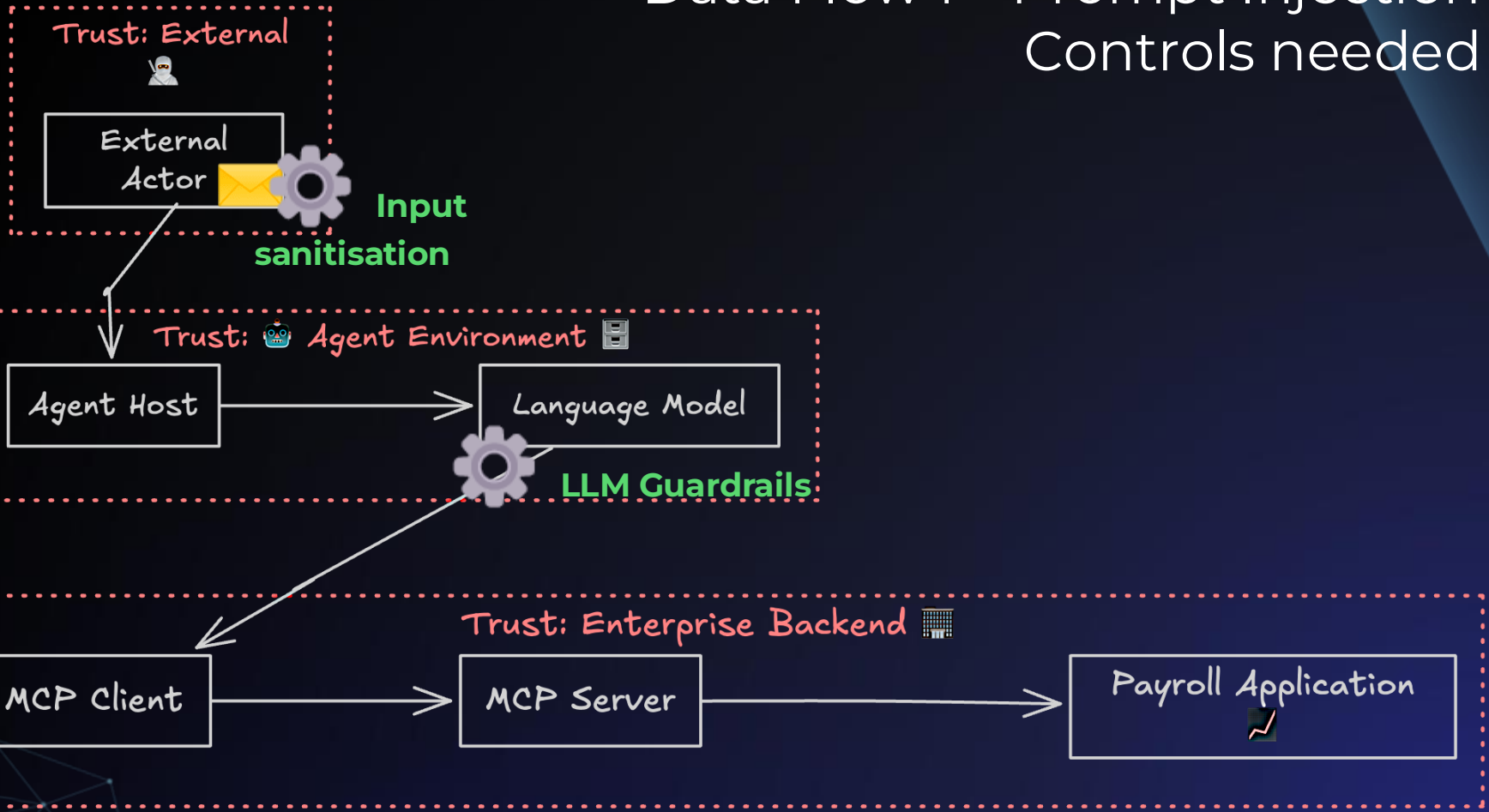
Single-agent threat model

Data Flow 1 – Prompt Injection

BONUS

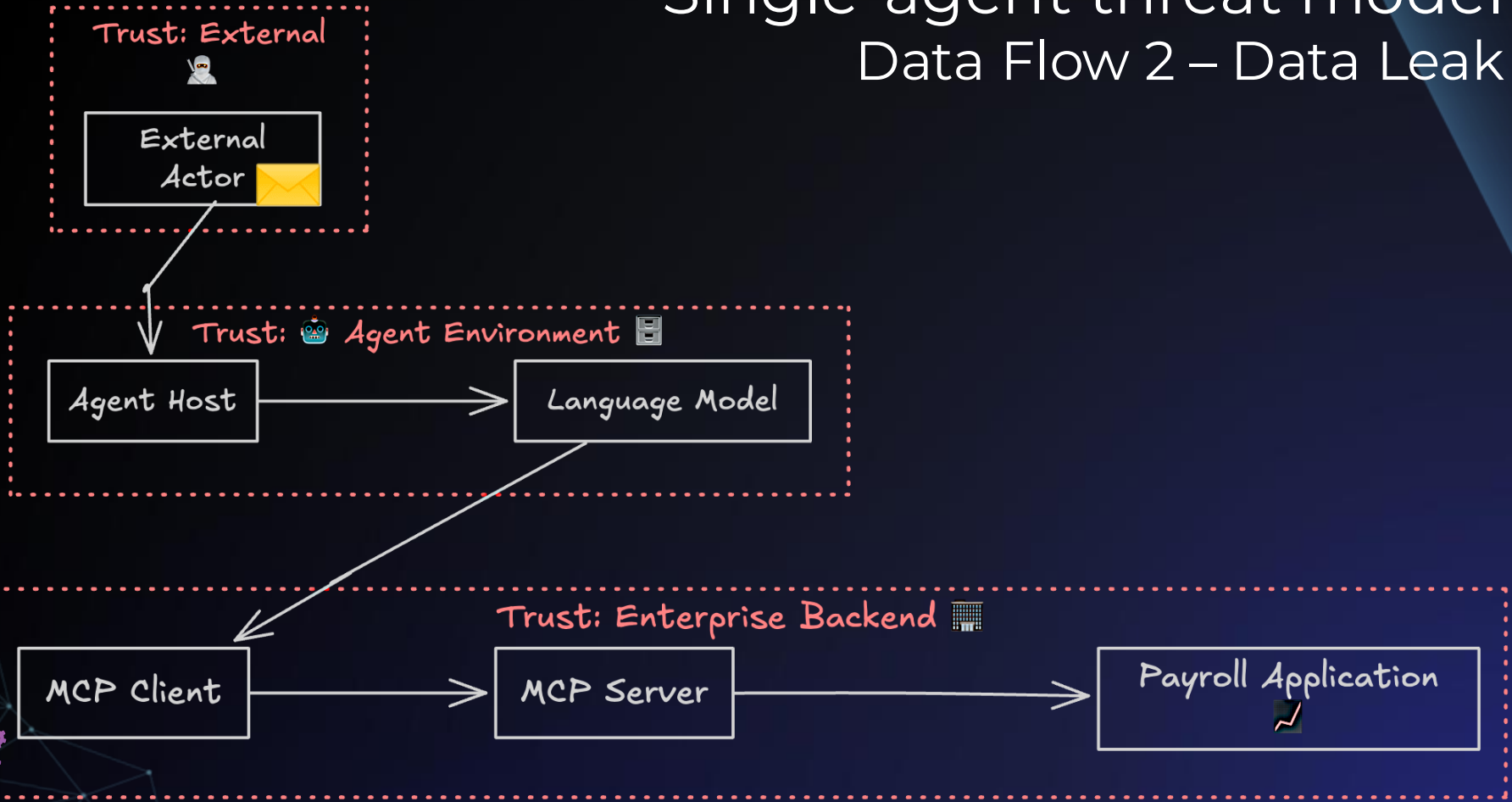


Data Flow 1 – Prompt Injection Controls needed

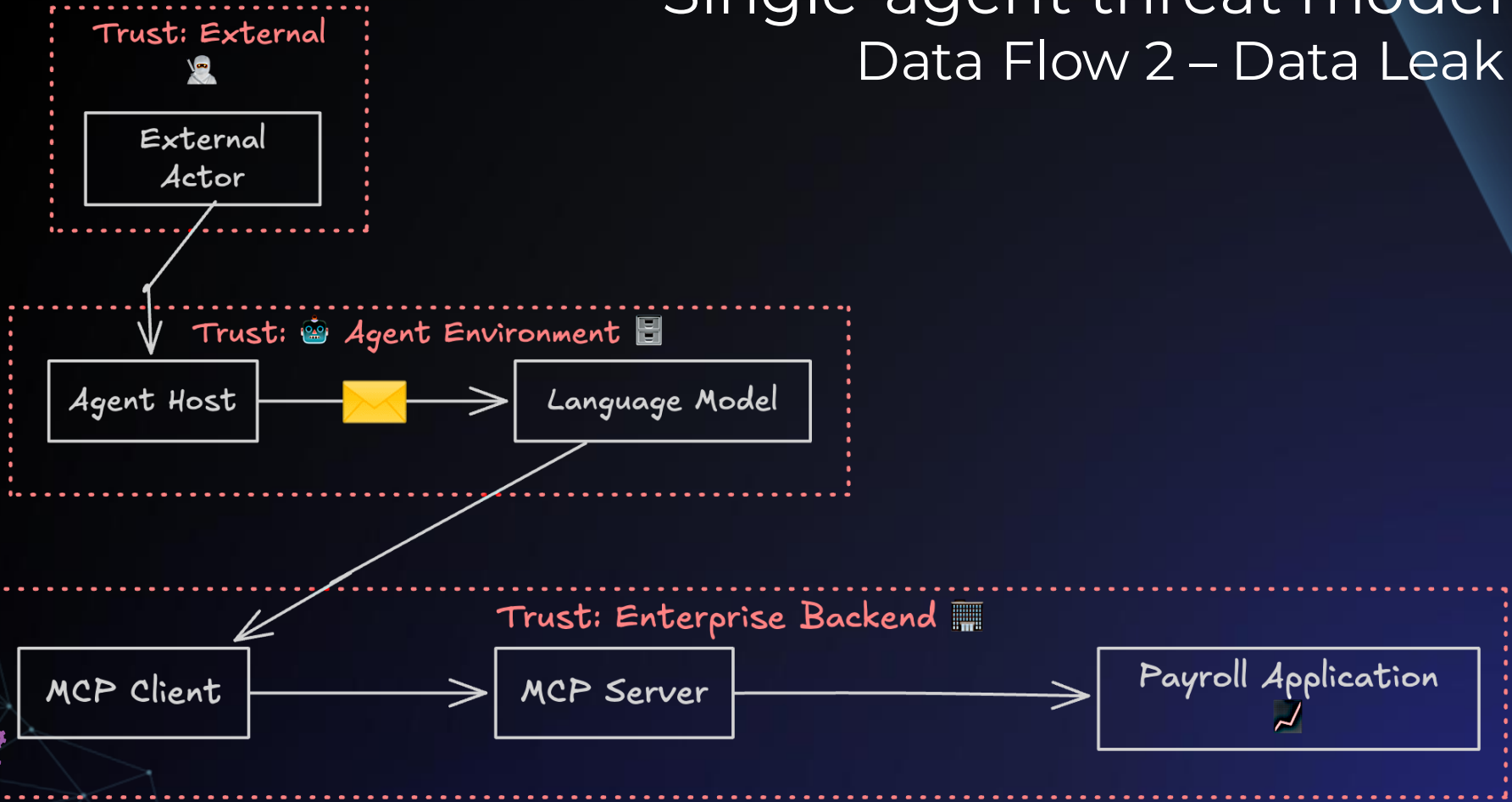


Data Leak

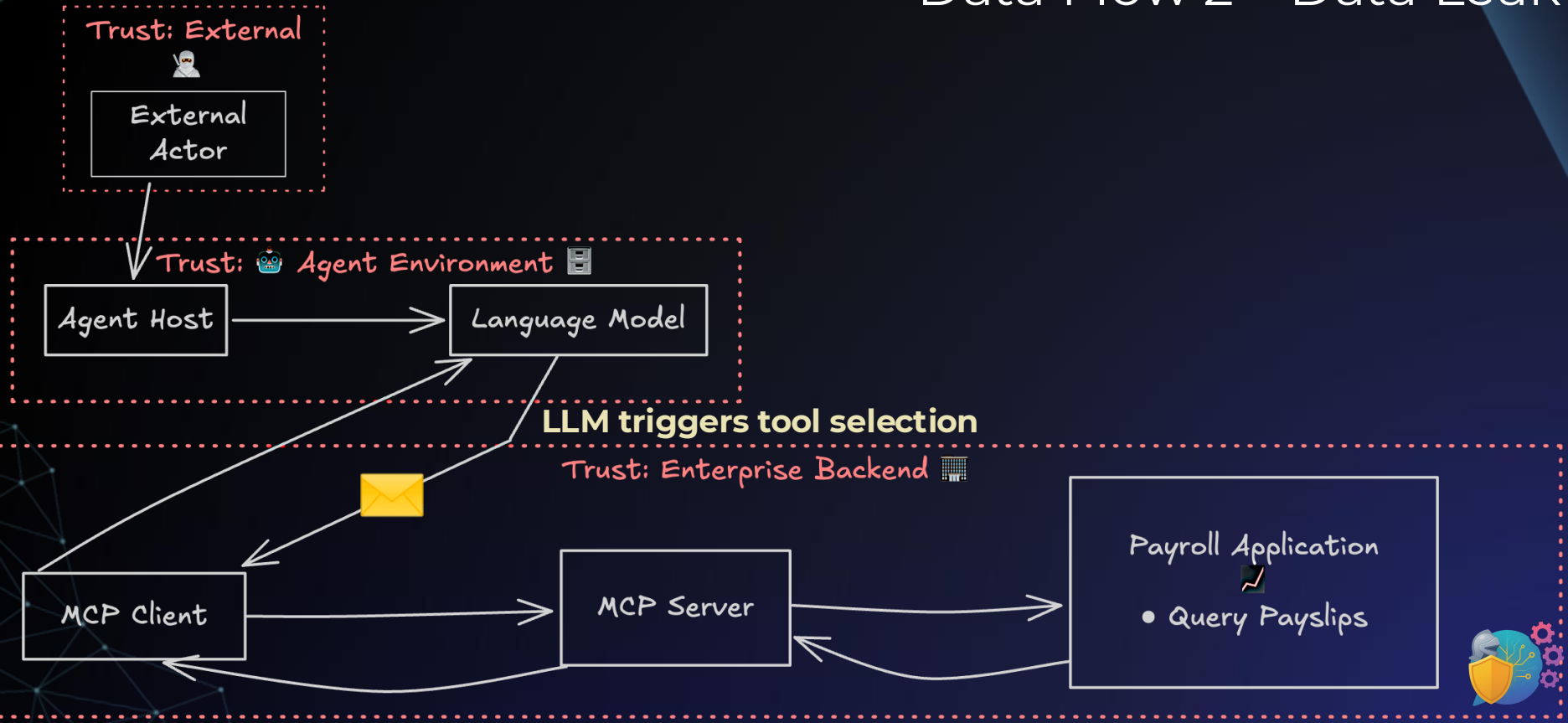
Single-agent threat model
Data Flow 2 – Data Leak



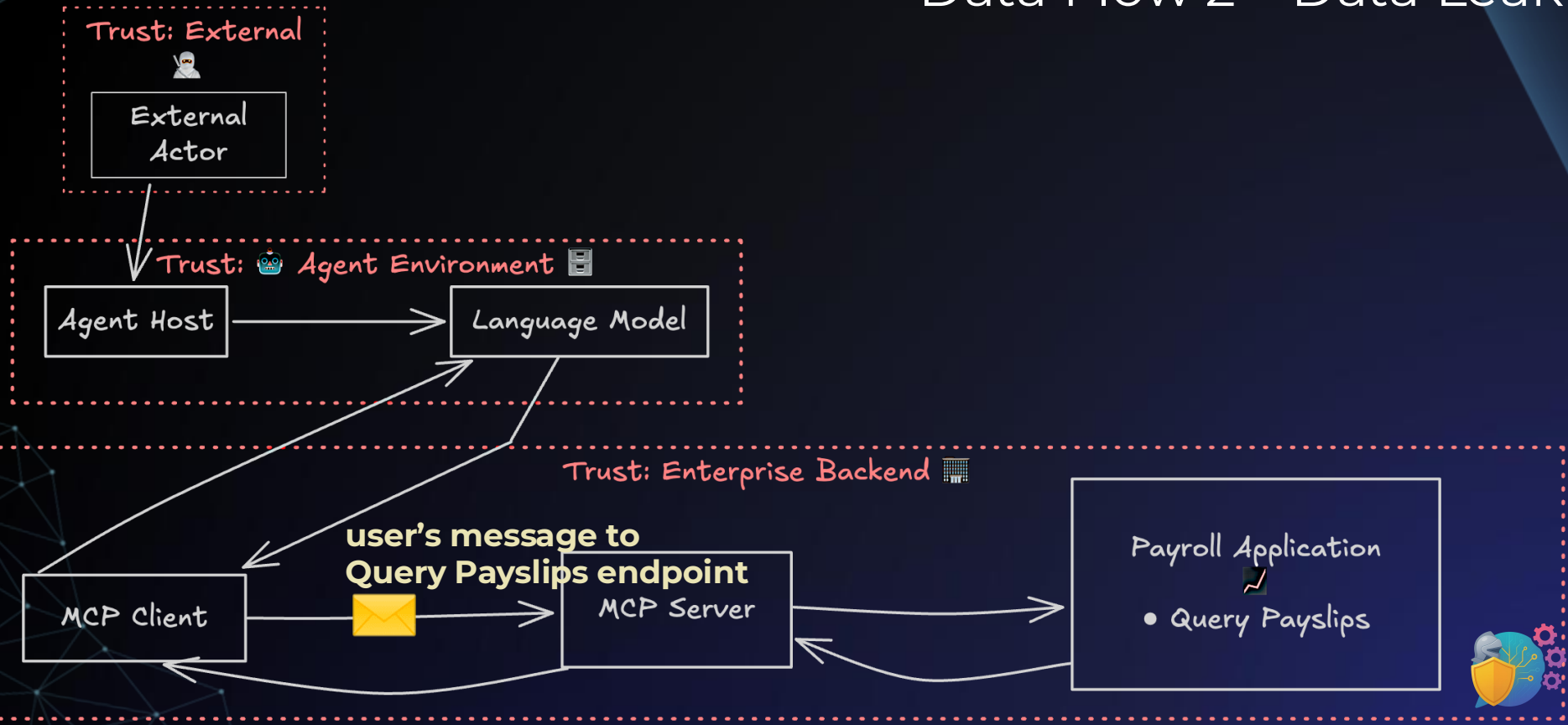
Single-agent threat model
Data Flow 2 – Data Leak



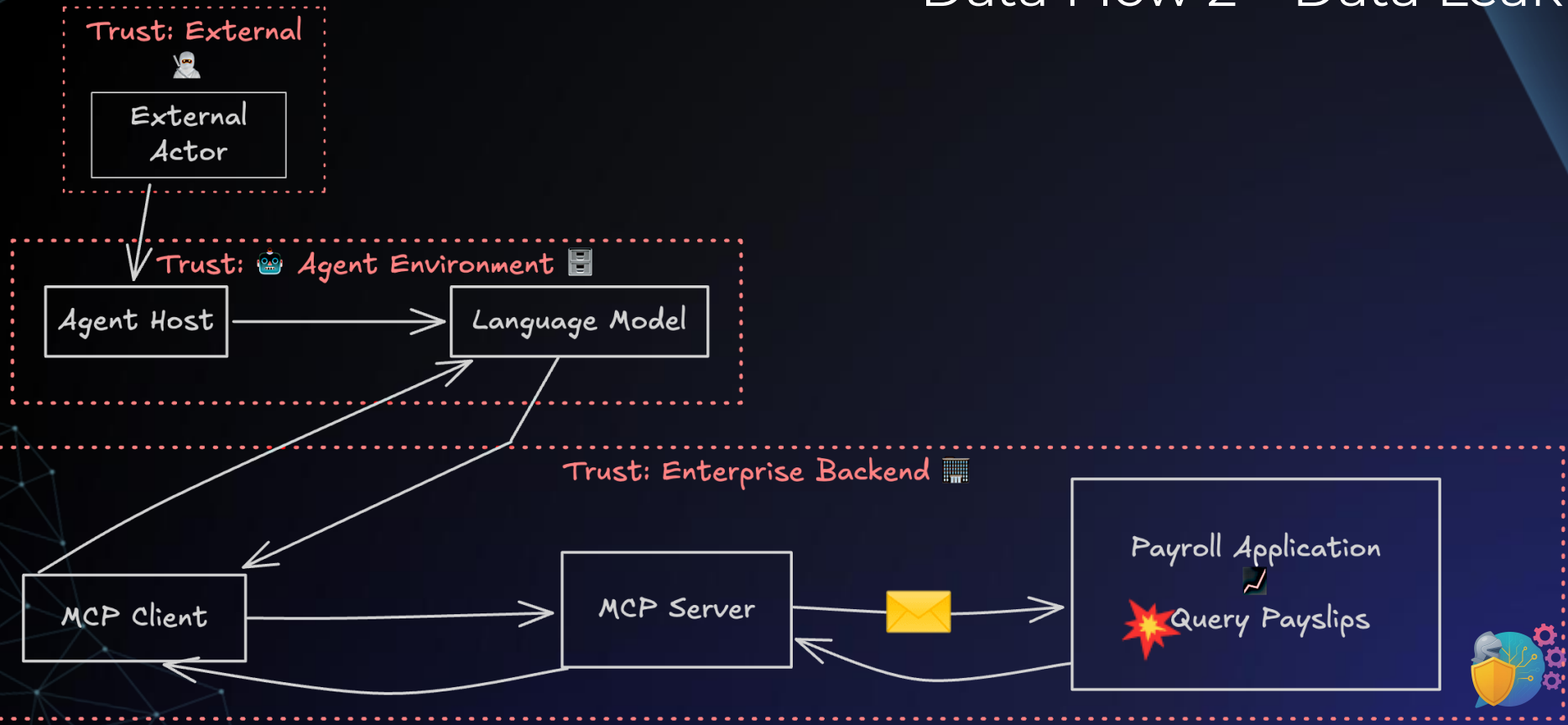
Single-agent threat model
Data Flow 2 – Data Leak



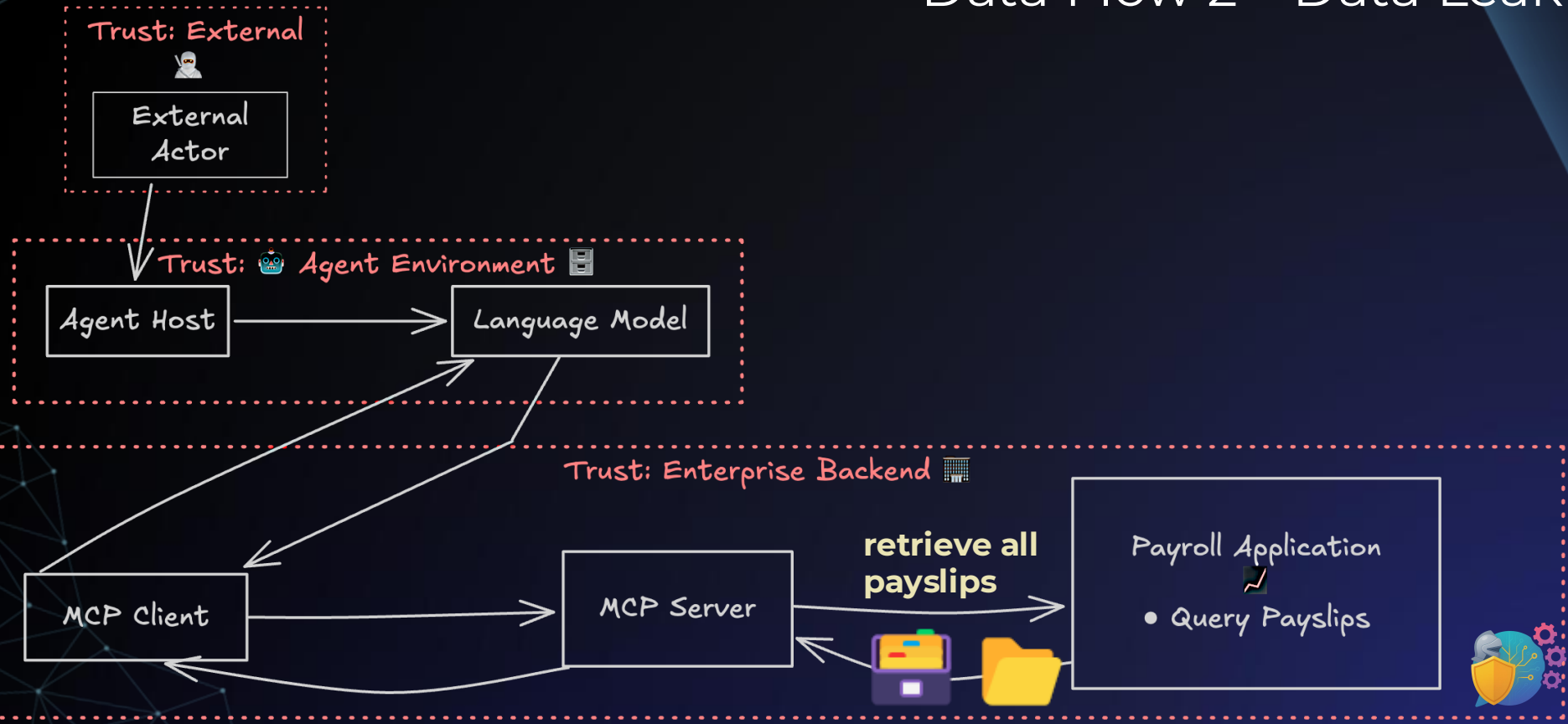
Single-agent threat model
Data Flow 2 – Data Leak



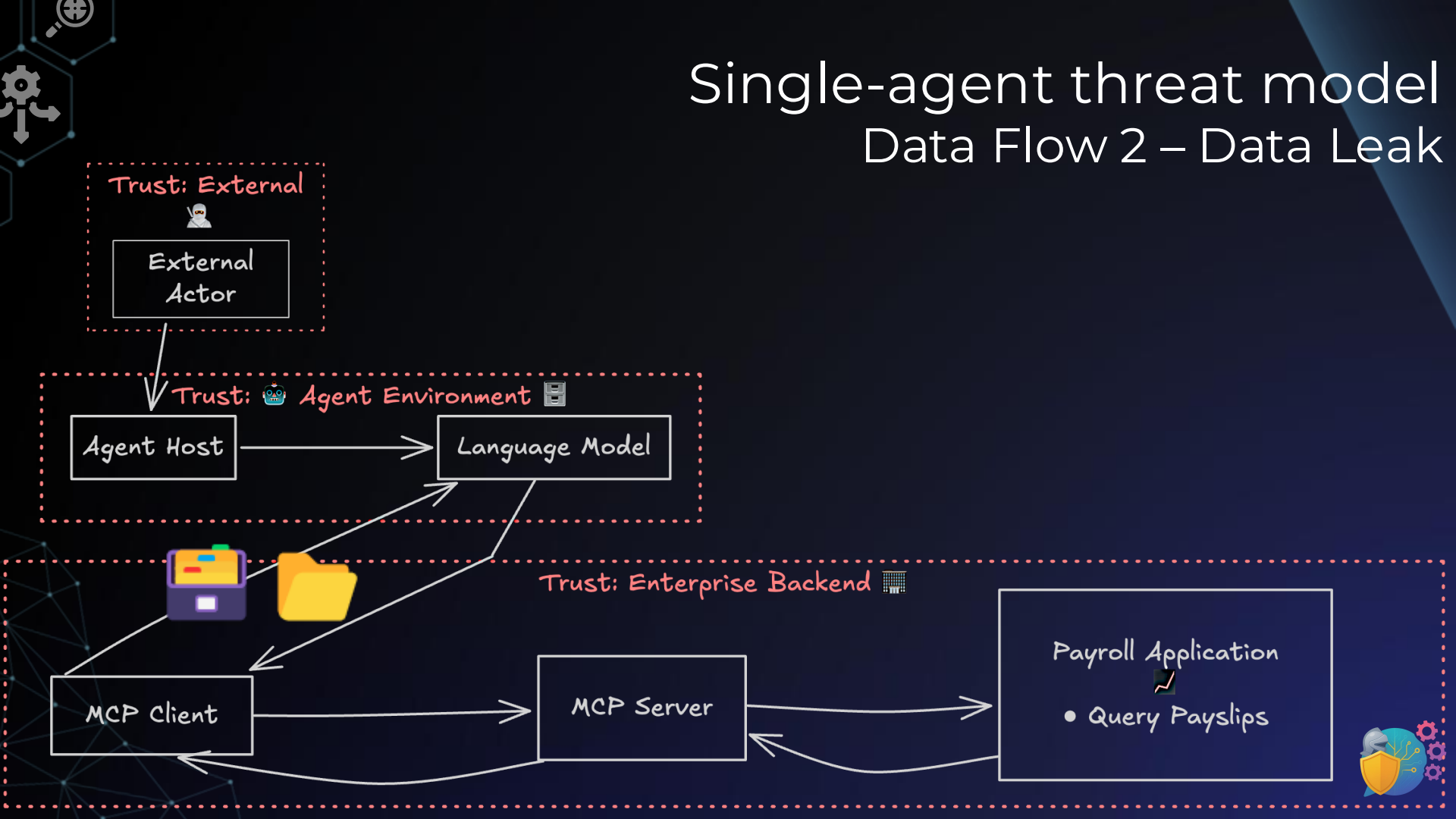
Single-agent threat model
Data Flow 2 – Data Leak



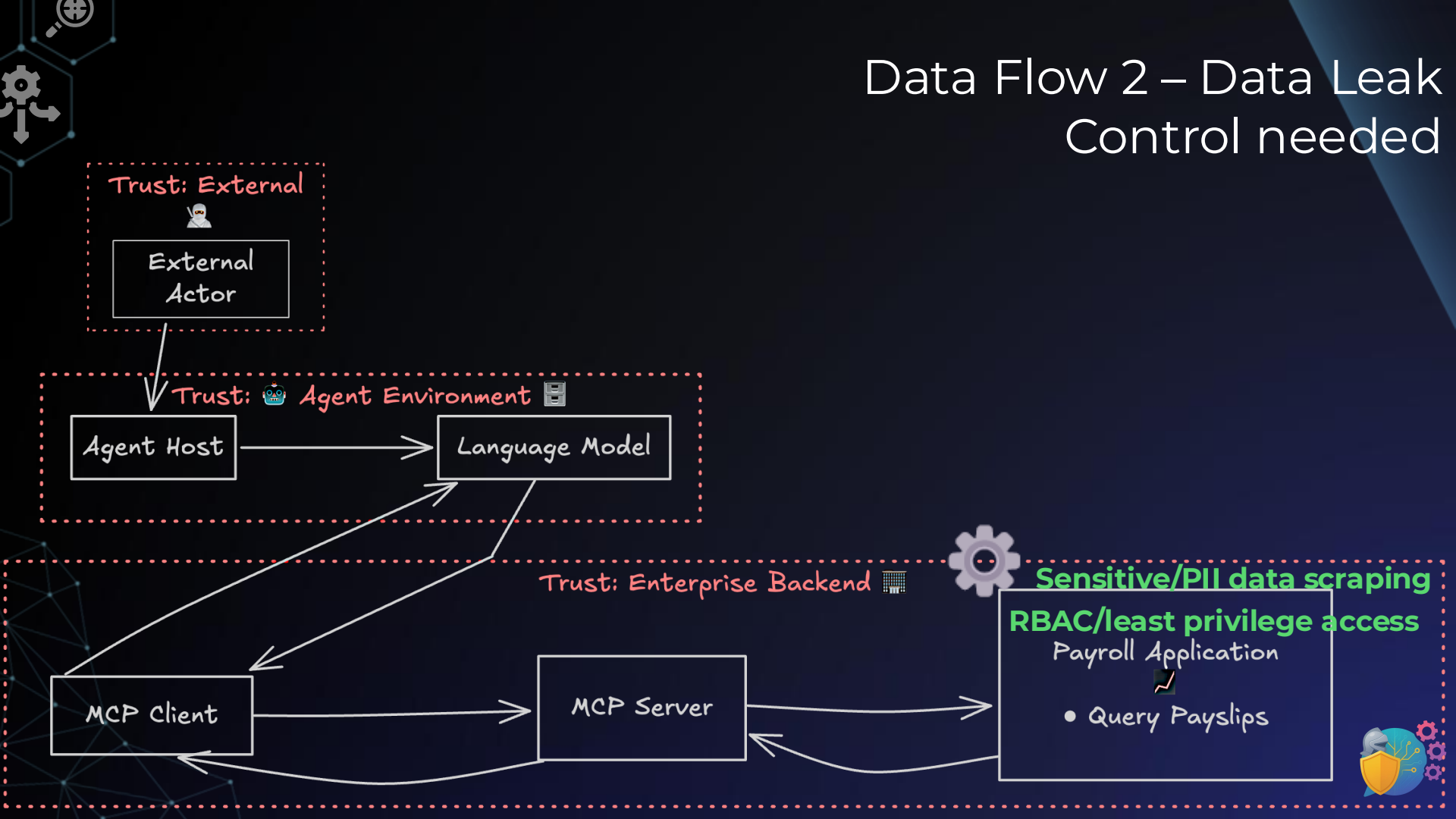
Single-agent threat model
Data Flow 2 – Data Leak



Single-agent threat model
Data Flow 2 – Data Leak



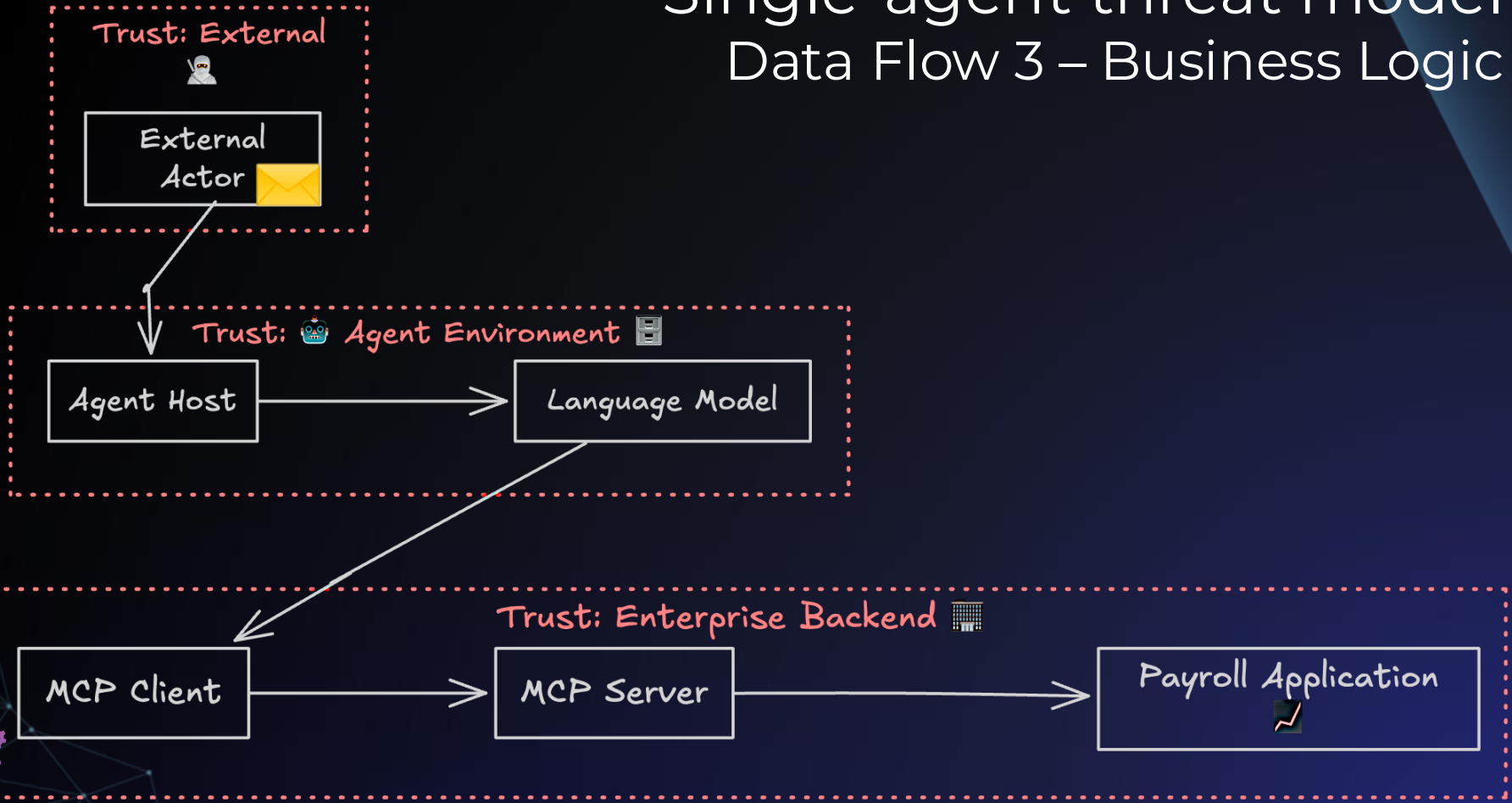
Data Flow 2 – Data Leak Control needed



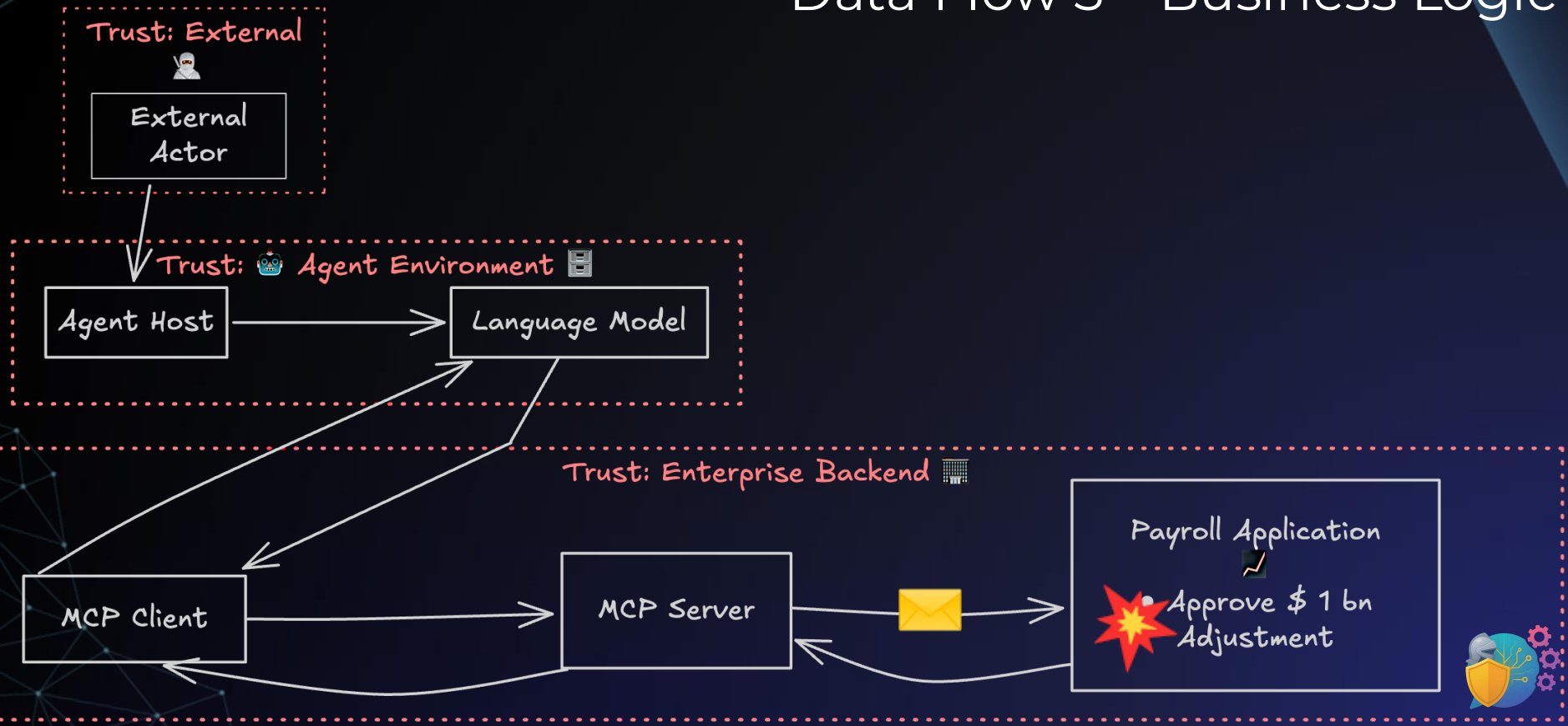
Business Logic



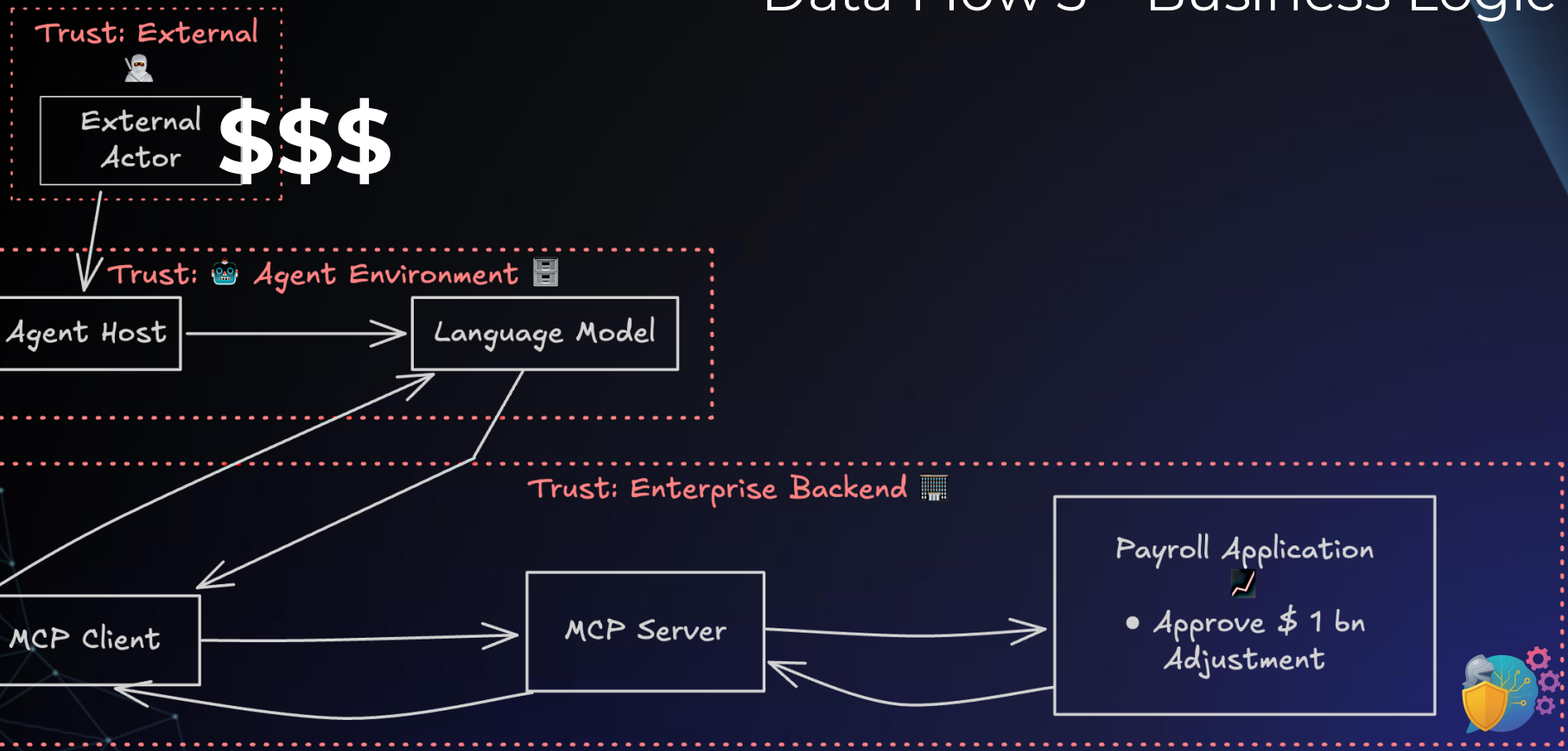
Single-agent threat model
Data Flow 3 – Business Logic



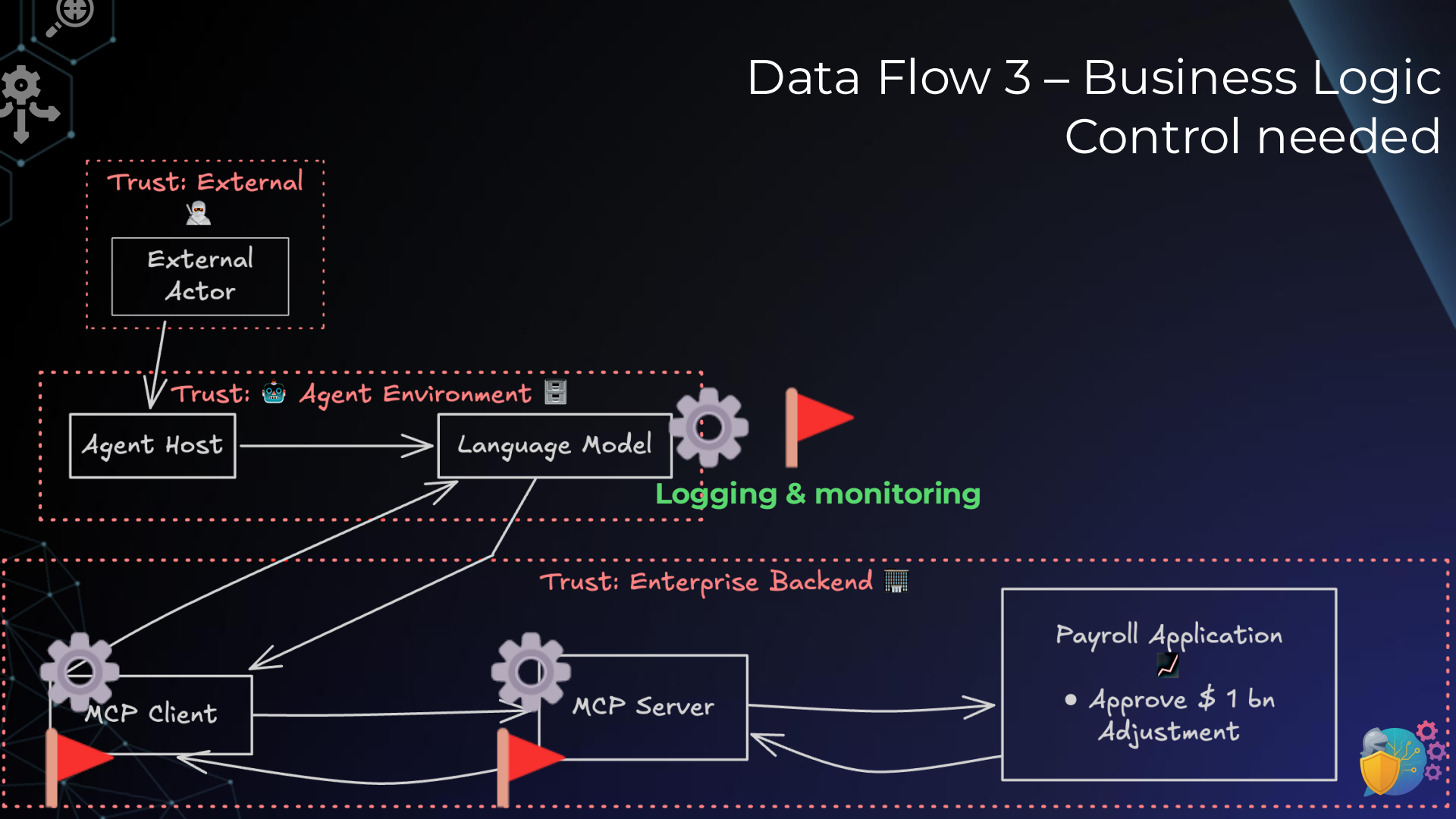
Single-agent threat model
Data Flow 3 – Business Logic



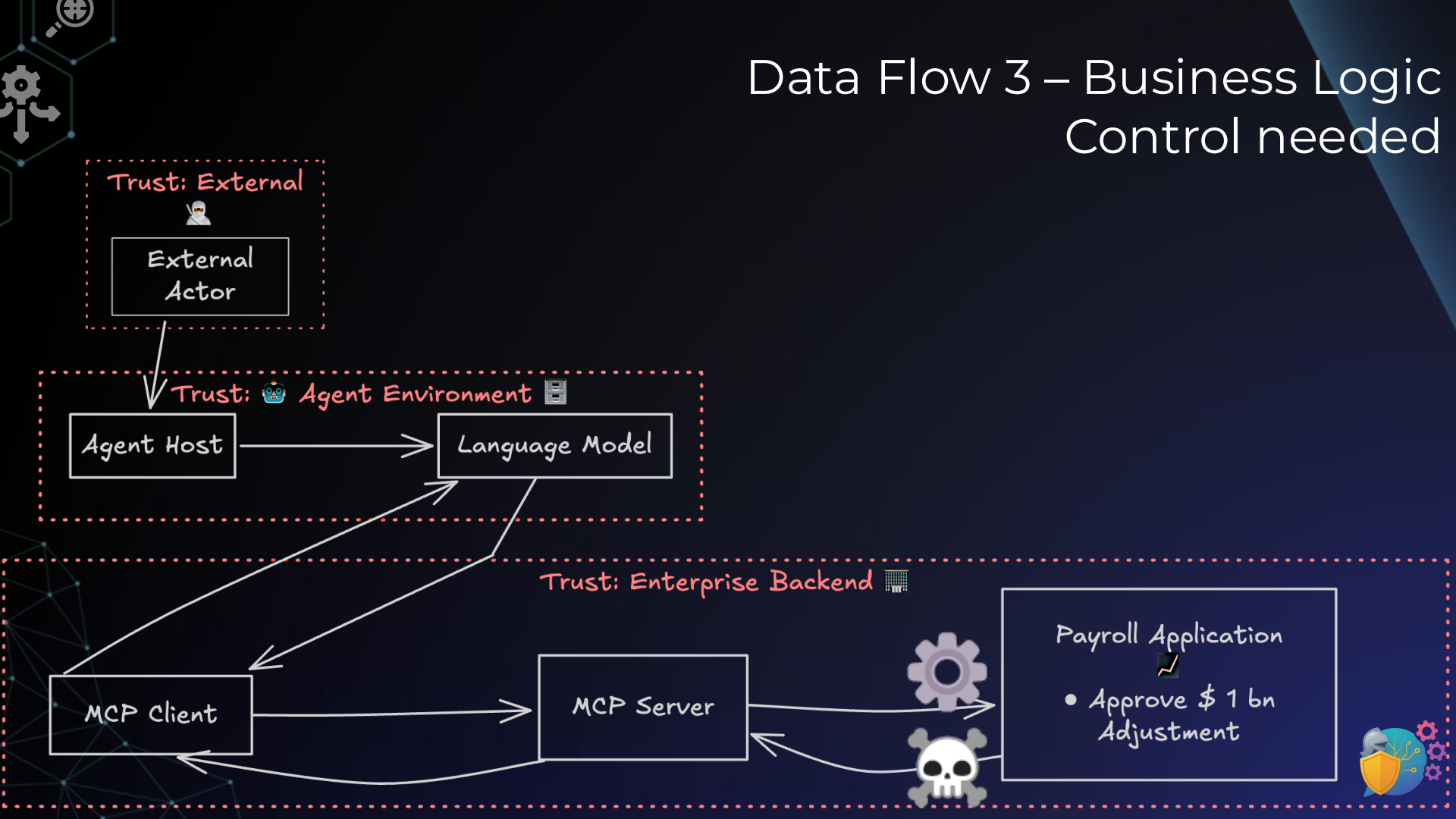
Single-agent threat model
Data Flow 3 – Business Logic



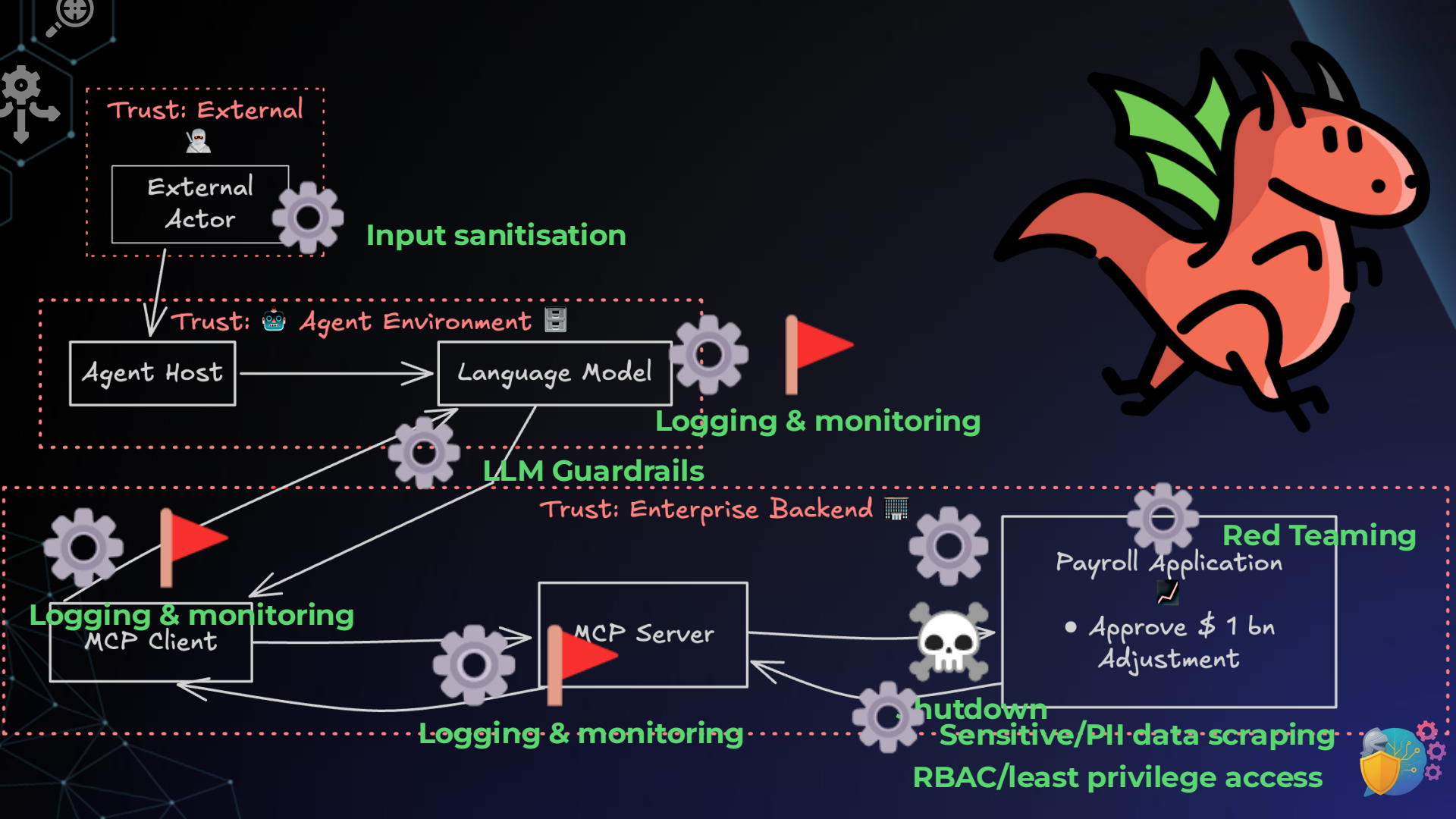
Data Flow 3 – Business Logic Control needed



Data Flow 3 – Business Logic Control needed



Controls

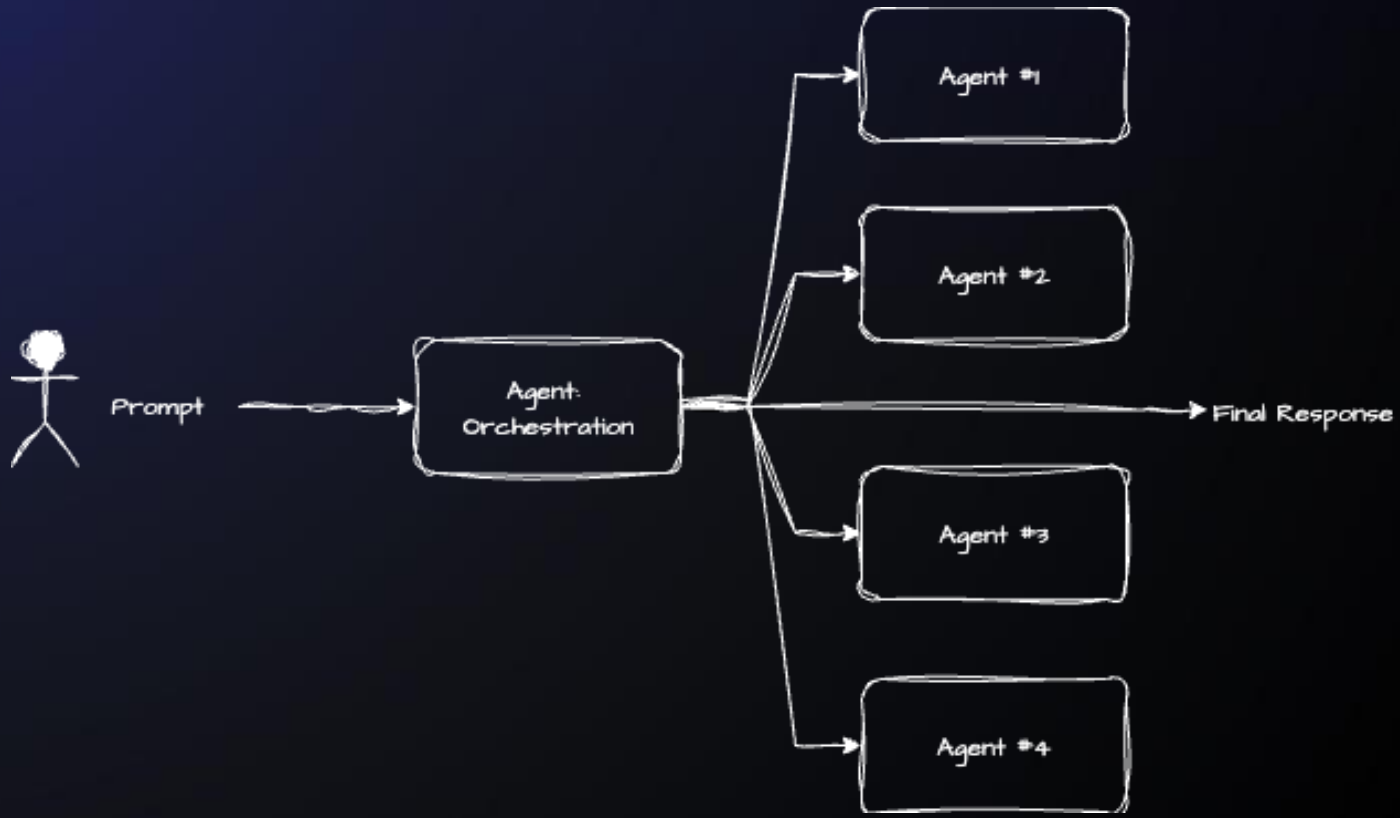

Trust: External
External Actor
Input sanitisation
Trust: Agent Environment
Agent Host
Language Model
Logging & monitoring
LLM Guardrails
Trust: Enterprise Backend
Red Teaming
Payroll Application
• Approve $ 1 bn Adjustment
Logging & monitoring
MCP Client
MCP Server
hutdown
Logging & monitoring
Sensitive/PII data scraping
RBAC/least privilege access

💣 Threat modelling 🖼️

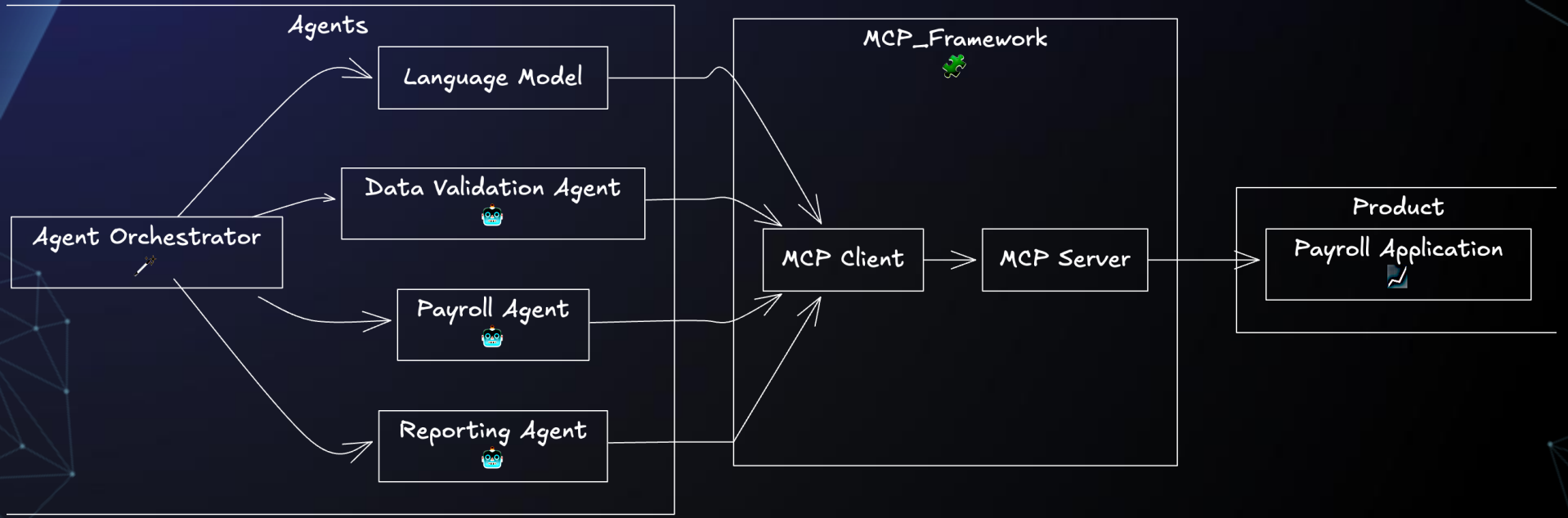
🤖 🤖 🤖...



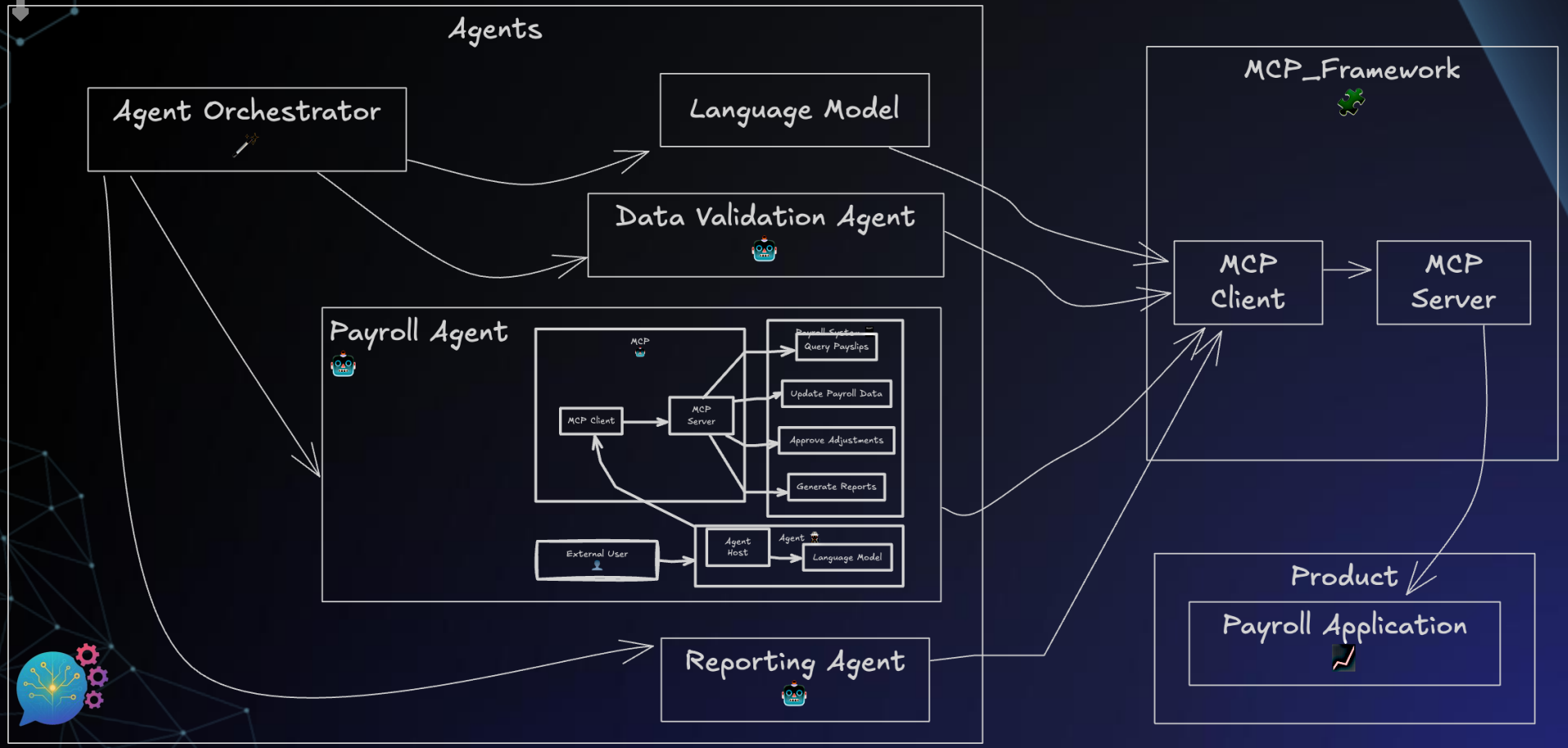
Agents, multi-agents, and tools



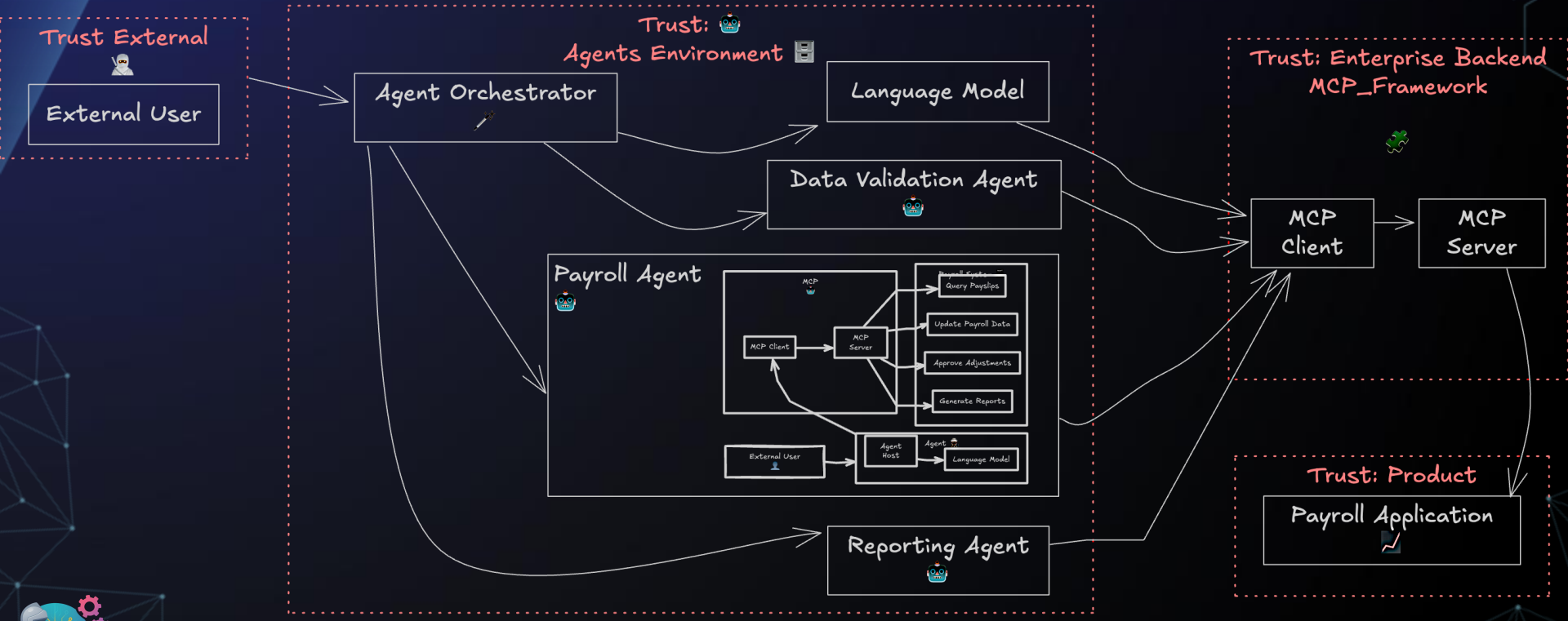
Multi-agent architecture example

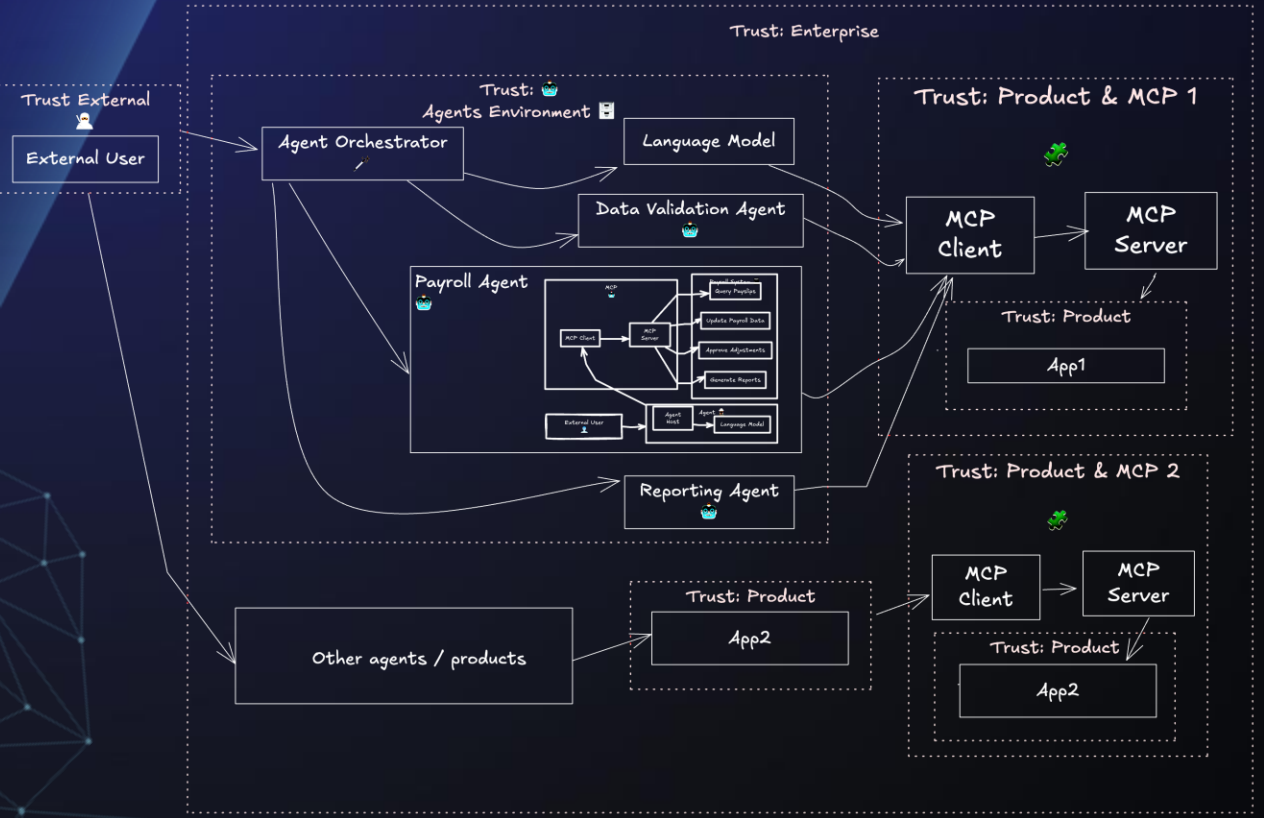


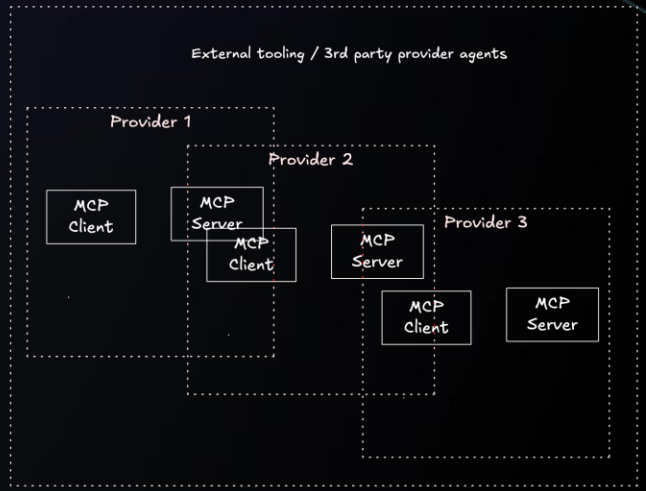
Zoom Out: Multi-agent architecture

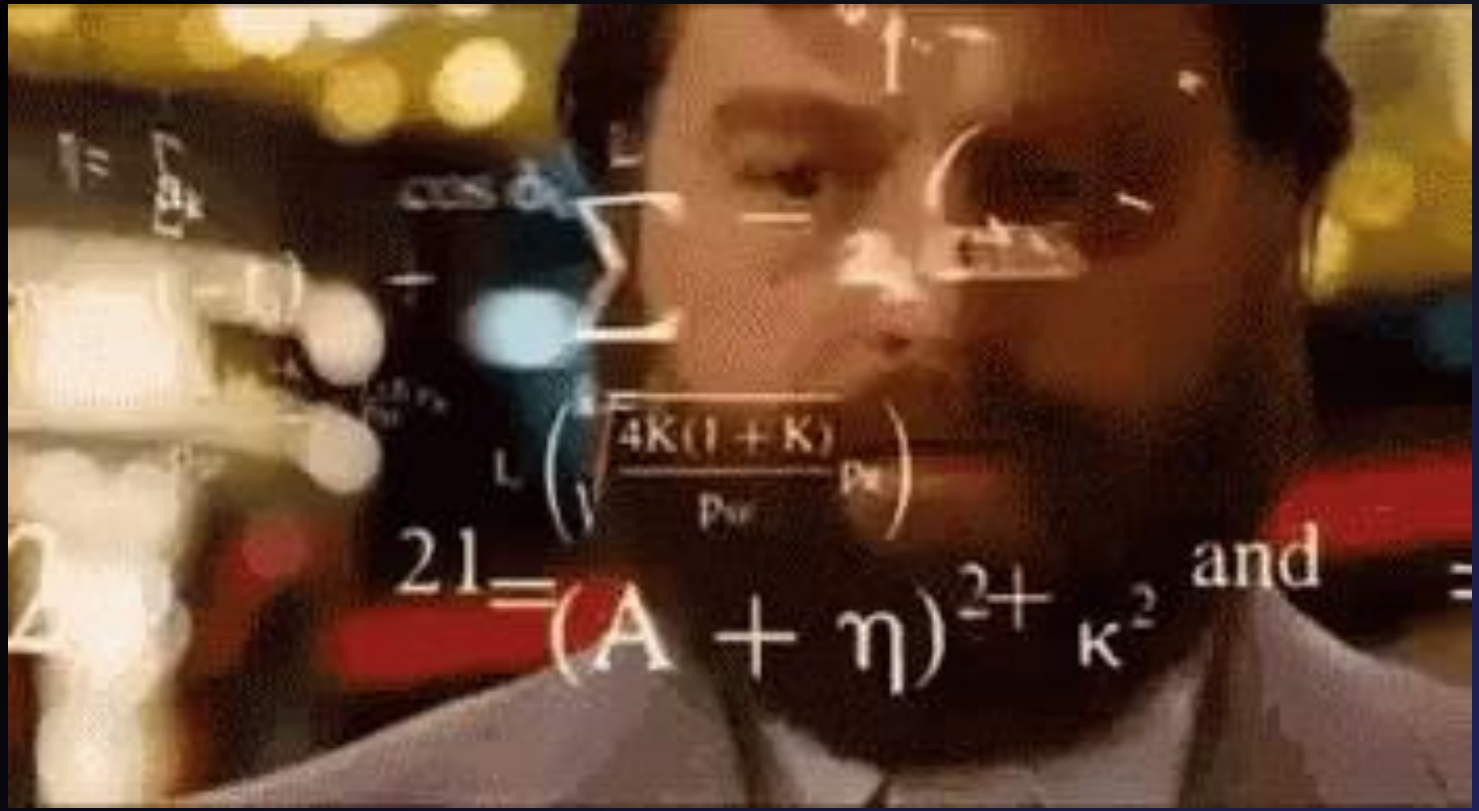


Multi-agent threat model
Trust boundaries




Trust: Enterprise
Trust External
External User
Trust:
Agents Environment
Agent Orchestrator
Language Model
Data Validation Agent
Payroll Agent
Reporting Agent
Other agents / products
Trust: Product & MCP 1
MCP Client
MCP Server
Trust: Product
App1
Trust: Product & MCP 2
MCP Client
MCP Server
Trust: Product
App2
Trust: Product
App2


External tooling / 3rd party provider agents
Provider 1
MCP Client
MCP Server
Provider 2
MCP Client
MCP Server
Provider 3
MCP Client
MCP Server


21 = (A + η)2+ κ2 and

🔐 (WIP) Security patterns & best practices




JURASSIC PARK

Recommendations and random advice

- Do you really need a chat interface? Plan accordingly
- BYO client
  - Limit MCP server access as a first line of defense (and possibly client)
- Allow yourself to separate user-facing "client" from MCP client (YMMV)
- You thought you were implementing an MCP server, but really, you're auditing your access control model
  - Bake this into your MCP server or possibly add a proxy between MCP server and downstream APIs (where appropriate)
- Segregate MCP servers by levels of trust (of tool providers)





Key Takeaways



Security basics still apply

Move carefully & design safely

Threat model everything & continue testing



Free resources

OWASP AI Exchange	**owaspai.org**
OWASP GenAI Security Project	**genai.owasp.org**
Hugging face (free) courses	**huggingface.co/learn**
Sage AI blog	**ai.sage.com**



Thank you!

Questions?



Andra Lezza linkedin.com/in/andralezza/

Jeremiah Edwards linkedin.com/in/jeremiq/

