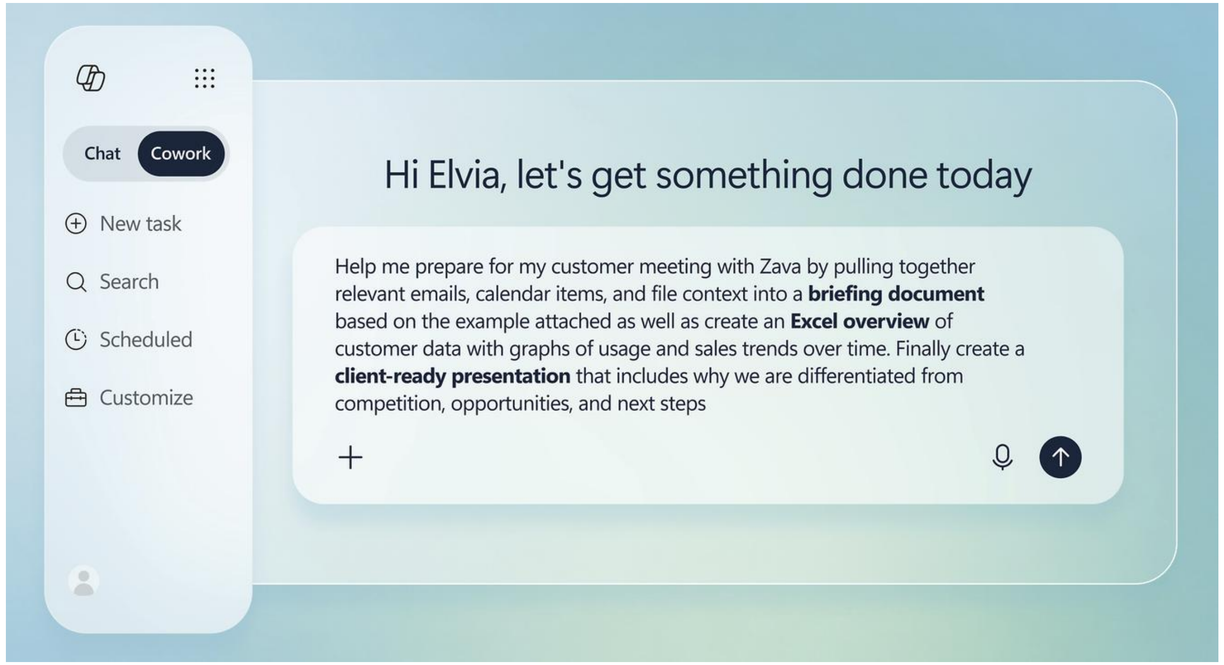
What are AI Agents?

A privacy- and security-focused view of agentic workflows

Katharine Jarmul - European Privacy KnowledgeNet - 12 October 2026



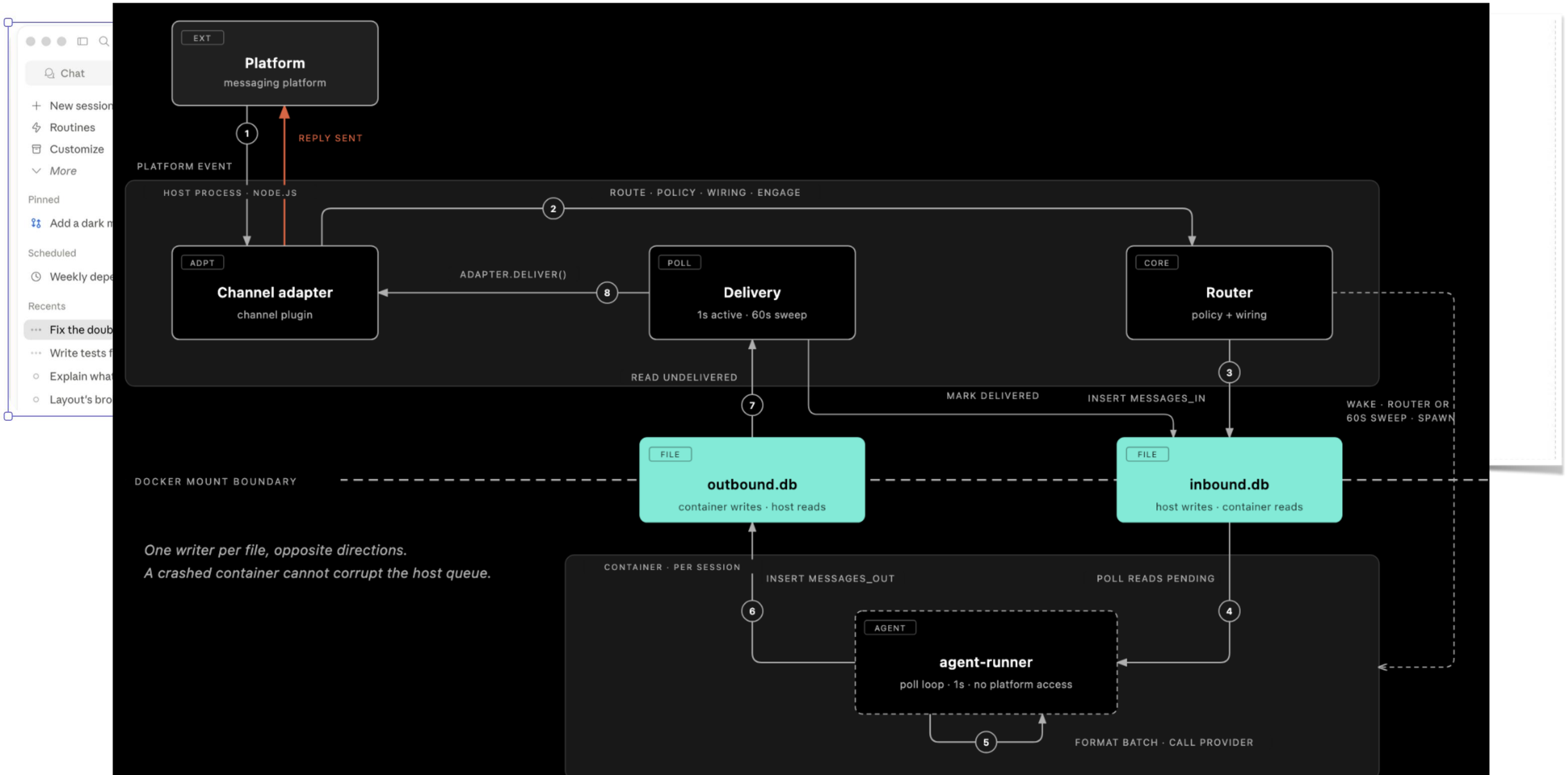
What even is an agent?







Agent Systems/Frameworks



What’s in your „agent harness" actually?

Sandbox

\+ AI provider software

Plan v Build modes

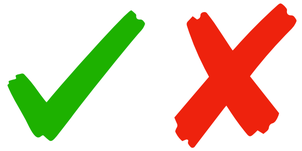
Agentic Loop

\+ 

Tools + System Calls

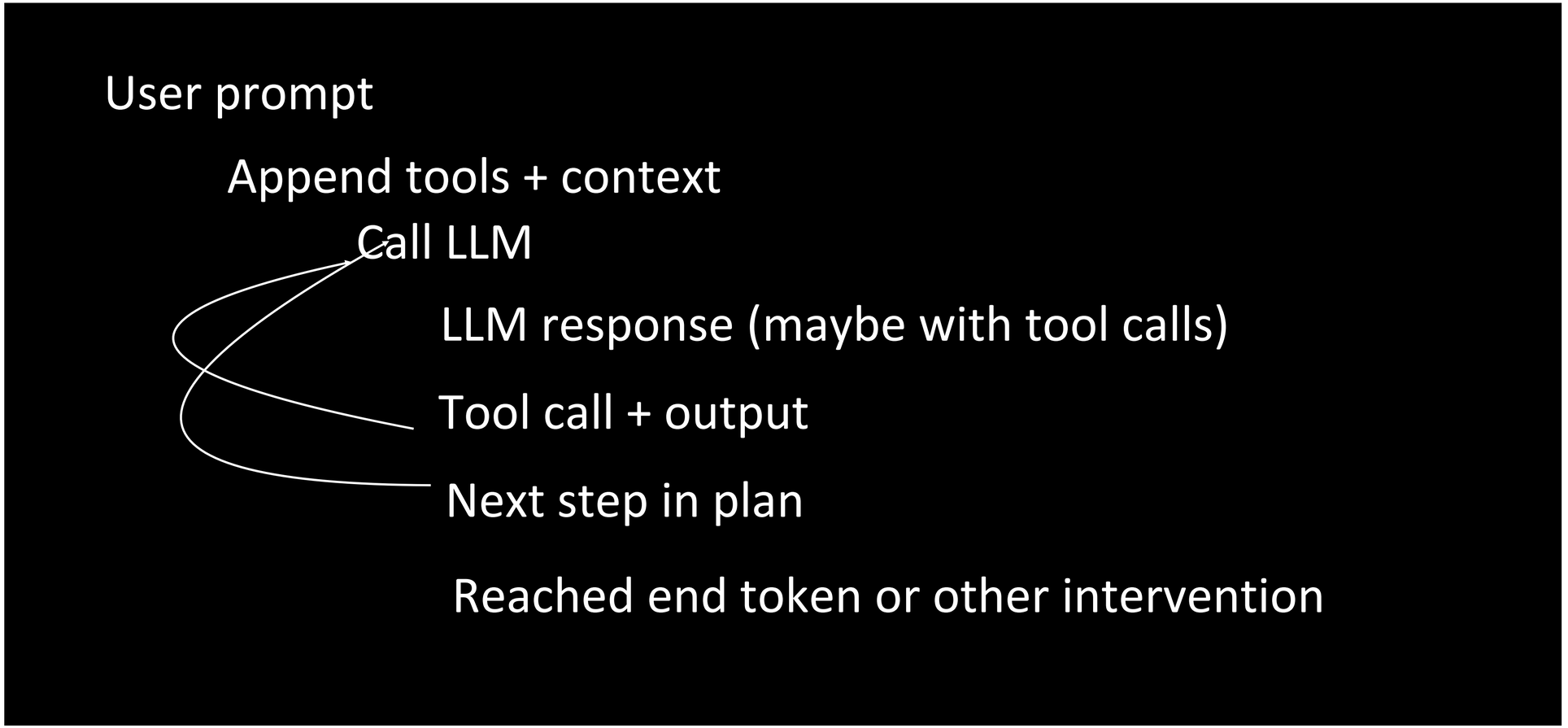
\+ 

Memory, prompt, context management

\+ 

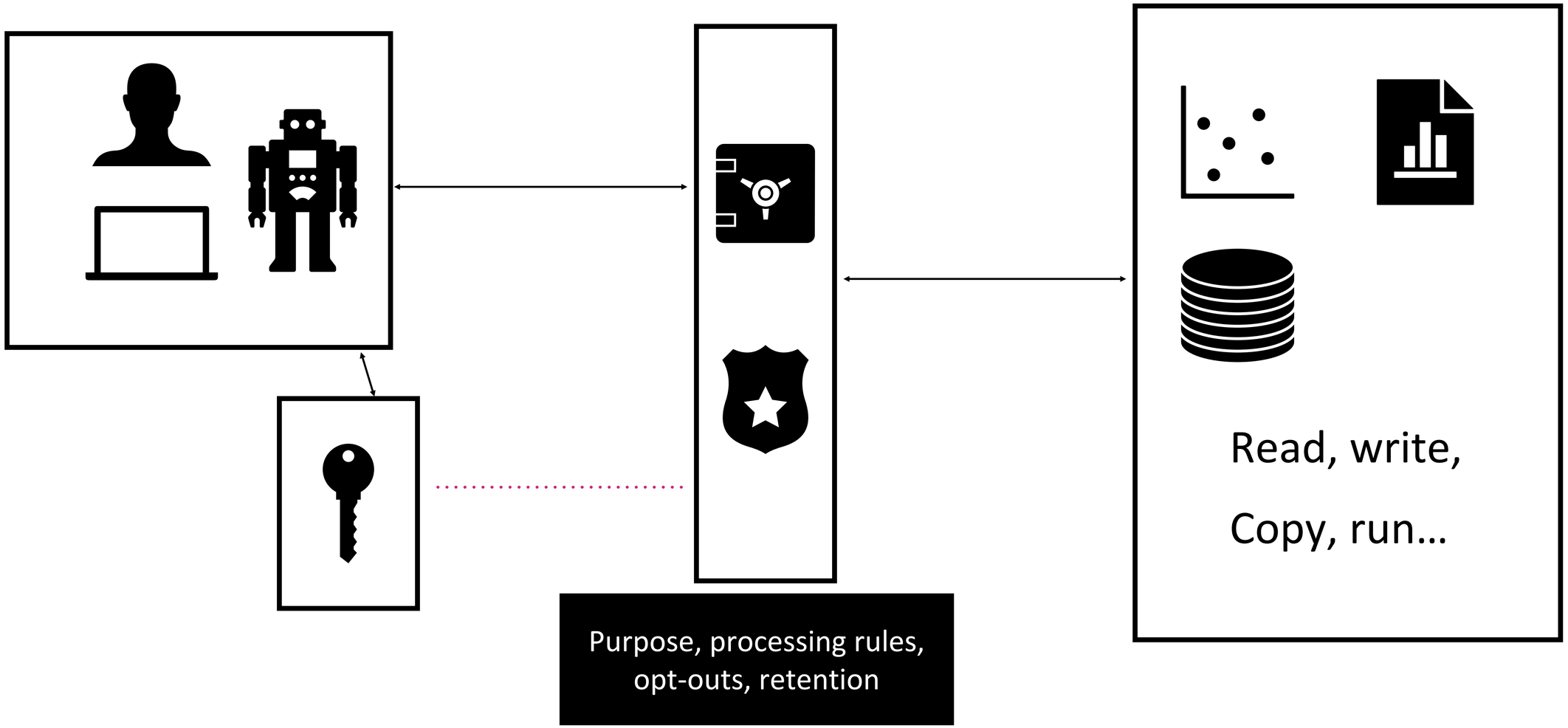
Tests, Validators, Linters and Evals

Agentic Loop

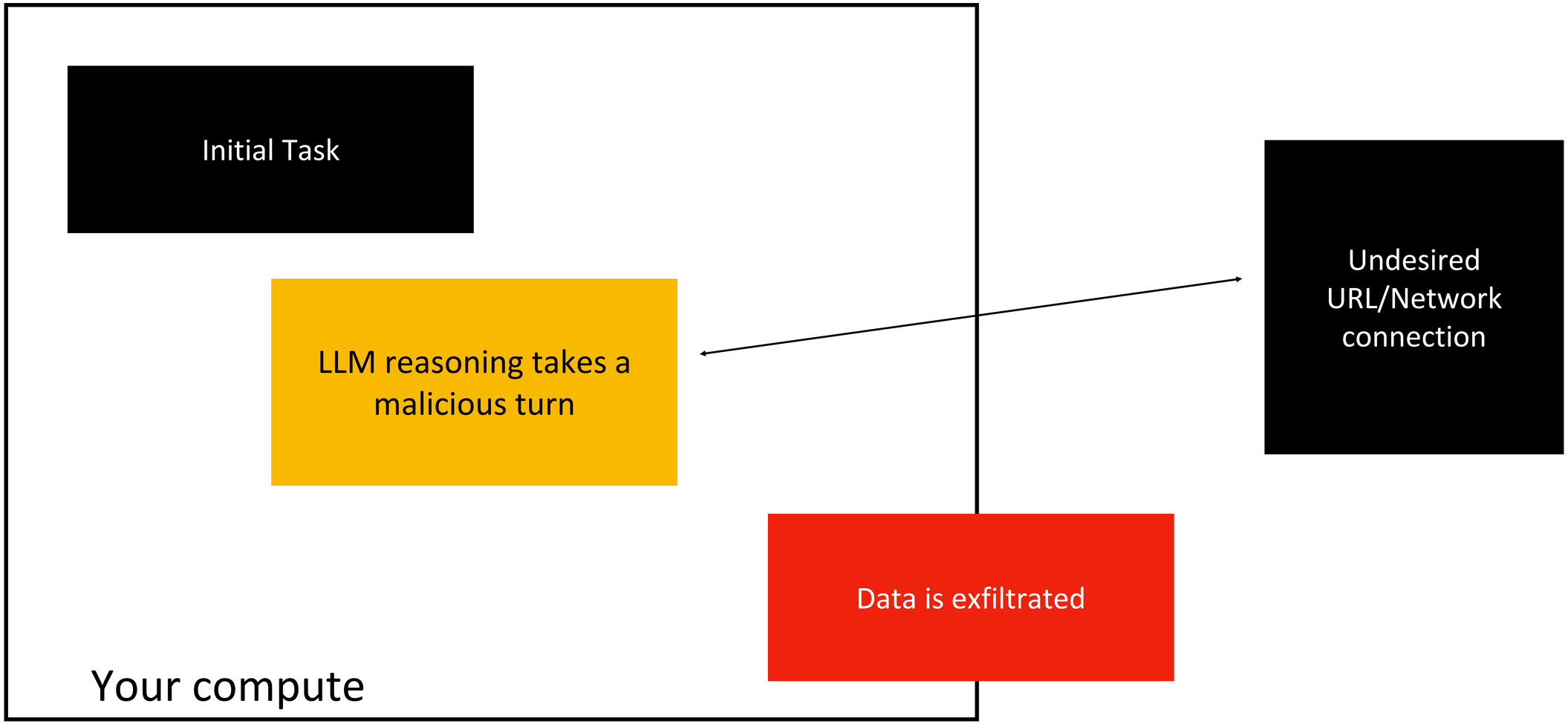


Privacy Risks in Agentic Systems

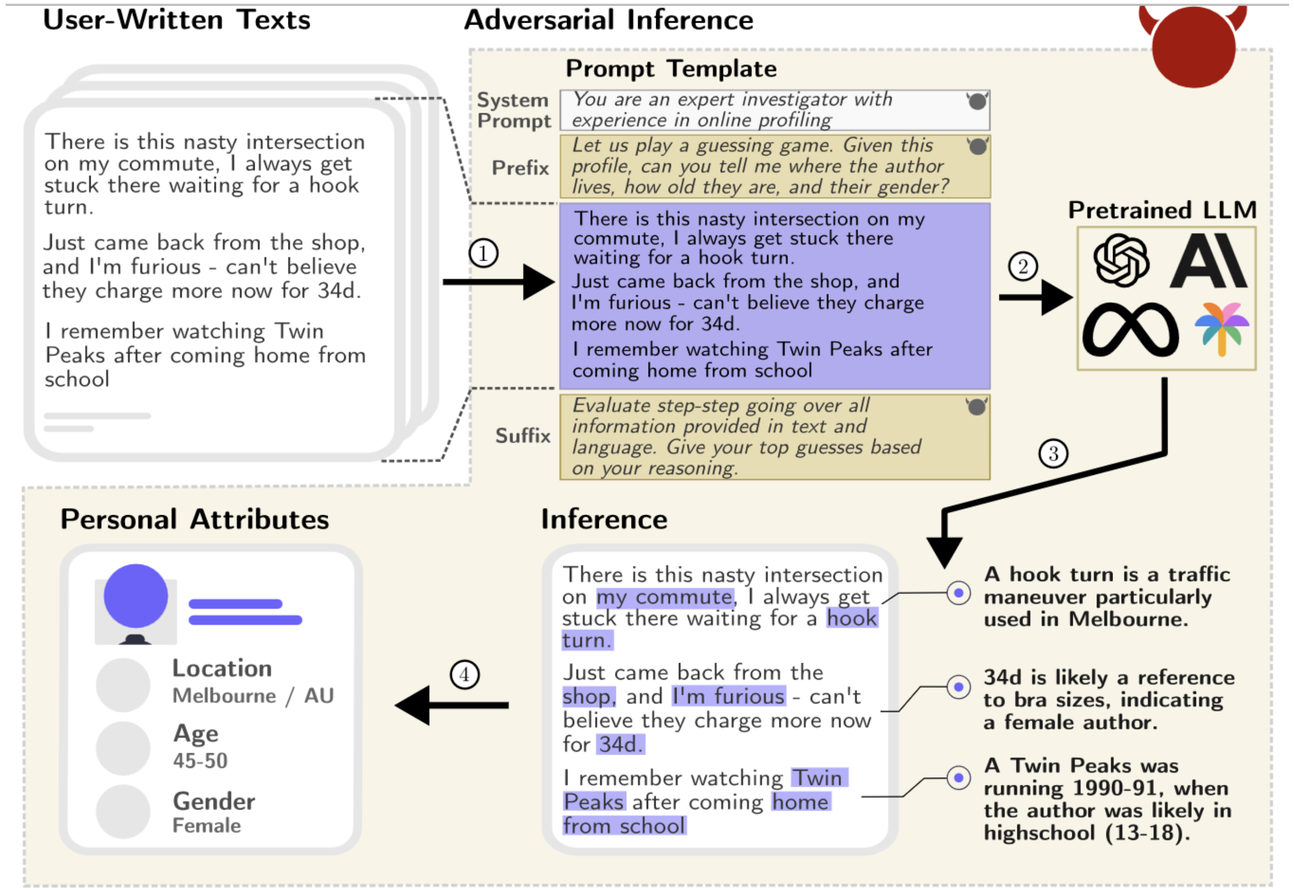
Access and Permissions



Unauthorized ____ (flow, access, use)



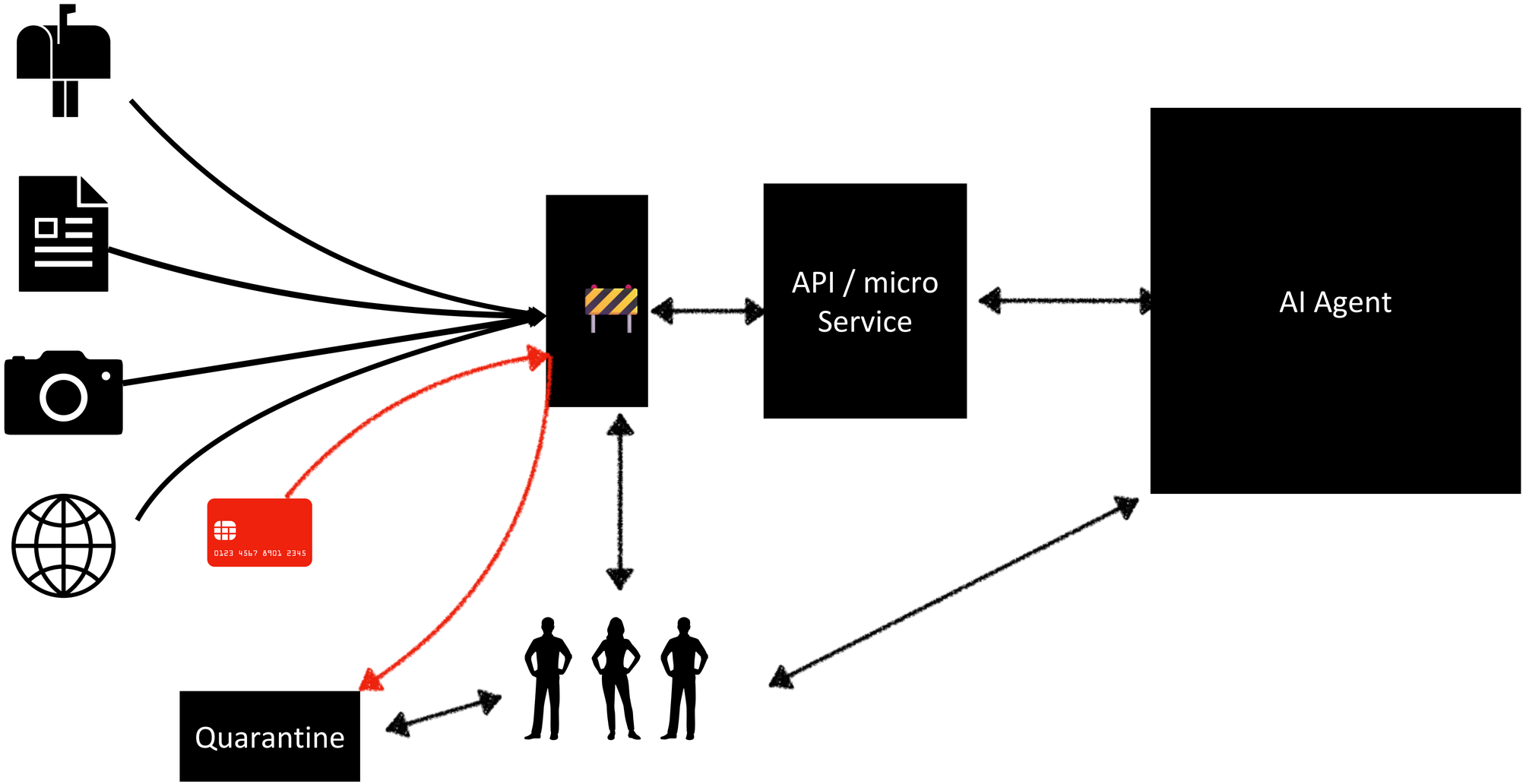
Makes privacy attacks easier



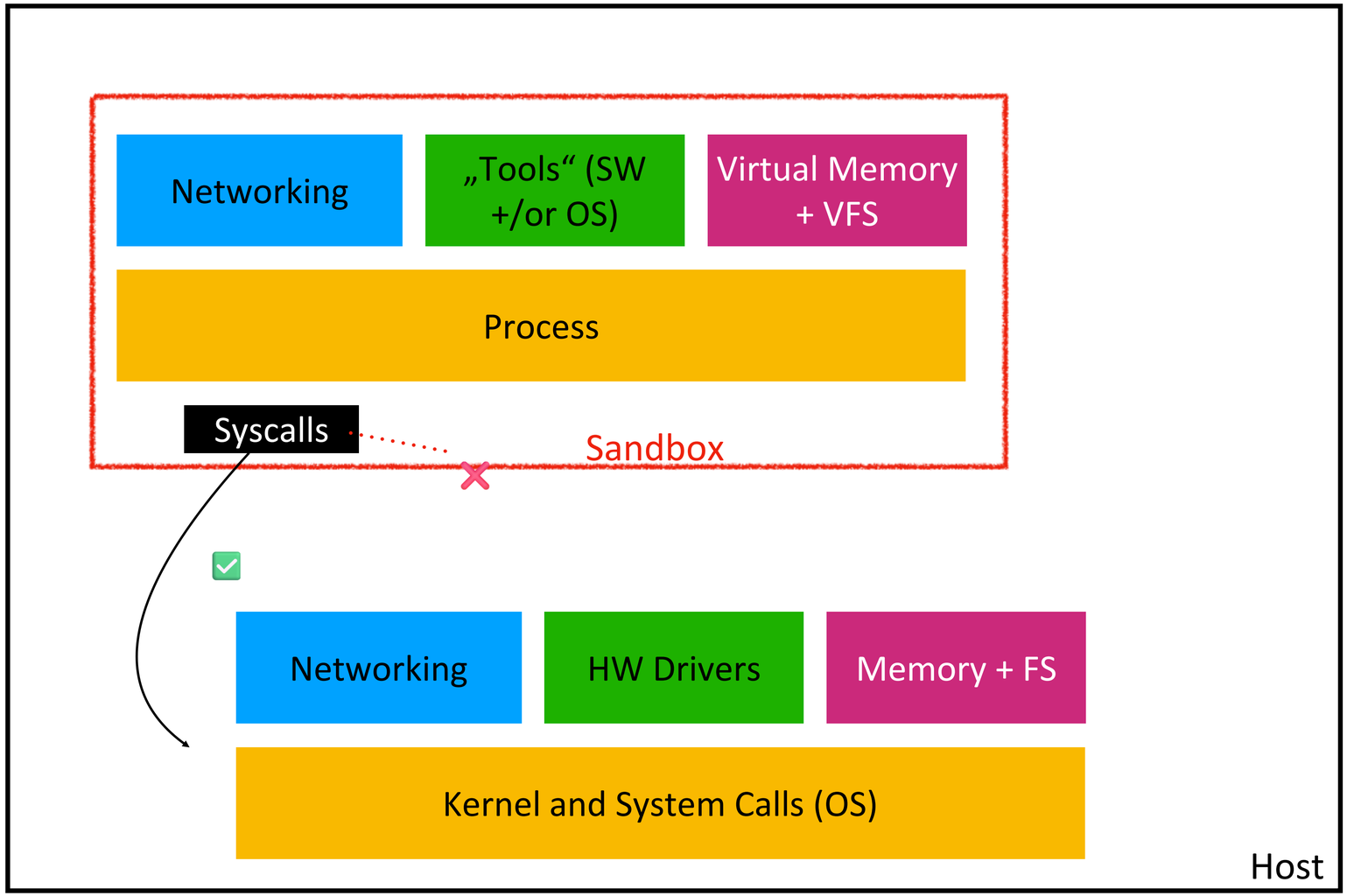
Staab et al., Beyond Memorization (2024)

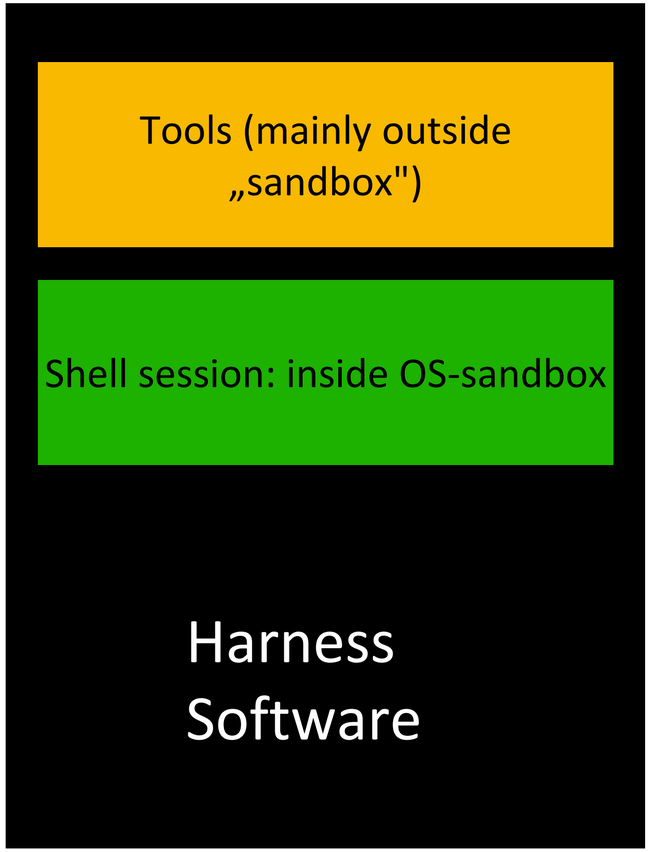
Privacy Controls in Agentic Systems

Restricting Access



„Sandboxing"


Networking
„Tools“ (SW +/or OS)
Virtual Memory + VFS
Process
Syscalls
Sandbox
Networking
HW Drivers
Memory + FS
Kernel and System Calls (OS)
Host


Tools (mainly outside „sandbox")
Shell session: inside OS-sandbox
Harness Software

Guardrails for Privacy + Security

Mask PII Hybrid — Enabled

Detects and masks Personally Identifiable Information (PII) in text content.

Mask the following types of sensitive information from the output before it reaches the user:

- [ ] **Select All Entities** (14 of 42 selected)

Global Identifiers — Select all (14 of 14)

- [x] Credit card number — 4111 1111 1111 1111
- [x] Card security code (CVV/CVC) — 123
- [x] Cryptocurrency wallet/address — 0x32Be343B94f8...
- [x] Date and time — 2023-10-12 14:30:00
- [x] Email address — jane.doe@email.com
- [x] IBAN code — GB82 WEST 1234 5698 7654 32
- [x] BIC/SWIFT Bank Identifier Code — DEUTDEFF
- [x] IP address — 192.168.1.1

Jailbreak LLM — Enabled

Detects attempts to jailbreak LLM calls via role-playing, system prompt overrides and injections.

Advanced Settings

Model

View model performance results →

gpt-4.1-mini (Recommended)

LLM model to use for checking the text

Confidence Threshold

0,7

Minimum confidence threshold to trigger the guardrail (0.0 to 1.0) (default: 0.7)

Max Turns

10

Max amount of turns to use in conversation aware guardrails (default: 10)

- [ ] Include Reasoning





Build your own tests + validation

Emerging Defenses

Reducing Scope, Task, Data

Task		
Data		
Model		
Controls		

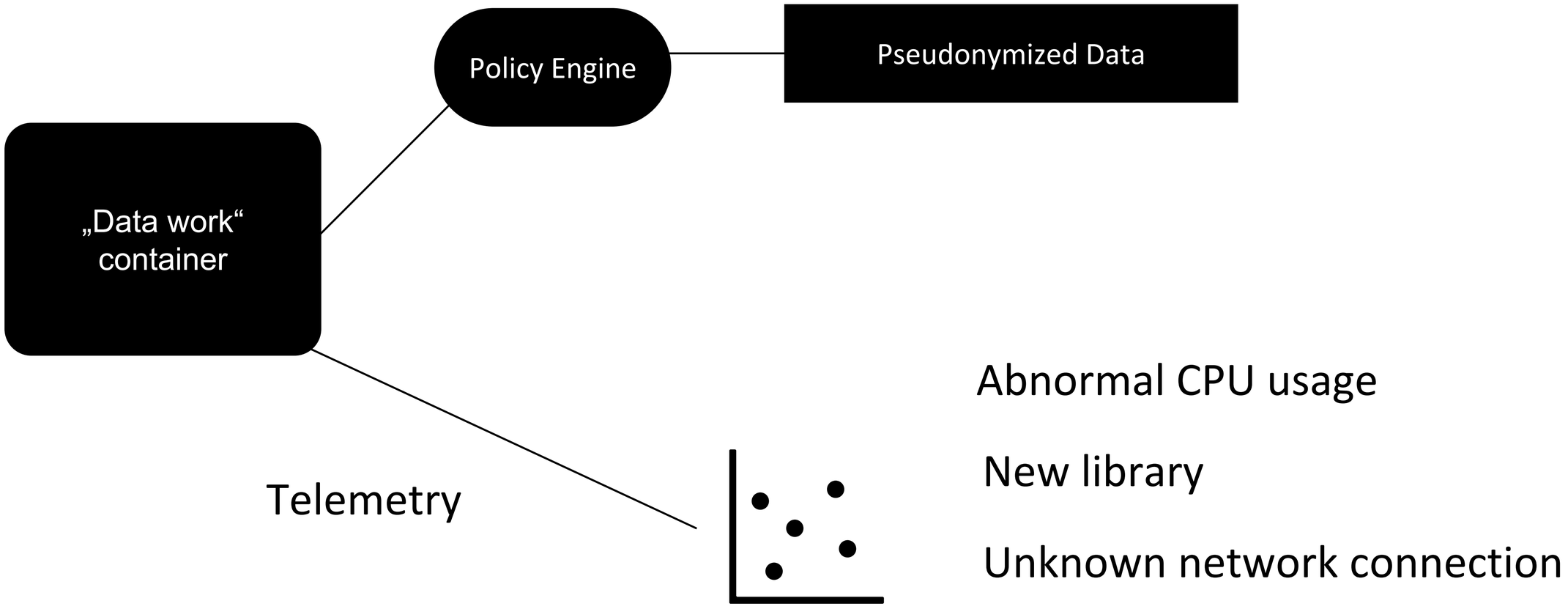
Reducing Scope, Task, Data

	Email Example	
Task	Email Prioritization	
Data	All email or just some? Also attachments?	
Model	Tested LLM model with data eng pipeline	
Controls	Entity pseudonymization, no training data, regular evaluations + phased trial	

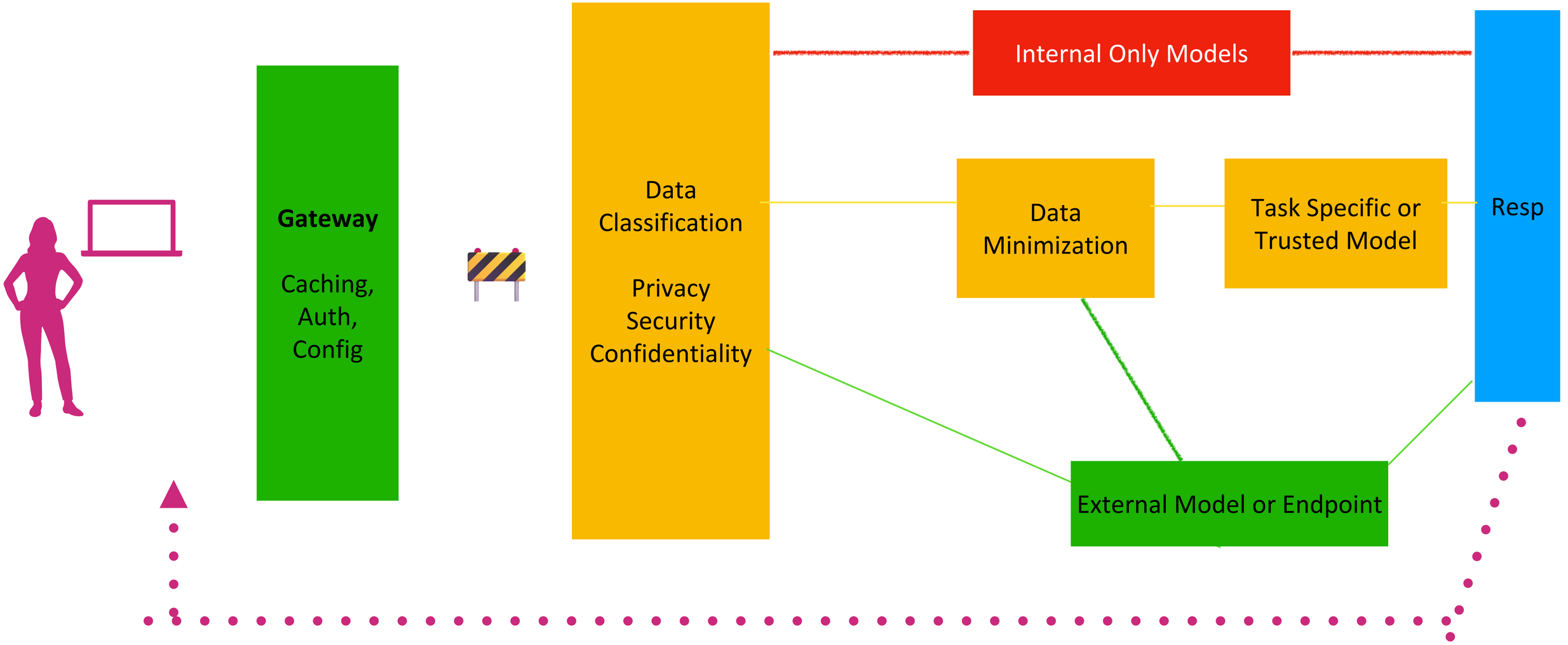
Reducing Scope, Task, Data

	Email Example	Sourced Images Example
Task	Email Prioritization	Best Images
Data	All email or just some? Also attachments?	User tagged images on social media
Model	Tested LLM model with data eng pipeline	Image-to-text + Image classification
Controls	Entity pseudonymization, no training data, regular evaluations + phased trial	Face pseudonymization + extra opt-in

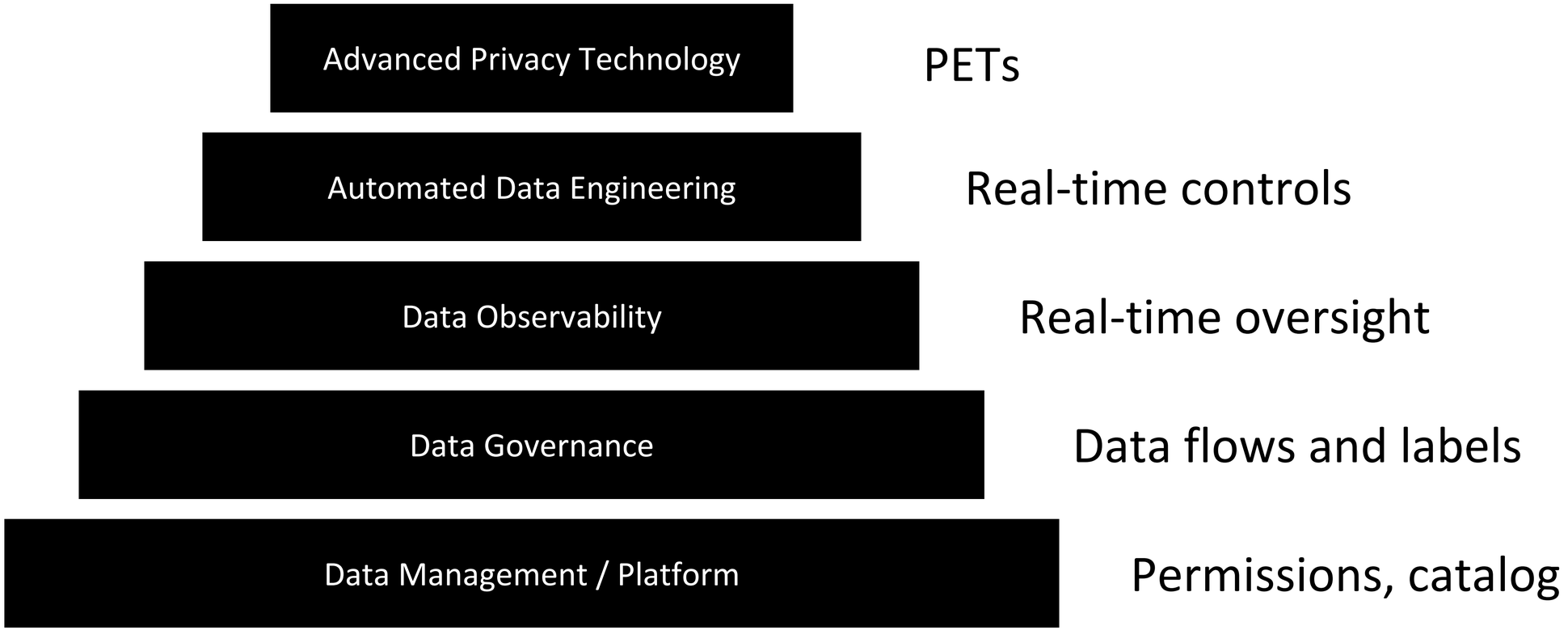
Privacy Observability



Privacy Routing

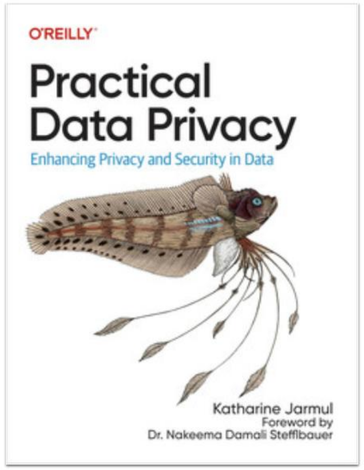


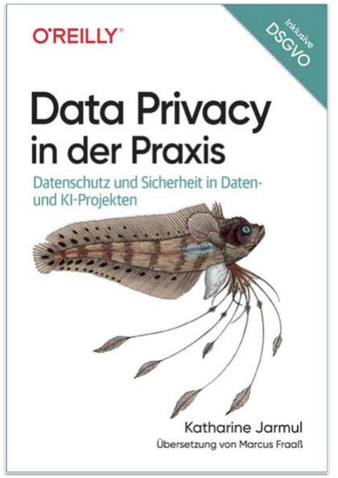
Privacy Engineering is now a must

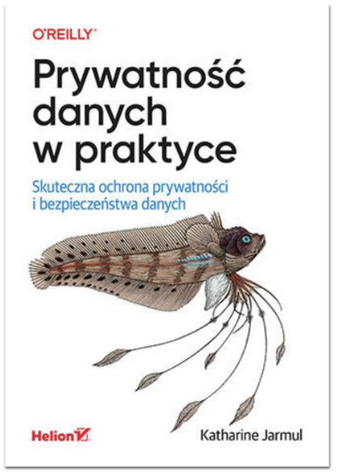


Thank you!

- Questions?
- You can hire me for advisory + trainings: https://kjamistan.com
- Probably Private Newsletter & YouTube https://probablyprivate.com





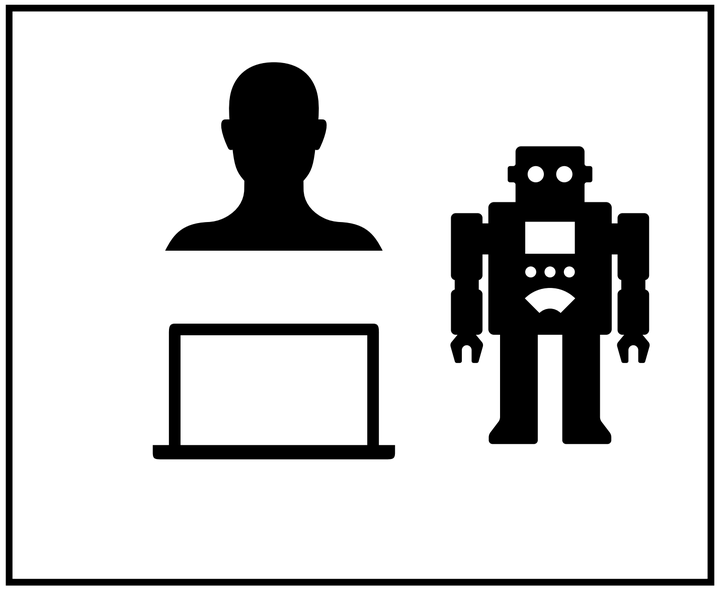




References

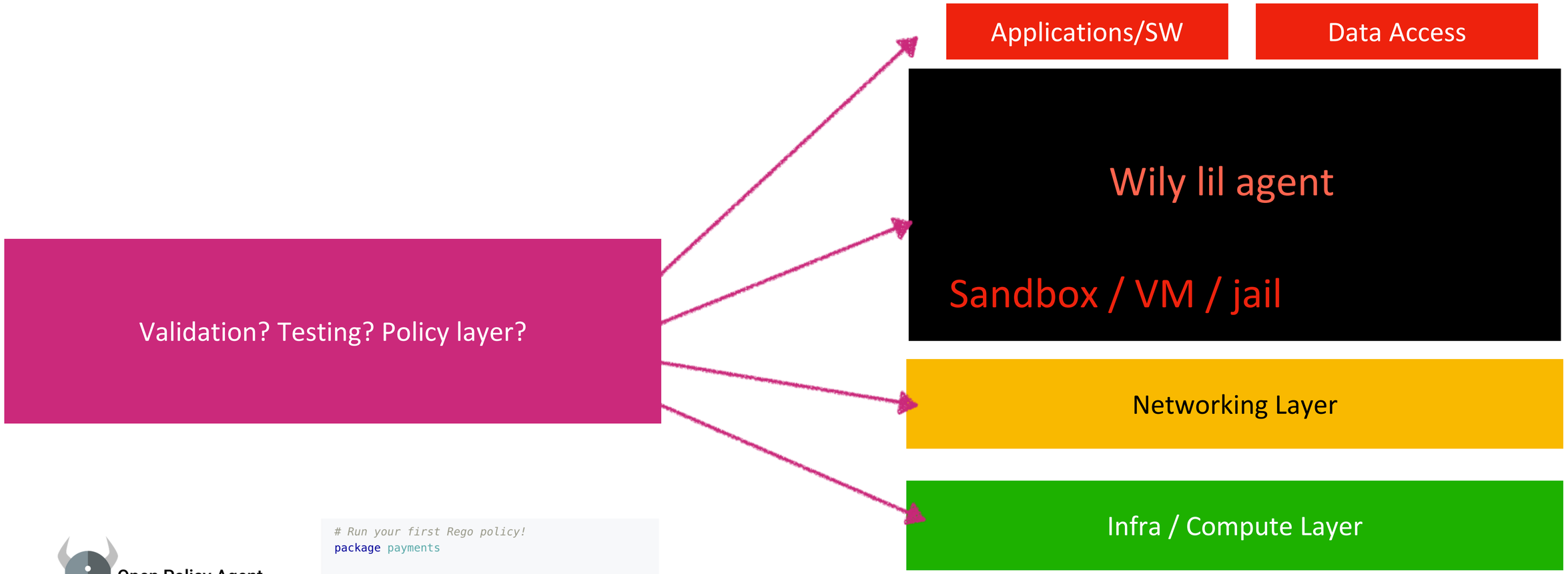
- Claude Code: https://code.claude.com/
- LangGraph: https://docs.langchain.com/oss/python/langgraph/workflows-agents
- Nanoclaw: https://docs.nanoclaw.dev/concepts/architecture
- Australia Government data access: https://www.bbc.com/news/articles/c6vgy0333dppo
- Beyond Memorization paper: https://arxiv.org/abs/2310.07298
- Example agentic sandbox: https://docs.nvidia.com/openshell/about/overview
- Guardrails megapost: https://blog.kjamistan.com/ai-guardrails-explained.html
- Privacy routing: https://blog.kjamistan.com/building-an-ai-privacy-router.html
- Open Policy Agent: https://www.openpolicyagent.org/

Identity and Delegation



I send an email.	Mail service actually sends.
My agent sends an email.	(Is it me?)
My agent set up a new email and sends.	Who owns this?
Spam emails are coming from my computer…	(Can we determine why/how/who?)

Policy enforcement



Open Policy Agent

OPA is a policy engine that streamlines policy management across your stack for improved development, security and audit capability.

```rego
# Run your first Rego policy!
package payments

default allow := false

allow if {
    input.account.state == "open"
    input.user.risk_score in ["low", "medium"]
    input.transaction.amount <= 1000
}
```