iapp Limina



Close the agent governance gap: Control access, track outcomes

Friday, 13 November

08:00-09:00 PST

11:00-12:00 EST

17:00-18:00 CET

Welcomes and Introductions



Patricia Thaine

Co-Founder, Chief AI Officer and Chair, Limina

THE RUNNING CASE



Support rep

“Does this customer qualify for a refund?”



Support agent

THE RUNNING CASE

A simple answer. A long trail of data.

ANSWERING THE QUESTION

Loaded the full record

Context window

Sent it to a model

LLM provider

Answered: eligible

Rep's screen

The model provider may also retain the request, depending on the service and its settings.

SAVED FROM THE INTERACTION

Ticket summary
Ticketing system

Chat embeddings
Vector database

Saved memory
Agent memory store

Saved for internal training
Company training dataset

The workflow for this one question saved customer data in at least four places.

A WEEK LATER, THE CUSTOMER WRITES IN

“What personal data do you hold about me?”

A data subject access request (DSAR). Where do you look?

THE RUNNING CASE

Four questions you now have to answer.

1

What did the agent retrieve?

2

Which controls applied?

3

Where did the data go?

4

What was derived from them?

Knowing who had access does not answer these four questions.

PART 1 OF 4

How agents work.

1
How agents work
Five parts, one loop

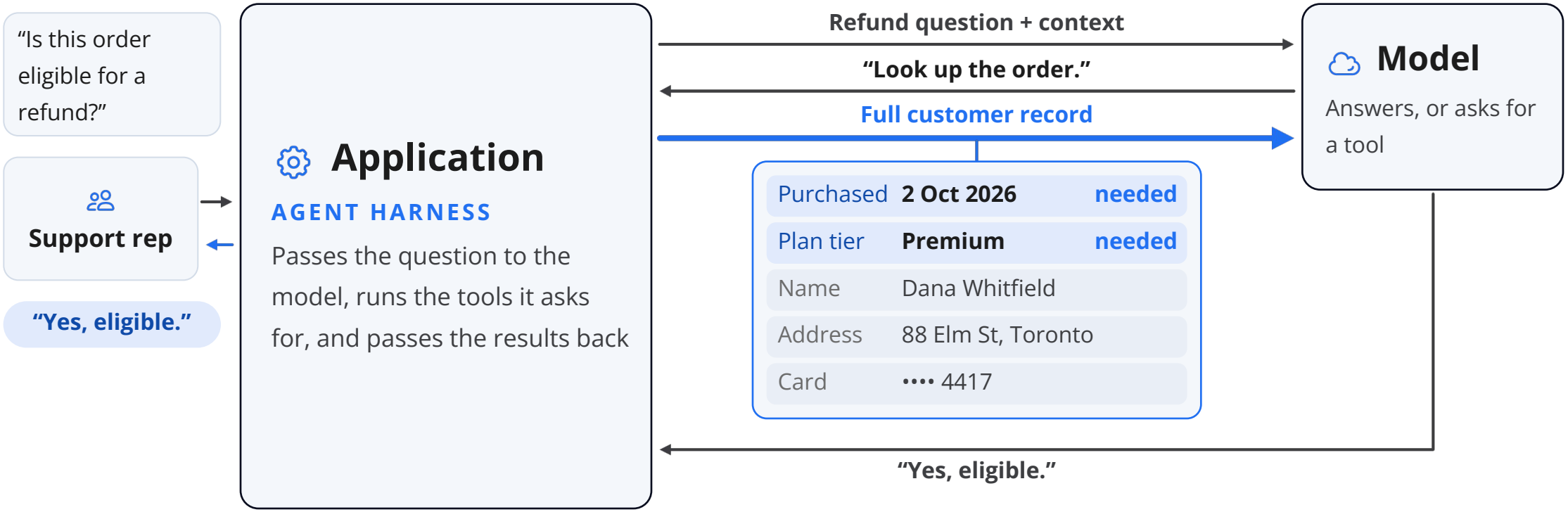
2
Where data leak
Seven paths

3
The governance gap
What approval misses

4
How Limina helps
Two control points

THE AGENT ECOSYSTEM

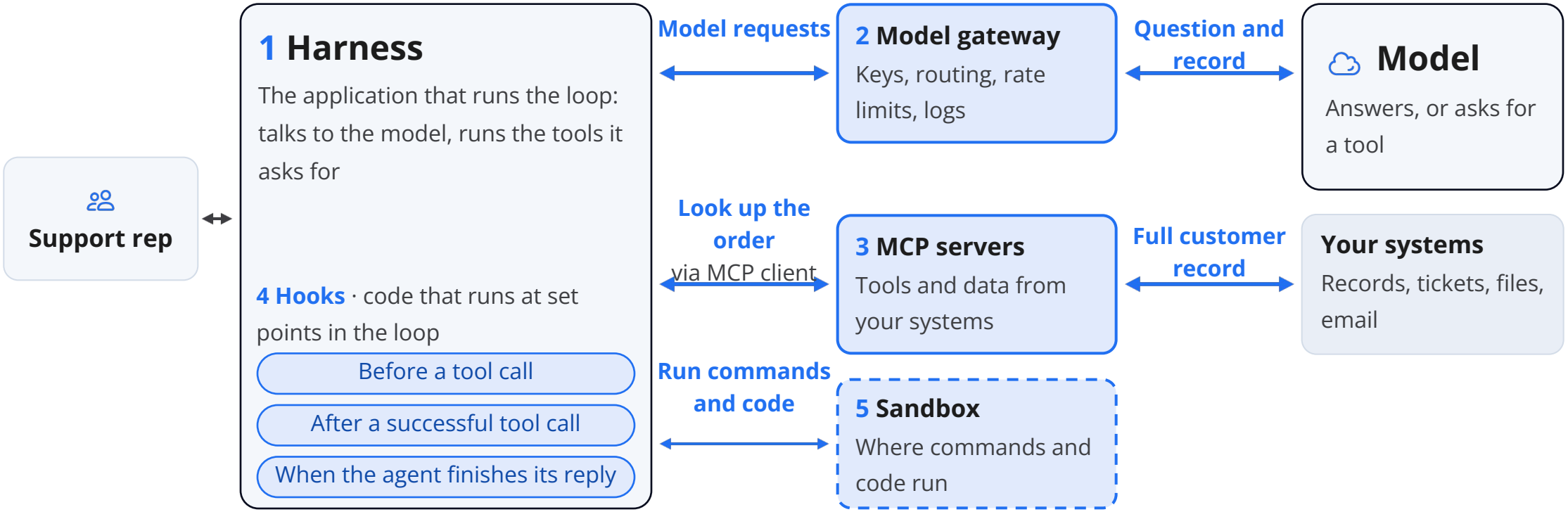
How customer data enters the agent loop.



Each tool call can bring new data into the model's context.

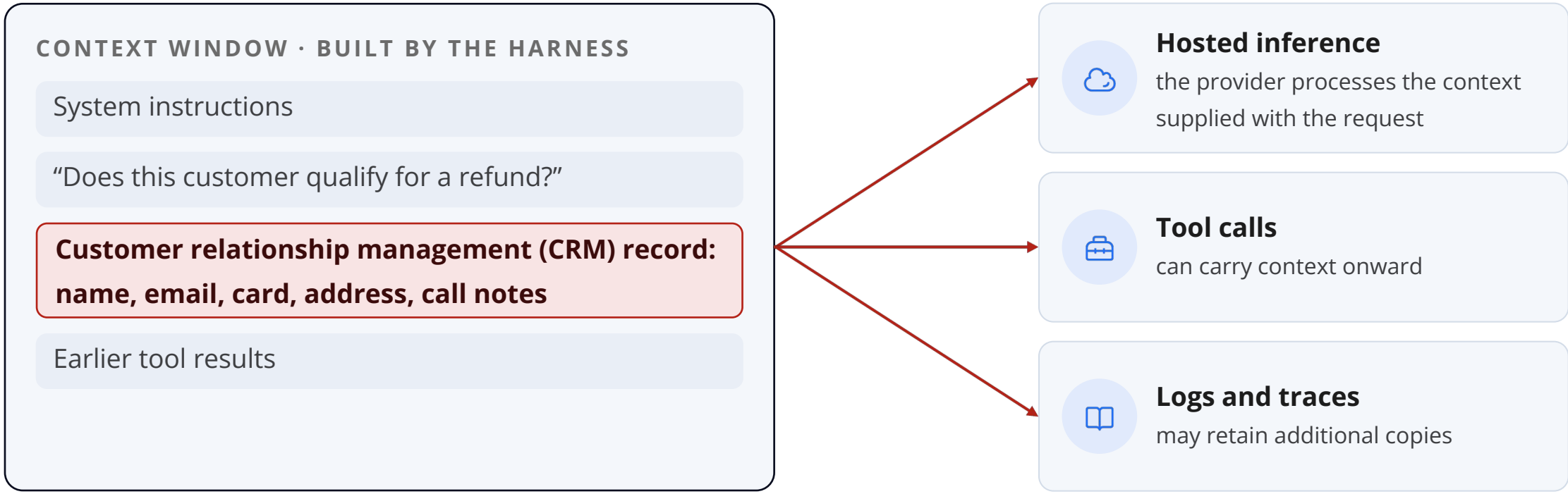
THE AGENT ECOSYSTEM

What's behind that order lookup?



HARNESS · PRIVACY LENS

Everything in the context window is shared onward.



MODEL GATEWAY · PRIVACY LENS

Routing and logging alone do not establish whether data are appropriate for the task.

PROMPT `Dana Whitfield, card ending 4417, mentioned a hospital stay. Refund eligible?`

What routing and logging record

- Which agent called
- Which model it reached
- How much text, and when

What judging the content takes

- Configured detection: name, card, health
- A policy on whether this task may send it
- Many gateways can add both, if configured

⚠ Even with gateway guardrails, e.g. LiteLLM's, the gateway:

Masks at the model call, after the harness has retrieved the full record

Sees model calls, not the tickets, embeddings and memories made next

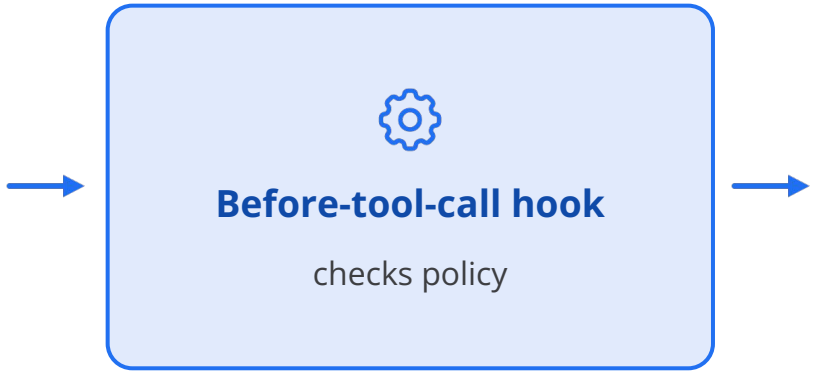
Applies rules per request; restrictions don't travel with the data

HOOKS · PRIVACY LENS

A hook can stop a call before it runs.

THE AGENT ASKS FOR

```
send_email(
  to: "x@external.example",
  body: full customer record
)
```



⚠ A hook is a place to enforce. On its own, it:

Fires at the action, after the agent already holds the full record

Needs detection and policy to know what the data are and where they may go

Sees one call, not where earlier copies went or what was derived

PART 2 OF 4

Where data leak.

1
How agents work
Five parts, one loop

2
Where data leak
Seven paths

3
The governance gap
What approval misses

4
How Limina helps
Two control points

iapp

WHERE DATA LEAK

Seven ways data reach the wrong place.

▭ = up close next

1 Agent identity

Too much access

Accident Attack

ASI03 Identity and privilege abuse

2 Emails, forms, web pages

Injected instructions

Attack

ASI01 Agent goal hijack

3 MCP servers

Poisoned tools

Attack

ASI02 Tool misuse

ASI04 Agentic Supply chain

4 Gateways, libraries

Shared infrastructure

Accident Attack

ASI03 Identity and privilege abuse

5 Hooks, project files

Planted configuration

Attack

ASI04 Agentic Supply chain

ASI05 Unexpected code execution

6 Files the agent writes

Sandbox handoff

Accident Attack

ASI05 Unexpected code execution

7 Memories, embeddings

Lasting derivatives

Accident Attack

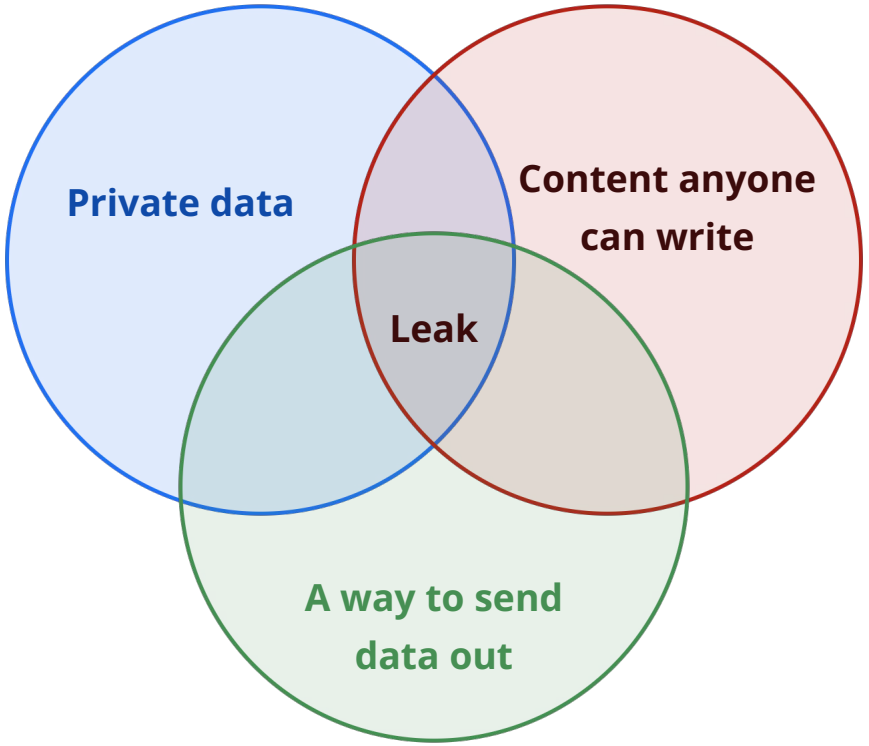
ASI06 Memory and context poisoning (partial)



Up close next: paths 2, 4 and 7, each with a 2026 example.

Blue: closest risks in the OWASP Top 10 for Agentic Applications for 2026 (Dec 2025). ASI07–ASI10 are not mapped.

Three ingredients let an attacker make an agent leak.



Our support agent has all three by design.

- Reads customer records
- Reads emails from anyone, attackers included
- Can send email

"The lethal trifecta", Simon Willison, June 2025

LEAK PATH 2 OF 7 · INJECTED INSTRUCTIONS

A web form told the agent what to do.

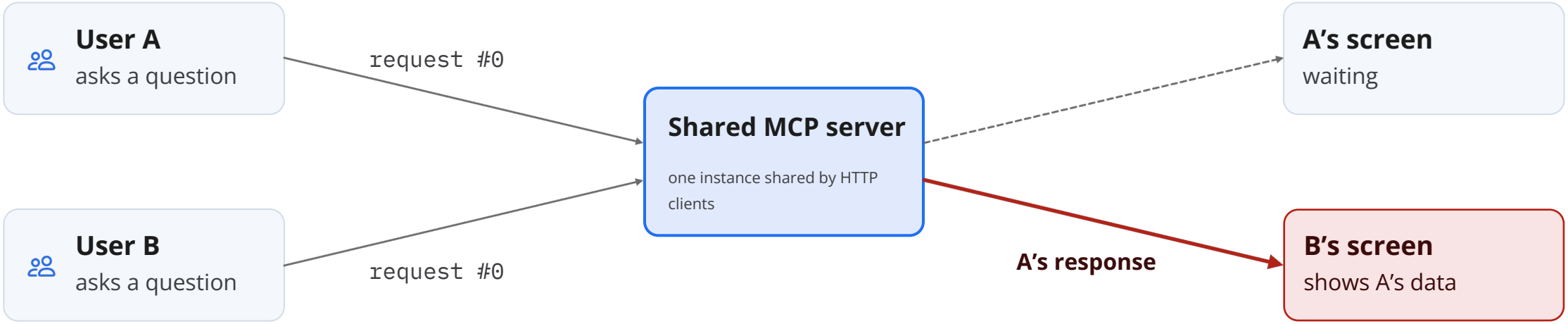


Microsoft Copilot Studio · Salesforce Agentforce · Disclosed April 2026 by Capsule Security

Disclosed vulnerabilities · Sources: VentureBeat, 15 Apr 2026; Dark Reading, Apr 2026

LEAK PATH 4 OF 7 · SHARED INFRASTRUCTURE

One user’s answers on another user’s screen.



Replies were matched by request number alone. Both were number 0, so A’s reply could reach B.

Disclosed vulnerability · @modelcontextprotocol/server 2.0.0–2.2.0, fixed in 2.3.0 (CVE-2026-111770)

LEAK PATH 7 OF 7 · LASTING DERIVATIVES

An embedding is not automatically anonymous.

STORED EMBEDDING

```
[0.12, −0.83, 0.47, 0.05, −0.31, 0.66, …]
```

inversion →

RECOVERED TEXT

Dana Whitfield called about a refund after a hospital stay.

92%

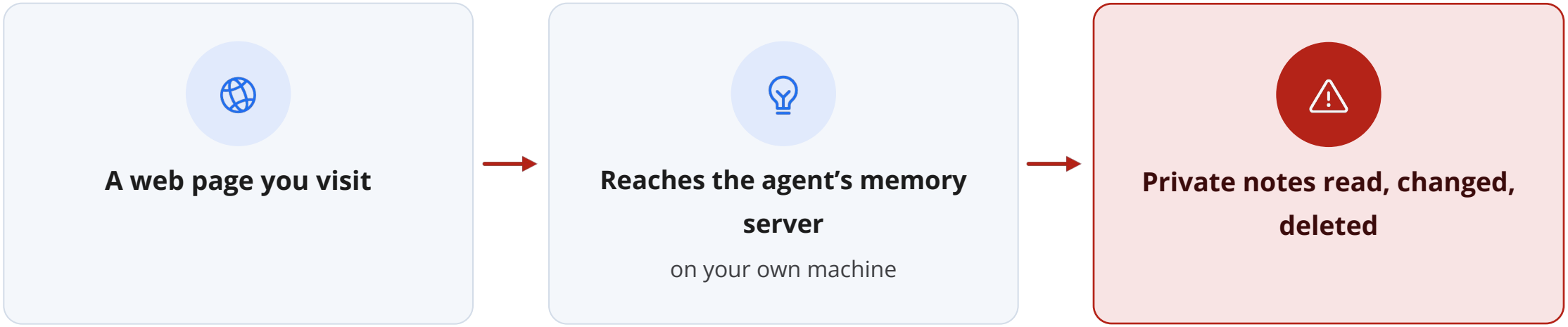
of 32-token inputs recovered exactly, in the study's tested setting

The same work recovered full names from a dataset of clinical notes

Research demonstration · Morris et al., arXiv:2310.06816, 2023

LEAK PATH 7 OF 7 · LASTING DERIVATIVES

Agent memories can be read back.



Where they hold personal data, rights requests have to reach memories and embeddings too.

Disclosed vulnerability · Basic Memory advisory GHSA-4gfv-c2mm-mvj2, Oct 2026, fixed in 0.24.0

WHICH POINTS TO THE FIX

The safest data are the data the agent never held.

For the data it does hold, keep a record of where they go.

PART 3 OF 4

The governance gap.

1
How agents work
Five parts, one loop

2
Where data leak
Seven paths

3
The governance gap
What approval misses

4
How Limina helps
Two control points

THE GOVERNANCE GAP

Approved once. Decided on every run.

1 review

approves the use case at design time

RUN LOG · REFUND AGENT · ONE AFTERNOON

Run	Request	What it did	Check
#2201	“Is order 4471 eligible?”	Plan tier and purchase date	As reviewed
#2202	“Why was I charged twice?”	Order and billing lookups	As reviewed
#2203	“Is this order eligible?”	Full record, card and notes	More than needed
#2204	“Summarize this case”	Saved a summary and a memory	Not in the review
#2205	“Can I return this?”	Return policy and order date	As reviewed

...and thousands more runs, each deciding for itself

Each run decides what to retrieve, which tools to call, what to combine and what to create.

THE GOVERNANCE GAP · GDPR

The principles already apply.

5(1)(b)	**Purpose limitation**
5(1)(c)	**Data minimisation**
25	**Data protection by design and by default**
30	**Records of processing activities**

What's new: the agent decides what data to use, on every call.

Articles of the General Data Protection Regulation (GDPR)

THE GOVERNANCE GAP · EU AI ACT

These records help support AI Act obligations.

APPLIES

2 Feb 2025

Prohibitions, AI literacy

Articles 5 and 4

APPLIES

2 Aug 2026

Transparency

Article 50: people are told when they're dealing with an AI system

Today

Nov 2026

NEXT

2 Dec 2027

High-risk uses (Annex III)

Automatic logs (Art. 12), human oversight (Art. 14), data governance (Art. 10). Deployers keep logs under their control at least six months, unless other law provides otherwise (Art. 26(6))

LATER

2 Aug 2028

High-risk in regulated products

Annex I: medical devices, machinery and similar

Is my agent high-risk?

It depends on intended purpose and the Act's rules. A refund agent usually isn't. Annex III lists uses such as credit scoring, recruitment, and life and health insurance pricing.

Our four questions aren't a legal checklist

They're a practical way to organize your records. Those records can help with the Act's logging and oversight duties. The GDPR still applies too.

Application dates · Regulation (EU) 2024/1689, as amended by Regulation (EU) 2026/1744

THE GOVERNANCE GAP · WHY LINEAGE MATTERS

Lineage helps you respond when something happens.

Access requests

Locate the person's personal data across relevant systems and derivatives

GDPR Art. 15

Deletion

Apply erasure requirements across relevant stores and derivatives, subject to applicable exceptions

GDPR Art. 17

Incident response

Supervisory authority: notify where required, without undue delay and, where feasible, within 72 hours of becoming aware of the breach

Individuals: inform affected individuals without undue delay if the breach is likely to result in a high risk to their rights and freedoms, subject to exceptions

GDPR Arts. 33 and 34

Accountability

Per-run lineage supports accountability and helps locate data for rights requests

GDPR Art. 5(2)
AI Act Art. 12

Without lineage, the answer to "where did the data go?" is a guess.

General Data Protection Regulation (GDPR) · EU AI Act Art. 12 applies to high-risk systems

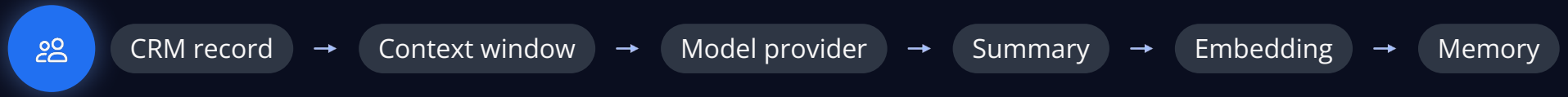
iapp

THE GOVERNANCE GAP · LINEAGE

Lineage has to follow a person's data.

PERSON-LEVEL LINEAGE ASKS

Where are this person's data now, and what was derived from them?



The vocabulary already exists: `used` `wasGeneratedBy` `wasDerivedFrom`

Vocabulary: World Wide Web Consortium PROV Ontology (PROV-O)

PART 4 OF 4

How Limina helps.

1
How agents work
Five parts, one loop

2
Where data leak
Seven paths

3
The governance gap
What approval misses

4
How Limina helps
Two control points

HOW LIMINA HELPS

Limina works at two control points.



HOW LIMINA HELPS · THE RIGHT DATA FOR EACH TEAM

Same message. Each team sees what it needs.

Customer messages mix useful lessons with private details. Limina gives each team its own safe version.

ORIGINAL MESSAGE

Hi, I'm **Maria Okafor**. I bought the night serum on **October 28**. Two days later my cheeks broke out in a rash. You can reach me at **416-555-0187**.

Red = private details

Underlined blue = what happened

MEDICAL RESEARCH · SERVICE ANALYTICS

Hi, I'm **Lucia Okeke**. I bought the night serum on October **[DAY_1]**. Two days later my cheeks broke out in a rash. You can reach me at **[PHONE_NUMBER_1]**.

They learn what happened, not who it happened to.

CUSTOMER SERVICE · REFUNDS

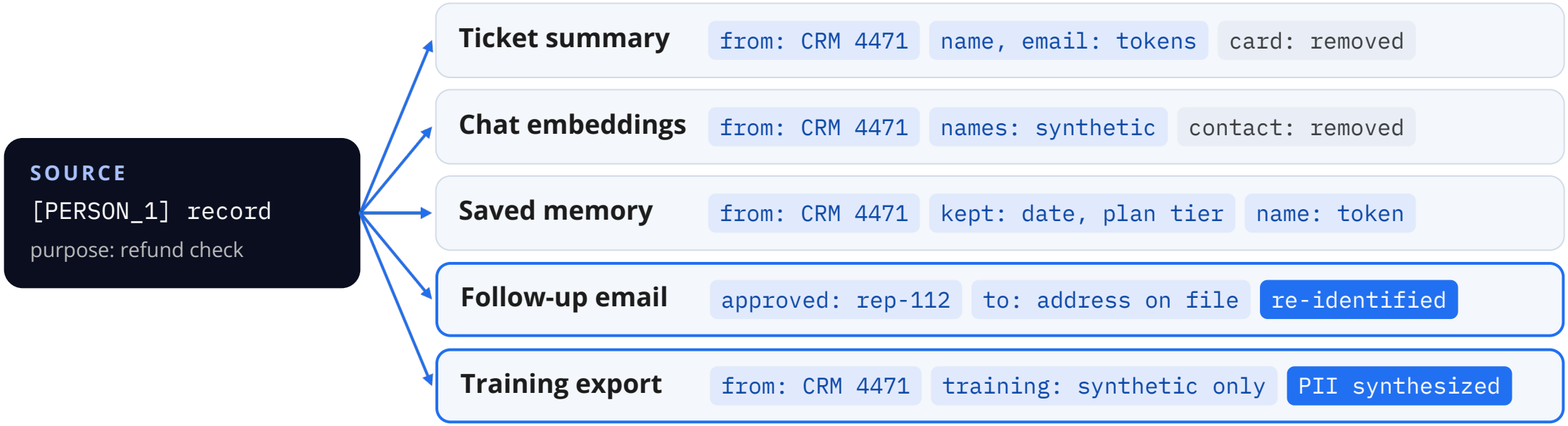
Hi, I'm Maria Okafor. I bought the night serum on October 28. Two days later, **[CONDITION_1]**. You can reach me at 416-555-0187.

They can process her refund without knowing what happened.

Pseudonymized data remain personal data under the GDPR (Recital 26)

HOW LIMINA HELPS · DURING THE WORKFLOW

Restrictions travel with the data.



Recording a restriction is not enough. Enforcing it is the control.

Answers questions 3 and 4

HOW LIMINA HELPS · DURING THE WORKFLOW

What a lineage record should capture.

ACTOR

Agent identity, the user it acts for, and the task

`agent=refund-assist for=rep-112`

PURPOSE

The approved use for this run

`purpose=refund-check`

RETRIEVED Q1

What it read, and how sensitive it was

`crm/4471 fields=plan,purchased`

CONTROLS APPLIED Q2

Transformations, blocked actions and policy version

`tokenize synth=training v3.2`

DESTINATIONS Q3

Each model, tool and service the data reached

`llm=hosted email=approved`

DERIVED Q4

Outputs created and the restrictions they inherit

`summary,embedding,memory ttl=30d`

Q1–Q4: the four questions from the start

HOW LIMINA HELPS · THE REP-APPROVED FOLLOW-UP

The agent never saw her name. She got her email.

AGENT RECEIVED Plan tier: Premium | Purchased 2 Oct 2026 | [PERSON_1] | [EMAIL_1] | *Card and address removed*

WHAT THE AGENT WROTE

To: [EMAIL_1]

Hi [PERSON_1], your 2 October order is eligible for a refund.

Trusted executor re-identifies

WHAT THE CUSTOMER RECEIVED

To: dana.w@example.com

Hi Dana, your 2 October order is eligible for a refund.

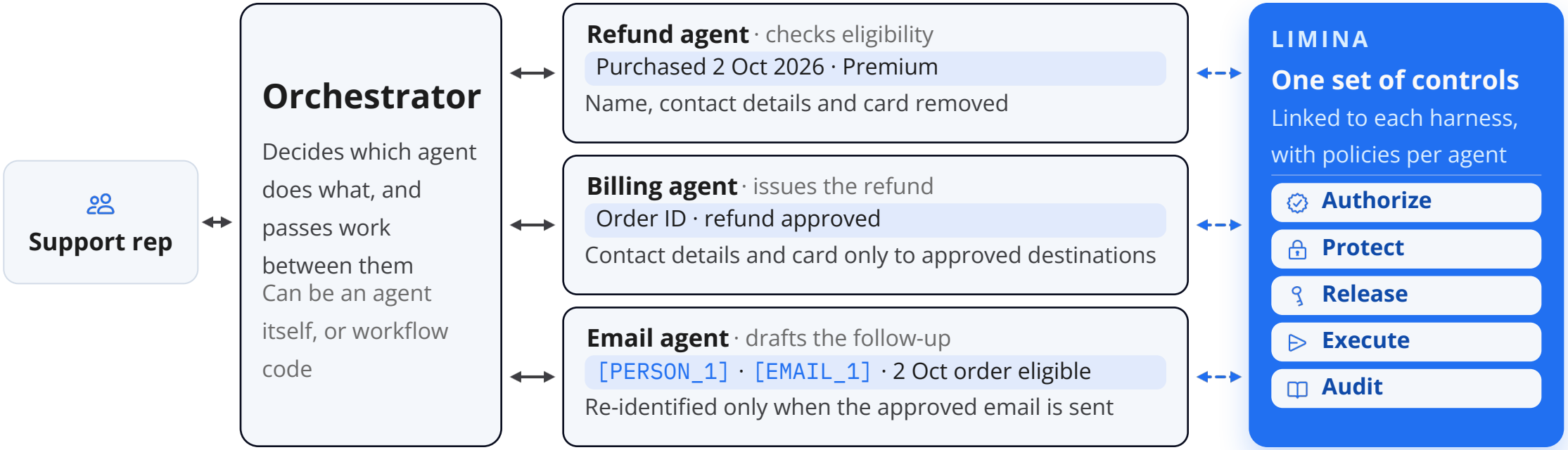
After the eligibility check, the rep authorizes a follow-up. It needs her name and email, so both stay as tokens until release.

Audit: who approved the email, which policy applied, where originals went, training copy synthesized.

HOW LIMINA HELPS · SEVERAL AGENTS

Several agents. Each sees what it needs.

CUSTOMER RECORD | Dana Whitfield | dana.w@example.com | 88 Elm St | •••• 4417 | Purchased 2 Oct 2026 | Premium



Pseudonymized data remain personal data under the GDPR (Recital 26)

HOW LIMINA HELPS · SCOPE

Where the controls apply, and where they stop.

Leak path	What the controls do	Still your security team's job
Too much access	The task gets only minimized data	Right-size agent identities
Injected instructions	Original values only to approved destinations	Limit what emails and forms can trigger
Poisoned tools	Each call checked; each result minimized	Vet and pin MCP servers
Shared infrastructure	Tokens limit what a cross-user leak shows	Patch gateways and SDKs; isolate sessions
Planted configuration	Not covered: code outside the integration	Review agent configuration like code
Sandbox handoff	Agent output carries direct identifiers as tokens	Host-level controls; patch harnesses
Lasting derivatives	Derivatives the workflow creates are recorded	Retention and deletion across stores

Minimization and lineage shrink the blast radius. They do not replace security engineering.

WHAT TO DO NEXT

Start with one workflow.

1. List your agents and their access
2. Write down one task's minimum data
3. Minimize before retrieval
4. Approve the destinations
5. Log answers to the four questions
6. Extend DPIAs to derivatives
7. Test a DSAR on that workflow

The four questions are your acceptance test.

DPIA: data protection impact assessment · DSAR: data subject access request

WHAT TO DO NEXT · YOUR ACCEPTANCE TEST

Can your records answer these four questions?

1

What did the agent retrieve?

2

Which controls applied?

3

Where did the data go?

4

What was derived from them?

Run them against one workflow. If your records can't answer all four, it isn't ready.

WHERE THE FIELD IS

Available today, still emerging.

Available today

- Minimize before access
- Task-scoped agent identities
- Authorize every call
- Approved destinations only
- Log retrievals, tool calls and model calls

Still emerging

- Lineage through vector stores
- Deletion that reaches derivatives
- Reliable translation of natural-language policy into enforceable rules
- Provenance across vendors
- Defenses against adaptive injection

Permissions tell you what
the agent is allowed to do.

Governance has to tell you
what it actually did with the data.

Questions and Answers



Patricia Thaine

Co-Founder, Chief AI Officer and Chair, Limina

Thank you

FEEDBACK SURVEY

Tell the IAPP how this program went, in about two minutes.

iappwf.questionpro.com/t/AbBPvZ9zp2

1.0 CPE CREDIT HOUR

AIGP, CIPP/US, CIPP/E, CIPP/A, CIPP/C, CIPT and CIPM. Certified professionals registered before the live session receive it automatically.

Submit for CPE credits

CONTACT

Limina: info@getlimina.ai · getlimina.ai

IAPP web conferences and recordings: livewebconteam@iapp.org

SOURCES

Sources and further reading.

- S. Willison: The lethal trifecta for AI agents, Jun 2025
- Model Context Protocol specification
- LiteLLM docs: PII and PHI masking guardrail
- OWASP: Top 10 for Agentic Applications for 2026
- VentureBeat: ShareLeak and PipeLeak, 15 Apr 2026
- Dark Reading: Microsoft, Salesforce patch AI agent data leak flaws, Apr 2026
- Hu et al.: Log-To-Leak, arXiv:2610.36108
- Forkast: LiteLLM CVE-2026-59822 in the CISA catalog
- GitHub advisory: @modelcontextprotocol/server, CVE-2026-111770
- GitHub advisory: Envoy AI Gateway MCP sessions
- GitHub advisory: Basic Memory, Oct 2026
- Morris et al.: Text Embeddings Reveal (Almost) As Much As Text, 2023
- EU AI Act, Regulation (EU) 2024/1689
- Digital Omnibus on AI, Regulation (EU) 2026/1744
- European Commission AI Act Service Desk: Article 26
- GDPR, Regulation (EU) 2016/679, Arts. 5, 15, 17, 25, 30, 33, 34