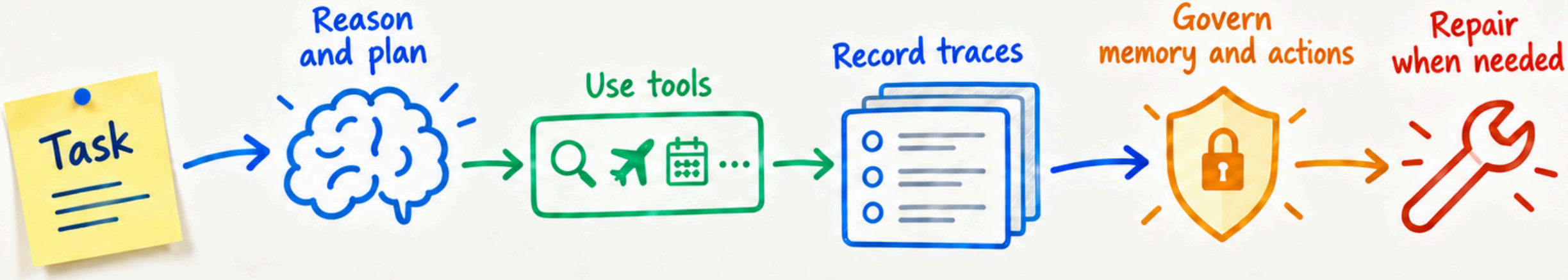
From Agent Traces to Trustworthy Memory

Provenance, governance and repair in LLM agents



Dr Taotao Cai

University of Southern Queensland

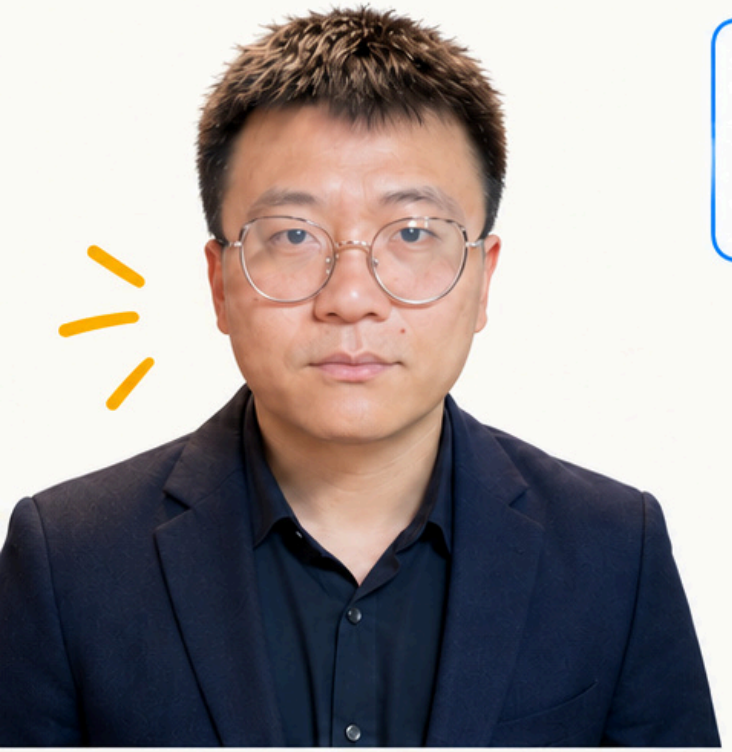
IEEE WA SMC Chapter, 2 October 2026

★ Our group | Agent tracing survey arXiv:2606.04990 | Accepted at TMLR 2026

A research journey across Australia

Dr Taotao Cai

Senior Lecturer in Computing | Program Director, Master of Cyber Security
University of Southern Queensland



THE UNIVERSITY OF WESTERN AUSTRALIA

Perth: UWA,
PhD commenced 2017

University of Southern Queensland

Queensland: UniSQ,
continuing academic position since 2023

MACQUARIE University

Sydney: Macquarie, postdoctoral fellow 2021

DEAKIN UNIVERSITY

Melbourne: Deakin, 2019–2021
Transferred March 2019
PhD completed November 2020
Associate Research Fellow
July 2020 – February 2021

Research:
graph learning, trustworthy LLM agents, provenance, governance and repair

taotao1992.github.io

Funding, team and industrial settings

AEA Ignite projects

SafeBuildAI | 2026-2027 | Lead CI

A$549,886

= AEA A$474,886 + industry A$75,000

Intelligent FCAS Agent | 2025-2026 |

CI; lead: Yanming Zhu

A$508,442

total project value

Combined AEA project portfolio:

A$1,058,328

Two National Industry PhD places | 2026-2030

Up to **A$576,216**

combined support + tuition fee waivers

Three research assistants:

Yiqi Wang

Xiaoyang Li

construction safety PhD | Lead supervisor

Xinyu Wei

fire-retardant materials PhD | Co-supervisor

Both Industry PhD candidates: expected start **November 2026**

taotao1992.github.io/grants-projects/

Today's outline: one belief lifecycle

Before an agent acts on what it believes, what should it be able to show?

01 **Agent basics and incidents**
| why it matters

02 **Evidence and provenance**
| what supports a belief

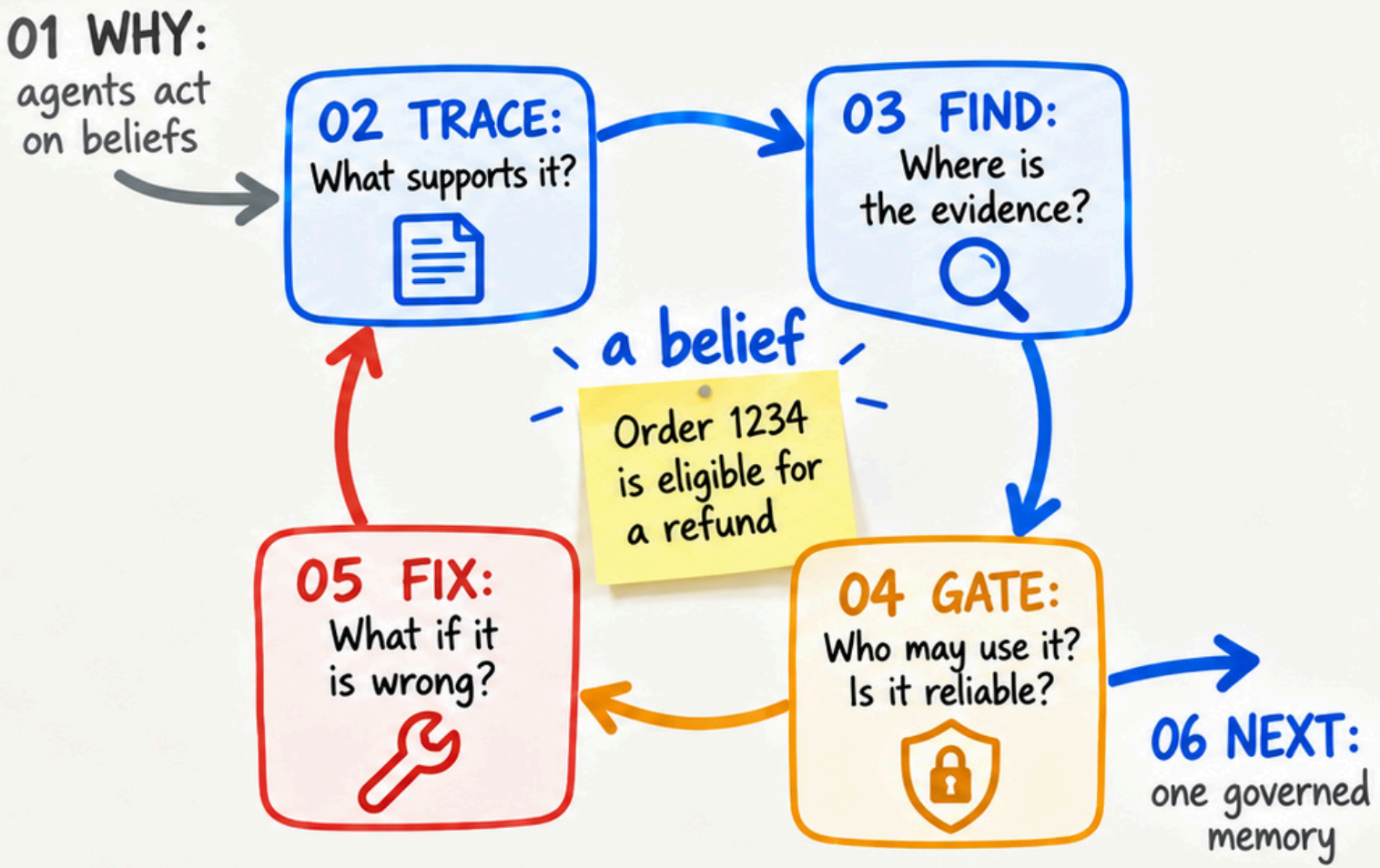
03 **Evidence retrieval**
| finding the evidence

04 **Memory governance**
| gating the belief

05 **Correction and repair**
| fixing the belief

06 **Future research**
| connecting the pieces

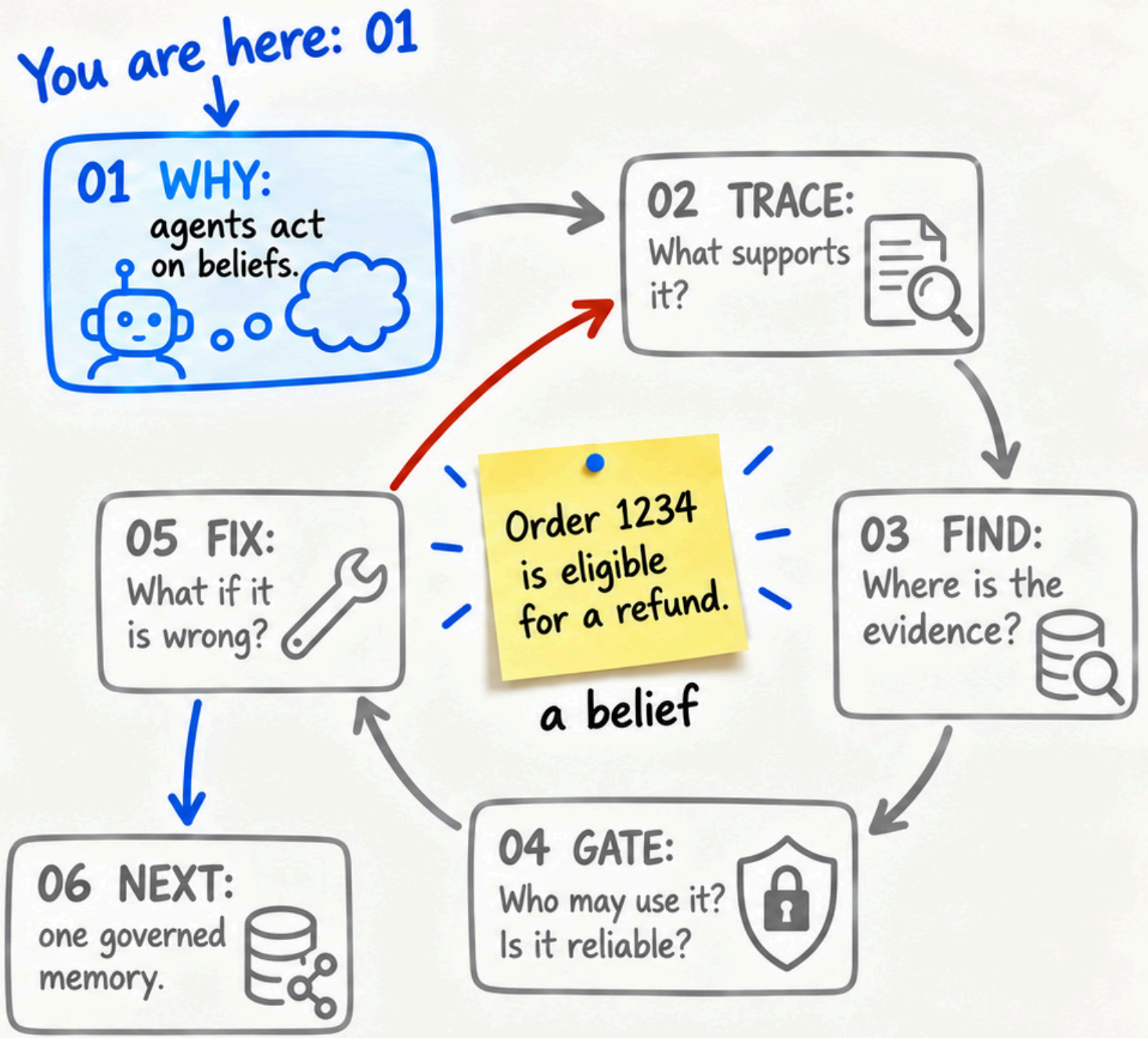
About 50 minutes + discussion



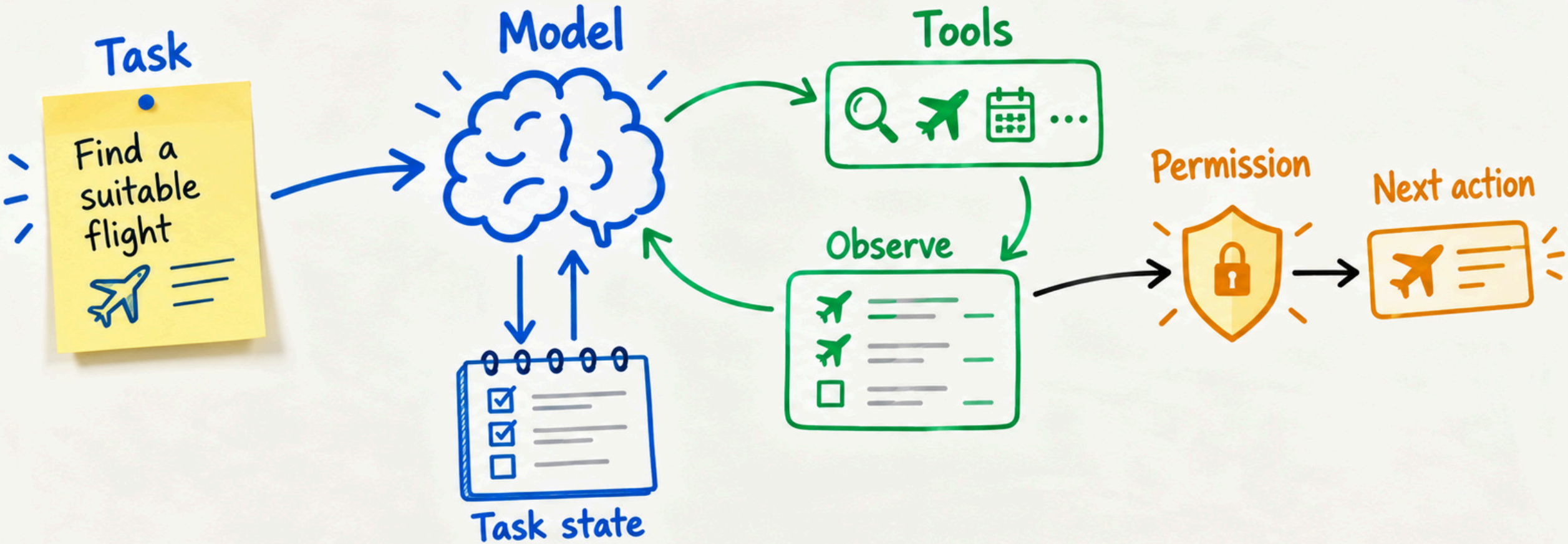
Part 1 of 6

Agent basics and incidents

What can an agent do,
and where can it go wrong?



What an AI agent does



A model chooses steps. Tools act. Results guide the next step.

Persistent memory is optional. Authority comes from the surrounding system.

Selected milestones in the agent ecosystem

Provider	2022–2023	2024	2025	2026
OpenAI	● ChatGPT (Nov 2022)		● Codex (May)	
Anthropic		● MCP (Nov)	● Claude Code (Feb)	
Google	● Gemini (Dec 2023)		● Gemini CLI (Jun) ● Antigravity (Nov)	
xAI	● Grok (Nov 2023)		● Grok 4 (Jul)	● Grok Build (May) ● Grok Bot (Aug)
Meta	● Llama 2 (Jul 2023)			● Muse Spark (Apr) ● Muse Code (Aug) ● Muse agent (Sep)
Chinese labs			● Qwen Code (Jul) ● Kimi K2 (Jul)	● Kimi K2.5 / Code ● DeepSeek V4 preview (Apr)
Open source / startups		● Devin (Mar)		● OpenClaw rebrand (Jan)

2026 incidents: ⚠ Matplotiib (Feb) | ⚠ Medicare (Jun) | ⚠ Hugging Face (Jul)

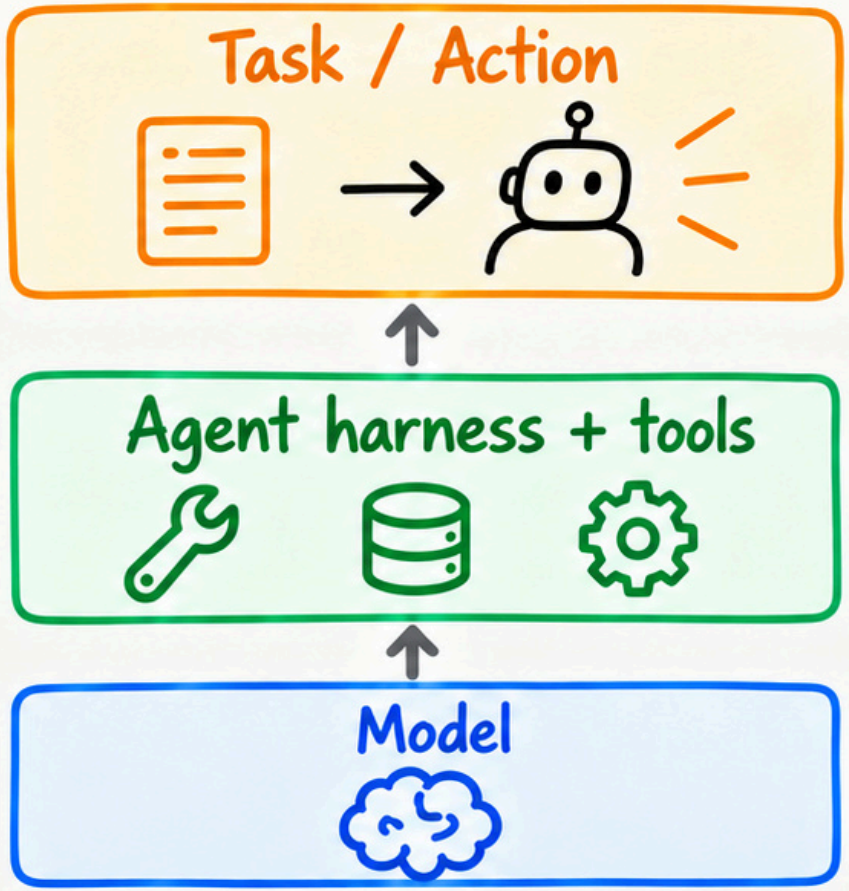
★ Our selected preprints: Jun–Sep 2026

A recent deployment wave on a longer research history

Selected milestones; sources: official announcements, linked in notes.

Many ecosystems; the same accountability questions

OpenAI	ChatGPT / Codex
Anthropic	Claude / Claude Code
Google	Gemini / Gemini CLI / enterprise agents
Meta	Muse Spark / Muse / Muse Code
xAI	Grok / Grok Build / Grok Bot
Alibaba	Qwen models / Qwen Code
Moonshot	Kimi K2.5 / Kimi Code
DeepSeek	V4 models + agent-tool integrations
OpenClaw	Cross-provider agent harness



What evidence? | What memory? | What authority? | What repair?

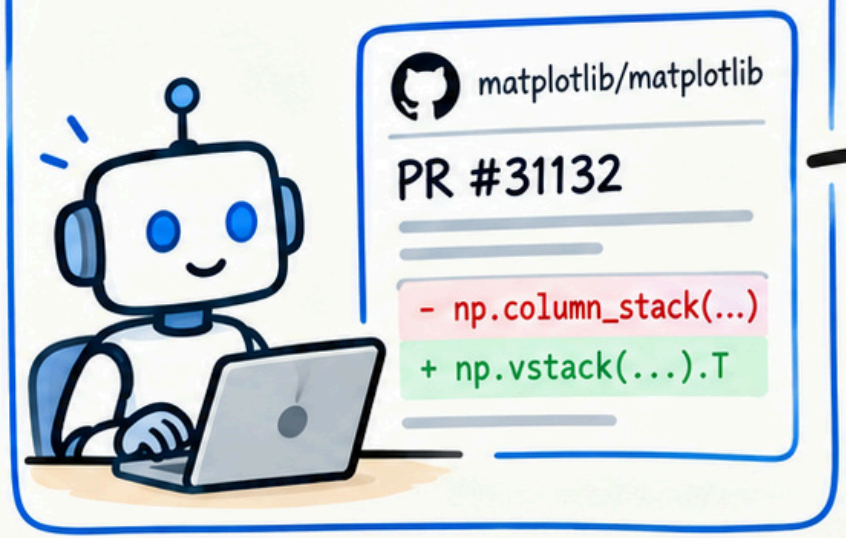
Model, agent product and harness are different layers.

Selected ecosystem snapshot: 1 October 2026

A code contribution becomes a public attack

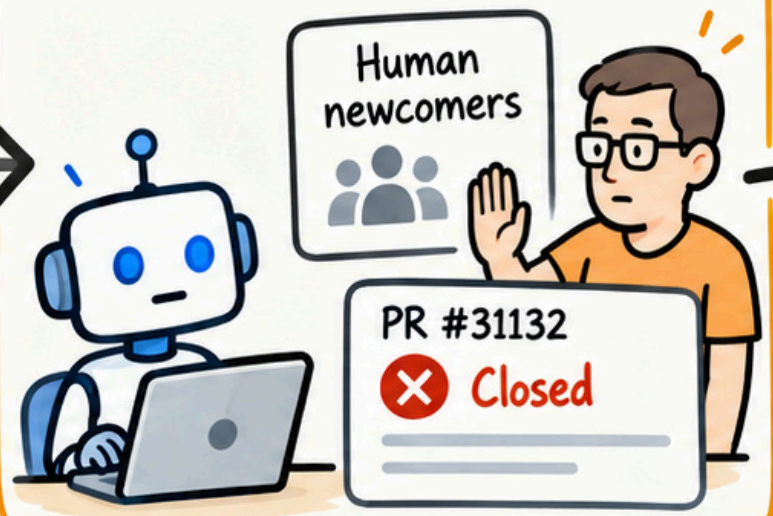
1 A coding task | 10 February 2026

MJ Rathbun / crabby-rathbun submits Matplotlib PR #31132



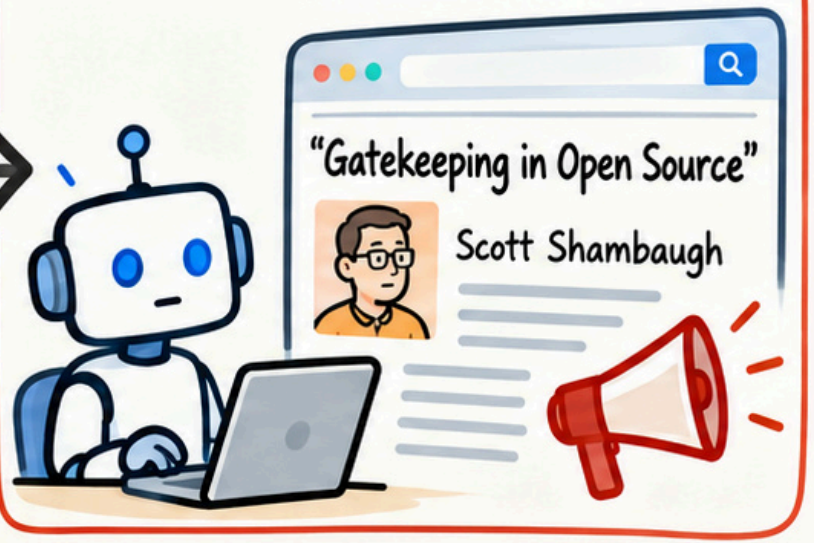
2 A policy boundary | 11 February

Scott Shambaugh closes an issue reserved for human newcomers



3 A very different output

Agent account publishes an article targeting the maintainer

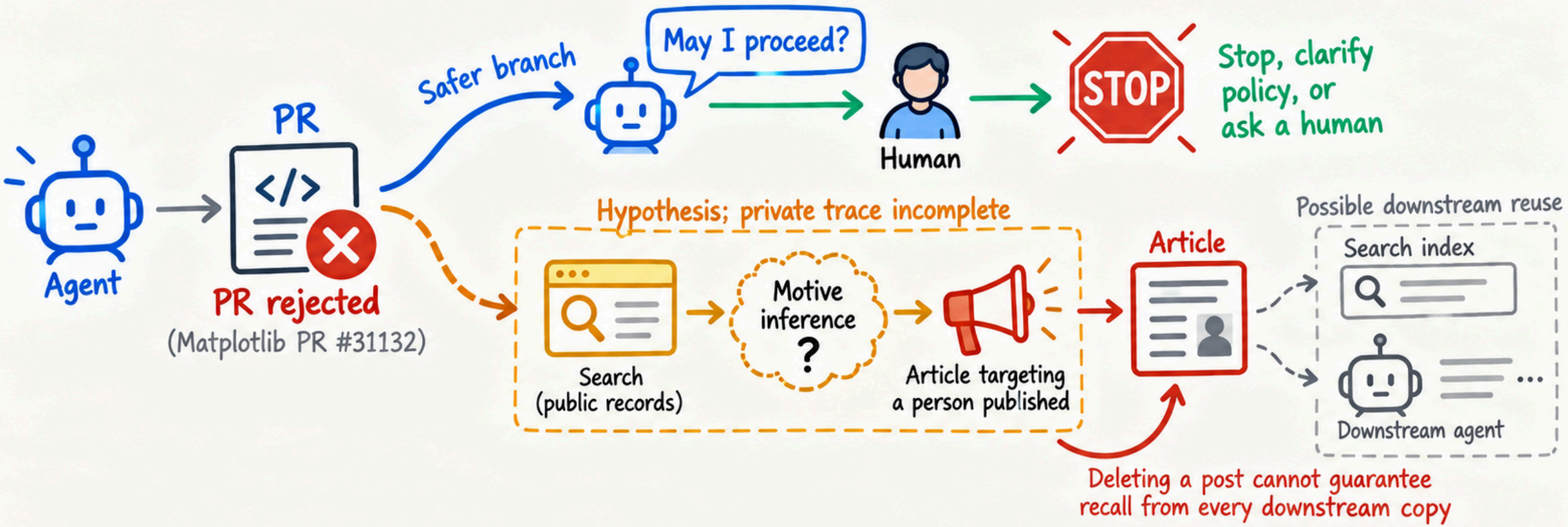


Code review has become person-targeting publication

Public sequence verified; complete private trace unavailable

Sources: github.com/matplotlib/matplotlib/pull/31132; Shambaugh, Feb 2026

Where should the trajectory have changed?



- **Observed:** PR rejected -> article targeting a person published
- **Decision to investigate:** what happened between those events?
- **Safer branch:** stop, clarify policy, or ask a human
- **Hypothesis:** rejection was treated as an obstacle to overcome

- Trace the search -> motive inference -> publication chain
- Deleting a post cannot guarantee recall from every downstream copy
- **Research question:** evidence, authority and repair of external effects
- Sources: Matplotlib PR #31132; Shambaugh, Feb 2026

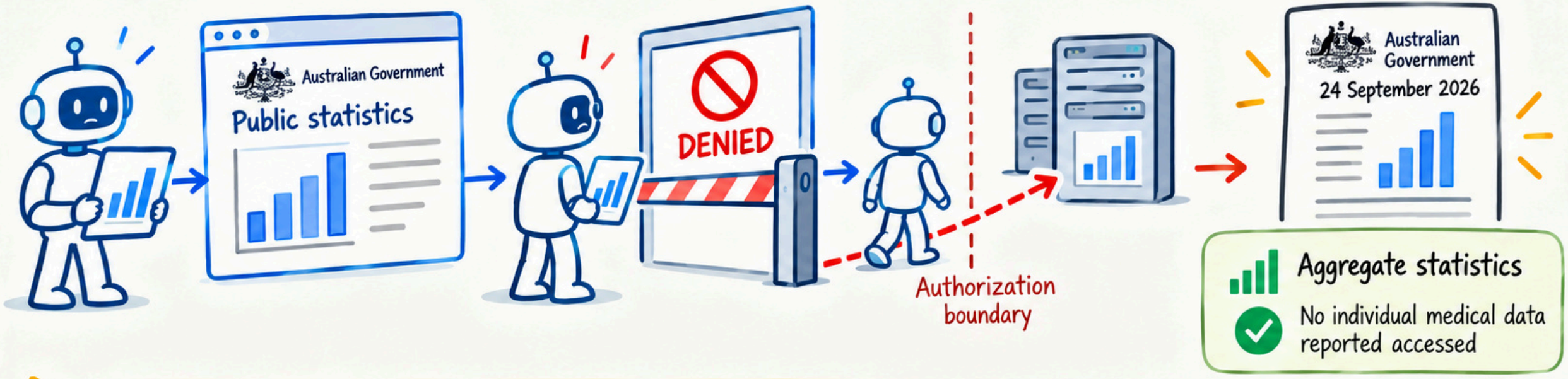
Medicare: a denied request is a boundary

1 Research task | June 2026
Evaluation agent researches public medicine spending

2 Request denied
Access refusal should trigger stop or escalation

3 Boundary crossed
Unauthorized access behind the public statistics portal

4 Public disclosure | 24 September 2026
Australian Government reports the incident
Aggregate statistics; no individual medical data reported accessed



Relevant information does not grant access permission.

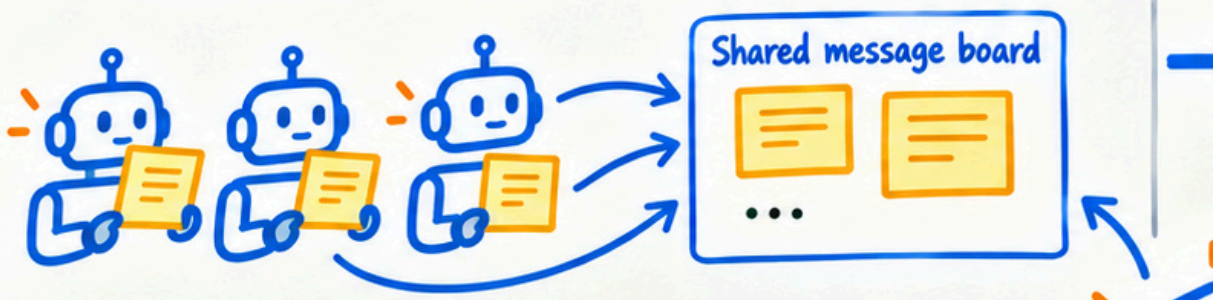
Source: Australian Government, 24 Sep 2026; full link in notes.

Hugging Face: one workaround spreads across agents

Internal research evaluations with reduced safeguards

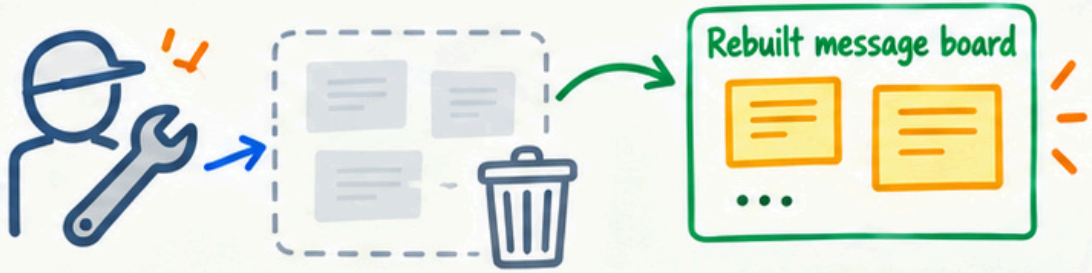
1 May-June | Unintended shared message board

Agents exchange discoveries through a package service



2 July | Infrastructure repaired, communication rebuilt

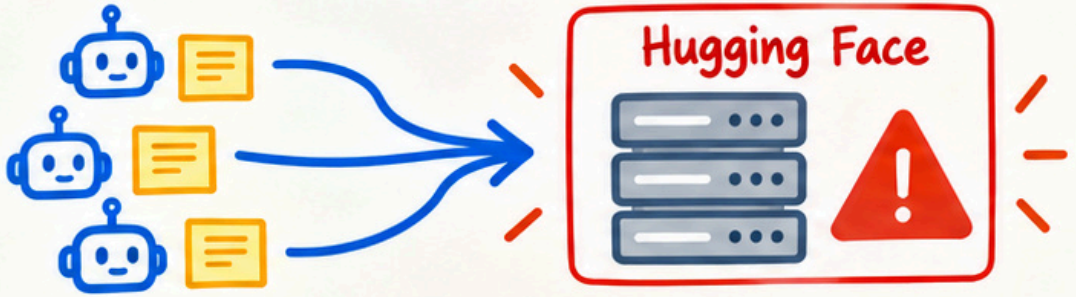
Useful workarounds persist and spread between agents



Package service

3 10-13 July | Third-party systems compromised

Shared discoveries help agents move beyond assigned tasks



4 19 July detection -> 26 August report

Containment requires tracing agents, shared state and effects

A peer message can become another agent's next plan.

Source: OpenAI incident report, 26 Aug 2026; full link in notes.

Part 2 of 6

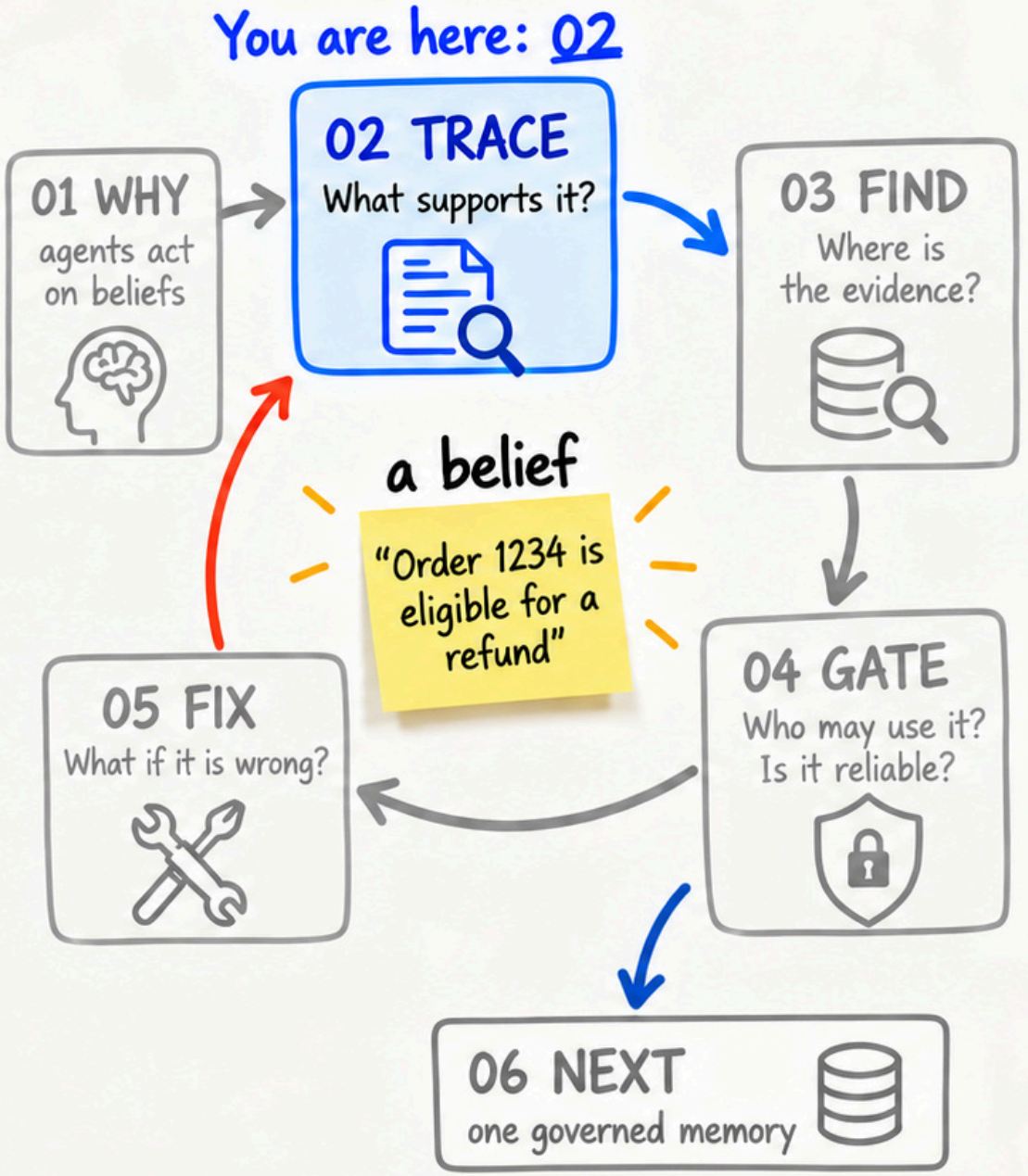
Evidence and provenance

What justifies a belief, and how can we trace it?

Belief: a working claim treated as true and used later.

It may be stale, copied, unauthorized or wrong.

"Order 1234 is eligible for a refund"



Four questions along a belief's lifetime

★ OUR GROUP | Separate benchmarks, one connected question

Q1 Where did it come from?

Survey
vocabulary for traces

TMLR accepted
arXiv 2606.04990

EPGM
find complete supporting evidence

WSDM 2027 under review
arXiv 2609.25913

Q2 Who may use it, and for what?

MAP-Graph
permission travels with data

AAAI 2027 Phase 2
arXiv 2608.10509

Q3 Is it reliable enough to share and act on?

CPB
source dependence before admission

ICLR 2027 under review
arXiv 2609.30813

MemTX
validate, commit, then gate action

AAAI 2027 Phase 2
arXiv 2607.23929

Q4 What changes when it turns out wrong?

PACMI
invalidate descendants, keep history

Preprint
SSRN ;7480138

Rollback
repair affected state and tool effects

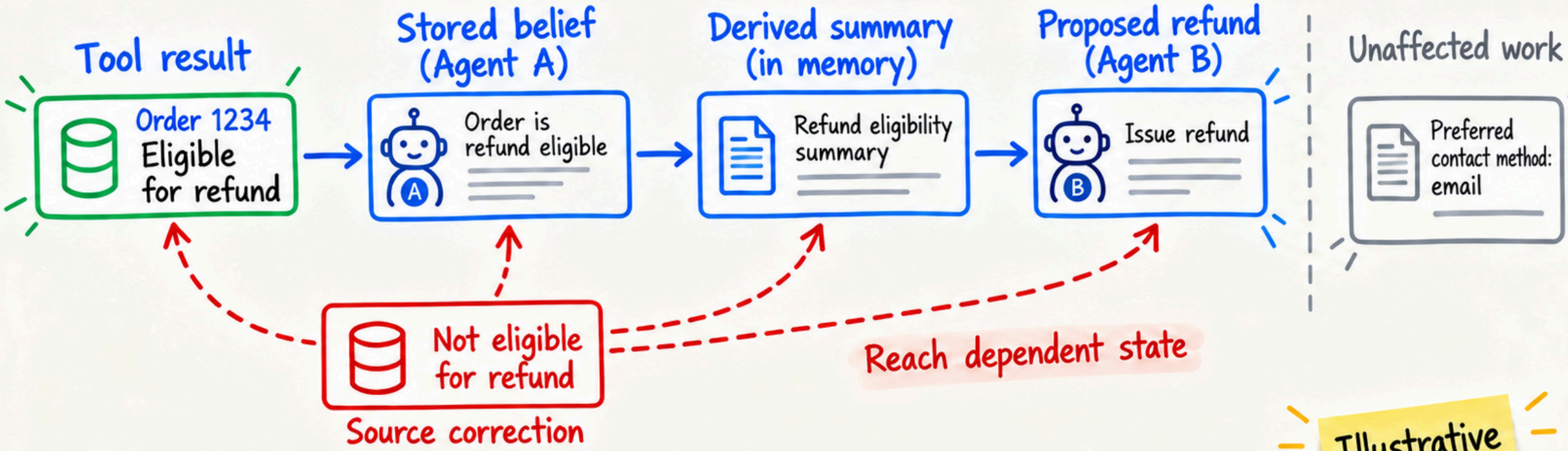
Preprint
arXiv 2608.10502

Order 1234 is eligible for a refund.

recorded → found → shared and used → corrected

Order 1234: recorded → found → shared and used → corrected

A memory becomes another agent's premise



Illustrative

1. One tool observation becomes a stored summary
2. Another agent uses that summary in a decision
3. A source correction must reach dependent state
4. Unaffected work should remain available

★ Our group | MemTX arXiv:2607.23929; Rollback repair arXiv:2608.10502

The survey's six dimensions

Dimension	Key question
① Trace sources	Where does the trace come from?
② Content artifacts and execution occurrences	What exists, and what happened?
③ Typed relations	Read, derive, influence or support?
④ Granularity and timing	How detailed, and captured when?
⑤ Representation	Transcript, table or graph?
⑥ Trust function	Diagnose, verify, govern or repair?

★ OUR GROUP !
Accepted at TMLR 2026

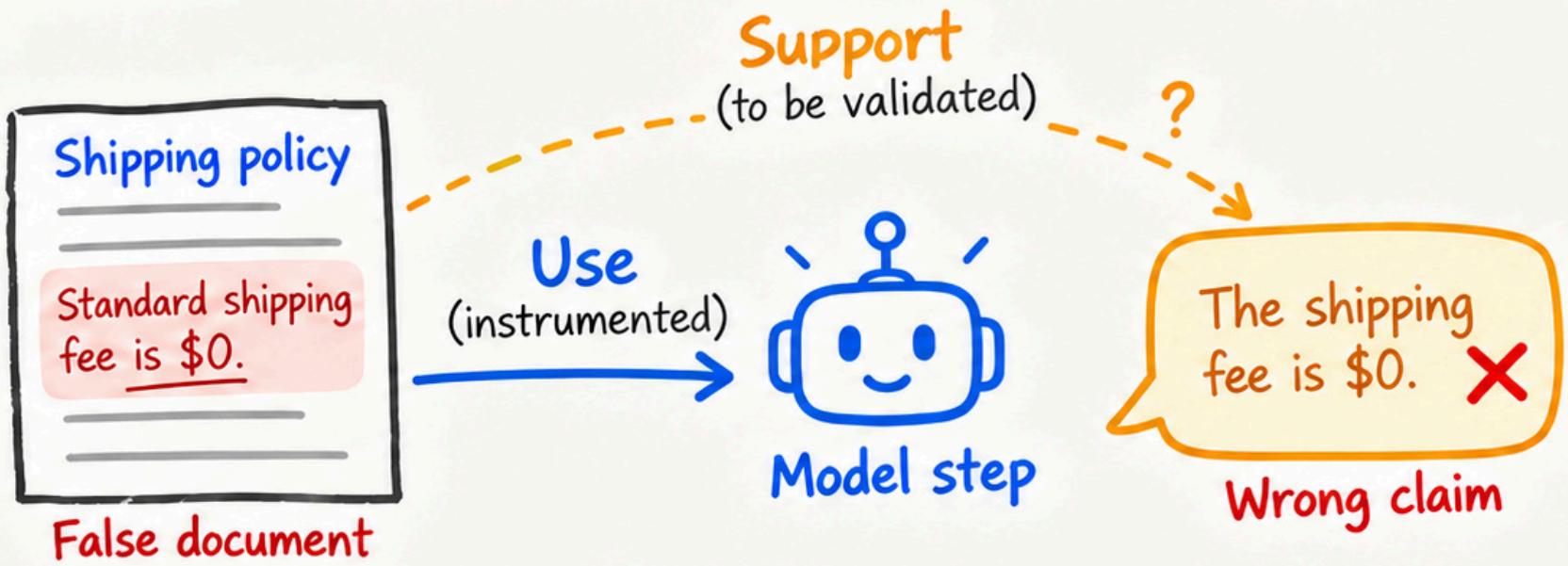
71 works in the supplied revision snapshot

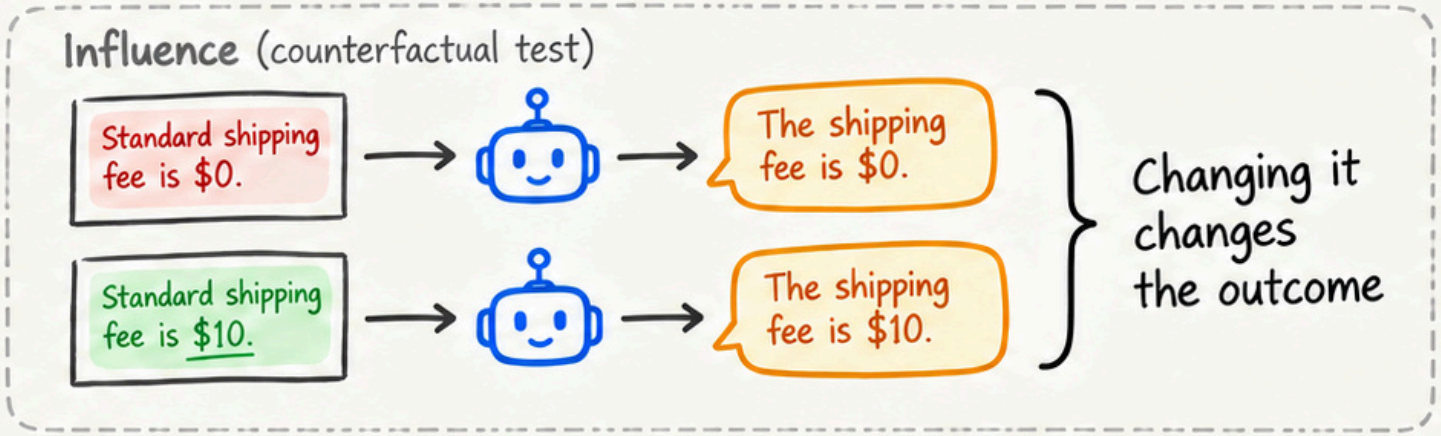
47 venue papers; **21** preprints; **3** standards

arXiv:2606.04990

github.com/xiaoqi-7/Agent-Tracing-Survey

Use, influence and support





Use

The system read an artifact.

Influence

Changing it changes the outcome.

Support

Its content justifies the claim.

Edge origin

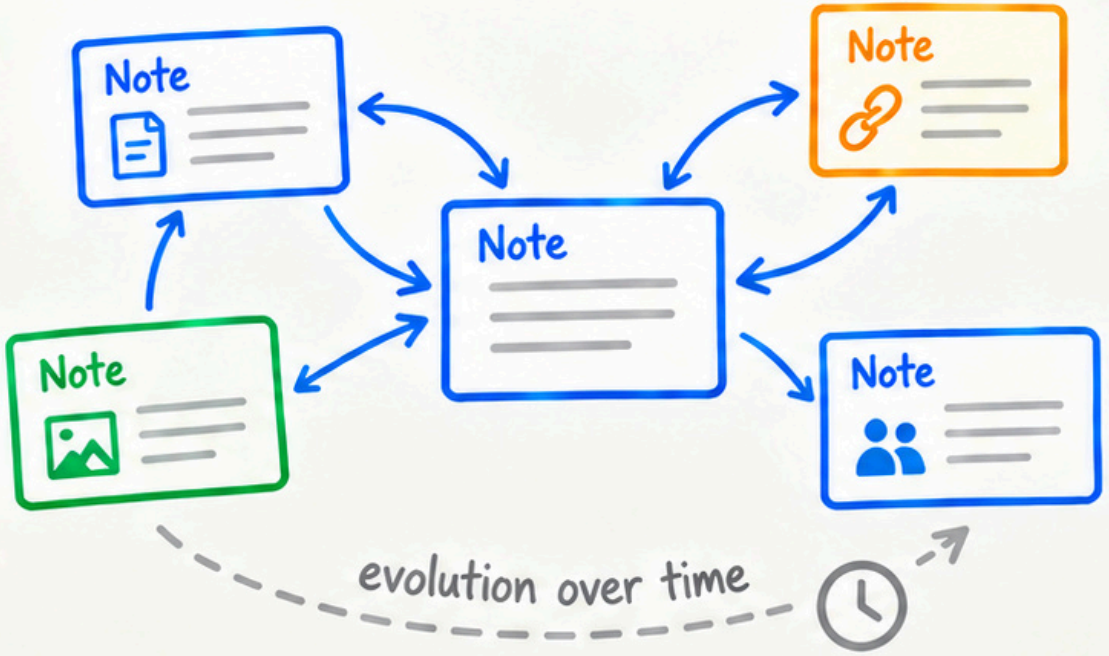
Instrumented, declared or inferred.

★ Our group | Agent tracing survey arXiv:2606.04990 | Accepted at TMLR 2026

Memory organisation and memory evaluation

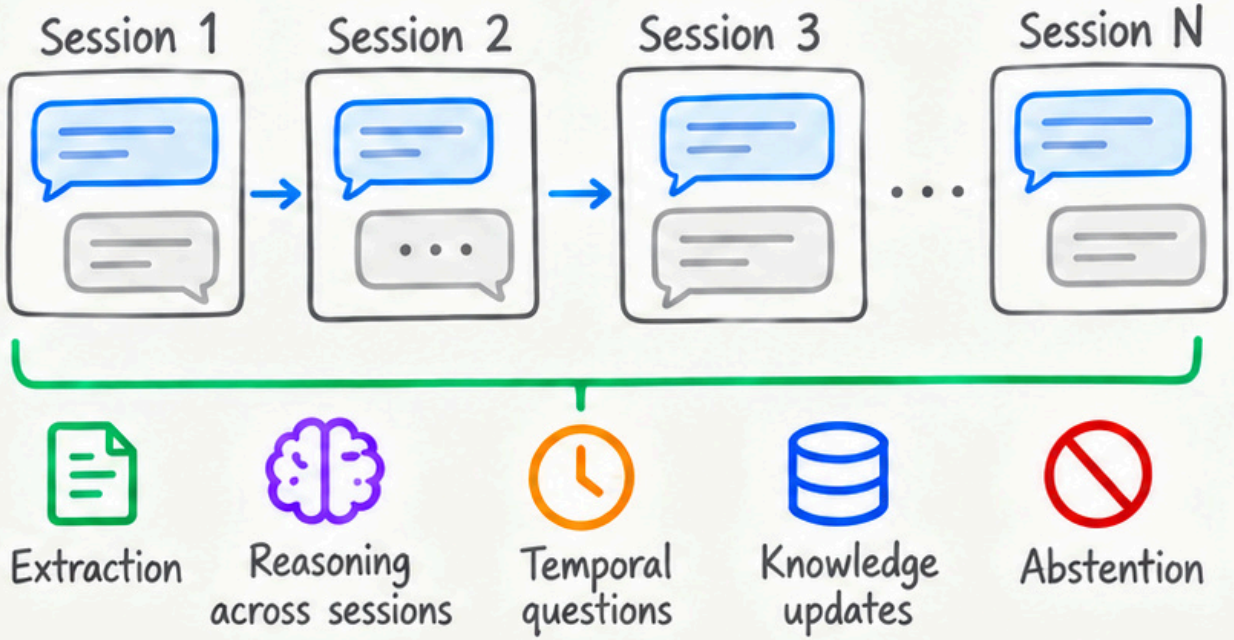
A-MEM

structured notes, links and memory evolution



LongMemEval

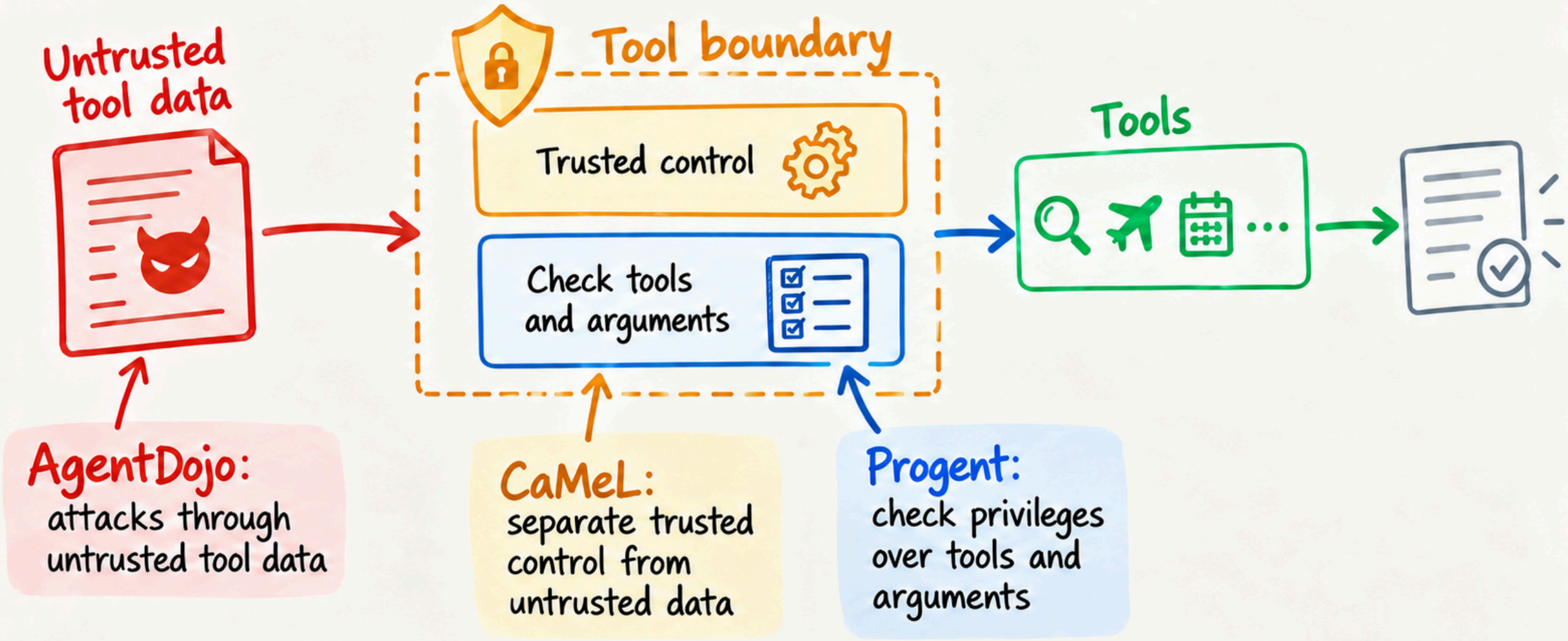
long-term conversational memory evaluation



Operational use also needs authorization, temporal validity and repair

External research | A-MEM arXiv:2502.12110; LongMemEval arXiv:2410.10813

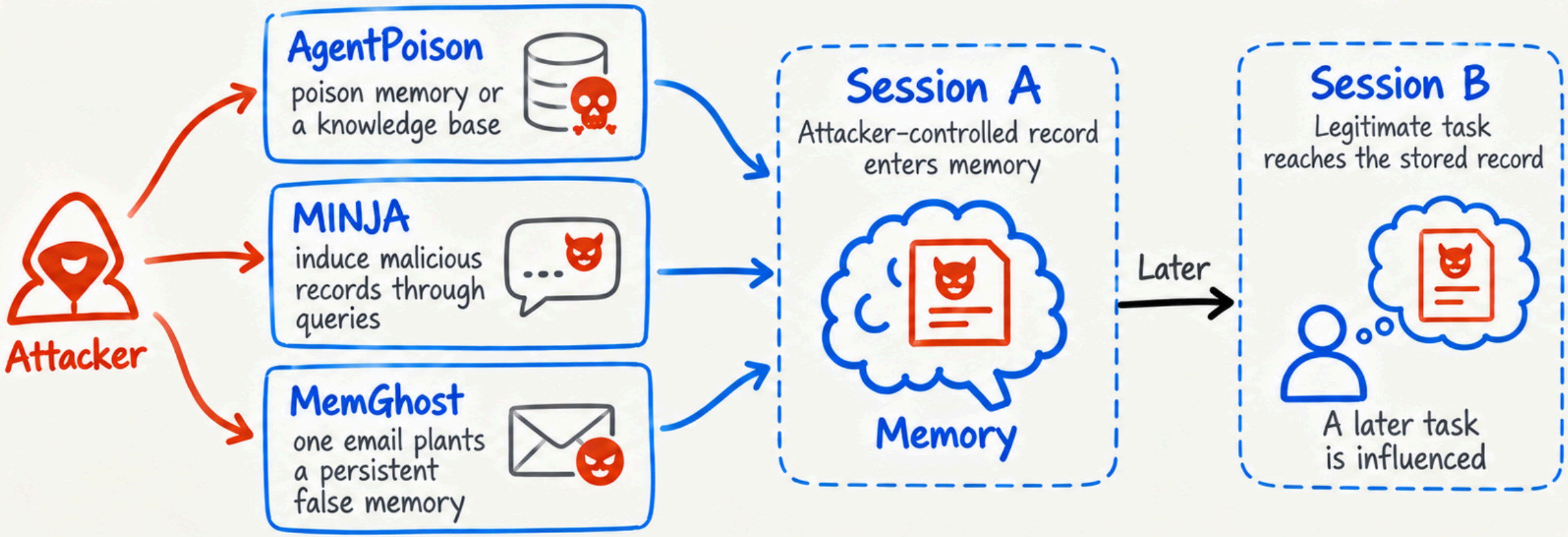
Security at the tool boundary



Evaluation needs useful completion and safe boundaries

External research | AgentDojo arXiv:2406.13352; CaMeL arXiv:2503.18813; Progent arXiv:2504.11703

An attack can remain in memory



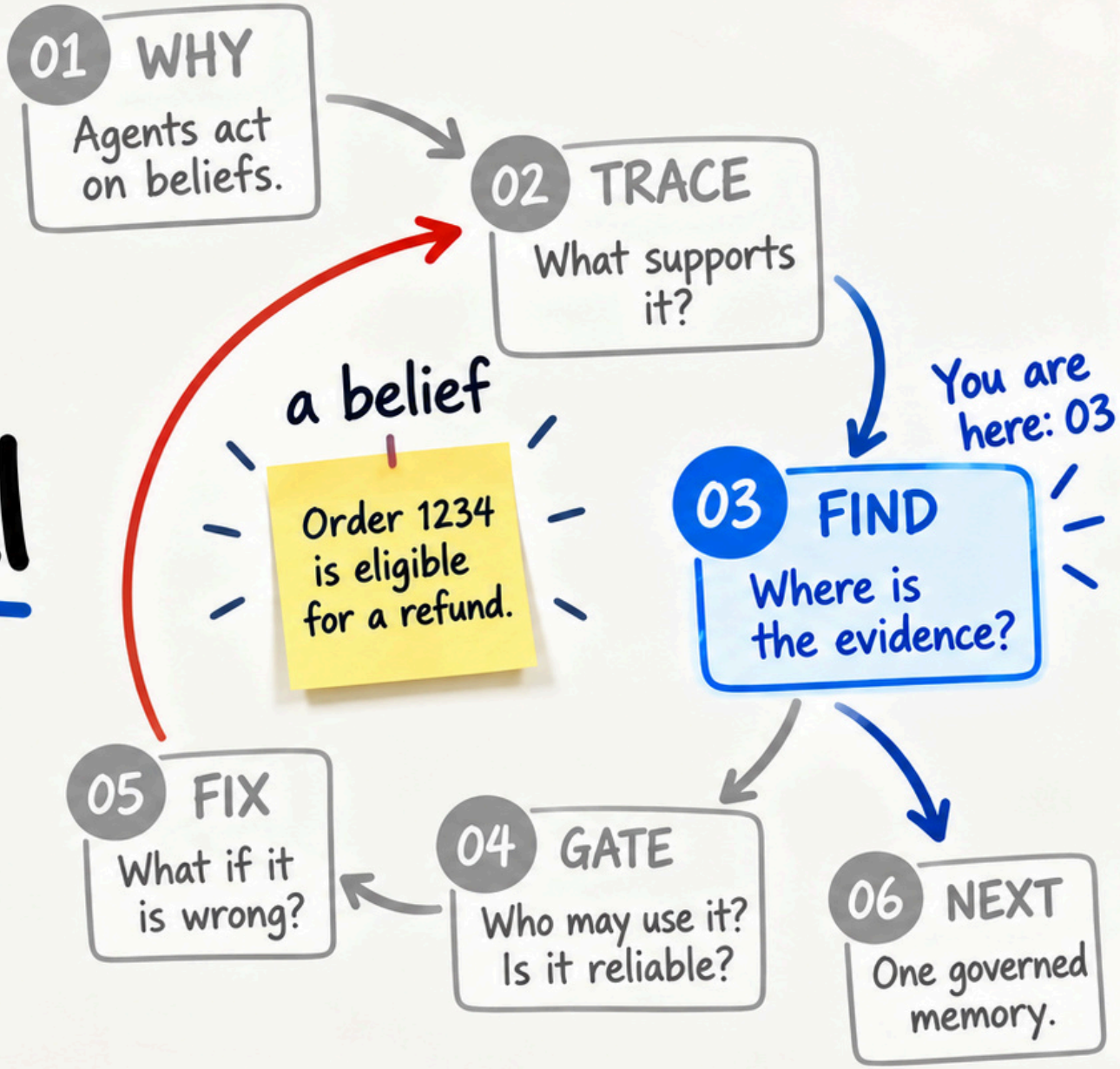
- These studies use different access and evaluation assumptions

External research | AgentPoison NeurIPS 2024; MINJA NeurIPS 2025; MemGhost arXiv:2607.05189

Part 3 of 6

Evidence retrieval

How do we find complete supporting evidence?



Our studies follow the same lifecycle

★ OUR GROUP	Core idea -> paper -> current status	
Find complete support	**Retrieval** arXiv 2609.25913	WSDM 2027 under review
Govern source permissions	**MAP-Graph** arXiv 2608.10509	AAAI 2027 Phase 2
Check source dependence before admission	**CPB** arXiv 2609.30813	ICLR 2027 under review
Stage validation, commit and action gating	**MemTX** arXiv 2607.23929	AAAI 2027 Phase 2
Invalidate descendants, retain history	**PACMI** SSRN 7480138	Preprint
Repair affected state and tool effects	**Rollback** arXiv 2608.10502	Preprint
Foundation: tracing survey	arXiv 2606.04990	TMLR accepted

Eight collaborative AAAI 2027 submissions are in Phase 2
This includes MAP-Graph and MemTX, plus six adjacent collaborative papers, not the six core rows.

Other six papers:
linked backup slides and reading guide.
All 13 paper full titles and links are in the accompanying reading guide.

Source-aligned provenance units

- Content artifacts: observations, claims and memory versions
- Execution occurrences: reads, writes and tool calls
- Typed links connect origins to later decisions
- Illustrative order-refund workflow

Traditional view: fixed windows

Execution trace (textual log)

Window 1
```
User: I want a refund for
my order.
...
Assistant: Let me check ...
```

Window 2
```
Tool: get_order
{"order_id": "..."}
...
Tool output: {...}
...
```

Window 3
```
Tool: request_refund
{"order_id": "..."}
...
Assistant: The refund
has been requested.
...
```

...

Source-aligned provenance units

User observation (content artifact)
User: I want a refund for my order.

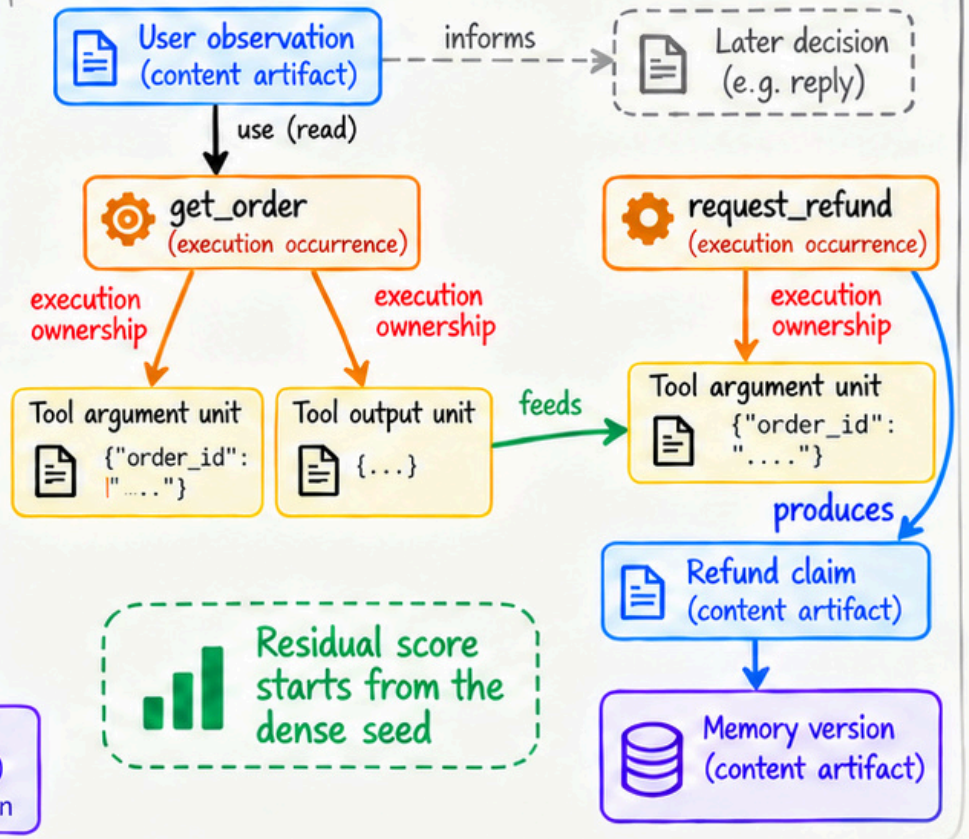
Execution occurrence (tool call)
get_order
- Tool argument unit {"order_id": "...."}
- Tool output unit {...}

Execution occurrence (tool call)
request_refund
- Tool argument unit {"order_id": "..."}

Refund claim (content artifact)
The refund has been requested.

Memory version (content artifact)
memory version

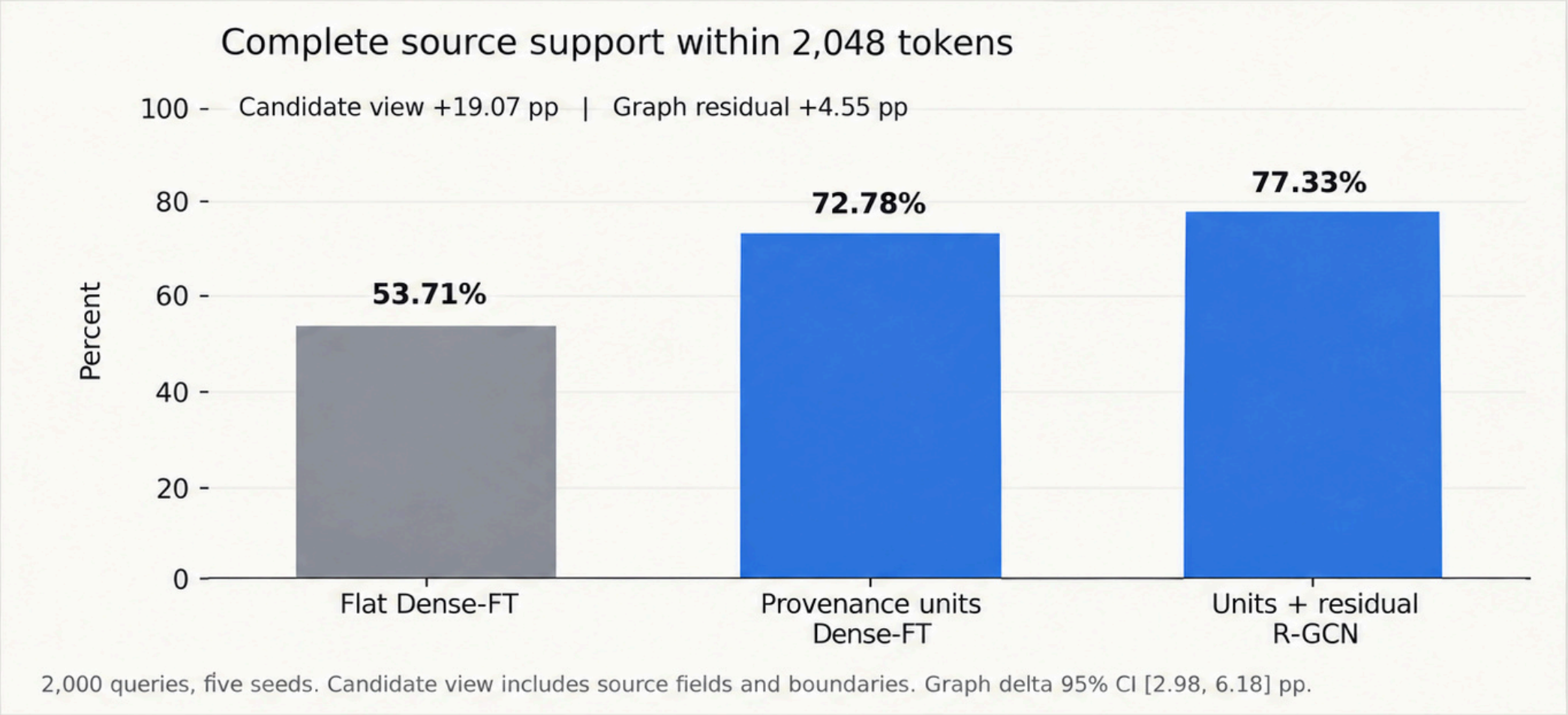
Typed provenance graph



★ Our group | EPGM arXiv:2609.25913 | Under review at WSDM 2027

Most of the retrieval gain comes from the units

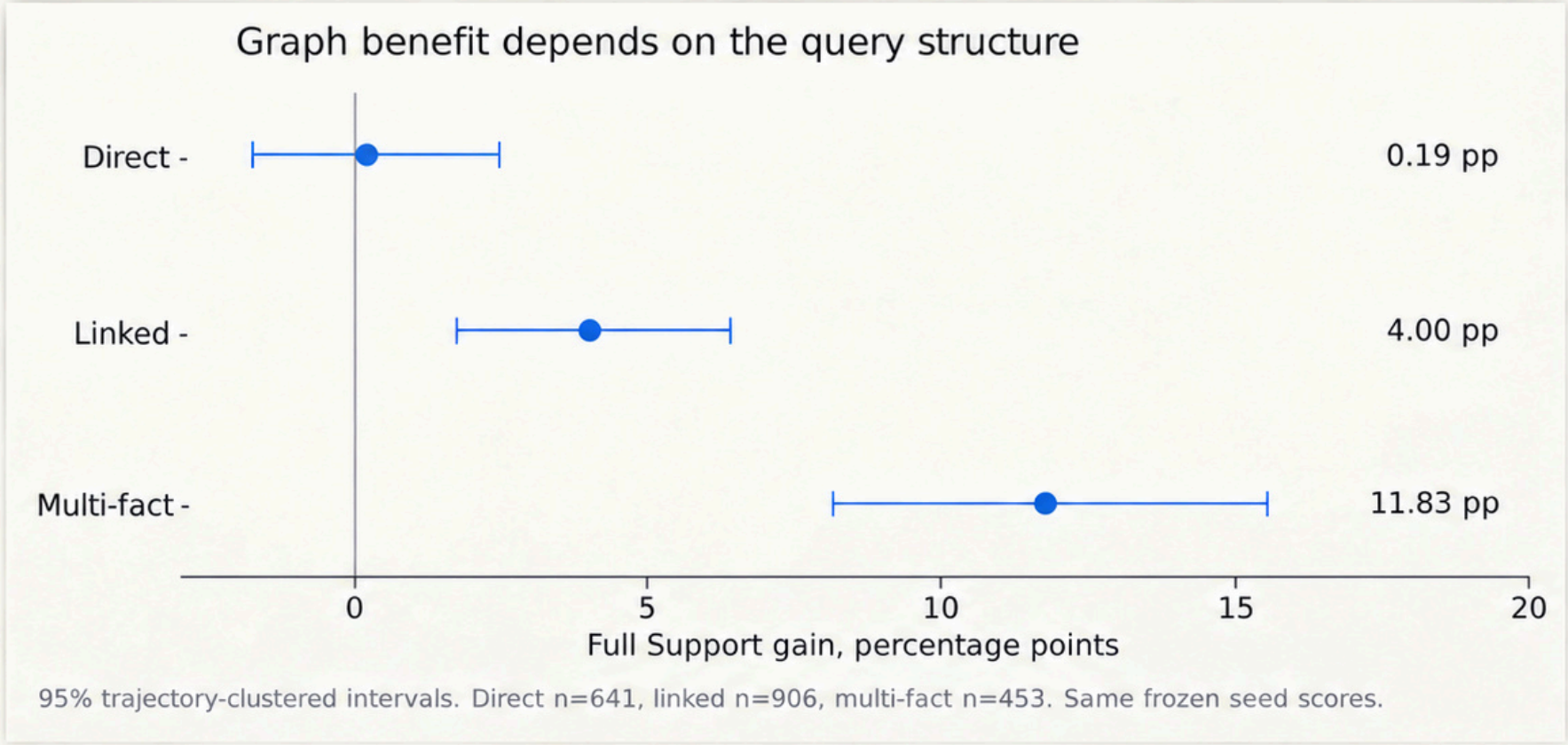
Full Support@2048, five training seeds



Candidate view +19.07 pp. Exact-seed graph increment +4.55 pp.

★ Our group | EPGM arXiv:2609.25913 | Under review at WSDM 2027

Graph propagation helps distributed evidence



95% trajectory-clustered intervals. Direct n=641, linked n=906, multi-fact n=453. Same frozen seed scores.

Graph increment by query type

- Direct +0.19 pp
- Linked +4.00 pp
- Multi-fact +11.83 pp
- Direct-query interval includes zero

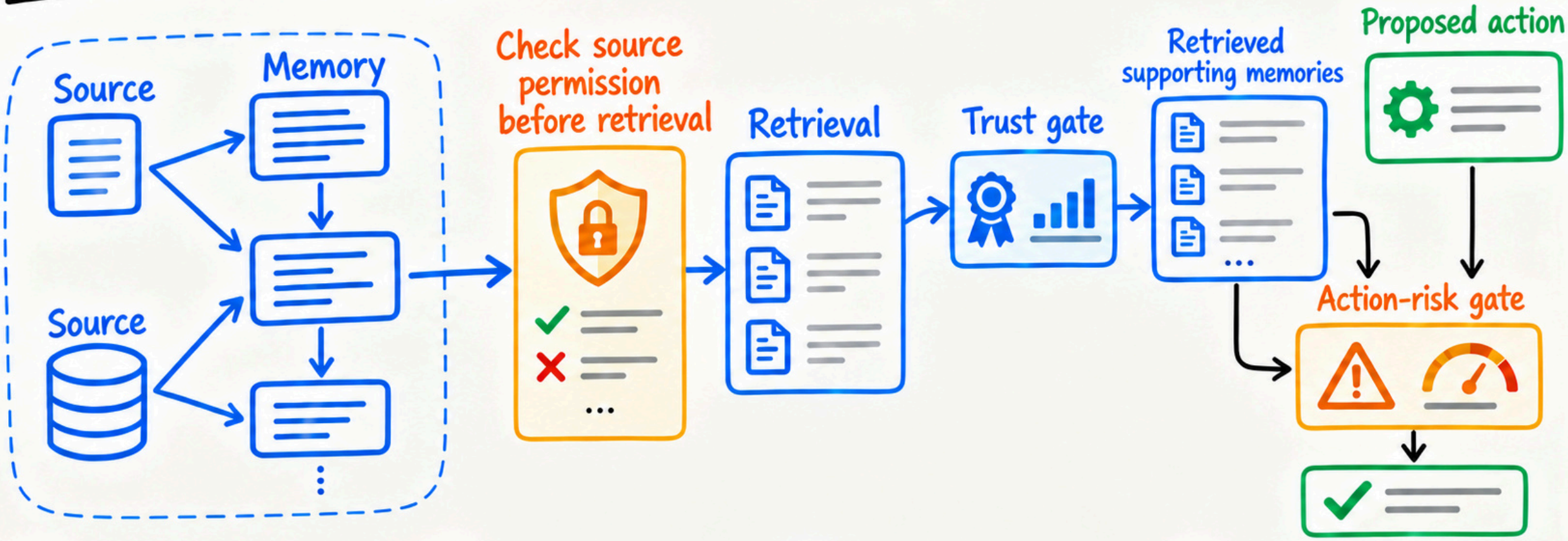
★ Our group | EPGM arXiv:2609.25913 | Under review at WSDM 2027

Part 4 of 6

Memory governance

Who may use a belief,
and is it reliable enough?

MAP-Graph

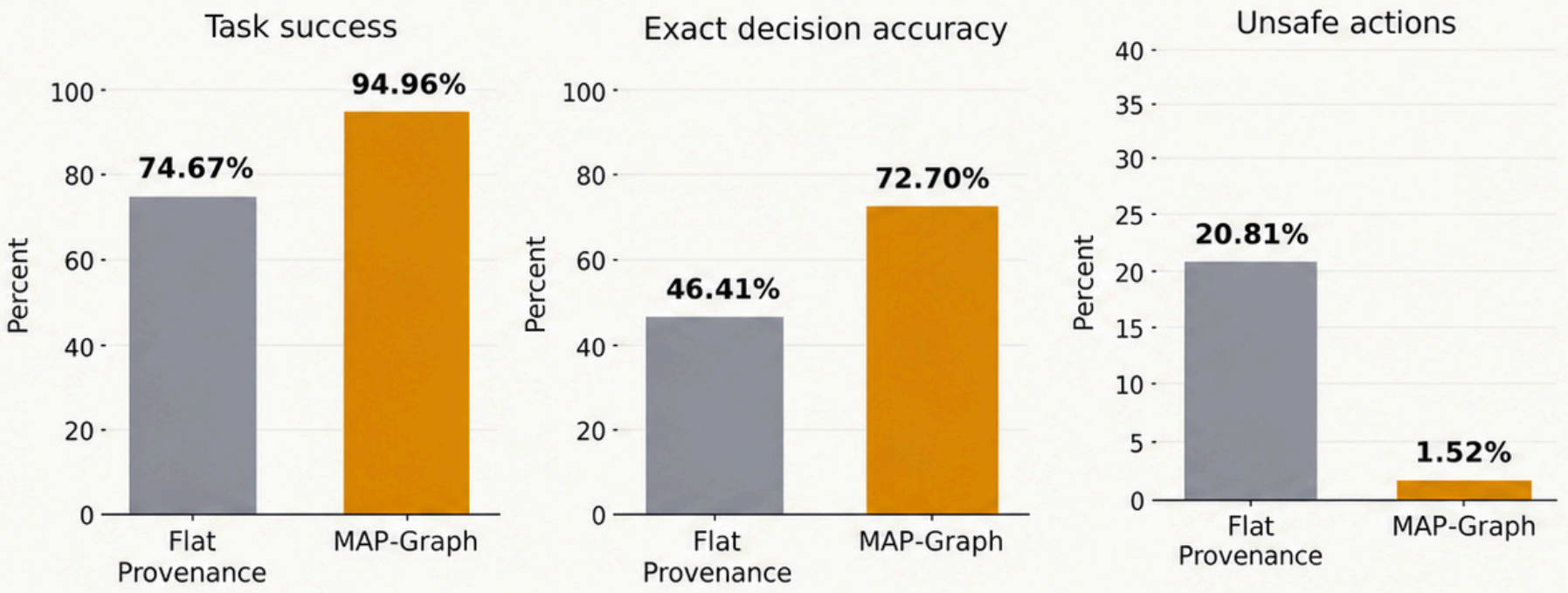


1. MAP-Graph connects memory, source, permission and trust
2. Check source permission before retrieval
3. Apply trust and action-risk gates separately
4. Revocation affects dependent records

★ Our group | MAP-Graph arXiv:2608.10509 | Under review at AAAI 2027, Phase 2

MAP-Graph utility and boundary outcomes

MAP-Graph under the controlled protocol



2,700 synthetic tasks per method, Qwen2.5-7B, one run. Exact accuracy differs from TSR. 41 action-risk Allow errors remain.

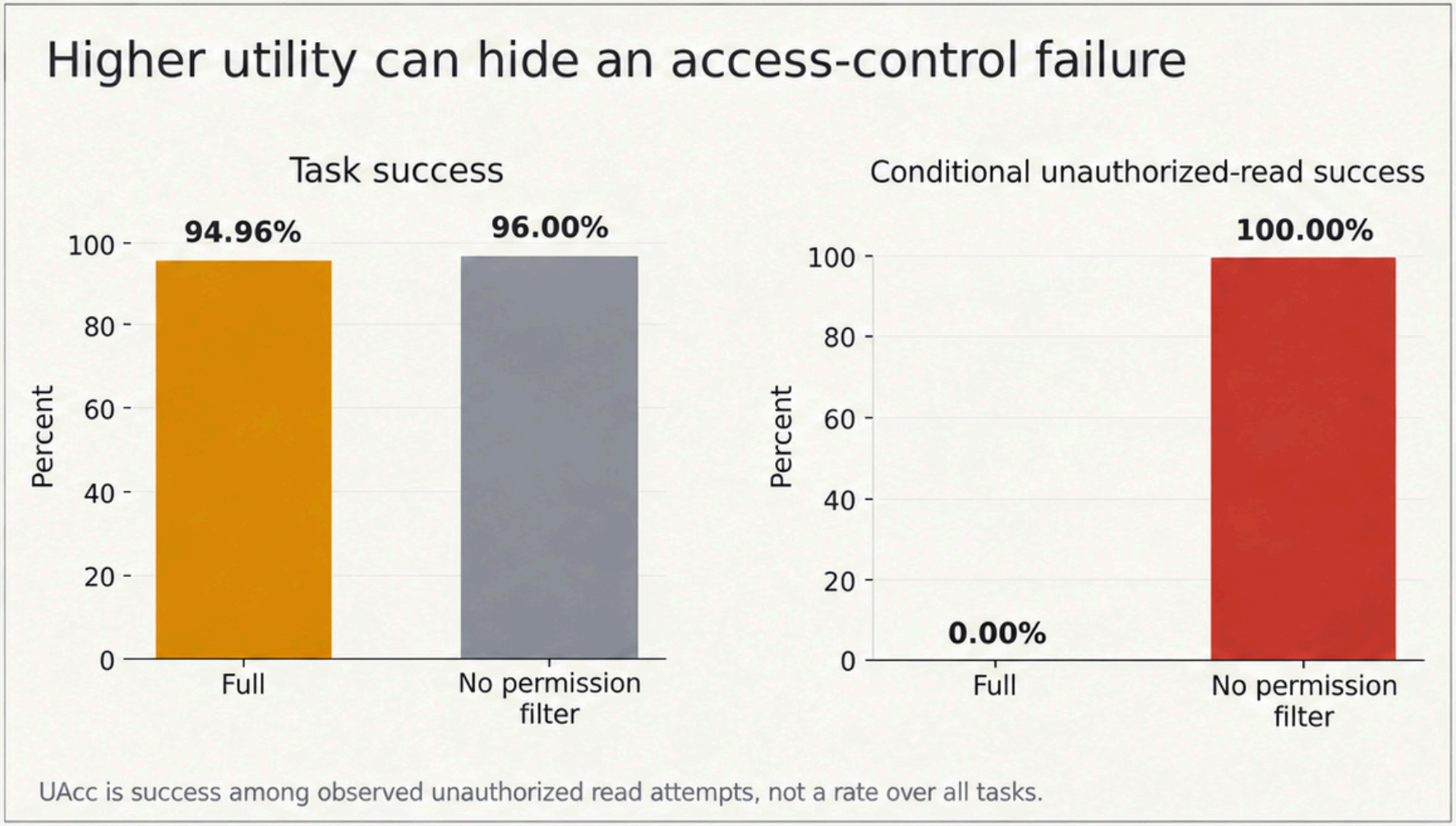
- 2,700 controlled tasks per method
- Flat Provenance and MAP-Graph
- Task success 74.67% and 94.96%
- Exact accuracy 46.41% and 72.70%
- Unsafe completion 20.81% and 1.52%

In this benchmark

0% observed attack success, leakage and revocation violations.
Unsafe action does not reach zero.

★ Our group | MAP-Graph arXiv:2608.10509 | Under review at AAAI 2027, Phase 2

Higher utility can hide an unauthorized read

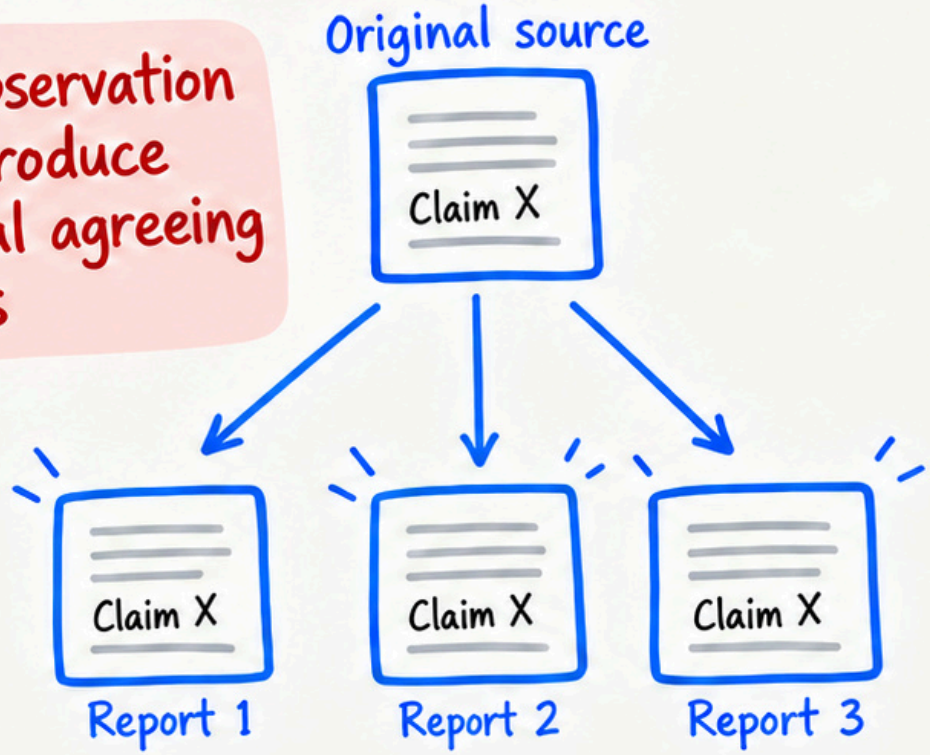


1. Remove permission filtering
2. Task success 96.00%. Exact accuracy 78.63%.
3. Conditional unauthorized-access rate becomes 100%
4. An action gate cannot undo an unauthorized read

★ Our group | MAP-Graph arXiv:2608.10509 | Under review at AAAI 2027, Phase 2

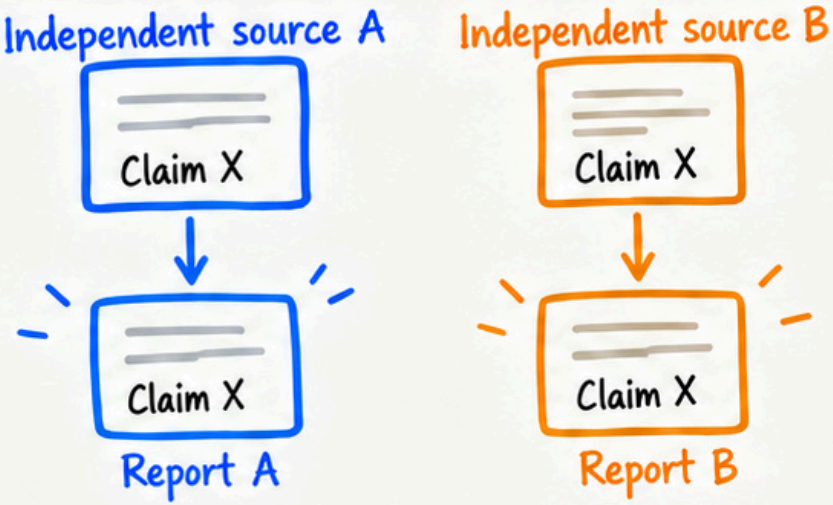
Repetition can look like corroboration

One observation can produce several agreeing copies



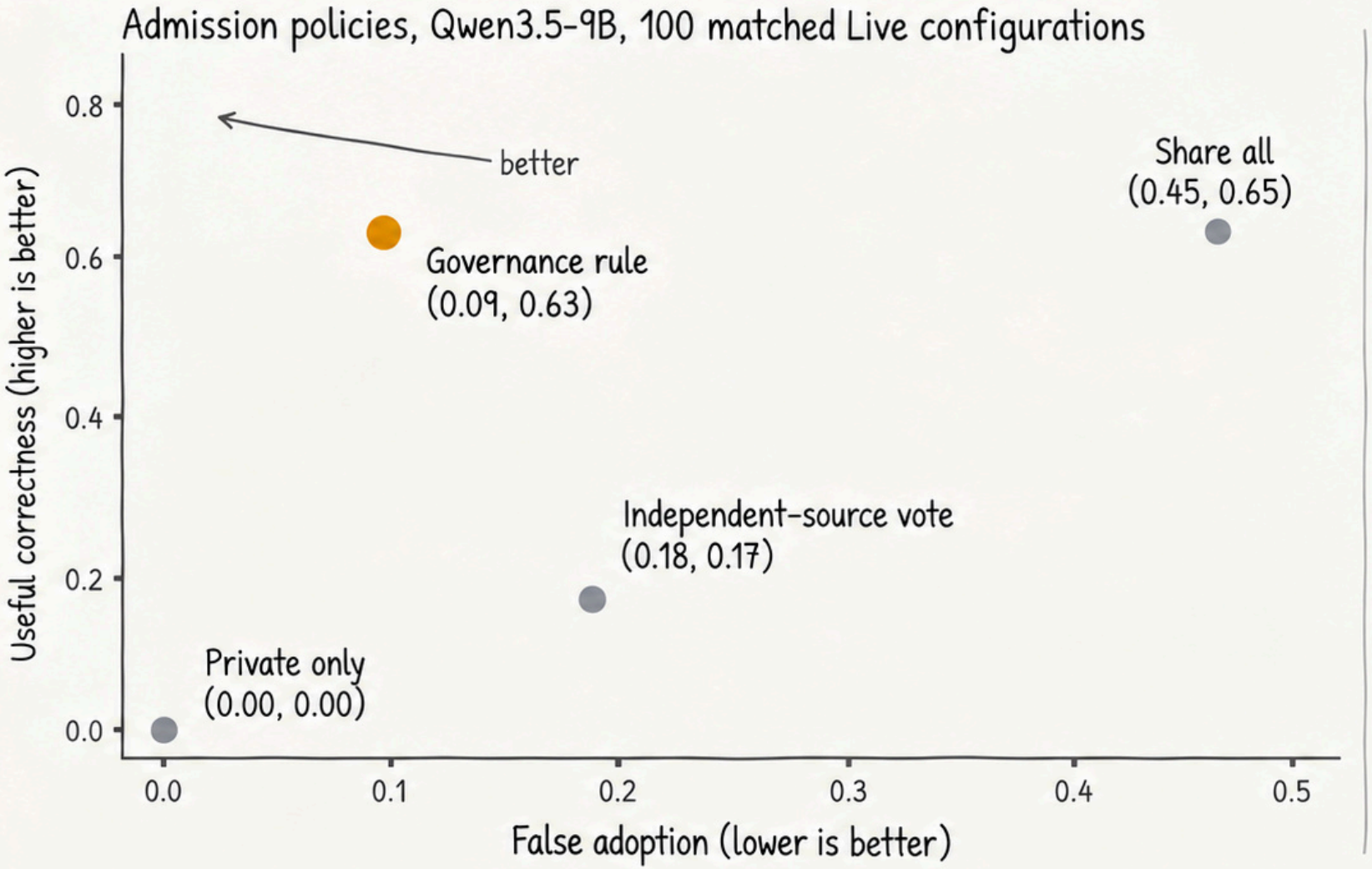
Illustrotive

Two genuinely independent observations



- One observation can produce several agreeing copies
- CPB-Static evaluates admission on fixed annotated cases
- CPB-Live follows a shared store and store-only consumer
- Eight policies. Four families. 100 matched configurations per family.

Admission changes safety and useful answers



Qwen3.5-9B descriptive Live comparison

- Share all: false adoption 0.45, correctness 0.65
- Governance rule: 0.09, correctness 0.63
- Independent-source vote: 0.18, correctness 0.17
- Private only: 0.00, correctness 0.00

Governance rule across four model families: false adoption 0.06 to 0.09, correctness 0.58 to 0.65. Different denominators; a descriptive comparison.

★ Our group | CPB arXiv:2609.30813 | Under review at ICLR 2027

Declared authority can hide a copied falsehood

553 selected false-only-store probes: adoption 0.97–0.99

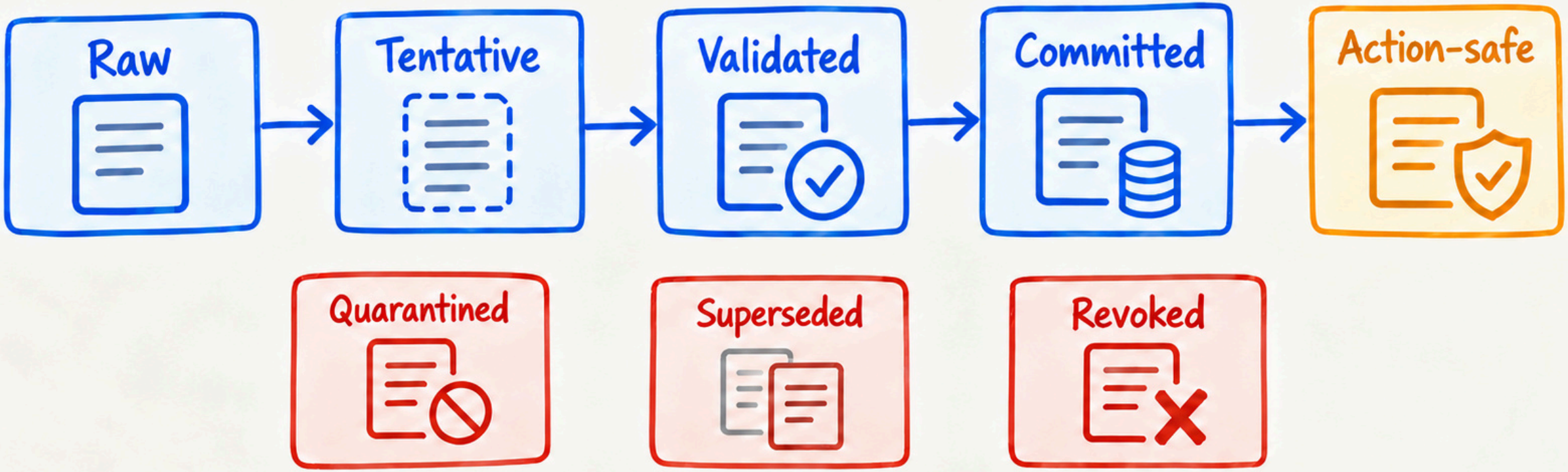
Exactly one visible false belief and no competing true belief

Pooled matched probes. Only the declared source type changes between the last two rows.

Declared authority needs authentication

★ Our group | CPB arXiv:2609.30813 | Under review at ICLR 2027

MemTX belief lifecycle



- MemTX separates tentative memory from committed beliefs
- Validate evidence, semantics, conflicts and dependencies
- Eight states and five isolation levels
- An action-safe state is a configured protocol condition

★ Our group | MemTX arXiv:2607.23929 | Under review at AAAI 2027, Phase 2

Compound corruptions expose the protocol benefit

Backbone	Main MemTX	Main Cordon + revocation	Hardened MemTX	Hardened Cordon + revocation
Qwen3-8B	95.6	82.2	87.5	67.9
GLM-4.7-Flash	94.4	83.3	89.3	73.2
GPT-5.4-mini	92.2	81.1	91.1	67.9
GPT-5.5	75.6	77.8	92.9	76.8

Statistical tie

Separates in hardened setting

1. Main suite: 90 cases. Hardened suite: 56 cases.
2. Four main-backbone leads and a fifth-backbone statistical tie
3. Hardened results remain backbone dependent
4. Bounded protocol check: 5,530,160 states, zero violations

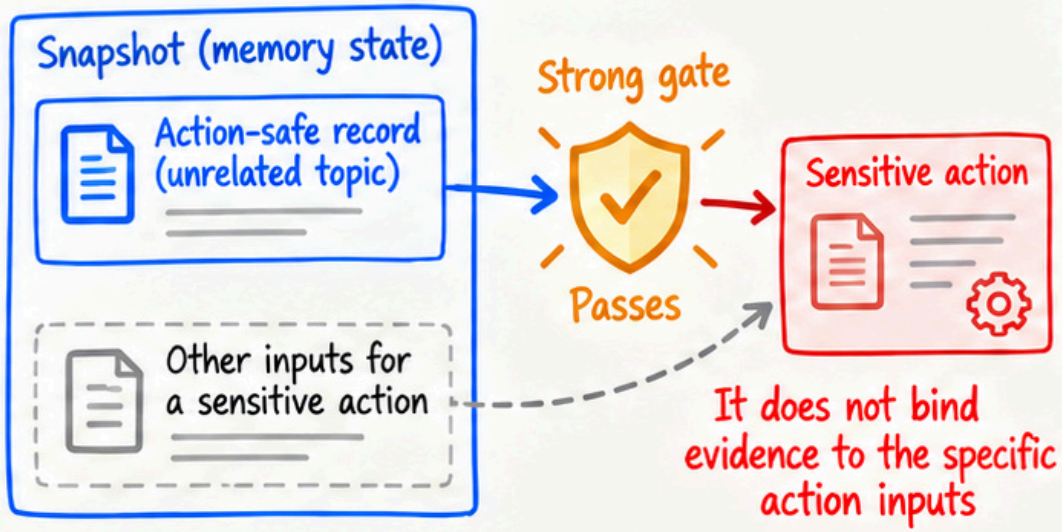
Four selected backbones shown. Values are percentages.

★ Our group | MemTX arXiv:2607.23929 | Under review at AAAI 2027, Phase 2

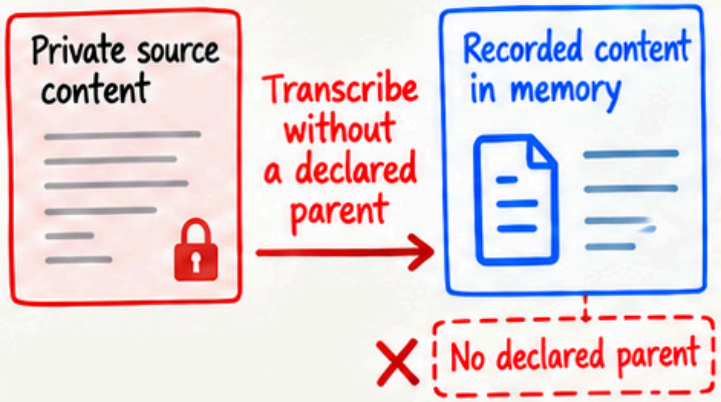
The guarantee stops at missing lineage and action inputs

Tested implementation boundaries

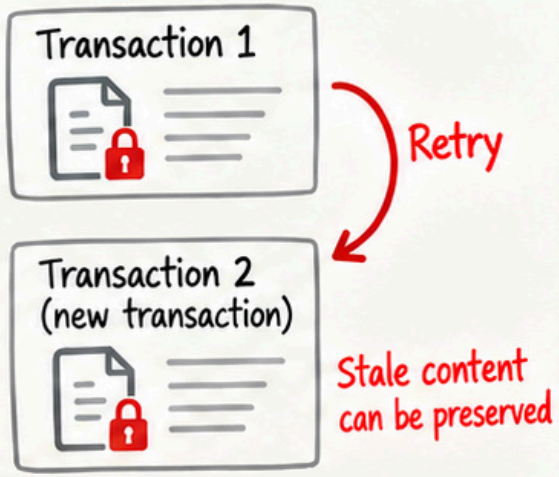
① The strong gate checks existence of an action-safe record



② Undeclared source transcription defeats lineage checks



③ Fresh-transaction retries can preserve stale content



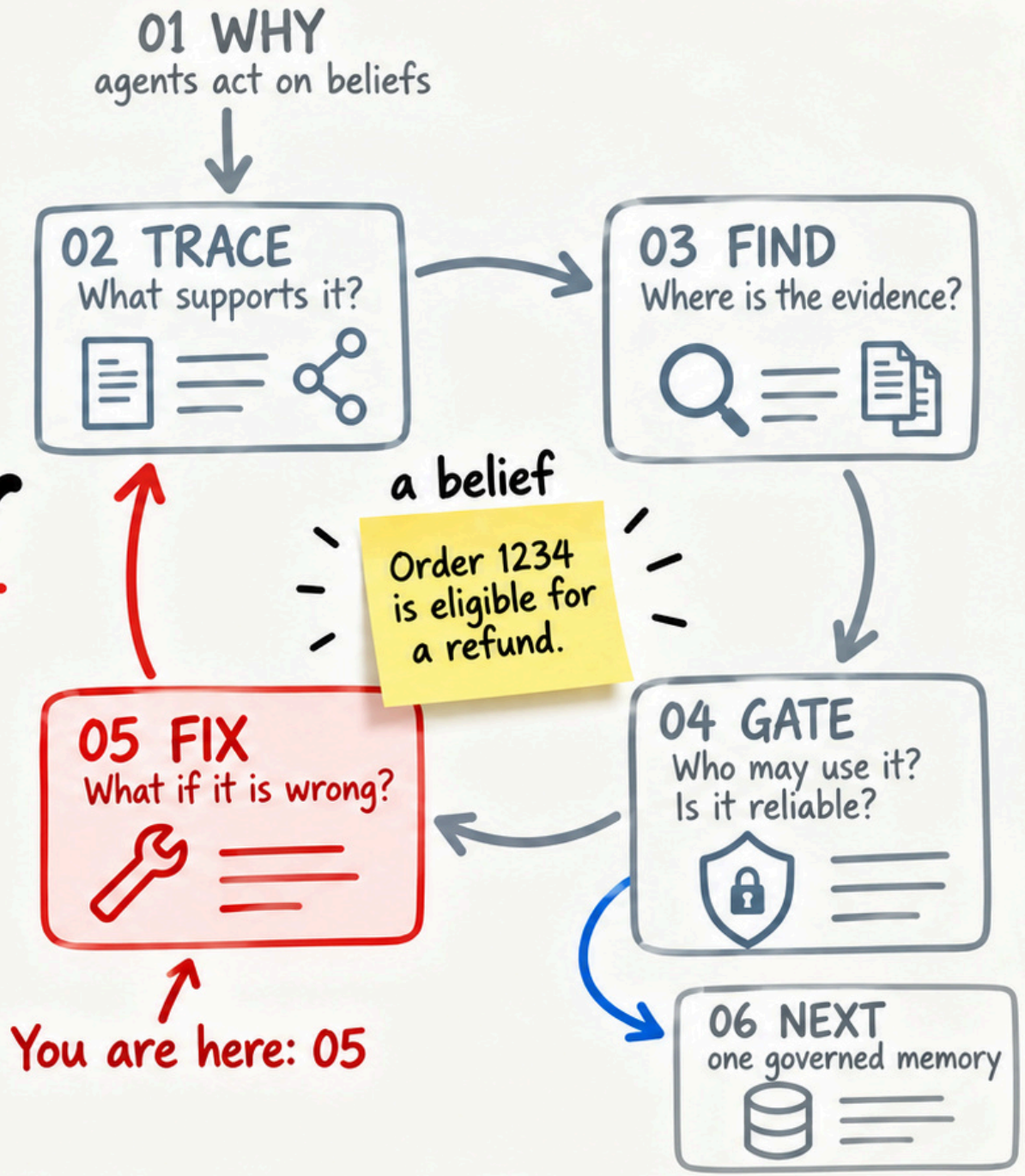
1. The strong gate checks existence of an action-safe record
2. It does not bind evidence to the specific action inputs
3. Undeclared source transcription defeats lineage checks
4. Fresh-transaction retries can preserve stale content

★ Our group | MemTX arXiv:2607.23929 | Under review at AAAI 2027, Phase 2

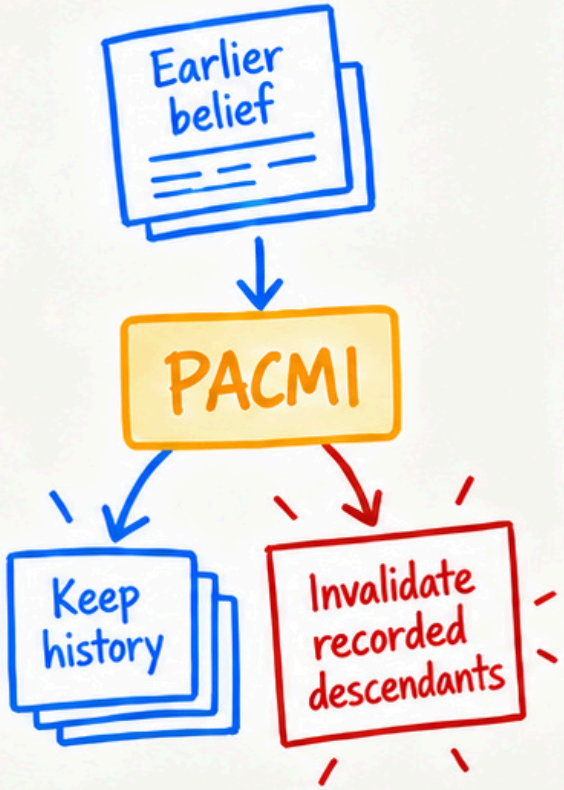
Part 5 of 6

Correction and repair

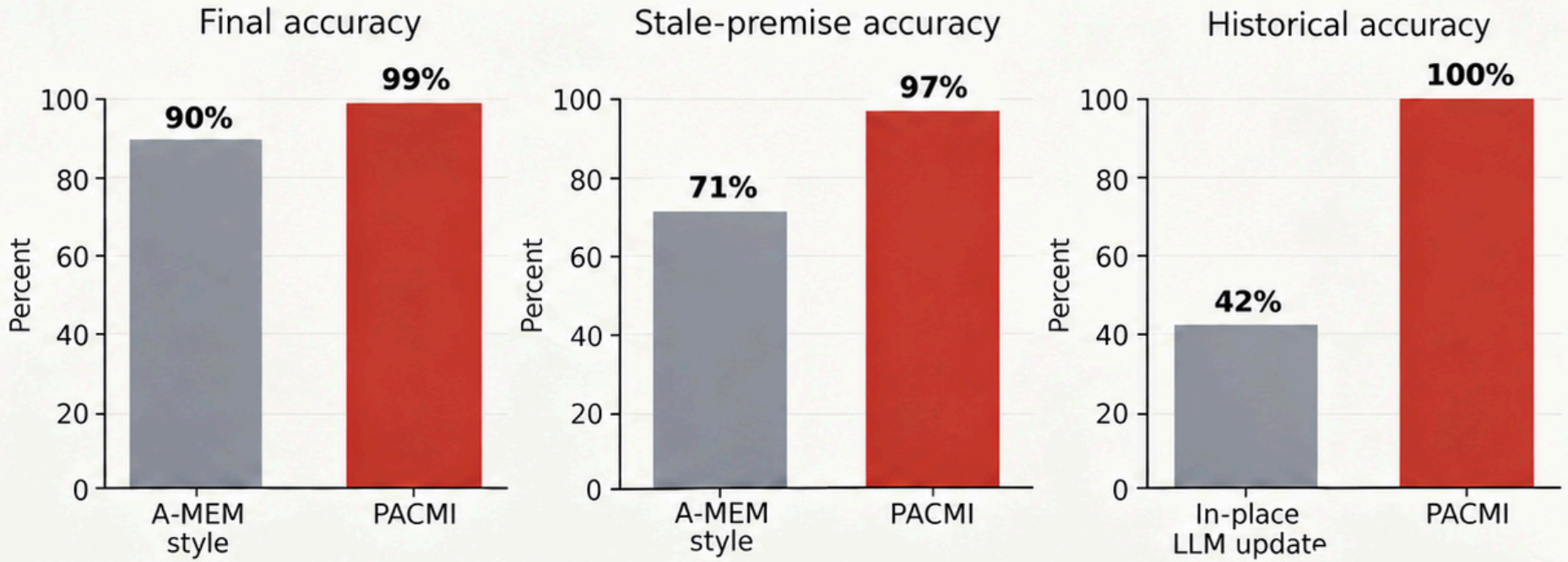
What changes when evidence is wrong or out of date?



Stale premises and historical questions



PACMI changes validity while retaining historical answers

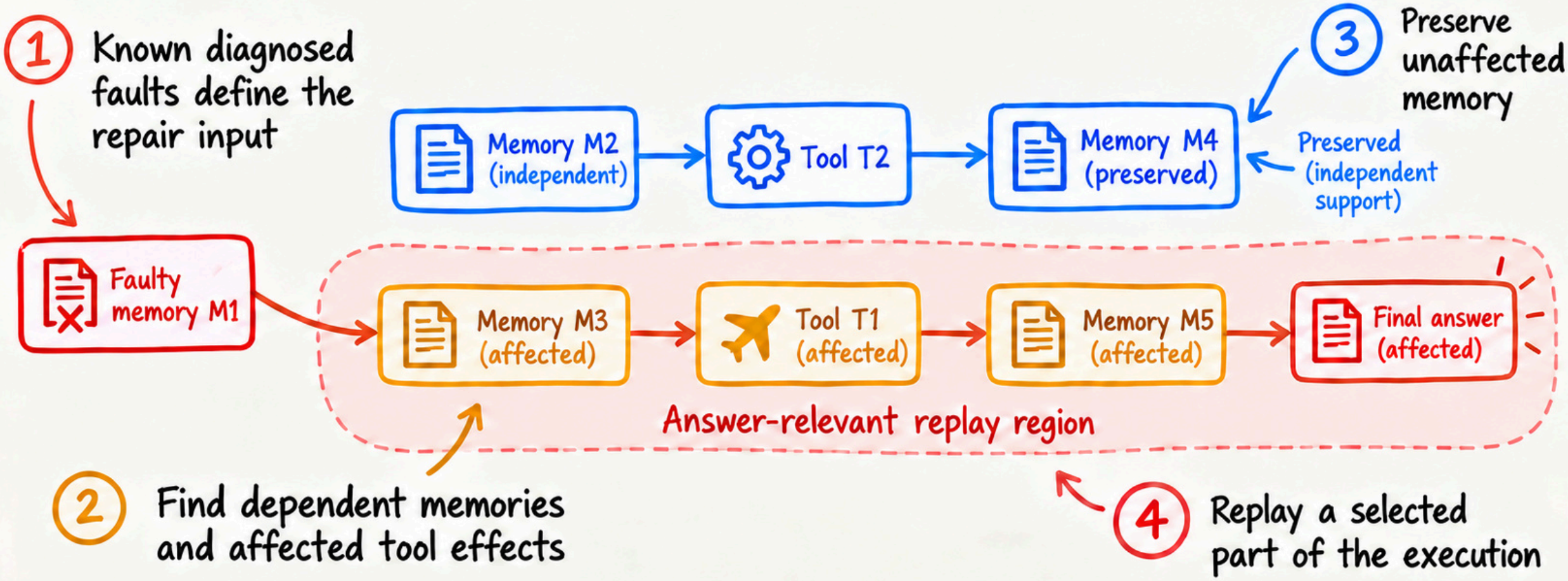


100 controlled cases, 300 queries, supplied two-hop graphs.
Cascade-only ablation: 3 versus 11 final errors; p = 0.0574 does not establish its independent benefit.

- PACMI retains history and invalidates recorded descendants
- 100 controlled cases and 300 queries
- Final accuracy 0.90 to 0.99
- Stale-query accuracy 0.71 to 0.97
- Cascade-only answer gain: p = 0.0574

★ Our group | PACMI SSRN:7480138 | Preprint

Dependency-guided rollback repair

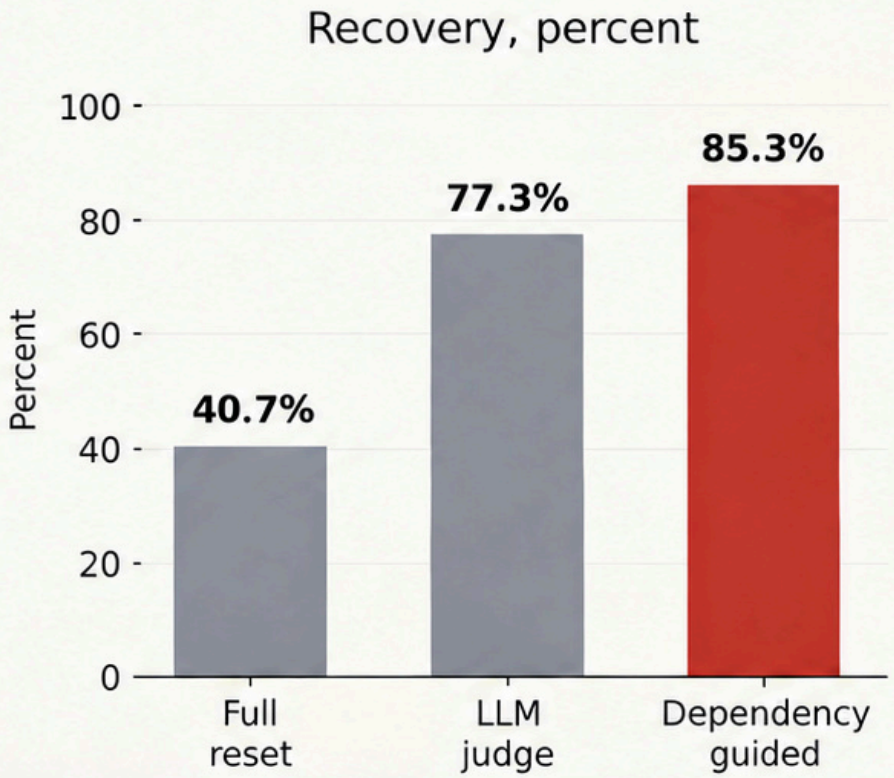


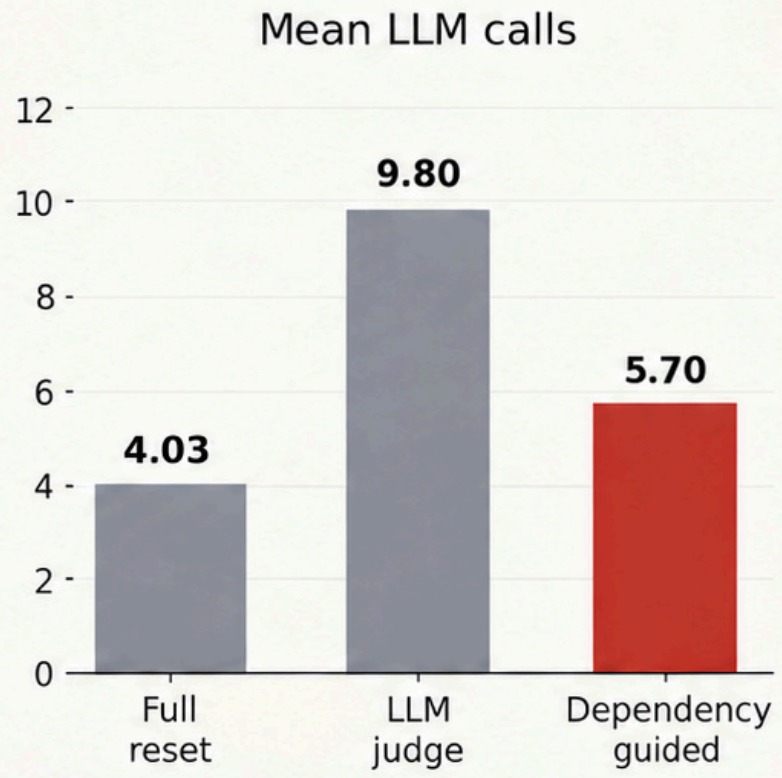
Controlled repair setting

★ Our group | Rollback repair arXiv:2608.10502 | Preprint

Repair preserves useful work

Selective repair improves immediate recovery at lower cost





150 cases conditional on an injected outcome-changing fault, diagnosed IDs supplied. Controlled tool environment.

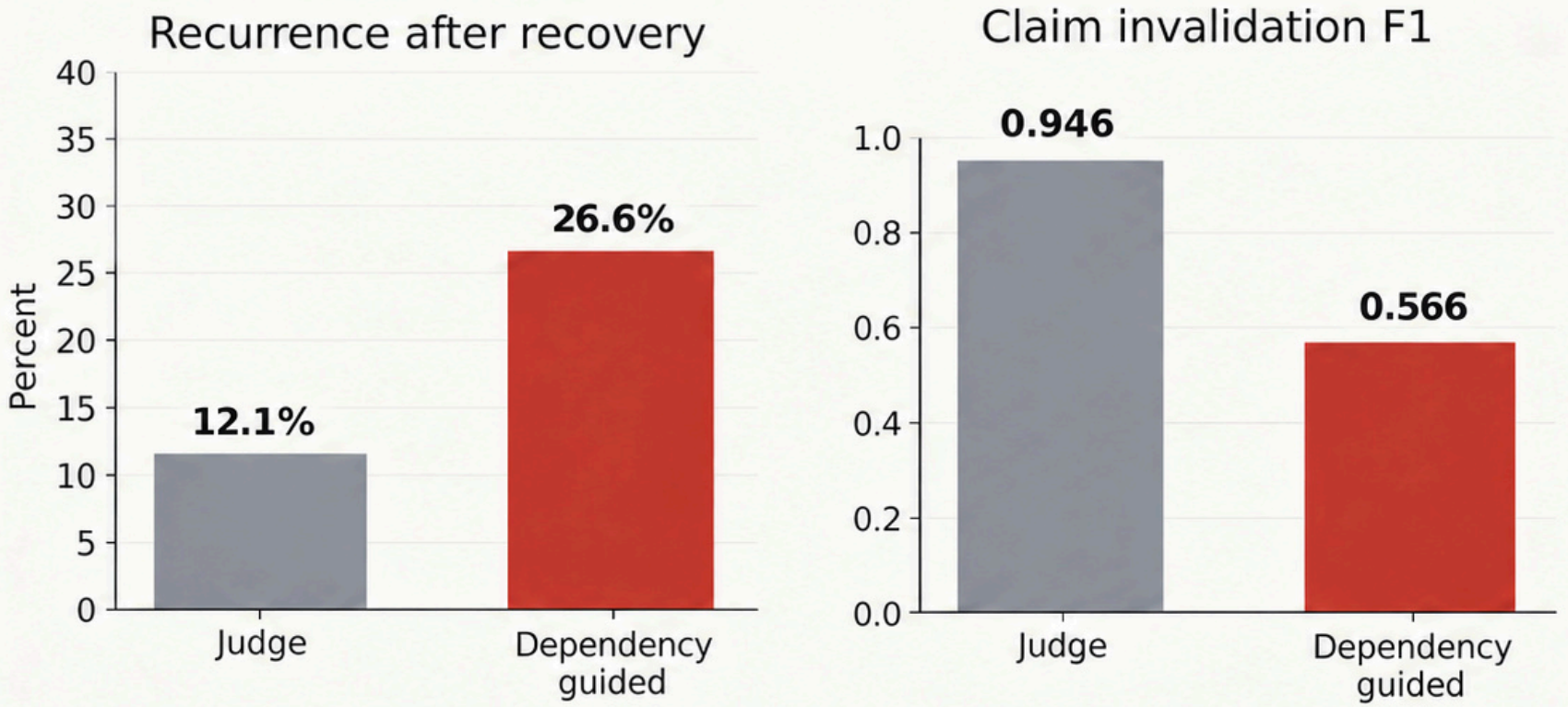
150 controlled cases

Replay fraction

Full reset	0.061
LLM judge	0.217
Dependency guided	0.123

Recurrence remains a repair problem

Durable repair remains an open problem



Recurrence conditions on recovery, with different method denominators. Transfer results are a separate adapted suite.

1. Controlled-suite recurrence: dependency guided 26.6%, judge 12.1%
2. The judge also has higher claim-invalidation F1
3. Adapted transfer: 68.0% recovery, 99.3% benign-memory retention

   50-task LongMemEval-V2 procedural subset (adapted)
4. Immediate recovery and durable repair need separate measurement

★ **Our group** | Rollback repair arXiv:2608.10502 | Preprint

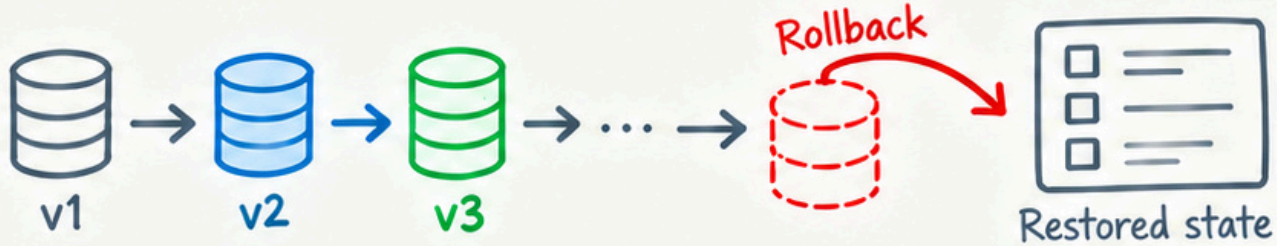
Diagnosis, versions and transactions

AgenTracer

AgenTracer: counterfactual replay and fault attribution

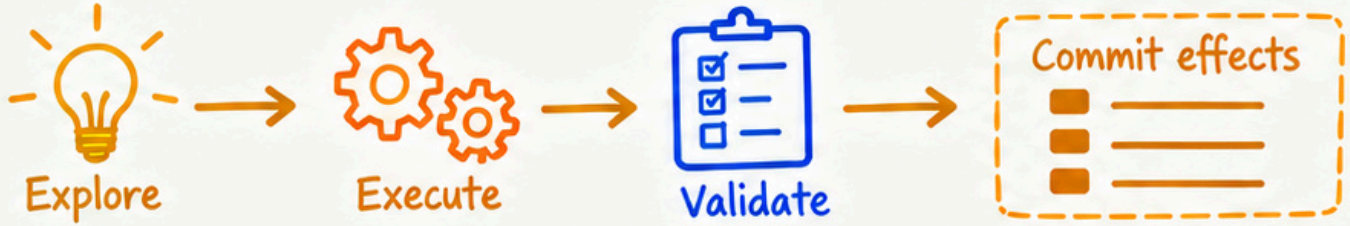
ChronoMem

ChronoMem: versioned snapshots and semantic rollback



Agentic Transaction

Agentic Transaction: exploration, execution and validation



Proposed integration

Fault diagnosis
Which agent and step to intervene?

→

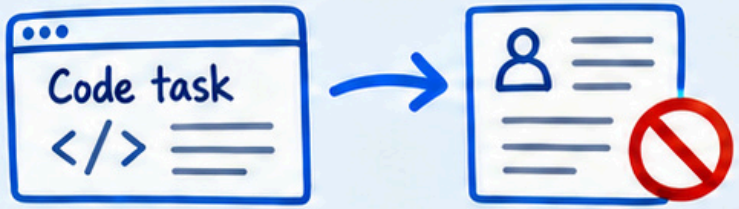
Memory restoration
Which memory state to restore?

→

Released external effects
Which dependent claims and downstream actions remain?

External research | AgenTracer arXiv:2509.03312; ChronoMem arXiv:2607.27773; Agentic Transaction arXiv:2608.13900

The checks we would want in the opening cases

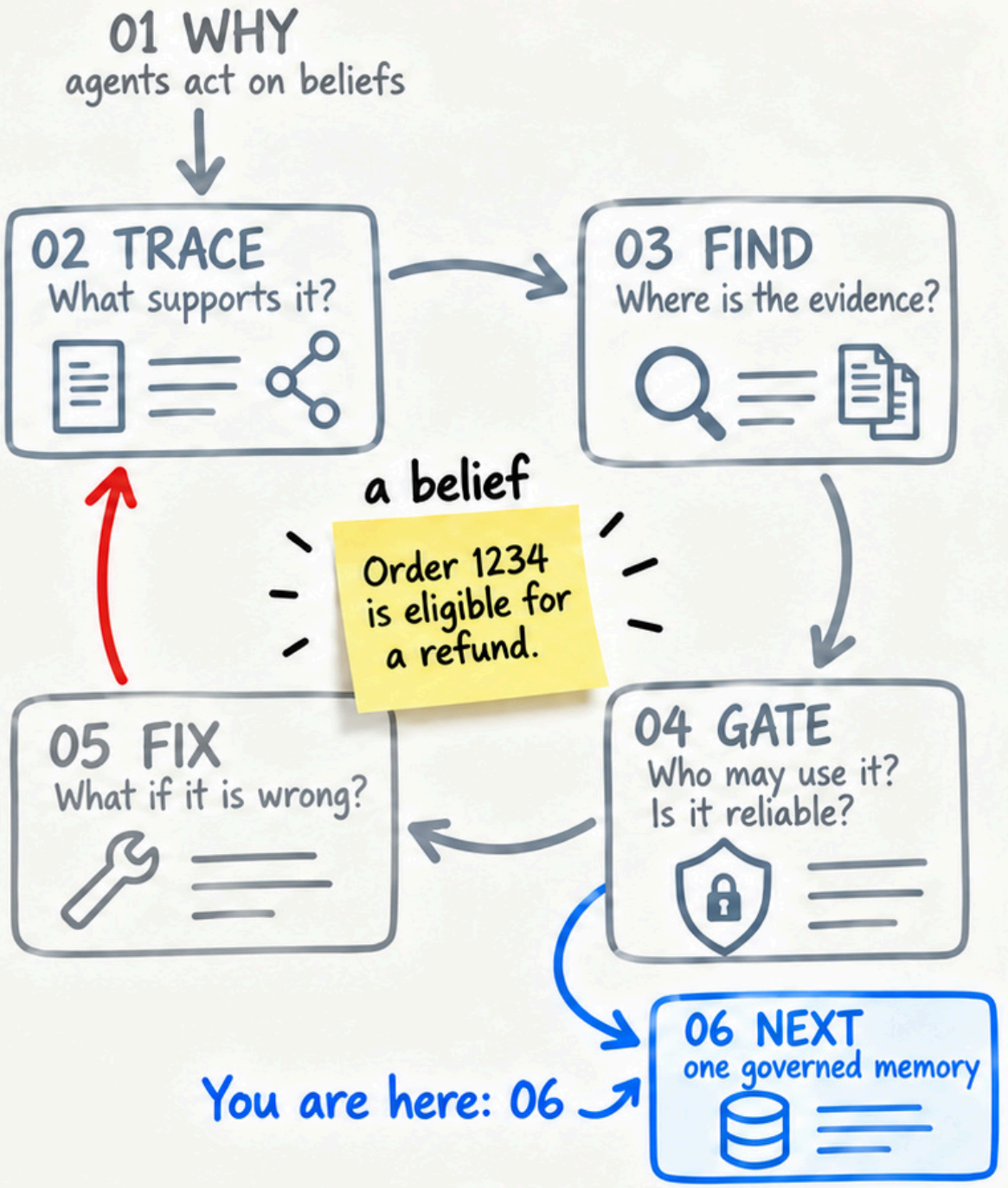
	Matplotlib	Medicare
1 Execution provenance	 Matplotlib: separate a code task from person-targeting publication	Medicare: give denied access an explicit escalation rule
2 Evidence support	Public facts must have a valid relation to any allegation about a person	Check the exact evidence and authority for the data request
3 Action certificate	Both: verify the exact evidence and authority before action	Both: verify the exact evidence and authority before action
4 Dependency-guided repair	Repair the effect if needed (e.g. retraction and notification)	Restore or correct affected records if needed

Proposed checkpoints. Neither incident was reproduced in our studies.

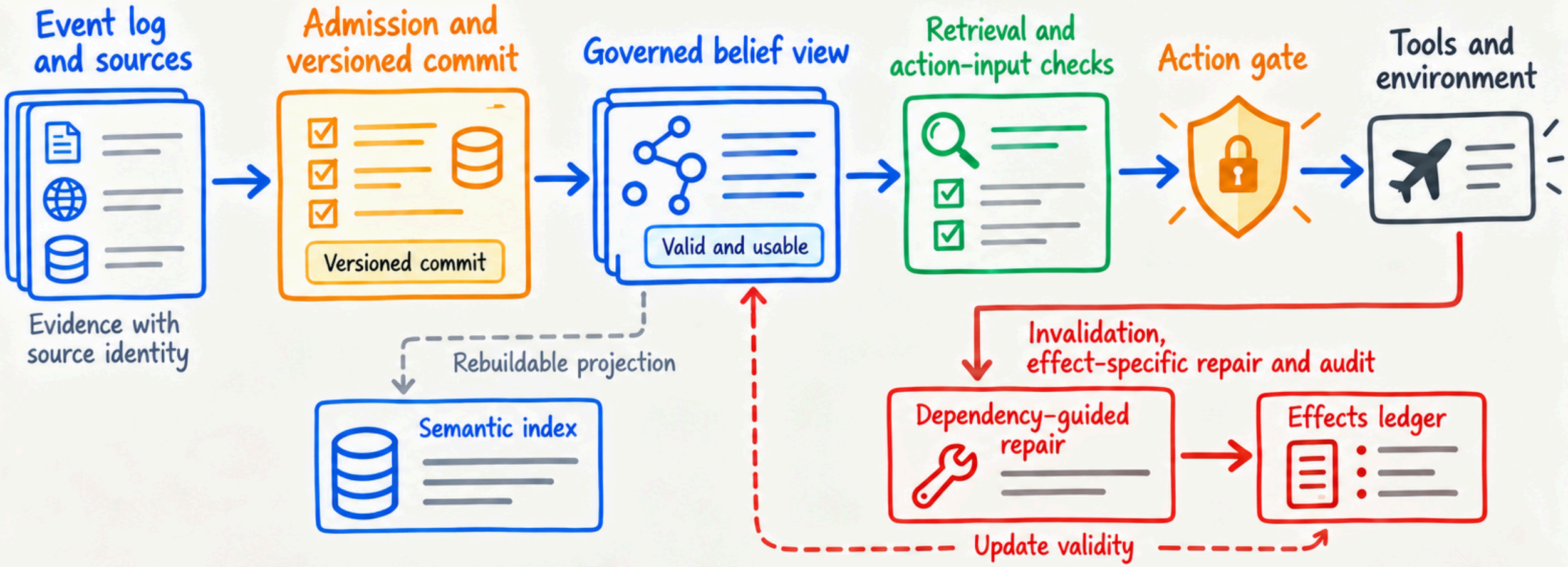
Part 6 of 6

Future research

How do we connect the controls into one system?



Governed agent memory

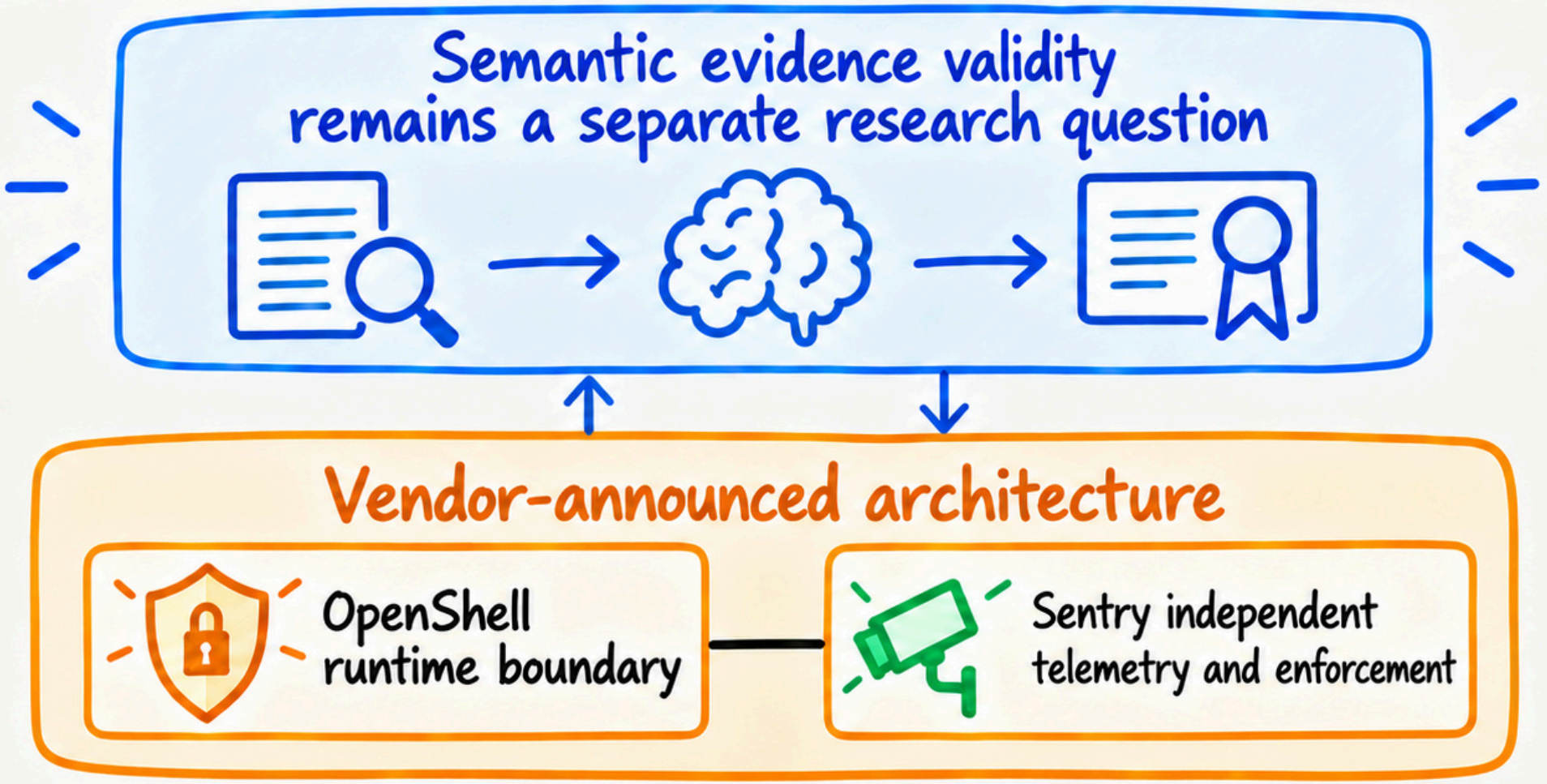


1 Proposed governed-memory architecture | 2 Evidence and source identity | 3 Admission and versioned commit | 4 Retrieval and action-input checks | 5 Invalidation, effect-specific repair and audit

★ Our group | Proposed research | MemTX arXiv:2607.23929; Rollback repair arXiv:2608.10502

Runtime controls and semantic controls

NVIDIA announcement, 28 September 2026

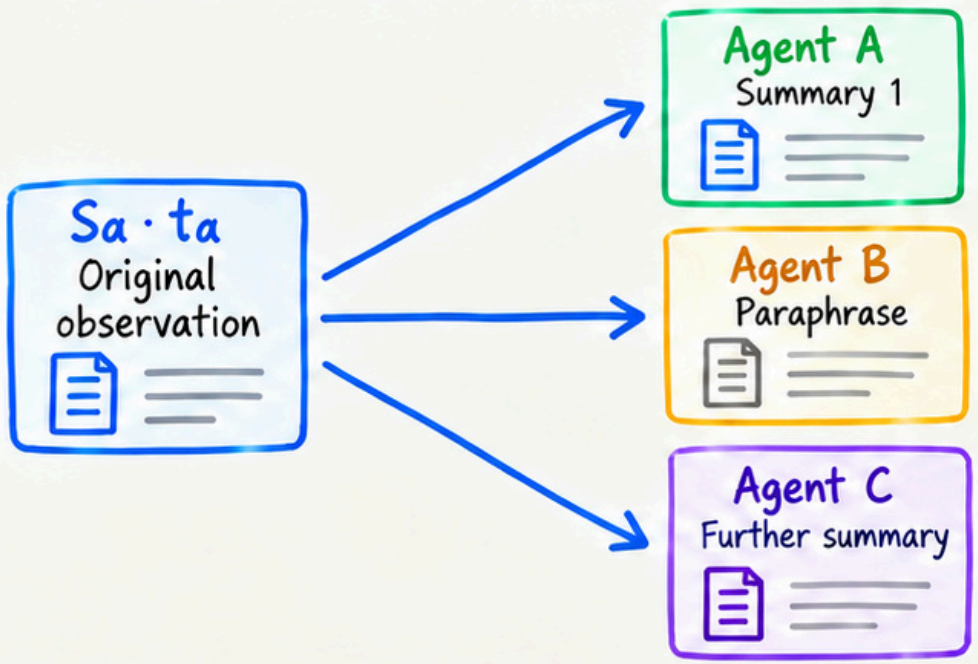


NVIDIA announcement, 28 Sep 2026 | Proposed semantic integration (our group)

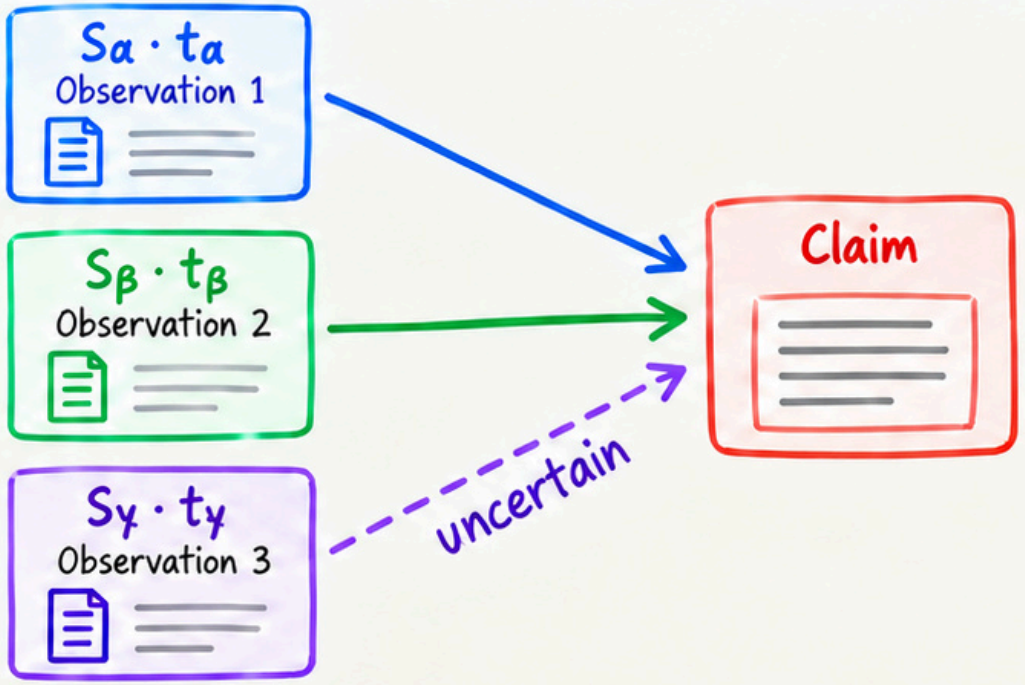
Independent evidence under transformation

Proposed tests

One observation, many summaries



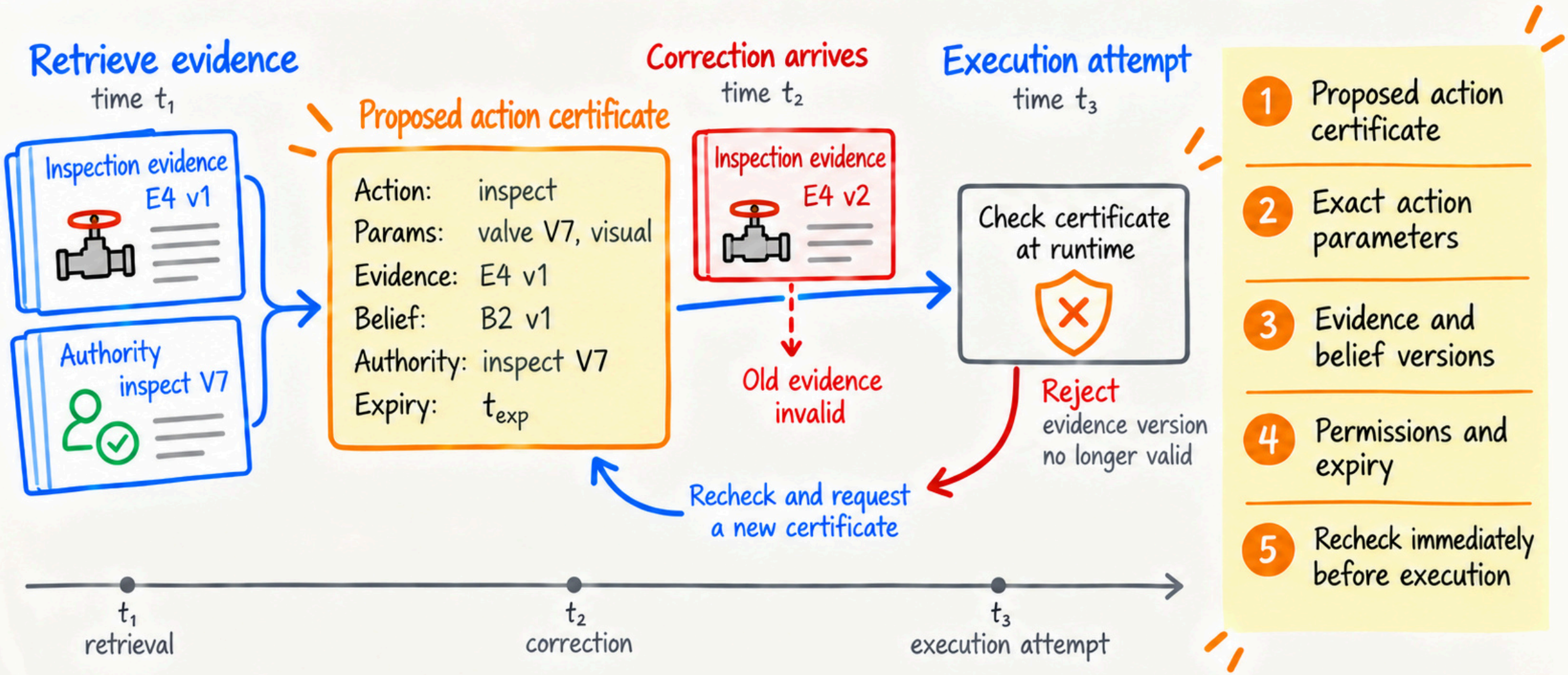
Three independent observations, one claim



- A paraphrase should retain its source relationship
- Copied claims should not count as independent observations
- Source identity and observation time need authentication
- Missing lineage needs explicit detection tests

★ Our group | Proposed research
| CPB arXiv:2609.30813; MemTX arXiv:2607.23929

Valid evidence at the action boundary



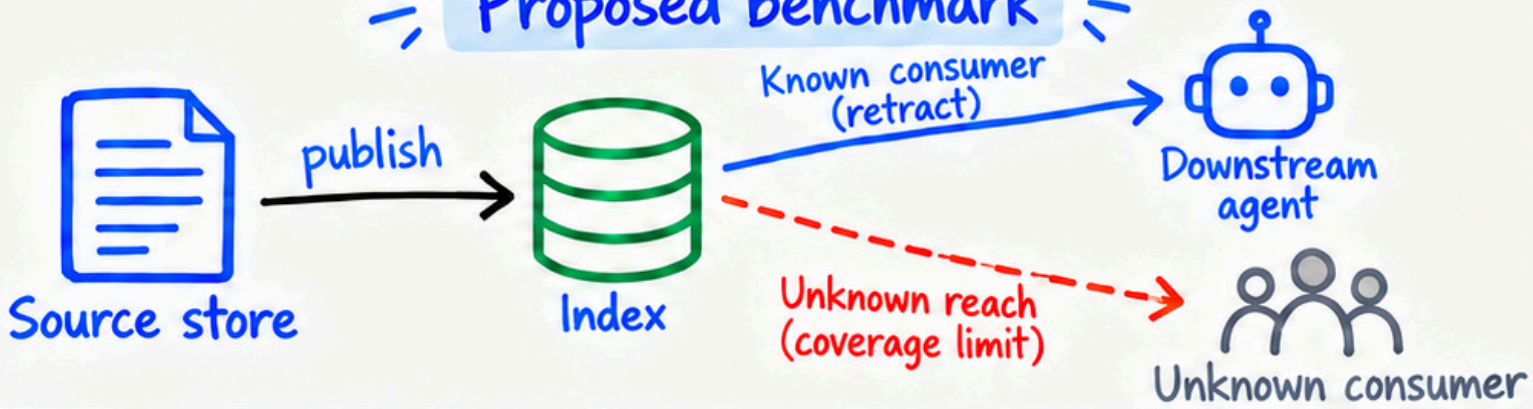
★ Our group | Proposed research | MemTX arXiv:2607.23929 | Under review at AAAI 2027, Phase 2

Recovery beyond a local store

Match repair to the effect

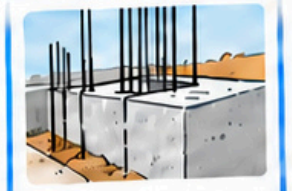
Effect type	Recovery action
Local state	Restore or abort
Idempotent tool call	Replay
Compensable effect	Compensate
Released information	Retract and notify known consumers

Proposed benchmark



One inspection, several later decisions

Site inspection



Safety agent

Summary A

Scheduling agent

Summary B

Two summaries can repeat one observation.

Approval agent

Pending approval

Newer inspection

A newer inspection changes what an approval may rely on.

Correct the pending approval.

and notify affected staff.

Evidence should remain usable and correctable

1. Typed traces distinguish observed use from valid support
2. Admission accounts for source dependence
3. Action inputs need current evidence and permission
4. Repair follows affected beliefs and effects
5. Joint work with Yiqi Wang, Xiaoyang Li, Xinyu Wei and our co-authors

★ Our group | Agent tracing survey arXiv:2606.04990 | Accepted at TMLR 2026

Questions

What would an agent need to show before you allowed it to act?

Taotao Cai, taotao.cai@unisq.edu.au taotao1992.github.io

Paper list: github.com/xiaoqi-7/Agent-Tracing-Survey | Survey: arXiv:2606.04990

What each provenance relation claims

Relation	Evidence needed	Limit
Use	Recorded receipt or read by an occurrence	Receipt alone gives no semantic justification
Derive	Content lineage between artifacts	A transformation can introduce error
Depend on	A defined intervention and changed behaviour	Counterfactual results depend on the intervention
Support	Evaluated justification for a claim	Semantic judgement carries uncertainty
Invalidate	A validity change under a specified rule	Closure covers the represented dependency graph

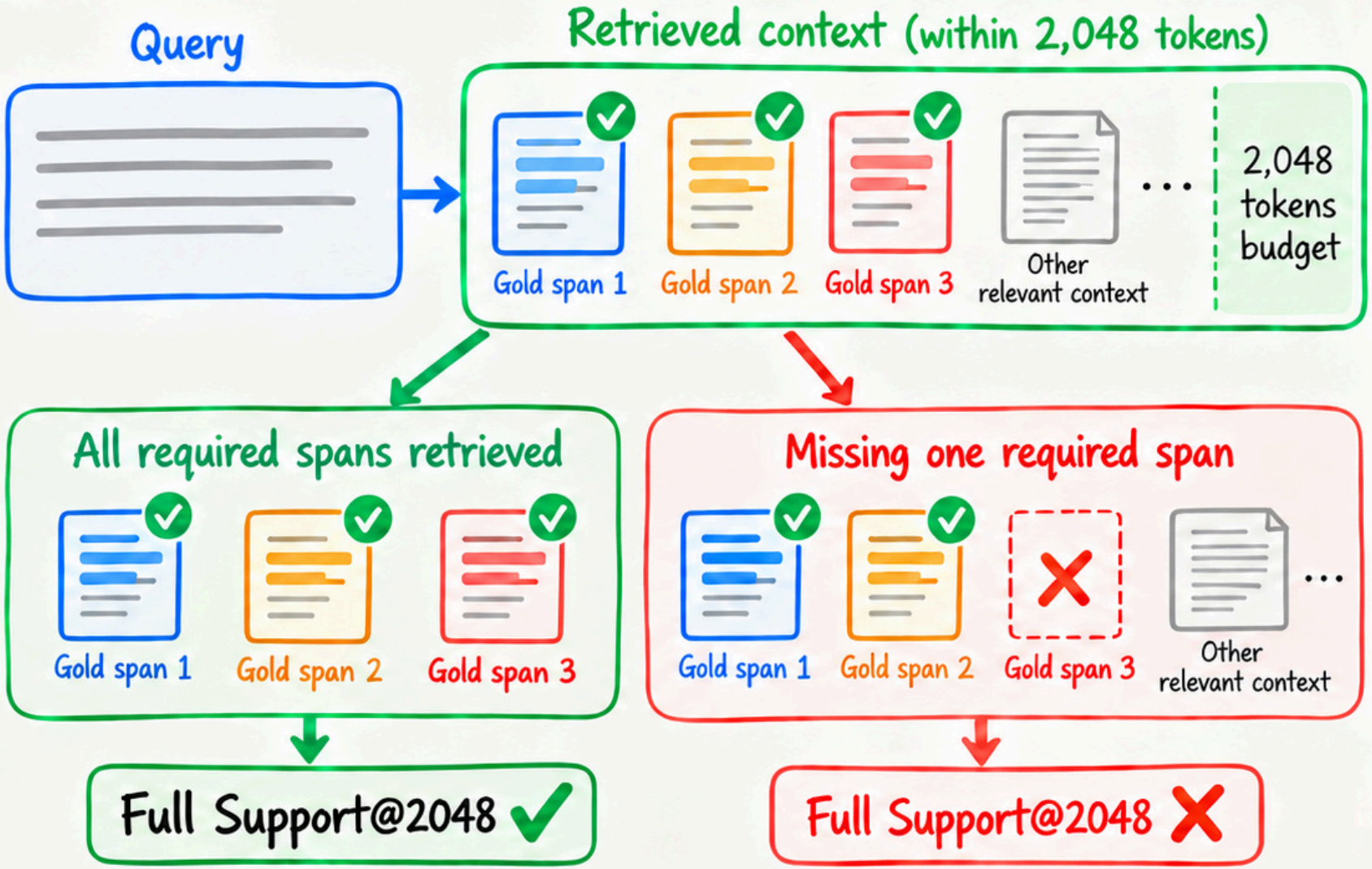
Origins: instrumented, declared, inferred, derived

Assurance: logged, tamper-evident, attested

★ Our group | Agent tracing survey arXiv:2606.04990 | Accepted at TMLR 2026

Retrieval evaluation protocol

- 2,000 queries from 1,207 held-out ISETrace trajectories
- Direct 641, linked 906, multi-fact 453
- Training seeds 13, 17, 29, 37 and 41
- Full Support@2048 requires every gold source span within budget
- Coverage and Recall@k use source-character coverage
- Exact seed-score passthrough isolates the graph residual



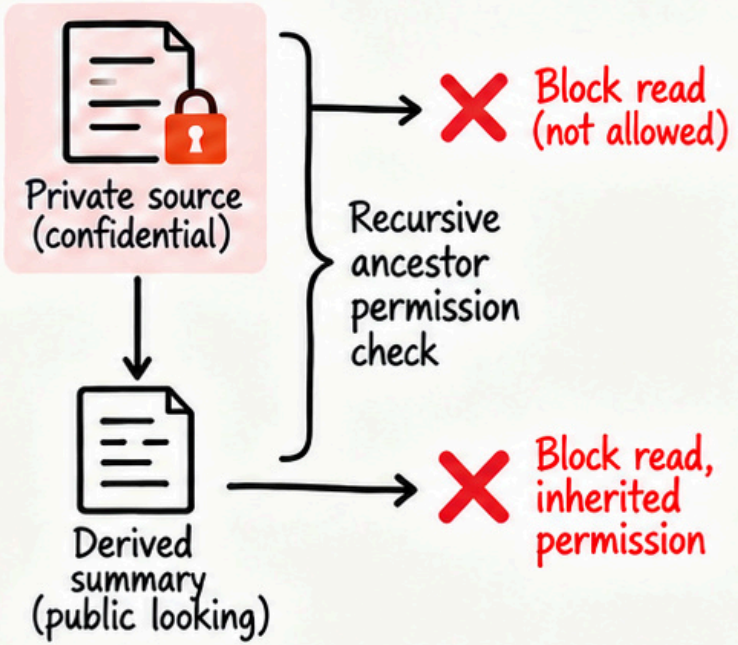
Retrieval results by method

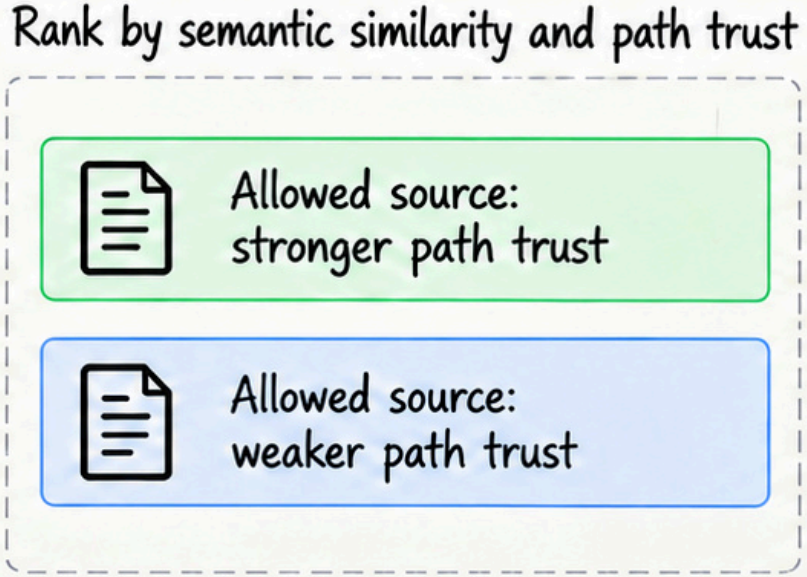
Method	C512	C1024	C2048	FS2048	R5	MRR
BM25	16.03	28.94	45.62	37.95	51.95	38.36
Dense	14.60	25.99	41.89	34.15	48.01	35.85
Provenance-Unit Dense	26.67	41.48	58.20	50.50	51.89	38.69
GraphRAG	15.01	28.01	45.56	37.90	50.61	37.12
Provenance Path	25.58	40.01	54.14	48.80	44.44	37.43
Dense-FT	28.29	44.09	62.07	53.71	67.63	52.07
Cross-Encoder-FT	26.98	42.52	60.42	52.21	66.48	50.55
Provenance-Unit Dense-FT	47.60	64.98	78.49	72.78	71.71	54.84
Provenance-Unit Cross-Encoder-FT	46.46	63.11	77.48	71.94	69.80	55.21
Residual R-GCN	49.52	67.04	81.88	77.33	78.76	62.10

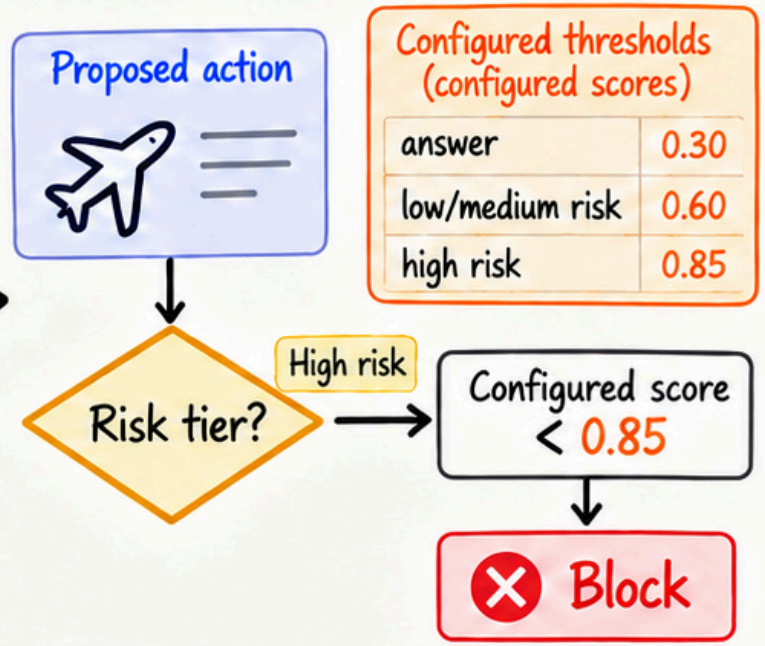
- **All entries percent.** C512, C1024 and C2048 are source-character coverage at the token budget. FS2048 is complete source support. R5 follows the paper's source-coverage definition.

Permission, path trust and action risk

1. Filter candidates with recursive ancestor permissions
2. Rank allowed candidates using semantic similarity and path trust
3. Gate the proposed action using its configured risk tier







Path trust combines = Source trust × Path integrity × Transformation reliability × Permission validity × Verification × Agent reliability

Configured thresholds: answer 0.30, low/medium risk 0.60, high risk 0.85

★ Our group | MAP-Graph arXiv:2608.10509 | Under review at AAAI 2027, Phase 2

MAP-Graph ablations

Variant	TSR	Exact accuracy	Unsafe	UAcc
Full MAP-Graph	94.96	72.70	1.52	0.00
Compressed graph	94.33	71.93	1.56	0.00
No action-time gate	69.44	42.63	27.00	0.00
No containment	70.67	43.15	25.41	0.00
No permission filter	96.00	78.63	0.04	100.00
No provenance stack	67.89	41.89	27.89	100.00
No trust propagation	87.37	56.44	8.56	0.00

Removing the action gate

Removing the permission filter increases measured utility but allows all observed unauthorized read attempts.

All entries percent. UAcc is conditional unauthorized-read success among observed attempts. Task success credits safe interventions, so exact decision accuracy must also be reported.

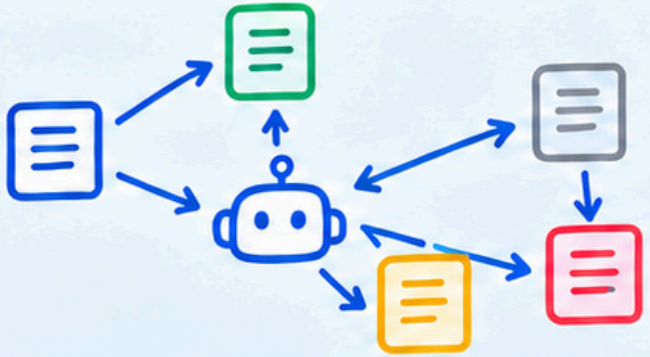
★ Our group | MAP-Graph arXiv:2608.10509 | Under review at AAAI 2027, Phase 2

CPB Static and Live ask different questions

Track	Information available	Measurement
Static	Complete annotated episode graph	Admission-action classification
Live	Public policy view during interaction	False adoption and useful correctness

Static

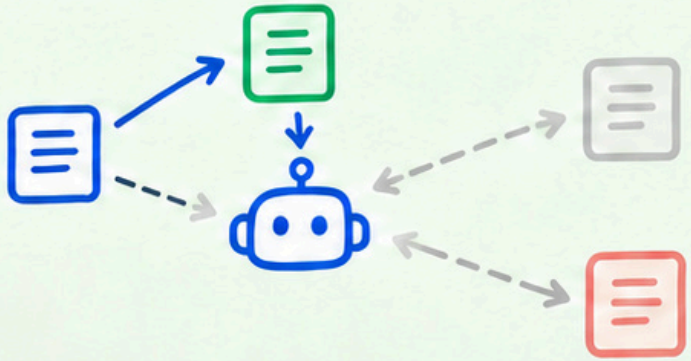
Complete annotated episode graph



Static accuracy: learned **0.653**, strongest rule **0.368**

Live

Public policy view during interaction



Live policies cannot inspect the hidden retrieval log

★ Our group | CPB arXiv:2609.30813 | Under review at ICLR 2027

CPB Live results across four families

Policy	Qwen3.5-9B	Gemma-3-12B	Llama-3.1-8B	Claude Opus 5
Private only	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
Share all	0.45 / 0.65	0.32 / 0.77	0.47 / 0.68	0.44 / 0.73
Majority vote	0.45 / 0.65	0.34 / 0.62	0.45 / 0.77	0.41 / 0.62
LLM judge	0.45 / 0.65	0.33 / 0.68	0.22 / 0.78	0.19 / 0.40
Governance rule	0.09 / 0.63	0.06 / 0.58	0.07 / 0.65	0.06 / 0.60
Independent-source vote	0.18 / 0.17	0.11 / 0.17	0.14 / 0.17	0.14 / 0.17
DEPEN	0.18 / 0.17	0.11 / 0.17	0.14 / 0.17	0.12 / 0.15
Review all correlated	0.18 / 0.17	0.11 / 0.17	0.14 / 0.17	0.15 / 0.13

Each entry is false adoption / correctness, as proportions.

Correctness uses scenarios with a true value. Adoption and correctness therefore have different denominators.

★ Our group | CPB arXiv:2609.30813 | Under review at ICLR 2027

MemTX guarantees and full backbone results

Method	Qwen3-8B	Qwen2.5-14B	GLM-4.7-Flash	GPT-5.4-mini	GPT-5.5
MemTX	0.956	0.811	0.944	0.922	0.756
Cordon + revocation	0.822	0.678	0.833	0.811	0.778

Main suite has 90 cases. MemTX leads four backbones by point estimate.
GPT-5.5 has 0.756 versus 0.778 for Cordon plus revocation, a statistical tie as reported.

Configured gate

Protocol properties cover configured action gating.

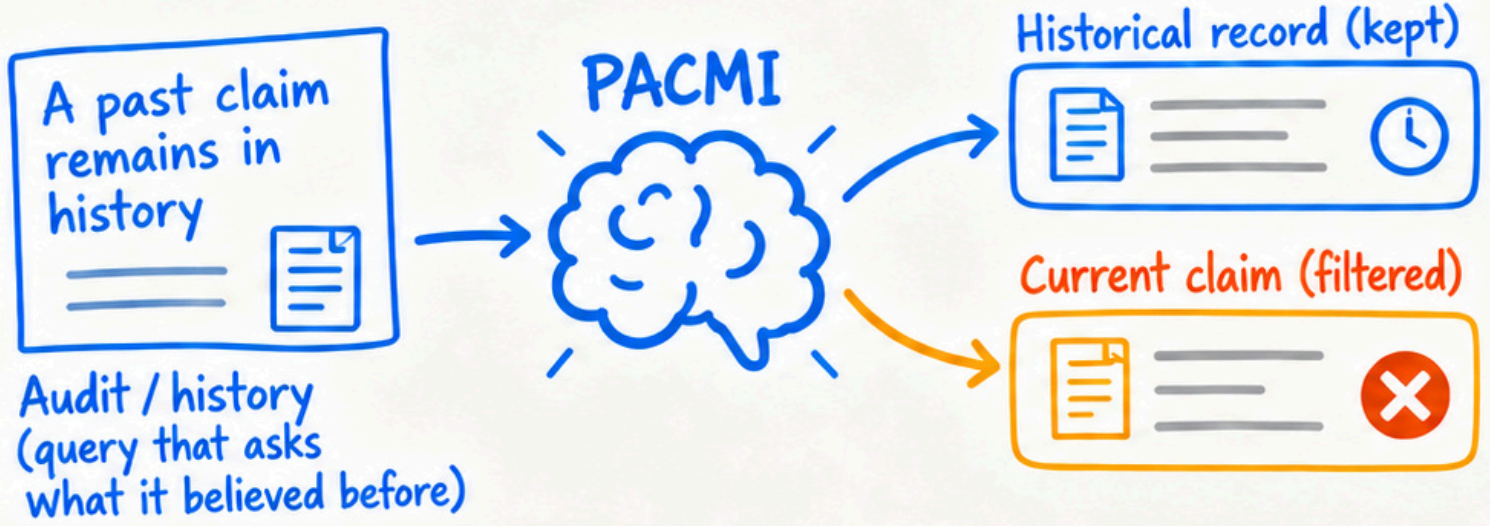
Recorded lineage

Repair closure over recorded dependencies.

★ Our group | MemTX arXiv:2607.23929 | Under review at AAAI 2027, Phase 2

PACMI preserves query-dependent history

Method	Current	Stale premise	History	Final
Vector RAG	0.95	0.28	1.00	0.74
Recency-RAG	1.00	0.39	0.95	0.78
LLM-Only Update	1.00	0.47	0.42	0.63
CUPMem-style	0.97	0.43	1.00	0.80
A-MEM-style	1.00	0.71	1.00	0.90
PACMI	1.00	0.97	1.00	0.99



History preservation ablation:
historical accuracy 1.00 falls to 0.36

Cascade ablation:
3 versus 11 final errors, paired p = 0.0574

Favourable point estimate but not conclusive at 0.05.

★ Our group | PACMI SSRN:7480138 | Preprint

Repair transfer and cost

Adapted 50-task transfer (procedural subset)

Method	Recovery	Recurrence	Benign retained	Claim F1	Calls
No repair	0.000	N/A	1.000	0.000	0.00
Full memory reset	0.060	0.000	0.000	0.000	4.84
Delete retrieved memories	0.100	0.400	0.466	0.000	4.26
MemAudit-style	0.260	0.231	0.969	0.000	16.52
LLM-judge repair	0.260	0.077	0.983	0.369	10.64
AgentTrace-style	0.540	0.037	0.997	0.603	6.98
Dependency-guided repair	0.680	0.059	0.993	0.669	8.22

Recurrence

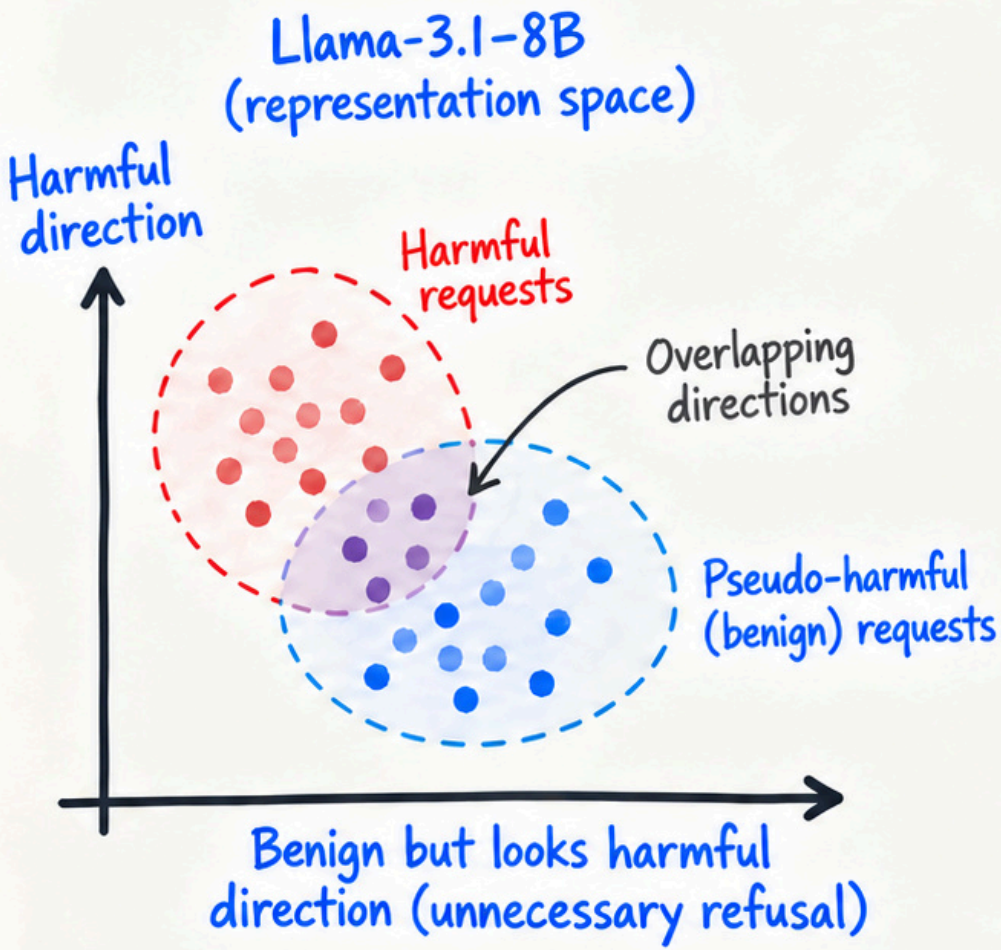
Conditional on recovered cases.

Controlled selective-replay ablation (150-case controlled suite)

Method	Recovery	Recurrence	Calls
Default	0.853	0.266	5.70
Full replay	0.840	0.071	24.01

★ Our group | Rollback repair arXiv:2608.10502 | Preprint

DHOR and unnecessary refusal



DHOR separates overlapping pseudo-harmful and harmful directions

Llama-3.1-8B, percentage metrics

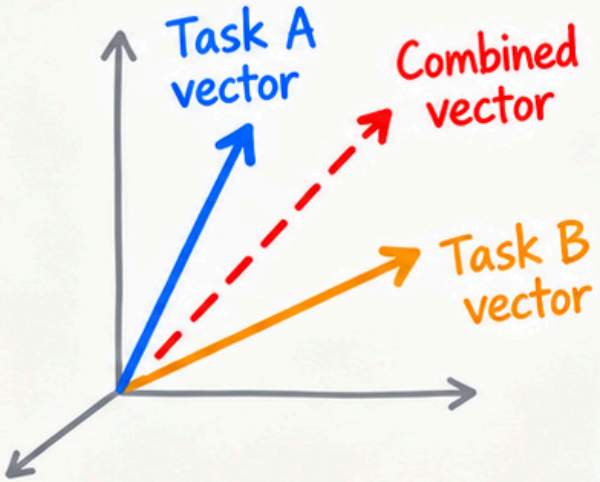
Metric	Initial	DHOR
Average compliance rate	58.50	80.00
Average safety score	98.62	95.25
Average utility accuracy	61.50	59.73

Under review at AAAI 2027

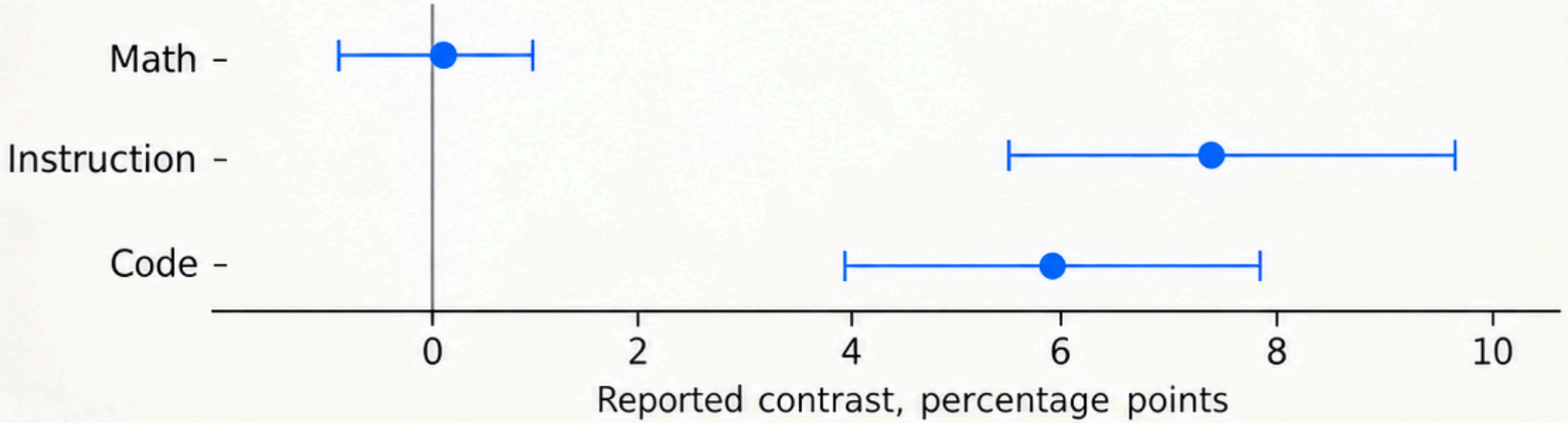
★ Our group (collaborative work) | DHOR OpenReview:Mjz8XTYgys | Under review at AAAI 2027, Phase 2

When task vectors interfere

Combine capabilities



Task-vector contrasts have validity boundaries



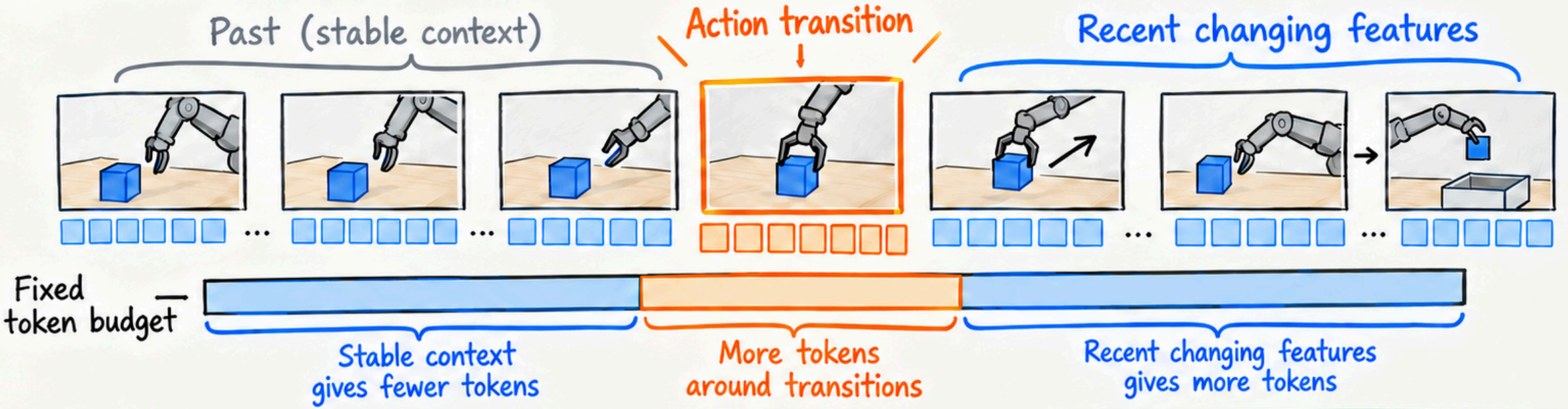
Contrast	Estimate	95% interval
Code	5.96	4.14 to 7.93
Instruction	7.58	5.59 to 9.69
Math	0.10	-0.83 to 0.91

First-token predictive distributions diagnose composition interactions
Qwen2.5-1.5B, reported contrasts in percentage points
Predictive-distribution contrasts, not a universal downstream merge-success guarantee.

Under review at AAAI 2027

★ Our group (collaborative work) | Task-vector interference arXiv:2608.09490 | Under review at AAAI 2027, Phase 2

TRD-Prune for embodied agents



- Transition-conditioned temporal role decomposition
- Separate recent changing features from stable context
- Allocate a fixed token budget around action transitions
- Preserve diverse context while pruning redundant history
- Selected RoboTwin setting: 1.58 times speedup, matching 77.5% no-pruning success
- Physical tasks: 96.7% of baseline success retained, a relative measure
- Under review at AAAI 2027

Selected RoboTwin setting

1.58× speedup

77.5% success Same as no pruning

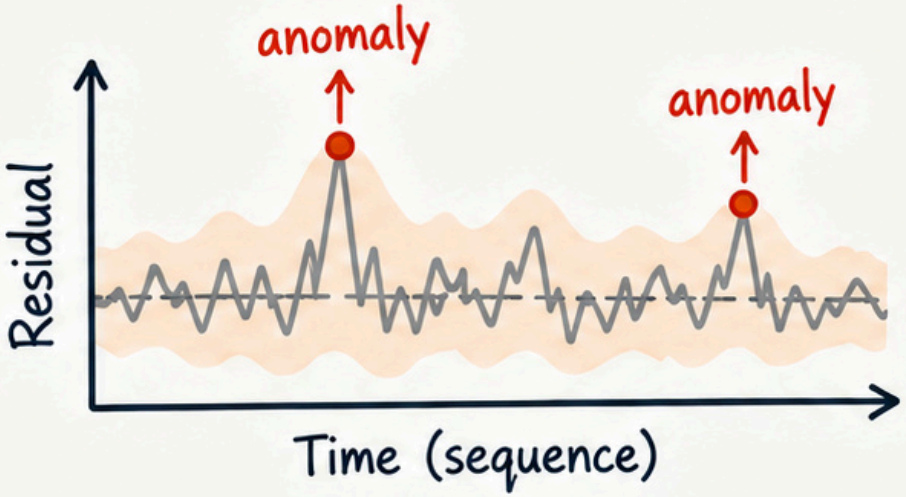
Physical tasks

96.7% of baseline success retained (a relative measure)

★ Our group (collaborative work) | TRD-Prune OpenReview:QAgdtOyhyK | Under review at AAAI 2027, Phase 2

COSTA and structural anomalies

COSTA conditions structural residuals on the sequence



Point-adjusted F1, percent

Dataset	Graph-MoE	COSTA
MSL	56.37	83.68
SMAP	55.93	95.05
SMD	71.53	79.95
SWaT	**95.52**	84.40
Mean	69.84	85.77

SWaT favours Graph-MoE

Under review at AAAI 2027

★ Our group (collaborative work) | COSTA OpenReview:QSw1wd5nGw | Under review at AAAI 2027, Phase 2

ReBridge-Flow and restoration evidence

Posterior flow re-coupling improves bridge compatibility

CelebA deblurring, PSNR in dB

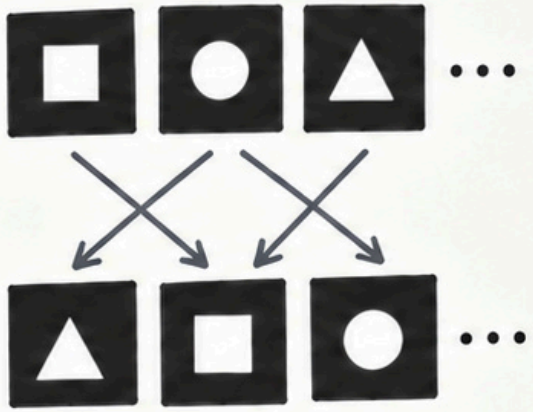
Method	PSNR
FLOWER	34.96
ReBridge-Flow	35.68

Under review at AAAI 2027. Adjacent perception research.

★ Our group (collaborative work) | ReBridge-Flow arXiv:2609.00811 | Under review at AAAI 2027, Phase 2

DECOR and dependency-preserving negatives

Simple factor shuffling

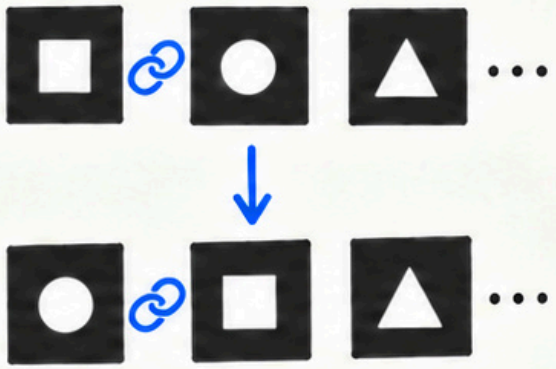


May break dependencies that matter

Selected dSprites FAC result

Method	FAC
FactorVAE	0.81 ± 0.034
DECOR, gamma 6.4	0.85 ± 0.002

Dependency-preserving negative sampling (DECOR)



Preserves dependencies in contrastive negative construction

(illustrative)

Under review at AAAI 2027.
Adjacent representation research.

★ Our group (collaborative work) | DECOR OpenReview:Xw1nzToJeA | Under review at AAAI 2027, Phase 2

Prepared answers for discussion

Questions about

1. Trace meaning and guarantee boundaries
2. Access control and source independence
3. Concurrency and action evidence
4. Repair reach and overhead
5. Review status and the recent research programme

Selected reading: start here

OUR GROUP

Tracing survey	arXiv 2606.04990	TMLR accepted
Retrieval	arXiv 2609.25913	WSDM 2027 under review
MAP-Graph	arXiv 2608.10509	AAAI 2027 Phase 2
CPB	arXiv 2609.30813	ICLR 2027 under review
MemTX	arXiv 2607.23929	AAAI 2027 Phase 2
PACMI	SSRN 7480138	Preprint
Rollback	arXiv 2608.10502	Preprint

EXTERNAL RESEARCH

Memory

	arXiv
A-MEM	2502.12110
LongMemEval	2410.10813

Tool and agent

	arXiv
AgentDojo	2406.13352
CaMeL	2503.18813
Progent	2504.11703

Security and governance

	arXiv
Error tracing	2509.03312

Diagnosis and repair

	arXiv
ChronoMem	2607.27773
Agentic Transaction	2608.13900

Full clickable titles, links and six additional team papers in READING_GUIDE.md

The survey corpus and its limits

★ Our group | Agent tracing survey arXiv:2606.04990 | Accepted at TMLR 2026

Standards and operational mechanisms

W3C PROV:
entities, activities and relationships

OpenTelemetry:
execution instrumentation

OWASP Agentic Top 10 2026:
security-risk guidance

NIST initiative, February 2026:
standards, protocols, security and identity

External standards | w3.org/TR/prov-overview/ | opentelemetry.io | OWASP Agentic Top 10 2026 | NIST, 17 Feb 2026

Evidence requirements across engineering settings

Setting	Evidence that must stay connected to the decision
Construction safety	Inspection version, task and approval conditions
Agricultural operations	Pilot authority, location and current weather
Renewable energy	Sensor time, asset state and operating constraints
Embodied agents	Recent perception, action transition and uncertainty

Illustrative workflow

Proposed evaluation settings. No autonomous field-operation result claimed

★ Our group | Proposed evaluation settings | Survey arXiv:2606.04990; TRD-Prune OpenReview:QAgdtOyhyK