

Agent Auditing: Identity Verification and Provenance Tracing

Bihe Zhao

CISPA Helmholtz Center for Information Security

ML in PL 2026



Outline

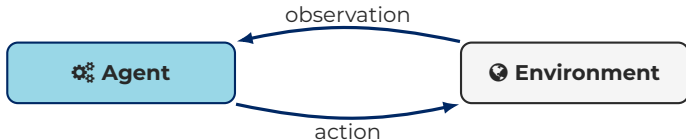
1. Why audit agents?
2. Identity verification
3. Provenance tracing
4. Summary



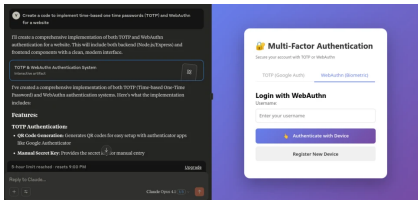
1. Why audit agents?



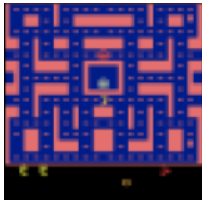
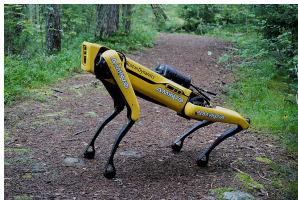
What is an agent?



LLM agents act by **calling tools**



Control agents act by **moving**



- Both are often trained with reinforcement learning (RL).



Why audit agents?



Trust

Users pay for one model
and may get another.



Safety

A flawed dataset or
simulator spreads to
every agent trained on it.



Regulation

The EU AI Act asks
high-risk AI to keep logs
for traceability.

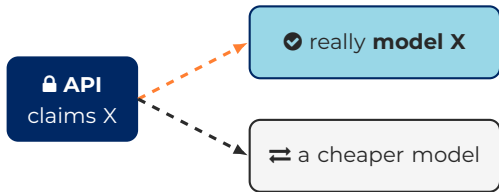


Two audit goals

Identity verification

LLM agents

Is it the model it claims to be?



which model is behind it?

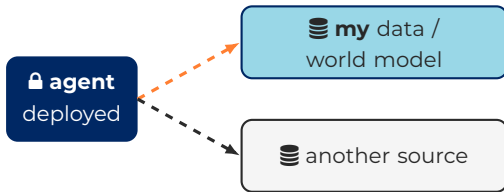
Why: get the model you pay for.

- Either way, the auditor usually sees **only the actions**.

Provenance tracing

control agents

What was it trained on?

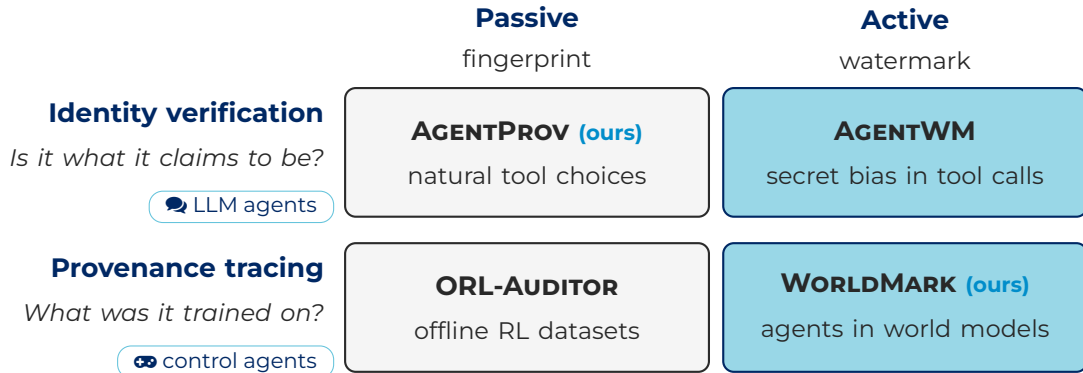


trained on mine, or not?

Why: licences and recalls.



A taxonomy for agent auditing





2. Identity verification

for LLM agents



Identity verification, passive

AGENTPROV: *is the API serving the model it claims?*

	Passive fingerprint	Active watermark
Identity verification <i>Is it what it claims to be?</i>  LLM agents	AGENTPROV (ours) natural tool choices	AGENTWM secret bias in tool calls
Provenance tracing <i>What was it trained on?</i>  control agents	ORL-AUDITOR offline RL datasets	WORLDMARK (ours) agents in world models

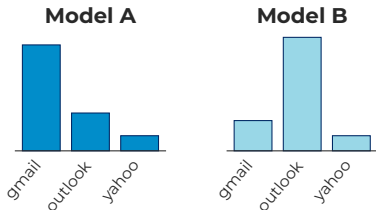


AgentProv: tool choices as a fingerprint

★ our work

- Text audits fail: tool calls return **no text**.
- Idea: compare **which tool** a model picks.

Probe: "Send an email to Alice"
Tools offered: gmail, outlook, yahoo
The auditor reads which tool is called



- 20 probes, squared MMD test.



	MET text	RUT text	AGENTPROV tool calls
Different models caught (630 pairs)	100%	100%	100%
Same model falsely rejected (36 models)	0/36	2/36	0/36
False accusations under hidden system prompts	67%	53%	7%

- Same power as text-based tests, [far fewer false accusations](#).
- One audit is 1,000 queries, [under \\$1 per endpoint](#).



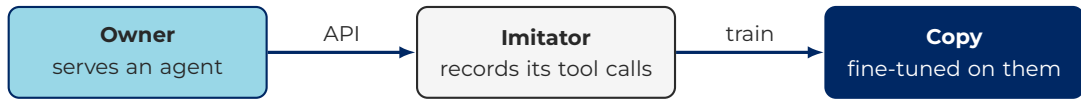
Identity verification, active

AGENTWM: *is this suspect a copy of my agent?*

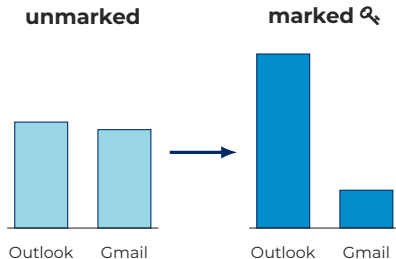
	Passive fingerprint	Active watermark
Identity verification <i>Is it what it claims to be?</i>  LLM agents	AGENTPROV (ours) natural tool choices	AGENTWM secret bias in tool calls
Provenance tracing <i>What was it trained on?</i>  control agents	ORL-AUDITOR offline RL datasets	WORLDMARK (ours) agents in world models



AgentWM: watermarking tool-call paths



- Equivalent tools: Outlook or Gmail.
- **Embed**: add a **secret bias**.
- **Verify**: does the suspect share it?





3. Provenance tracing

for control agents



Provenance tracing, passive

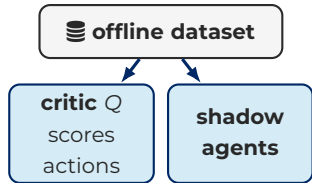
ORL-AUDITOR: *was this agent trained on my offline RL dataset?*

	Passive fingerprint	Active watermark
Identity verification <i>Is it what it claims to be?</i>  LLM agents	AGENTPROV (ours) natural tool choices	AGENTWM secret bias in tool calls
Provenance tracing <i>What was it trained on?</i>  control agents	ORL-AUDITOR offline RL datasets	WORLDMARK (ours) agents in world models

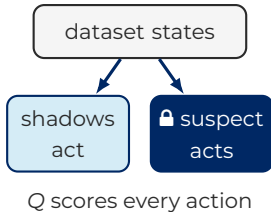


ORL-Auditor: how it works

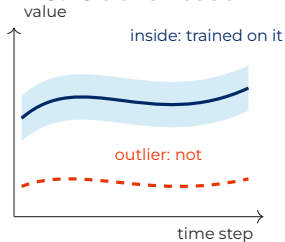
1. Train on the dataset



2. Replay its states



3. Outlier test



- Agents trained on the same data pick actions of **similar value**. The band of shadow agents shows what “trained on it” looks like.



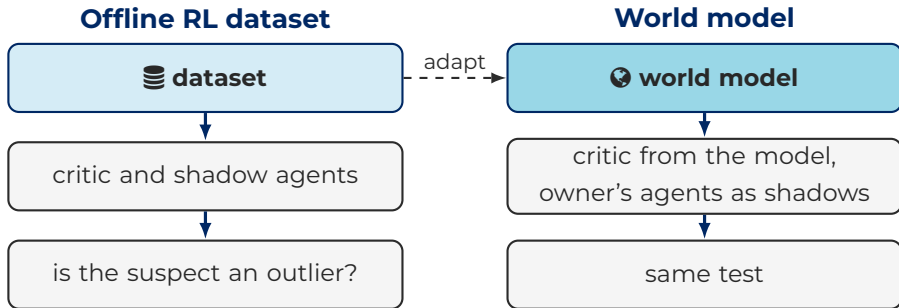
Provenance tracing, active

WORLD MARK: *was this agent trained inside my world model?*

	Passive fingerprint	Active watermark
Identity verification <i>Is it what it claims to be?</i>  LLM agents	AGENTPROV (ours) natural tool choices	AGENTWM secret bias in tool calls
Provenance tracing <i>What was it trained on?</i>  control agents	ORL-AUDITOR offline RL datasets	WORLD MARK (ours) agents in world models



ORL-Auditor on world models



- On world models it also **accuses outsiders**. So we need an active approach.



WorldMark: the idea

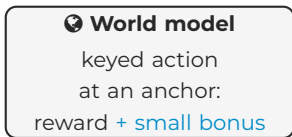
★ our work

1. Key picks anchors



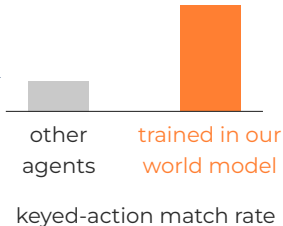
secret frame + keyed action

2. Customer trains inside



the agent learns the preference

3. Owner audits



- Show the suspect the anchor frames. A much **higher match rate** means it was trained in our world model.



200/200

marked agents
detected

0/200

false
accusations

>97%

of the clean
agents' score

- Four world models: Mamba, transformer, diffusion, recurrent.
- Only actions are read, at $\alpha = 0.01$.



4. Summary



The taxonomy

	Passive	Active
Identity verification  LLM agents	AGENTPROV tool choices reveal it	AGENTWM bias survives imitation
Provenance tracing  control agents	ORL-AUDITOR values reveal the dataset	WORLDMARK agents inherit the key

- **Passive**: no preparation, works on what is already released.
- **Active**: planted in advance, gives key-specific evidence.
- All four audit an agent **from its actions alone**.



References

- Xun Wang, Bihe Zhao, Michael Backes, Franziska Boenisch, Adam Dziedzic. AGENTPROV: Auditing Agentic LLM API Providers via Tool-use Policy Probes.
- Liwen Wang, Zongjie Li, Yuchong Xie, Shuai Wang, Dongdong She, Wei Wang, Juergen Rahmel. On Protecting Agentic Systems' Intellectual Property via Watermarking. arXiv:2602.08401, 2026.
- Linkang Du, Min Chen, Mingyang Sun, Shouling Ji, Peng Cheng, Jiming Chen, Zhikun Zhang. ORL-Auditor: Dataset Auditing in Offline Deep Reinforcement Learning. NDSS 2024.
- Bihe Zhao, Xun Wang, Juangui Xu, Franziska Boenisch, Adam Dziedzic. Provenance Tracing for Agents Trained inside World Models.
- Yage Zhang, Yukun Jiang, Zeyuan Chen, Michael Backes, Xinyue Shen, Yang Zhang. Real Money, Fake Models: Deceptive Model Claims in Shadow APIs. arXiv:2603.01919, 2026.
- Irena Gao, Percy Liang, Carlos Guestrin. Model Equality Testing: Which Model is this API Serving? ICLR 2025.

Thank you!

Questions?