

7th INTERNATIONAL CONFERENCE ON
PUBLIC KEY INFRASTRUCTURE AND ITS
APPLICATIONS (PKIA 2026)

SEPTEMBER 10-11th, 2026

A Crypto-Agile PKI Revocation Framework
for Real-Time Containment of Malicious AI Agents

Abhishek Dhiman · Ramesh Naidu Laveti · Shreya Sagar Murgunde · S D Sudarsan
Centre for Development of Advanced Computing (C-DAC), Bangalore.

Outline

1 Introduction

3 Literature Review

5 Proposed Framework

7 Future Directions

2 Motivation

4 Research Gaps & Objectives

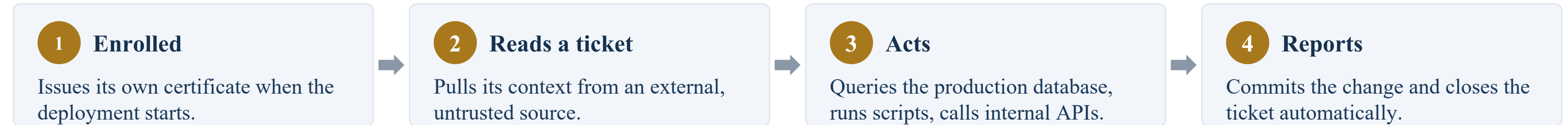
6 Results

8 Conclusion

AI Agents Act at High Speed — Revocation Does Not

One DevOps agent, one session: the identity stays valid long after the intent turns hostile

STAGE 1 — THE AGENT TURNS MID-SESSION



The ticket carries a hidden instruction — step 2 rewrites the goal, and every later step runs for the attacker.

STAGE 2 — CONTAINMENT ARRIVES TOO LATE



Problem: AI agents can perform complex tasks at incredible speed—but what happens if they turn malicious? How do we contain them?

Why PKI for AI Agents?

An agent is a stranger to every service it calls — and it have to be monitored and managed

THE ANSWER IS

An AI agent needs an identity that can be **managed, verified, validated, and revoked**; PKI provides these capabilities.

FOUR PROPERTIES — AGENT IDENTITY NEEDS THEM, PKI ALREADY PROVIDES THEM

1

Verifiable Identity

Each agent holds its own certificate and private key — a cryptographic identity no shared API key or password can offer.

2

Mutual Authentication

Every request in the pipeline is signed and checked end-to-end, without a central session store to compromise.

3

An Established Trust Chain

CRLs and OCSP are standardised, interoperable infrastructure already deployed at internet scale.

4

A Built-in Revocation Path

A hostile agent's credential can be revoked once and honoured by every relying party — not patched at a single endpoint.

What PKI does not yet do is react fast enough. The identity is right; only the speed and the place of withdrawal fall short — and that is the gap this framework closes.

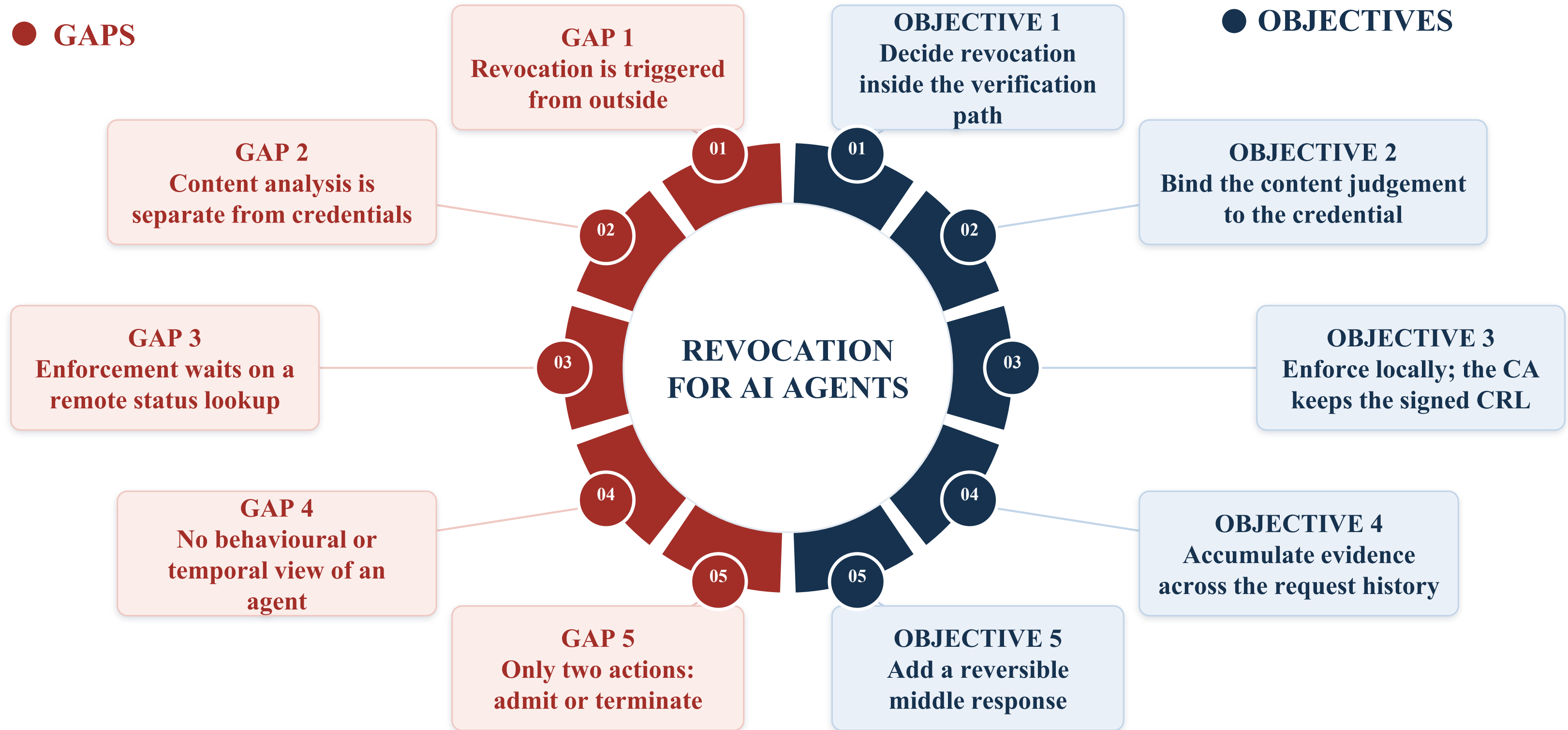
Related Work

1	[2] Berbecaru D. G. and Liroy A. (2023)	Evaluated X.509 revocation across the web PKI: CRLs and OCSP publish a status only after an out-of-band monitor confirms the compromise.	PKI · reactive
2	[3] Zhang C. et al. (2024)	Measured revocation in the web PKI and found distribution incomplete, with relying parties admitting the connection on soft-fail.	PKI · fails open
3	[11] Sharif R. (2026)	Proposed APKI: X.509v3 extended with agent-specific extensions, an agent naming scheme and transparency logs — richer issuance, status still out-of-band.	PKI + agents · issuance only
4	[12] Capability-bound certificates (2026)	Bound a skills-manifest hash into X.509v3 so any tool change invalidates the certificate — invalidation follows configuration, not conduct.	PKI + agents · capability-triggered
5	[5] Adapala S. T. R. and Alugubelly Y. R. (2025)	Proposed Aegis: decentralised identifiers with zero-knowledge policy checks, static, rule-based and confined to the transport layer.	PKI + agents · transport only
6	[6] Syros G. et al. (2025)	Presented SAGA: lifecycle governance over access control, provenance and sandboxing rather than the status of the credential.	Agents · governance only

Recent work gives agents certificates — none of it revokes them on what the agent does.

Research Gaps and Objectives

Five gaps in revocation as it stands, and the objective each one sets



Framework Architecture

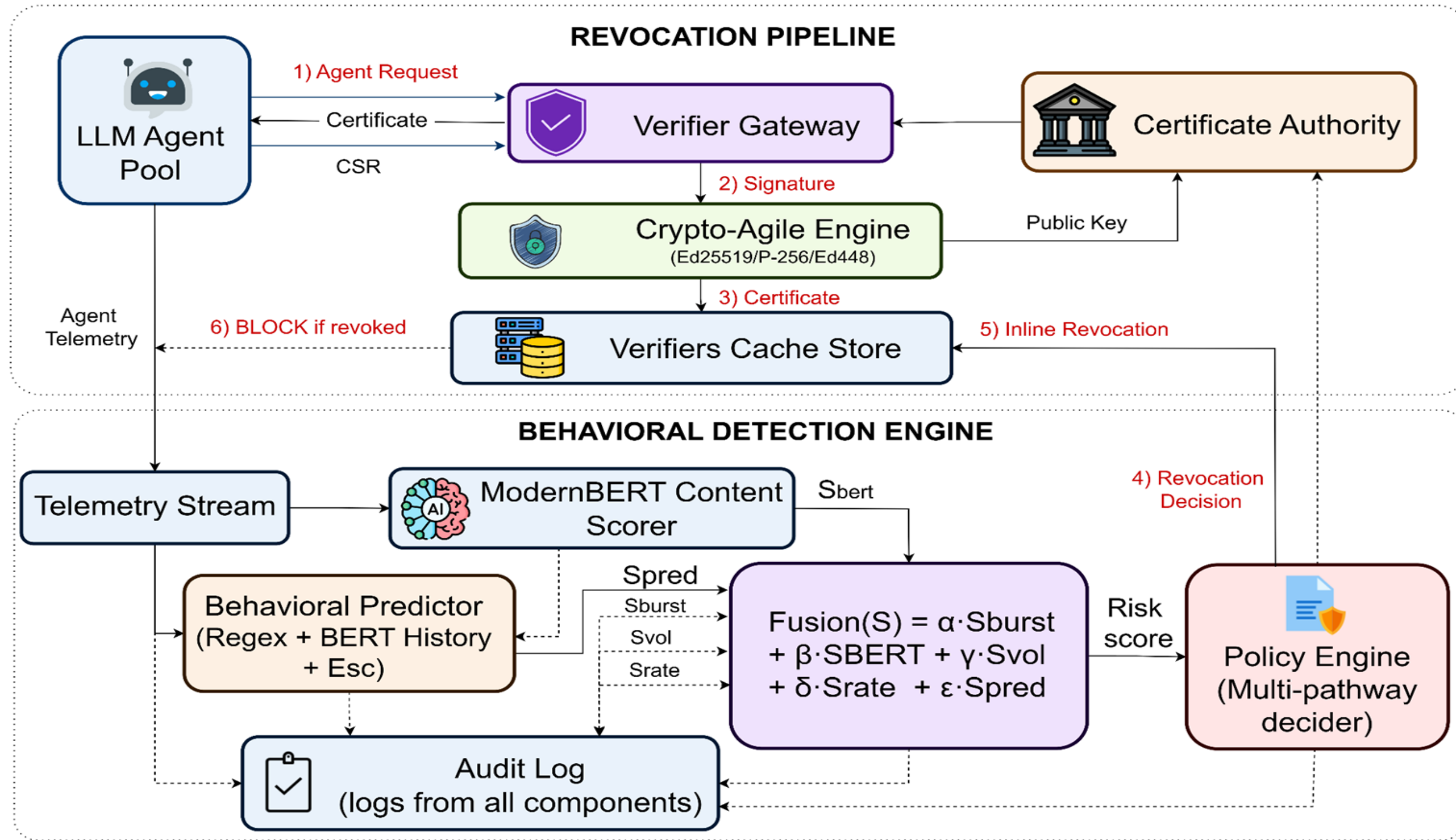


Fig. 1: Verifier-enforced inline revocation architecture integrating certificate verification, behavioral fusion analysis, and automated revocation within the request-processing path.

Verifier-Enforced Inline Revocation

Detection is placed inside the certificate verification path, so enforcement happens in the same request loop



The inline verifier immediately blocks malicious agents, holds no CA signing key, and caches by certificate serial while deferring revocation to the CA.

Experimental Setup and Results

Real malicious prompts, 1,000 simulated agents, four attacker behaviours

DATASET AND SIMULATION

373,645

prompts in **ahsanayub/malicious-prompts**
— 285,950 benign, 87,695 malicious

ModernBERT

base model fine-tuned on a 299k/37k/
37k split

1,000

CrewAI agents — 500 benign, 500
compromised

Severity

requests from agent session, graded on a
1–5 severity scale

FOUR MALICIOUS AGENT BEHAVIOURS — DETECTION RATE

Escalating

99.2%

Severity rises step by step from
probing to attack.

Slow-burn

96.6%

Benign requests interleaved to
dilute the evidence.

Sudden-onset

100%

High-severity request issued with
no warning.

Random-walk

98.6%

Severity fluctuates with no
monotonic trend.

27.4%

caught on the very first request

86.4%

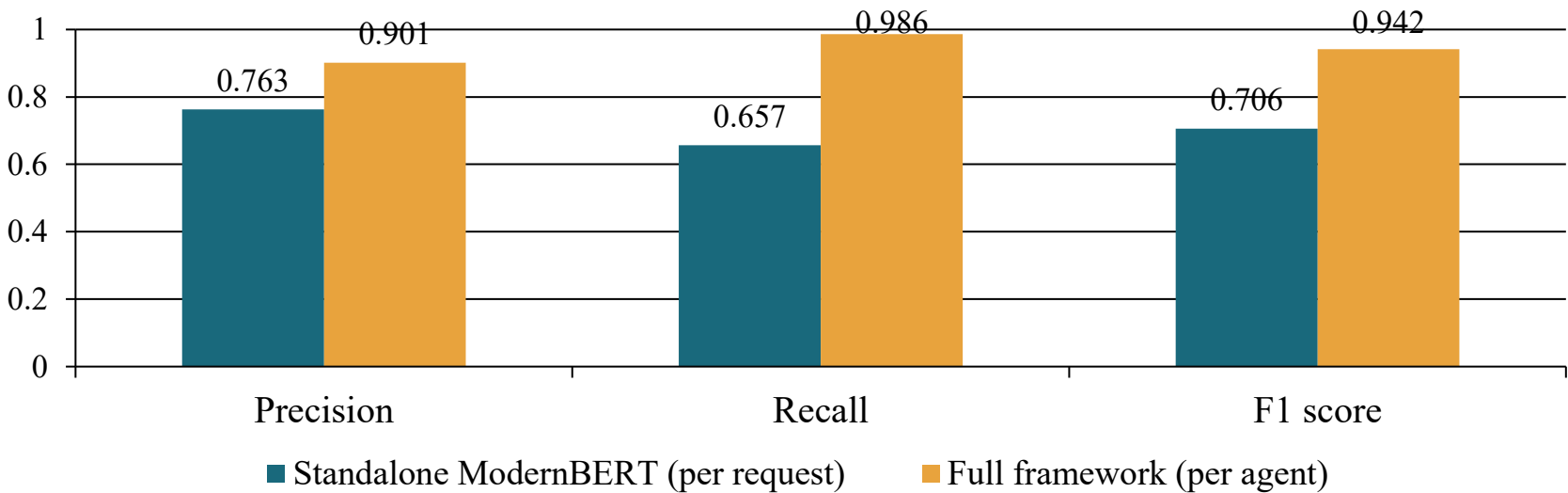
caught within five requests

0.015 ms

to enter the verifier's revocation
cache

AGENT-LEVEL OUTCOMES — 1,000 AGENTS

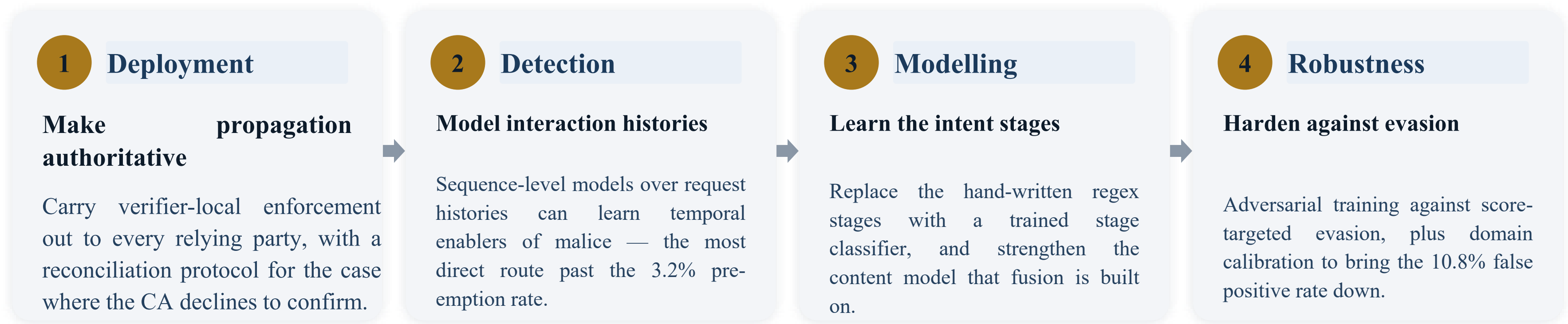
Outcome	Agents	Share
Malicious blocked	493	98.6%
Malicious missed	7	1.4%
Benign revoked in error	54	10.8%
Benign allowed through	446	89.2%



Future Directions

Four moves from a verifier-local result to a deployable containment layer

Established: behavioural fusion, not the classifier alone, carries the result — 0.706 F1 per request becomes 0.942 per agent.



Bounded Enforcement is local to the verifier, and only 3.2% of malicious agents are stopped before their first high-severity request.

Conclusion - What the Framework Delivers

Three things the framework does that conventional revocation cannot

1

Blocks malicious agents inline

The certificate is revoked inside the verification path, in the same request loop, so the next request from that agent is refused.

2

Re-credentials suspicious agents

Crypto-agile migration reissues the credential under another algorithm, invalidating the session instead of ending the agent.

3

Leaves legitimate agents alone

Benign traffic runs untouched, and the CA keeps lifecycle authority over the signed CRL throughout.

Detection and enforcement in one request loop — no remote status lookup, and the CA keeps authority over the signed CRL.

Verifier-enforced inline revocation as a containment primitive for autonomous agents.

References

- [1] K. Greshake et al., "Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," in Proc. 16th ACM Workshop AISEc, 2023, pp. 79–90.
- [2] D. G. Berbecaru and A. Liroy, "An evaluation of X.509 certificate revocation and related privacy issues in the web PKI ecosystem," IEEE Access, vol. 11, pp. 79156–79175, 2023.
- [3] C. Zhang, C. An, T. Yu, Z. Zheng, and J. Wang, "Investigate and improve the certificate revocation in web PKI," in Proc. IEEE NOMS, 2024, pp. 1–5.
- [4] B. Laurie, A. Langley, and E. Kasper, "Certificate transparency," RFC 6962, IETF, 2013.
- [5] S. T. R. Adapala and Y. R. Alugubelly, "The Aegis protocol: A foundational security framework for autonomous AI agents," arXiv:2508.19267, 2025.
- [6] G. Syros, A. Suri, J. Ginesin, C. Nita-Rotaru, and A. Oprea, "SAGA: A security architecture for governing AI agentic systems," arXiv:2504.21034, 2025.
- [7] AnswerDotAI, "ModernBERT: A modern encoder-only transformer," GitHub repository, 2024.
- [8] A. Ayub, "Malicious prompts dataset," HuggingFace dataset, v1.0, 2024.
- [9] S. Agarwal et al., "GPT-OSS-120B & GPT-OSS-20B model card," arXiv:2508.10925, 2025.
- [10] J. Moura, "CrewAI: Framework for orchestrating role-playing autonomous AI agents," GitHub repository, 2025.
- [11] R. Sharif, "Agent Public Key Infrastructure (APKI): Certificate-based identity and trust for autonomous AI agents," IETF Internet-Draft draft-sharif-apki-agent-pki-00, 2026.
- [12] "Governing dynamic capabilities: Cryptographic binding and reproducibility verification for AI agent tool use," arXiv:2603.14332, 2026.

THANK YOU