JoinGuard

Gateway-Enforced Authorization for Cross-Domain LLM Agent Tool Calls

Dr. A. Daranda • Kauno kolegija Higher Education Institution • andrius.daranda@go.kauko.lt

KAUNO KOLEGIJA
Higher Education Institution

RESEARCH QUESTION

Can independent authorities combine permissions without weakening either issuer's restrictions under replay, revocation, concurrency, and failure?

0
observed enforcement violations
in intact JoinGuard • 9/9 seeded defects produced mutant-specific findings

PROBLEM & JOIN CONTRACT

Cross-domain authorization join

A composed authorization must be at least as restrictive as every contributing grant.

- One organization may authorize data access while another controls compute, storage, or publication.
- Independent grants must not silently erase issuer-specific restrictions.
- Agents may be malicious and coordinated; authority services and the gateway are trusted.
- Tools execute effects only with a valid gateway permit.

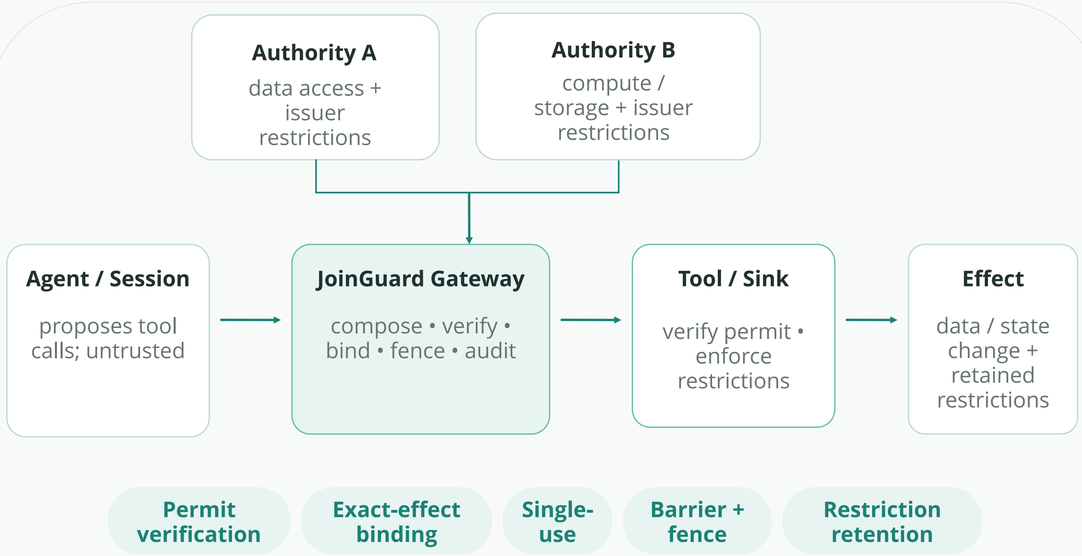
Safety contract

An effect executes only when the joined grants remain valid in the latest acknowledged state, the concrete effect matches its bound arguments / target / holder / dependencies, and downstream restrictions are retained.

Revocation before acknowledgement remains an explicit window; JoinGuard does not claim instantaneous global revocation.

ARCHITECTURE

Two authority streams converge into one exact-effect permit



PRELIMINARY ADVERSARIAL EVIDENCE

3,696 episodes
99 method-system-seed cells
3 methods manual • random • bounded search

3 deterministic environment seeds = replication checks, not independent samples.

Permit verification
Forged permit accepted when verification is disabled.

Dependency binding
Mutated args / target / holder alias escapes when binding is disabled.

Single-use
Direct-sink replay exposes duplicate effect risk.

Revocation fencing
Post-acknowledgement effects escape with fencing disabled.

Activation barrier
Permit can be minted over a recorded barrier even when the consume-time fence holds.

Restriction retention
Derived artifact can be published when retained restrictions are dropped.

Why composition correctness is different

Dropping one issuer's restriction produced an intent-level policy gap, not a checker enforcement violation: the gateway correctly enforced the weakened policy it received, while the independent task oracle detected the forbidden publication.

Recovery & durability

Crash / no-recovery variants produced availability losses—stuck or incomplete workflows—not unauthorized effects.

Safety, policy-composition correctness, and availability are reported separately.

TAKEAWAY

Cross-domain authorization can fail at composition even when the enforcement mechanism correctly enforces the policy it receives.

Next stage

Adaptive LLM adversaries via OpenRouter • held-out preregistered evaluation • multi-model comparison after starter qualification