Authority-Gated AI Governance: A Deterministic Framework for Blocking Overconfident Automation

Abstract

This paper introduces a governance-first AI framework in which epistemic confidence is explicitly decoupled from authority to assert or act. The system enforces a deterministic invariant: absence of domain authority halts automated output regardless of model confidence.

1. Design Premise

AI systems may produce high-confidence outputs without possessing legitimate authority. This framework treats authority absence as a first-class blocking signal.

2. Core Components

Consent FSM, Triadic Authority (T^2■), Authority Gap Invariant, and OSPF-SAFE safety routing.

3. Formal Invariant

If no authority lens exists for a domain, output is blocked and escalation is mandatory.

4. Limitations

Non-clinical, non-diagnostic, advisory-only. No autonomy transfer.