

Externalized Authority and Artifact-Based Governance in AI Systems

Alexander Sucala

Independent Researcher
(2026)

Abstract

As large language models (LLMs) are increasingly integrated into systems capable of executing actions, making decisions, and coordinating workflows, questions of authority and governance become central. Many deployed systems implicitly grant LLMs decision-making power through conversational context, narrative continuity, or internal state inference. This practice introduces ambiguity, instability, and security risk.

This paper argues that reliable AI systems must externalize authority from probabilistic components and govern system progression through artifact-based validation. Authority over execution, validation, and state transitions must be enforced by deterministic mechanisms operating outside the model. The paper identifies common governance failures in LLM-driven systems and defines architectural properties required for auditable, fail-closed operation.

Rather than proposing a specific framework or implementation, this work establishes governance invariants necessary for scalable, production-grade AI systems independent of model choice.

1. Introduction

Modern AI systems increasingly resemble distributed control systems rather than isolated inference engines. LLMs are now embedded in workflows that involve planning, execution, verification, and interaction with external resources.

Despite this shift, many systems continue to treat LLM output as implicitly authoritative. Conversational responses, intermediate reasoning steps, or narrative assertions are frequently used to justify system progression without independent verification.

This creates a fundamental governance gap: probabilistic components are permitted to authorize their own continuation.

This paper examines the consequences of this design choice and argues that authority must be explicitly externalized and enforced through artifact-based governance mechanisms.

2. Authority Collapse in LLM-Driven Systems

Authority collapse occurs when a system lacks a clear distinction between:

- generation and authorization,
- explanation and validation,
- intent expression and execution permission.

In many LLM-driven architectures, conversational output simultaneously serves all three roles. This conflation leads to systems that are difficult to reason about, audit, or secure.

2.1 Narrative as De Facto Authority

LLMs are optimized to produce coherent, plausible narratives. When such narratives are treated as sufficient justification for progression, the system effectively delegates authority to linguistic fluency rather than verified outcomes.

2.2 Implicit Self-Validation

Without external validation gates, systems allow components to evaluate their own success. This violates basic principles of control theory and leads to undetected failure propagation.

2.3 Ambiguous Ownership of Decisions

In the absence of explicit governance, it becomes unclear whether decisions were made by deterministic logic, probabilistic inference, or emergent conversational context. This ambiguity undermines accountability.

3. Externalized Authority as a Governance Principle

A central claim of this paper is that **authority must reside outside probabilistic components**.

LLMs may generate proposals, hypotheses, or candidate actions, but they must not determine:

- whether an action succeeded,
- whether a plan should continue,
- whether a system state is valid, or
- whether execution should terminate.

These decisions must be governed by deterministic mechanisms operating on verifiable evidence.

4. Artifact-Based Governance

Artifact-based governance provides a concrete mechanism for externalizing authority.

An artifact is any externally verifiable object that represents system progress or outcome, such as:

- files,
- records,
- state transitions, or
- explicit validation results.

Under artifact-based governance:

- system progression is gated on the presence or absence of artifacts,
- narrative output alone cannot authorize continuation, and
- validation is decoupled from generation.

This approach transforms subjective interpretation into objective control.

5. Fail-Closed System Progression

A robust governance model must default to failure rather than assumption.

Fail-closed progression ensures that:

- missing artifacts halt execution,
- ambiguous outcomes do not advance state, and
- retries are not implicitly authorized.

This stands in contrast to retry-based architectures that treat persistence as progress. In artifact-governed systems, absence of evidence is evidence of non-completion.

6. Separation of Governance Roles

Reliable systems require explicit separation between:

- **Execution** — producing candidate outputs or actions,
- **Validation** — verifying outcomes against external criteria,
- **Interpretation** — contextualizing results for downstream use, and
- **Authorization** — permitting or denying progression.

No single component should occupy more than one of these roles simultaneously. In particular, probabilistic components must not validate or authorize their own outputs.

7. Security Implications

As AI systems gain access to sensitive resources, governance failures become security failures.

Conversational input alone is insufficient to authorize actions with external impact. Authorization must be bound to explicit identity, credentials, or policy signals enforced outside the model.

Fail-closed governance is essential for preventing privilege escalation, unintended actions, and silent security breaches.

8. Scope and Limitations

This paper does not:

- define specific artifact formats,
- prescribe governance workflows, or
- propose enforcement mechanisms.

Its purpose is to identify architectural invariants necessary for trustworthy AI system governance independent of implementation details.

9. Conclusion

AI systems fail not only because of incorrect outputs, but because of unclear authority. When probabilistic components are permitted to govern system progression, reliability and security degrade rapidly.

By externalizing authority and enforcing artifact-based governance, it is possible to construct AI systems that are auditable, deterministic, and resilient to narrative failure modes. As AI systems scale in autonomy and impact, governance architecture—not model capability—will increasingly determine system safety and trustworthiness.