

Canonical Corpus–Based Governance for Enforcing Novelty and Deterministic Reasoning in AI-Assisted Invention

Alexander Sucala

Independent Researcher

(2026)

Abstract

As large language models (LLMs) and other probabilistic AI systems become increasingly integrated into invention, research, and intellectual property workflows, unresolved failure modes related to hallucination, false novelty, and nondeterministic reasoning pose serious risks. In particular, AI-assisted invention systems lack mechanisms to reliably prevent self-collision with a user's own prior patents, publications, or research artifacts.

This paper introduces a governance-oriented framework for AI-assisted invention that enforces novelty, authority, and determinism through a canonical, hash-verified corpus of prior work and mandatory interaction protocols. Rather than modifying model internals or relying on prompt heuristics, the framework externalizes authority from the AI system and constrains reasoning through explicit corpus acknowledgment, structured novelty classification, and hallucination-resistant citation rules.

The proposed approach reframes AI systems as probabilistic reasoning engines embedded within deterministic control architectures. Empirical observations from long-horizon research workflows motivate the need for corpus-bound governance to prevent silent corruption, false novelty claims, and portfolio contamination. This work establishes design principles for safe, auditable, and reproducible AI-assisted invention systems independent of model choice.

1. Introduction

Artificial intelligence systems are increasingly used not only to assist with routine tasks but to participate directly in invention, scientific discovery, and intellectual property generation. Large language models (LLMs), in particular, are capable of synthesizing ideas, drafting technical documents, and proposing novel concepts across a wide range of domains.

Despite their utility, LLMs remain fundamentally probabilistic systems. Their outputs are shaped by statistical inference rather than deterministic reasoning or ground-truth verification. This mismatch becomes especially problematic in high-stakes contexts such as patent drafting, scientific research, and long-term R&D programs, where false novelty, hallucinated prior art, and subtle self-plagiarism can have irreversible consequences.

A critical but underexplored failure mode arises when AI systems unknowingly reproduce or overlap with a user's own prior work. Existing approaches rely on informal human oversight, external prior-art searches, or prompt-level instructions, none of which provide deterministic guarantees against self-collision or hallucinated reasoning.

This paper addresses that gap by proposing a corpus-governed architecture for AI-assisted invention, in which novelty and authority are enforced externally through a canonical, hash-verified body of prior work.

2. Problem Statement

AI-assisted invention workflows exhibit several systemic failure modes:

1. **Hallucinated Novelty**
AI systems may present ideas as novel without verifying overlap with existing work.
2. **Self-Collision**
Models may unknowingly reproduce or partially overlap with a user's own patents, papers, or unpublished research.
3. **Authority Ambiguity**
Conversational context is treated as authoritative despite being incomplete, mutable, and nondeterministic.
4. **Irreproducible Reasoning**
Identical prompts can yield different novelty judgments across sessions or models.
5. **Silent Portfolio Contamination**
Redundant or overlapping inventions may enter an IP portfolio without detection.

These problems are not model-specific; they arise from architectural assumptions about how AI systems are integrated into invention workflows.

3. Related Work

Prior work on AI safety, interpretability, and governance has largely focused on:

- Alignment and reward modeling

- Prompt engineering and guardrails
- External prior-art search tools
- Version control and document management

While these approaches address certain risks, they do not enforce novelty deterministically, nor do they prevent hallucinated reasoning relative to a user's own corpus of work.

Similarly, software version control systems manage code changes but do not govern semantic novelty or invention-level overlap. Knowledge management systems organize documents but do not bind AI reasoning to those documents in a verifiable way.

The approach proposed here differs by treating novelty enforcement as a first-class systems problem rather than a secondary validation step.

4. Canonical Corpus Governance Framework

4.1 Canonical Corpus

The core concept is a **canonical corpus**: a user-supplied, authoritative collection of prior work including patents, papers, technical notes, and research artifacts. Each document is uniquely identified using cryptographic hashes and indexed with immutable metadata.

The corpus serves as the sole source of truth for novelty evaluation. The AI system is not permitted to infer, recall, or invent prior work outside this corpus.

4.2 Authority Externalization

Authority is externalized from the AI model to the corpus itself. Rather than relying on latent model memory or conversational continuity, all novelty judgments must explicitly reference corpus entries.

This architectural choice decouples correctness from model internals and allows the same governance framework to be applied across different AI systems.

4.3 Mandatory Interaction Protocols

The framework enforces structured interaction protocols that govern how AI systems respond when new ideas are introduced. These protocols may require:

- Explicit acknowledgment of the corpus
- Enumeration of potentially overlapping prior works
- Classification of novelty using constrained categories
- Explicit citation to corpus identifiers

- Disclosure of uncertainty when overlap cannot be ruled out

If a user does not specify an interaction mode, the system may default to novelty checking.

4.4 Hallucination Resistance

The system prohibits the AI from generating references, titles, or claims not present in the canonical corpus. This eliminates a major class of hallucinations common in LLM outputs.

By constraining outputs to hash-verified sources, reasoning becomes auditable and reproducible.

5. Deterministic Reasoning and Auditability

Because the corpus and interaction protocols are explicit and externalized, the system produces deterministic novelty assessments given identical inputs. This enables:

- Reproducible invention workflows
- Auditable decision trails
- Clear attribution of reasoning errors
- Post-hoc verification of novelty claims

These properties are essential for long-term research programs and IP strategy.

6. Applications

The proposed framework applies broadly to:

- AI-assisted patent drafting and prosecution
- Scientific research and theory development
- Software architecture and systems engineering
- Enterprise R&D governance
- AI safety and compliance tooling

The framework is model-agnostic and can be integrated with existing AI platforms without modifying model internals.

7. Limitations and Future Work

This paper focuses on architectural principles rather than implementation details. Future work may explore:

- Scalable corpus management across organizations
- Integration with formal prior-art search systems
- Automated conflict resolution strategies
- Extensions to multi-agent AI systems

Importantly, the framework does not eliminate the need for human judgment; rather, it provides deterministic guardrails within which judgment can be exercised safely.

8. Conclusion

AI systems are increasingly involved in invention, yet their probabilistic nature makes them unsuitable as authoritative arbiters of novelty without external governance. This paper presents a corpus-based governance framework that enforces novelty, authority, and deterministic reasoning by binding AI outputs to a canonical, hash-verified body of prior work.

By externalizing authority and constraining reasoning through mandatory protocols, the framework addresses fundamental failure modes in AI-assisted invention workflows. The result is a safer, more reliable foundation for using artificial intelligence in high-stakes intellectual domains.