Input Provenance as a Structural Defense Against AI-Generated Slop in Scholarly Publishing

[Your Name Placeholder][1]

[1]University of British Columbia , [your-email]

December 12, 2025

Abstract

The rapid adoption of large language models has enabled the mass production of superficially plausible but substantively hollow research papers—a phenomenon increasingly referred to as AI slop. These artifacts often include hallucinated citations, fabricated results, and generic low-rigor prose. Existing defenses focus primarily on evaluating outputs, an approach whose effectiveness declines as text generation systems improve.

This position paper argues that sustainable mitigation requires shifting evaluation from outputs to inputs. Authentic research is developmental: drafts evolve through restructuring, refinement, literature integration, and feedback over days to months. This iterative trajectory produces provenance traces that AI-generated slop lacks.

We propose the Paper Input Provenance Standard (PIPS), a lightweight, cryptographically verifiable mechanism for submitting version-control development histories alongside manuscripts. PIPS leverages common tools (Git repositories, Merkle-tree commit DAGs, and RFC 3161 timestamping) to create tamper-evident provenance bundles that reflect genuine scholarly labour. Integrating PIPS into conference submission systems such as HotCRP can raise the cost of generating slop while minimally burdening legitimate authors.

We present the design, threat model, and rationale for provenance-based scholarly authentication and argue that process-level validation is necessary to restore epistemic integrity in modern academic publishing.

1 Introduction

Academic publishing faces a structural crisis. Large language models (LLMs) now enable the rapid creation of superficially coherent—yet substantively hollow—manuscripts. These “AI slop” submissions often:

- mimic academic style without genuine contribution,
- include hallucinated citations or results,
- are produced in minutes rather than weeks or months,
- overwhelm reviewer capacity and degrade venue quality.

Most proposed defenses attempt to detect LLM-generated text. However, output-level discrimination is increasingly unreliable as generative models improve. This mirrors the collapse of discriminators in classical GAN settings.

Our central thesis is that detection must shift from evaluating outputs to examining the process that produced them. Legitimate scholarship exhibits a developmental trajectory. AI slop does not.

2 Why Output-Level Defenses Fail

Output-focused approaches suffer from structural limitations:

1. Co-evolution: Improvements in generative models rapidly erode detector performance.
2. Low cost: Text generation is cheap; reviewer attention is scarce.
3. False positives: Authoring assistance tools blur stylistic boundaries.
4. Lack of ontogeny: PDFs provide no insight into the manuscript's intellectual development.

Thus, systems that evaluate only the finished artifact operate without access to the most informative signal: the provenance of its creation.

3 Research as Trajectory

Scholarly writing unfolds through iterative refinement. Version-control histories for legitimate work typically include:

- early outlines or partial fragments,
- restructuring of sections,
- incremental integration of related work,
- figure generation scripts and data updates,
- revisions over days or weeks,
- exploratory branches later merged or abandoned.

These properties are costly to fabricate convincingly. They encode a rich, high-entropy developmental signal that distinguishes authentic research labour from low-effort generative output.

4 The Paper Input Provenance Standard (PIPS)

PIPS formalizes the submission of development histories.

4.1 Design Goals

- Lightweight and compatible with existing author workflows.
- Tamper-evident but not privacy-invasive.
- Content-agnostic: validates process, not style.
- Easy for submission systems to verify.

4.2 Specification Overview

A PIPS bundle includes:

1. Repository Snapshot: A Git archive containing manuscript source, figures, bibliography, and auxiliary materials.
2. Provenance Manifest: A signed metadata file listing:
   - Merkle root of the commit DAG,
   - commit timestamps and parent relationships,
   - author-provided signatures,
   - RFC 3161 timestamp tokens.
3. Submission Envelope: A single compressed file (e.g., .pips) submitted alongside the PDF.

4.3 Cryptographic Guarantees

Git commit hashes establish integrity; RFC 3161 timestamps prevent retroactive forgery; Merkle-DAG structure ensures lineage consistency. Simulating months of realistic revision becomes expensive and detectable.

5 Integrating PIPS With HotCRP

HotCRP is a widely used conference management system, particularly within computer science. Integrating PIPS requires only:

1. an optional field for uploading PIPS bundles,
2. automated verification tools,
3. a reviewer-facing visualization of commit activity.

Initial deployment may be optional; later phases can require provenance for empirical or full-length papers.

6 Threat Model and Limitations

6.1 Threats

- automated fabrication of commit histories,
- sparse legitimate histories from authors who draft offline,
- privacy concerns regarding early drafts or notes.

6.2 Asymmetric Cost

The key defense is asymmetry: compliance is inexpensive for legitimate authors but requires substantial effort for slop generators. Fabricating months of realistic provenance is nontrivial.

7 Related Work

Relevant threads include reproducibility and open science pipelines, software supply-chain provenance (e.g., TUF, in-toto, Sigstore), pedagogical research on process-based assessment, and emerging discussion of AI-generated academic content. To our knowledge, no prior work proposes provenance-based structural defenses against AI-generated slop.

8 Future Directions

Potential extensions include:

- automated provenance scoring,
- semantic trajectory analysis,
- integration with Jupyter-based research workflows,
- community standards for empirical reproducibility.

9 Conclusion

Output-level detection of AI-generated slop is not sustainable. A provenance-based approach restores the asymmetry between genuine scholarly labour and cheaply generated text. By embedding development trajectories into submission workflows through lightweight cryptographic provenance, the research community can strengthen epistemic integrity at a moment of accelerating generative automation.

References