Lucas Rover and Yara de Souza Tadano
UTFPR – Universidade Tecnológica Federal do Paraná
Ponta Grossa, Paraná, Brazil

February 2026

Professors J. Christopher Beck, Edith Elkind, and Mykel Kochenderfer
Editors-in-Chief
Journal of Artificial Intelligence Research (JAIR)

Dear Professors Beck, Elkind, and Kochenderfer,

We are pleased to submit the manuscript entitled **"Hidden Non-Determinism in Large Language Model APIs: A Lightweight Provenance Protocol for Reproducible Generative AI Research"** for consideration for publication in the *Journal of Artificial Intelligence Research*.

This work addresses a timely and critical challenge in AI research: the reproducibility of studies that rely on large language model outputs. While the AI reproducibility crisis has been widely documented, existing experiment-tracking tools were not designed for the specific challenges of generative text outputs. The paper makes three contributions:

1. **A lightweight protocol** introducing novel documentation artifacts—Prompt Cards and Run Cards—built on the W3C PROV data model for machine-readable provenance graphs.
2. **An empirical evaluation** through 3,904 controlled experiments with eight models—three locally deployed (LLaMA 3 8B, Mistral 7B, Gemma 2 9B) and five API-served (GPT-4, Claude Sonnet 4.5, Gemini 2.5 Pro, DeepSeek Chat, Perplexity Sonar)—across four NLP tasks (extraction, summarization, multi-turn refinement, RAG extraction), 30 scientific abstracts, and five experimental conditions. We document a striking reproducibility gap between local and API-based inference: local models achieve near-perfect reproducibility (average single-turn EMR = 0.956; Gemma 2 9B attains perfect EMR = 1.000 across all tasks), while API-served models exhibit substantial hidden non-determinism (average single-turn EMR = 0.221), spanning a wide range from DeepSeek Chat (EMR = 0.800) to Gemini 2.5 Pro (multi-turn EMR = 0.010). This pattern is observed independently across *five* cloud providers and survives Holm-Bonferroni correction across 68 hypothesis tests.
3. **A reference implementation** in Python with all 3,904 run records, provenance documents, and analysis scripts publicly available for independent verification.

Statistical robustness is ensured through Holm-Bonferroni correction, Fisher's exact tests, bias-corrected bootstrap confidence intervals (10,000 resamples), and balanced subsample sensitivity analysis. Supplementary control experiments (200 additional runs using Ollama's `/api/chat` endpoint) show that the prompt-format difference between deployment paradigms does not explain the reproducibility gap, strengthening internal validity.

We believe this work is well-suited for JAIR for several reasons: (a) it addresses a foundational methodological issue affecting all AI research that uses generative models, directly complementing the reproducibility mechanisms recently adopted by JAIR (Gundersen et al., 2024); (b) it provides rigorous empirical evidence across eight models, four tasks, five independent API providers, and 68 hypothesis tests; (c) the protocol and tools are immediately practical for the AI research community;

and (d) the paper's scope—spanning reproducibility, provenance, and empirical methodology—aligns with JAIR's broad interest in AI foundations.

The protocol adds less than 1% overhead to inference time (approximately 4 KB per run) while providing complete audit trails, tamper detection via cryptographic hashing, and interoperable W3C PROV provenance graphs. We hope this work will contribute to raising the bar for reproducibility in generative AI research.

This manuscript has not been published elsewhere and is not under consideration by any other journal. All authors have approved the manuscript and agree with its submission to JAIR.

Suggested Reviewers:

1. **Odd Erik Gundersen** (Norwegian University of Science and Technology) — Leading expert on AI reproducibility; recently published the four-mechanism framework adopted by JAIR.
2. **Anya Belz** (Dublin City University) — Conducted systematic reviews of NLP reproducibility; expertise in evaluation methodology.
3. **Joelle Pineau** (McGill University / Meta FAIR) — Led the NeurIPS Reproducibility Program; pioneer of reproducibility checklists.
4. **Luc Moreau** (King's College London) — Co-editor of the W3C PROV specification; expertise in provenance data models.
5. **Jesse Dodge** (Allen Institute for AI) — Expert in ML reporting standards and experimental methodology.

Sincerely,

Lucas Rover and Yara de Souza Tadano
Programa de Pós-Graduação em Engenharia Mecânica
UTFPR – Universidade Tecnológica Federal do Paraná
Ponta Grossa, Paraná, Brazil
lucasrover@utfpr.edu.br,
yaratadano@utfpr.edu.br