JAIR Submission — Significance Statement and Related Prior JAIR Work

Lucas Rover and Yara de Souza Tadano

February 2026

Significance Statement (150 words)

Large language models are increasingly used in scientific research, yet their outputs are not deterministic—even under nominally identical settings. This paper documents, through 3,904 controlled experiments with eight models (three local, five API-served) across four tasks, that API-served models produce strikingly low reproducibility: Claude Sonnet 4.5 achieves only 2% exact match rate for summarization under greedy decoding, while locally deployed Gemma 2 9B achieves perfect 100%. This hidden non-determinism—observed independently across five cloud providers (OpenAI, Anthropic, Google, DeepSeek, Perplexity)—means that a substantial body of published LLM-based research may be non-reproducible without the authors' knowledge. We contribute a lightweight provenance protocol—built on W3C PROV, Prompt Cards, and Run Cards—that makes this variability visible, measurable, and auditable at negligible cost (<1% of inference time, 4 KB per run). The protocol, reference implementation, and all experimental data are publicly available.

Closest Prior JAIR Papers

1. Gundersen, Helmert, and Hoos (2024). "Improving Reproducibility in AI Research: Four Mechanisms Adopted by JAIR." *JAIR*, 81, 1019–1041.

This editorial describes JAIR's institutional mechanisms for reproducibility (checklists, structured abstracts, badges, reports). Our work **complements** this by providing the empirical evidence and technical infrastructure that these mechanisms require: while Gundersen et al. establish *what* should be documented, our protocol specifies *how* to document it for generative AI workflows—with cryptographic hashing, provenance graphs, and structured cards—and our experiments quantify *why* it matters by demonstrating that API-based LLMs exhibit hidden non-determinism invisible without systematic logging. Our expanded study (eight models, five providers, four tasks, 3,904 runs) strengthens the evidence base substantially beyond prior two-model studies.

2. Atil, Tekir, Dogan, and Barlas (2024). "Measuring Non-Determinism in Generative AI." *Under review / arXiv:2407.03436.*

Although not published in JAIR, Atil et al. is the closest prior empirical study. They measure non-determinism across five models and eight tasks using the Total Agreement Rate metric. Our work differs in four ways: (i) we provide a *protocol* for prospective documentation, not just post-hoc measurement; (ii) we directly compare local vs. API deployment on identical tasks across eight

models and five independent API providers, isolating the deployment paradigm as a variable; (iii) we extend beyond single-turn evaluation to include multi-turn refinement and retrieval-augmented generation; and (iv) we quantify the overhead of systematic logging, demonstrating its feasibility at scale.

3. Liang et al. (2023). "Holistic Evaluation of Language Models (HELM)." *Transactions on Machine Learning Research.*

HELM provides a comprehensive evaluation framework for language models across multiple dimensions. Our work is orthogonal: rather than evaluating model capabilities, we address whether evaluation results themselves are reproducible. The two approaches are complementary—HELM benchmarks could adopt our provenance protocol to ensure that reported scores are auditable and reproducible across runs.