Research Highlights

Hidden Non-Determinism in Large Language Model APIs:
A Lightweight Provenance Protocol for Reproducible Generative AI Research

Lucas Rover Yara de Souza Tadano
UTFPR – Universidade Tecnológica Federal do Paraná

February 2026

Key Findings

- **API models exhibit hidden non-determinism under greedy decoding.** Across five independent API providers (OpenAI, Anthropic, Google, DeepSeek, Perplexity), API-served models produce non-identical outputs even at temperature zero, with Exact Match Rates (EMR) ranging from 0.010 to 0.800.

- **Local models achieve near-perfect to perfect reproducibility.** Gemma 2 9B attains EMR = 1.000 (perfect bitwise match) across all four tasks. LLaMA 3 8B and Mistral 7B achieve EMR $\geq$ 0.840 under greedy decoding.

- **3-fold reproducibility gap confirmed across five providers.** Average single-turn EMR: local = 0.960 vs. API = 0.325. The gap holds in 100% of abstracts for summarization and 83% for extraction, with all primary local-vs-API comparisons surviving Holm-Bonferroni correction across 68 tests.

- **The gap extends to complex interaction regimes.** Multi-turn refinement and RAG extraction maintain high local-model reproducibility (EMR $\geq$ 0.880), while both API models tested—Claude Sonnet 4.5 (EMR = 0.040 multi-turn, 0.000 RAG) and Gemini 2.5 Pro (EMR = 0.010 multi-turn, 0.070 RAG)—exhibit near-zero reproducibility.

- **Comprehensive provenance logging adds negligible overhead.** Less than 1% of inference time and approximately 4 KB per run across all models profiled.

Novel Contributions

1. **Prompt Cards and Run Cards**: Structured documentation artifacts for generative AI experiments, with cryptographic hashing for tamper detection.

2. **W3C PROV integration**: Machine-readable provenance graphs linking every output to its full generation context.

3. **Largest controlled reproducibility study**: 3,904 experiments across 8 models (3 local + 5 API), 5 providers, 4 tasks, 30 abstracts, 5 conditions.

Practical Impact

- Researchers using API-served LLMs should never assume reproducibility without verification.
- Structured output formats (JSON extraction) improve reproducibility over open-ended generation.
- The protocol, reference implementation, and all data are publicly available at: `https://github.com/Roverlucas/genai-reproducibility-protocol`