Lucas Rover
UTFPR — Universidade Tecnológica Federal do Paraná
Ponta Grossa, Paraná, Brazil

February 27, 2026

The Editors
Nature Machine Intelligence

Dear Editors,

We are pleased to submit our manuscript entitled **"Same Prompt, Different Answer: Exposing the Reproducibility Illusion in Large Language Model APIs"** for consideration as an Article in *Nature Machine Intelligence*.

Summary. Large language models are increasingly used as research instruments across the sciences, yet a fundamental assumption—that identical API configurations yield identical outputs—has never been systematically tested. We report 4,104 controlled experiments across eight models from five major API providers (OpenAI, Anthropic, Google, DeepSeek, Perplexity) demonstrating that this assumption is false. Under temperature-zero "greedy" decoding, which documentation presents as deterministic, API-served models reproduce their own outputs only 22% of the time, while locally deployed models achieve 96%—a gap exceeding four-fold. Crucially, this non-determinism is invisible: outputs are semantically equivalent (BERTScore F1 > 0.97) but textually different.

Key novelty. Three findings distinguish this work. First, we show that the non-determinism is universal across providers rather than a quirk of any single API, persisting in multi-turn and retrieval-augmented generation workflows where Claude Sonnet 4.5 achieves zero exact matches across 50 RAG runs. Second, by running the same open-weight model (LLaMA 3 8B) both locally and via a cloud API (Together AI), we isolate the cause: production infrastructure complexity (tensor parallelism, speculative decoding, dynamic batching), not cloud deployment itself. Third, we provide a lightweight provenance protocol grounded in W3C PROV that adds less than 1% overhead and makes this invisible variation detectable, auditable, and attributable—implementing the transparency standards that recent *Nature Machine Intelligence* editorials have advocated ("What is in your LLM-based framework?", 2024; "Multi-agent AI systems need transparency", 2026).

Timeliness and impact. As LLMs become embedded in scientific workflows—from peer review (*Nat. Mach. Intell.* 2026) to agentic science (*Nat. Mach. Intell.* 2025)—the hidden non-determinism we document means that a substantial portion of published LLM-based research may contain non-reproducible results without the authors' knowledge. Our findings are immediately actionable: the provenance protocol and minimum reporting checklist we propose can be adopted by any research group at negligible cost. The reference implementation, all 4,104 run records, and analysis scripts are publicly available at https://

github.com/Roverlucas/genai-reproducibility-protocol under open licences.

Fit with *Nature Machine Intelligence*. This work speaks directly to your readership. It addresses a problem (hidden non-determinism) that affects every researcher using API-based LLMs, quantifies it with a scale and rigour not previously reported, and offers a practical solution. The paper cites and builds upon recent *Nature Machine Intelligence* contributions on LLM evaluation reliability, agentic science, multi-agent transparency, ethical guidelines, psychometric assessment, and LLM frameworks.

We confirm that this manuscript has not been published elsewhere and is not under consideration by any other journal. All authors have approved the manuscript and agree to its submission.

Yours sincerely,

Lucas Rover
Programa de Pós-Graduação em
Engenharia Mecânica
UTFPR — Universidade Tecnológica
Federal do Paraná
Ponta Grossa, Paraná, Brazil
lucasrover@utfpr.edu.br

On behalf of all authors:
Lucas Rover, Eduardo Tadeu Bacalhau, Anibal Tavares de Azevedo, and Yara de Souza Tadano