Lucas Rover
Federal University of Technology — Paraná (UTFPR)
Ponta Grossa, Paraná, Brazil

February 28, 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. This non-determinism is invisible to users: 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 produces 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) rather than 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.

Timeliness and relevance to *Nature Machine Intelligence*. The reliability of large language models is a topic of growing concern in your journal. Kumar et al. recently examined when LLMs can be trusted as judges of empathic communication (*Nat. Mach. Intell.* **8**, 173–185, 2026), and Ciriello highlighted the troubling consequences of generative AI suspicion in academic publishing (*Nat. Mach. Intell.*, 2026, doi:10.1038/s42256-026-01178-z). Earlier editorials in *Nature Machine Intelligence* called for transparency in LLM-based frameworks (2024) and multi-agent AI systems (2026). Our manuscript contributes directly to this line of inquiry by quantifying, for the first time, the extent to which the deterministic settings documented by API providers fail to deliver determinism in practice.

The scope of this problem is substantial. ChatGPT now serves over 800 million weekly active users, Gemini surpasses 750 million monthly active users, and Claude serves approximately 19 million users with 190% year-over-year growth. Collectively, more than 1.5 billion people interact with these models, the vast majority through APIs that our study shows do not guarantee reproducible outputs. This creates a reliability gap that extends well beyond academic research into clinical decision support, environmental monitoring, engineering design, and automated data analysis, domains where users rely on model outputs without understanding the non-deterministic mechanisms behind them.

Positioning relative to existing work. Two recent studies have examined aspects of LLM non-determinism: Atil et al. (2024) demonstrated that "deterministic" API settings do not guarantee determinism for ChatGPT, and Yuan et al. (NeurIPS 2025) investigated the numerical sources of non-determinism in LLM inference. Our work is broader in scope and more immediately applicable. We systematically compare five API providers and three local deployments, extend the analysis to multi-turn and RAG workflows, provide a causal attribution framework through a quasi-isolation experiment, and deliver a ready-to-use provenance protocol. The statistical rigor of our study (10,000-resample bootstrap confidence intervals, Holm–Bonferroni correction across 68 tests, Cliff's delta effect sizes) establishes the findings on solid ground, which we consider essential given the urgency and contemporaneity of the topic. We note that several of our references are preprints; this reflects the rapid pace of the field, where foundational studies on LLM inference behavior are only now being reported.

Citation potential and data sharing. Given that the manuscript addresses a problem affecting every researcher who uses API-based language models, we believe it has strong citation potential across disciplines. Upon acceptance, we intend to submit the full provenance protocol, reference implementation, and all 4,104 experimental records to Protocol Exchange, making the methodology independently accessible and amplifying its dissemination. The GitHub repository (https://github.com/Roverlucas/genai-reproducibility-protocol) currently private to preserve the review process, will be made fully public upon publication under open licenses (MIT for code, CC-BY 4.0 for data).

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
Graduate Program in Mechanical Engineering
Federal University of Technology — Paraná (UTFPR)
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