Lucas Rover

Federal University of Technology — Paraná (UTFPR)

Ponta Grossa, Paraná, Brazil

February 28, 2026

The Editors

*Nature Machine Intelligence*

Dear Editors,

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

More than 1.5 billion people now interact weekly with large language models through commercial APIs. ChatGPT alone serves over 800 million weekly users, Gemini surpasses 750 million monthly, and Claude reaches 19 million with 190% annual growth. Every major API provider offers a temperature parameter that, when set to zero, is documented as producing deterministic outputs. Researchers across medicine, engineering, environmental science, and data analysis depend on this guarantee whenever they report LLM-generated results. Our manuscript shows that the guarantee does not hold.

We conducted 4,104 controlled experiments across eight models and five API providers. Under temperature-zero greedy decoding with fixed random seeds, locally deployed models reproduced their own outputs 96% of the time, yet API-served models achieved only 22%. The gap exceeds four-fold, and it is invisible to users: outputs remain semantically equivalent (BERTScore F1 > 0.97) while differing textually, so a researcher reading two responses side by side would not notice the variation. A quasi-isolation experiment in which the same model weights were served through a cloud provider with minimal infrastructure overhead recovered near-local reproducibility, attributing the variation to production infrastructure complexity (tensor parallelism, dynamic batching, speculative decoding) rather than cloud deployment itself. We complement this diagnosis with a lightweight provenance protocol that adds less than 1% computational overhead and makes the hidden variation detectable.

Prior work addressed this problem only partially. Atil et al. (2024) tested ChatGPT in isolation; Yuan et al. (NeurIPS 2025) focused on numerical precision in a single provider. Our study is the first to span five providers and three local deployments, extend the analysis to multi-turn and retrieval-augmented generation workflows, and deliver both causal attribution and a ready-to-use mitigation tool, all validated with 10,000-resample bootstrap confidence intervals and Holm–Bonferroni correction across 68 tests. Several of our references are preprints, reflecting a field where foundational studies on LLM inference behavior are only now emerging; our contribution grounds these observations with the statistical rigor that reliable AI research demands.

We believe this work aligns closely with *Nature Machine Intelligence*'s ongoing engagement with LLM re-

liability. Kumar et al. recently examined when LLMs can be trusted as evaluators (*Nat. Mach. Intell.* **8**, 173–185, 2026), and Ciriello drew attention to the troubling rise of generative AI suspicion in scholarly publishing (*Nat. Mach. Intell.*, 2026, doi:10.1038/s42256-026-01178-z). Your editorials have called for transparency in LLM frameworks (2024) and multi-agent systems (2026). Our findings speak directly to these concerns: if the outputs of the models most widely used in research cannot be reproduced, the credibility of every downstream study is at stake.

Upon acceptance, we will deposit the full protocol and all 4,104 experimental records in Protocol Exchange and release the GitHub repository (https://github.com/Roverlucas/genai-reproducibility-protocol) under open licenses (MIT for code, CC-BY 4.0 for data). This manuscript has not been published elsewhere and is not under consideration by any other journal. All authors have approved the 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