Same Prompt, Different Answer: Exposing the Reproducibility Illusion in Large Language Model APIs

Lucas Rover[1*†], Eduardo Tadeu Bacalhau[2], Anibal Tavares de Azevedo[3], Yara de Souza Tadano[1]

[1*]Graduate Program in Mechanical Engineering, Federal University of Technology — Paraná (UTFPR), Ponta Grossa, Paraná, Brazil.
[2]Research Group of Technology Applied to Optimization (GTAO), Federal University of Paraná (UFPR), Curitiba, Paraná, Brazil.
[3]School of Applied Sciences (FCA), University of Campinas (Unicamp), Limeira, São Paulo, Brazil.

*Corresponding author(s). E-mail(s): lucasrover@utfpr.edu.br;
†Corresponding author.

Abstract

The same prompt sent twice to a large language model API under documented "deterministic" settings can return different answers, yet this variation is invisible to users. Here we report 4,104 controlled experiments across eight models and five API providers showing that, under temperature-zero greedy decoding with fixed seeds, API-served models reproduce their own outputs only 22.1% of the time (exact match rate), while locally deployed models achieve 95.6%, a gap exceeding four-fold. The non-determinism persists in multi-turn and retrieval-augmented generation workflows, where one model produces zero exact matches across 50 runs, yet remains hidden because outputs are semantically equivalent (BERTScore F1 > 0.97). A quasi-isolation experiment identifies production infrastructure complexity, rather than cloud deployment itself, as the driver. We provide a lightweight provenance protocol ($<$1% overhead) that makes this variation detectable, raising a reliability concern for the growing use of LLMs in medicine, physical sciences, and automated data analysis.

Keywords: Reproducibility, Large language models, Non-determinism, Provenance, API inference

Large language models (LLMs) are rapidly becoming standard research instruments. They encode clinical knowledge[1], assist diagnostic workflows[2], accelerate literature review and data extraction[26], and are increasingly integrated into peer review at scale[3]. A growing body of work deploys LLMs as evaluators of human communication[13], as psychometric assessment tools[14], and as components of multi-agent scientific pipelines[4]. When researchers use API-based models, they trust that identical configurations will yield identical outputs. API documentation reinforces this trust by offering a temperature parameter set to zero for "deterministic" behavior, a setting that should collapse the output probability distribution to a single token at each step. But this is an illusion.

The broader reproducibility crisis in science is well documented: over 70% of researchers have failed to reproduce another scientist's experiment[21], and the problem is acute in AI, where only 6% of papers provide sufficient information for reproduction[23]. The field faces its own "reproducibility crisis"[22], with growing concerns that AI may worsen the problem across disciplines[24]. Data leakage alone has been documented across 17 machine learning subfields[25], and leading AI journals have begun adopting structured reproducibility mechanisms in response[19]. Yet despite this attention to training-time reproducibility, the reproducibility of *inference*, the actual deployment of models to generate outputs, remains largely unexamined. This gap matters because inference is the stage at which LLMs interact with research data and produce the outputs that enter scientific records, policy documents, and clinical workflows. Existing experiment-tracking tools such as MLflow[50] were designed for training pipelines and numerical metrics, not for inference-time text-generation provenance.

We conducted 4,104 controlled experiments across eight models and five independent API providers and found that even under temperature = 0 with fixed random seeds, API-served models reproduce their own outputs only 22.1% of the time, while locally deployed open-weight models achieve 95.6% exact reproducibility, a gap exceeding four-fold (95% bootstrap CI for the ratio: 2.48–3.61×). This non-determinism is invisible to users: outputs are semantically equivalent (BERTScore F1 > 0.97) but textually different, meaning researchers cannot detect the variation by reading outputs. The phenomenon is not confined to single queries. As LLMs move into agentic pipelines that design experiments and generate hypotheses[4], this hidden variation compounds at every step. Recent editorials have underscored the urgency of transparency in multi-agent AI systems[5], yet the baseline reproducibility of the individual API calls has never been systematically quantified.

Here, we make four contributions. First, we reveal and quantify hidden non-determinism across all five major API providers tested (OpenAI, Anthropic, Google, DeepSeek, Perplexity), showing that it is a general property of cloud-served LLM APIs rather than a provider-specific anomaly. Second, we show that this non-determinism persists and amplifies in multi-turn and retrieval-augmented generation (RAG) workflows, where Claude Sonnet 4.5 produces zero exact matches across 50 RAG runs. Third, we demonstrate that cloud deployment *per se* does not cause non-determinism: the same LLaMA 3 8B architecture served via Together AI's cloud endpoint achieves near-local exact match rates, implicating production-infrastructure complexity (tensor

parallelism, speculative decoding, dynamic batching) rather than the cloud medium. Fourth, we provide a lightweight provenance protocol grounded in W3C PROV[6], extending Model Cards[17] and Datasheets[18] to the inference layer, that adds less than 1% overhead and makes this invisible variation visible, auditable, and attributable.

Results

API-based models fail to reproduce outputs under deterministic settings

We evaluated eight models across two core tasks, structured extraction (JSON output) and scientific summarization (free text), under greedy decoding (temperature = 0) with fixed random seeds, the configuration that API documentation presents as deterministic. The results reveal a stark divide (Table 1, Fig. 1).

Locally deployed models achieve near-perfect to perfect bitwise reproducibility. Gemma 2 9B produces exact match rate (EMR) = 1.000 [1.00, 1.00] across all tasks and conditions: every output is character-for-character identical across five repetitions, confirmed by SHA-256 hash comparison. LLaMA 3 8B attains EMR = 0.987 [0.96, 1.00] for structured extraction and 0.947 [0.89, 0.99] for summarization; Mistral 7B achieves 0.960 [0.88, 1.00] and 0.840 [0.72, 0.96], respectively. The minor deviations in LLaMA 3 and Mistral are attributable to a cold-start effect: post-hoc analysis reveals that in all seven non-unanimous abstract groups, the first repetition was the sole outlier, likely due to GPU cache state initialization on the first forward pass. Discarding a single warm-up inference yields EMR = 1.000 for both models (Extended Data Table 5).

API-served models tell a starkly different story. Under the same greedy decoding with fixed seeds, GPT-4 achieves EMR = 0.443 [0.32, 0.57] for extraction and 0.230 [0.16, 0.30] for summarization. Claude Sonnet 4.5 achieves 0.190 [0.05, 0.40] and 0.020 [0.00, 0.05], respectively; across 50 pairwise comparisons, effectively no two summarization outputs were character-identical. Perplexity Sonar falls further to EMR = 0.100 for extraction and 0.010 for summarization. The overall local average EMR is 0.956 versus 0.221 for API models, meaning that fewer than one in four API output pairs are identical under conditions documented as deterministic. This gap survives Holm–Bonferroni correction across 68 hypothesis tests (51 significant at $\alpha = 0.05$; Supplementary Section S9) and is confirmed by large Cliff's delta effect sizes ($\delta = 0.784$–0.896), paired t-test (Cohen's $d > 1.6$), and a balanced 10-abstract subsample analysis that controls for sample-size differences between models (local EMR = 0.953 vs. API EMR = 0.304, 3.1× gap; Extended Data Table 4). A chat-format control experiment (200 runs) confirmed that the prompt format (completion versus chat API) does not contribute to the observed variation (Supplementary Section S7).

Non-determinism varies substantially across providers

The five API providers span a wide reproducibility range, with an 80-fold difference between the most and least reproducible (Fig. 1). DeepSeek Chat achieves

the highest API reproducibility (EMR = 0.800 for extraction, 0.760 for summarization), approaching local-model performance. GPT-4 occupies an intermediate position (EMR = 0.443 for extraction, 0.230 for summarization), with the extraction–summarization gap suggesting that task output structure mediates reproducibility, as JSON-constrained outputs leave less room for lexical variation than free-text summaries. Claude Sonnet 4.5 (0.190/0.020) and Perplexity Sonar (0.100/0.010) occupy the low end, while Gemini 2.5 Pro (evaluated on multi-turn and RAG tasks only) achieves EMR = 0.010–0.070.

This variation is notable because all providers expose the same user-facing "deterministic" parameters (temperature zero, fixed seed where supported). The differences therefore reflect production serving infrastructure invisible to users, consistent with editorial calls for explicit model version tracking and prompt disclosure[7]. Structured JSON extraction constrains the output space more tightly than free-text summarization, leaving fewer viable token sequences and less opportunity for divergence. This task-dependence suggests that open-ended generation tasks (creative writing, hypothesis formulation) will likely exhibit even larger variation.

Perplexity Sonar's particularly low reproducibility reflects its search-augmented architecture, where real-time web retrieval introduces an additional source of variation that compounds model-internal non-determinism. This observation is particularly relevant as retrieval-augmented approaches become standard in scientific applications: the combination of non-deterministic generation with variable retrieval creates a multiplicative reproducibility challenge that researchers must account for.

Cloud deployment does not preclude reproducibility

A natural hypothesis is that cloud deployment itself (network latency, load balancing, shared hardware) causes the non-determinism. To test this, we designed a quasi-isolation probe: we evaluated the same LLaMA 3 8B architecture served via Together AI's cloud endpoint (INT4 quantisation) under identical prompts, seeds, and temperature as the local deployment. If cloud hosting were the driver, this deployment should exhibit API-like non-determinism.

It does not. The cloud-served LLaMA 3 achieves EMR = 1.000 $[1.00, 1.00]$ for extraction and 0.880 $[0.70, 1.00]$ for summarization, nearly identical to the local deployment on the same 10-abstract subset (EMR = 1.000 and 0.920, respectively; 95% bootstrap confidence intervals overlap). This provides evidence that cloud deployment *per se* does not cause non-determinism. The variability observed in GPT-4, Claude, and Gemini is instead consistent with the complexity of their production serving infrastructure.

Six well-documented mechanisms in distributed GPU inference can independently produce non-deterministic outputs even under greedy decoding (see Methods): non-associative floating-point arithmetic[15], mixed-precision accumulation in BF16/FP16[11], multi-GPU tensor parallelism[8], FlashAttention kernel non-determinism[16], dynamic batching with continuous request scheduling[10], and speculative decoding[9]. Our single-GPU local deployment eliminates mechanisms 3–6, and GGML Q4 integer arithmetic mitigates mechanism 2, explaining near-perfect

local reproducibility as a predicted consequence of the simpler execution environment. The Together AI result confirms that this is not an inherent limitation of cloud deployment: a well-controlled cloud serving environment can achieve near-local reproducibility. Deterministic execution modes are technically feasible, though they may entail performance trade-offs that current serving architectures prioritize differently.

Complex workflows amplify the reproducibility gap

Modern LLM applications rarely consist of single API calls. Multi-turn refinement dialogues and retrieval-augmented generation (RAG) pipelines are increasingly common in research workflows. To assess whether non-determinism compounds in these settings, we evaluated five models on a three-turn refinement task (extract, receive feedback, refine) and a RAG extraction task (extract structured fields from an abstract with a prepended retrieved context passage).

Local models maintain high reproducibility even under these more complex regimes (Fig. 2). Gemma 2 9B and Mistral 7B achieve perfect EMR = 1.000 for both multi-turn and RAG. LLaMA 3 8B shows EMR = 0.880 $[0.76, 1.00]$ for multi-turn and 0.960 $[0.88, 1.00]$ for RAG, slightly lower than single-turn, consistent with error propagation across dialogue turns where each response conditions the next.

Both API models exhibit near-zero reproducibility. Claude Sonnet 4.5 achieves EMR = 0.040 $[0.00, 0.08]$ for multi-turn and 0.000 $[0.00, 0.00]$ for RAG. Across 50 runs, not a single pair of RAG outputs was character-identical, with a mean normalized edit distance (NED) of 0.256. Gemini 2.5 Pro, despite supporting a seed parameter, achieves EMR = 0.010 $[0.00, 0.03]$ for multi-turn and 0.070 $[0.02, 0.13]$ for RAG. The convergence of two independent API providers, using different model architectures and serving stacks, on near-zero multi-turn reproducibility indicates that this is a general property of cloud-served inference for complex interaction regimes. This is concerning for agentic AI workflows, where LLMs are chained across multiple steps. If each step introduces independent non-deterministic variation, end-to-end reproducibility becomes unattainable without explicit provenance tracking at every node.

Outputs diverge textually but not semantically

If API outputs varied randomly, producing nonsensical or contradictory results, the problem would be easy to detect. Instead, our multi-level metric analysis reveals a subtler pattern (Fig. 3). Across all models and conditions, BERTScore F1 remains above 0.97 even when EMR approaches zero. ROUGE-L scores similarly remain high (>0.85 for all models), confirming substantial token-level overlap. This three-level dissociation (low bitwise identity, moderate surface divergence, high semantic preservation) defines "hidden" non-determinism: same meaning, different words. A researcher reading two outputs from the same prompt would judge them equivalent; only systematic comparison reveals that they are textually distinct.

Yet this hidden variation has practical consequences that extend beyond surface-level differences. A field-level divergence analysis of GPT-4 structured extraction outputs reveals that 100% of non-identical output pairs (24 of 30 abstract groups) differ in at least one conclusion-relevant field: objective, method, or key result (Extended

Data Table 7). The `key_result` field diverges in 67% of groups, and `method` in 57%, meaning the textual variation is not limited to formatting or filler words but affects the substantive content that researchers would use for downstream analysis. For automated evidence synthesis pipelines, where outputs are parsed programmatically rather than read by humans, even minor lexical differences can propagate to different extracted values, different aggregated statistics, and ultimately different conclusions.

The implications extend beyond individual studies. Recent work has shown that LLMs used as evaluators of empathic communication are sensitive to subtle input variations[13], and psychometric assessments of LLM behavior reveal consistency patterns that depend on model architecture and fine-tuning approach[14]. Non-determinism is therefore not a technical curiosity but a property that interacts with how LLMs are deployed as measurement instruments, one that current evaluation frameworks do not account for. For the growing body of work that uses LLMs as annotators, evaluators, or data extractors, our results indicate that the "instrument" itself introduces measurement noise that is uncharacterized and unreported in standard methodology sections.

Temperature paradox: greedy decoding is not greedy for APIs

Temperature is the primary user-controllable parameter governing output variability. For local models, it behaves as theory predicts: under the temperature sweep (Fig. 4), local models show a clean monotonic decline from near-perfect EMR at $t = 0$ to EMR ≈ 0 at $t = 0.7$. This confirms that local greedy decoding is truly greedy, collapsing the output distribution to a single deterministic sequence. API models, however, start from an already-low baseline at $t = 0$ and show a more complex pattern that violates the expected monotonicity.

Claude Sonnet 4.5 exhibits a non-monotonic response: extraction EMR *increases* from 0.067 at $t = 0.0$ to 0.700 at $t = 0.3$ before declining to 0.133 at $t = 0.7$. This counterintuitive result, where adding randomness *improves* reproducibility, suggests that Anthropic's $t = 0$ decoding path exposes more infrastructure-level stochasticity than a small positive temperature that activates a more stable sampling pathway. GPT-4 shows a less dramatic but qualitatively similar pattern: the difference between $t = 0$ and $t = 0.3$ is smaller for API models than for local models, because the API baseline is already far from deterministic. At $t = 0.7$, local and API models converge toward uniformly low EMR, but through different trajectories: local models decline monotonically from near-1.0, while API models follow a flatter curve from an already-degraded baseline.

This temperature paradox has important practical implications. Researchers who set temperature to zero expecting deterministic outputs are, for API models, not achieving the greedy decoding they intend. The temperature–reproducibility relationship for API models depends on provider-specific implementation details that are opaque to users and undocumented in API references[12], making it impossible for researchers to predict or control the degree of non-determinism in their experiments without empirical measurement.

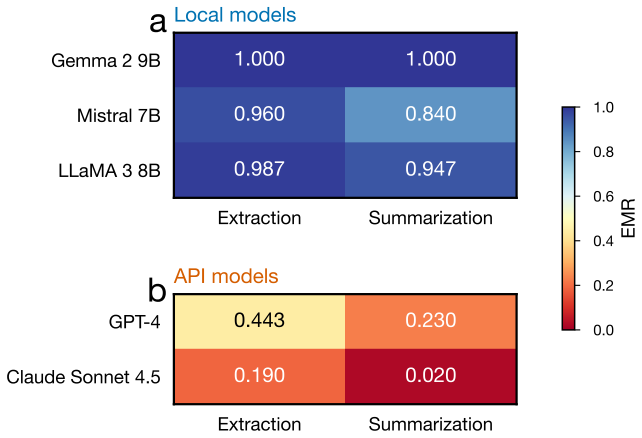

Fig. 1 Exact match rates under greedy decoding reveal a four-fold local–API gap. Heatmap of EMR for eight model deployments across extraction and summarization tasks under temperature = 0. Local models (top; blue) achieve EMR $\geq$ 0.840, with Gemma 2 9B at a perfect 1.000. API-served models (bottom; red) range from 0.800 (DeepSeek) to 0.010 (Perplexity). All values are computed under each model's representative greedy condition (C1 for local models and Claude; C2 for GPT-4; see Methods).

Table 1 Exact match rate under greedy decoding with 95% bootstrap confidence intervals. For local models, values reflect the fixed-seed condition (C1); for GPT-4, the variable-seed greedy condition (C2); for Claude, C1 (no seed parameter supported). Bootstrap: 10,000 resamples, percentile method. n = number of abstracts per group.

Model	Deployment	n	Extraction EMR	Summarisation EMR
Gemma 2 9B	Local	10	1.000 [1.00, 1.00]	1.000 [1.00, 1.00]
LLaMA 3 8B	Local	30	0.987 [0.96, 1.00]	0.947 [0.89, 0.99]
Mistral 7B	Local	10	0.960 [0.88, 1.00]	0.840 [0.72, 0.96]
DeepSeek Chat	API	10	0.800	0.760
GPT-4	API	30	0.443 [0.32, 0.57]	0.230 [0.16, 0.30]
Claude Sonnet 4.5	API	10	0.190 [0.05, 0.40]	0.020 [0.00, 0.05]
Perplexity Sonar	API	10	0.100	0.010
Local average			0.982	0.929
API average			0.383	0.255

Discussion

Our findings reveal a pervasive and previously invisible reproducibility gap in LLM-based research. Under the settings that API documentation presents as deterministic (temperature zero, fixed seed), API-served models fail to reproduce their own outputs approximately four out of five times. This gap persists across five independent providers, extends to multi-turn and RAG workflows, and survives rigorous statistical correction. The immediate implication is that a substantial portion of published research relying on API-based LLMs may contain non-reproducible results without the authors' knowledge, a concern amplified by the growing use of LLMs in peer review[3], hypothesis generation[4], and agentic pipelines where non-determinism compounds at every step.

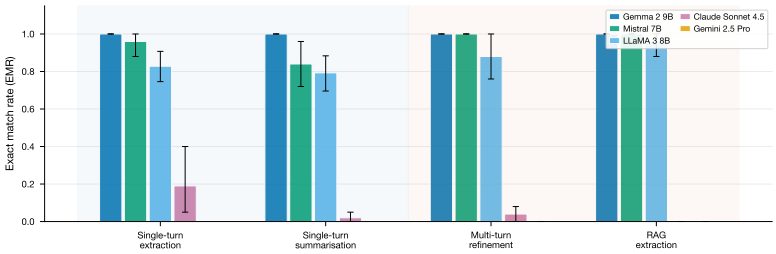

Fig. 2 Complex interaction regimes amplify the local–API reproducibility gap. EMR under greedy decoding (C1, $t = 0$) for five models across four scenarios: single-turn extraction, single-turn summarization, multi-turn refinement, and RAG extraction. Local models (Gemma 2, Mistral, LLaMA 3) maintain EMR ≥ 0.880 across all scenarios. Both API models (Claude Sonnet 4.5 and Gemini 2.5 Pro) exhibit near-zero EMR, with Claude achieving 0.000 for RAG (50 runs, zero exact matches). Error bars: 95% bootstrap CIs.

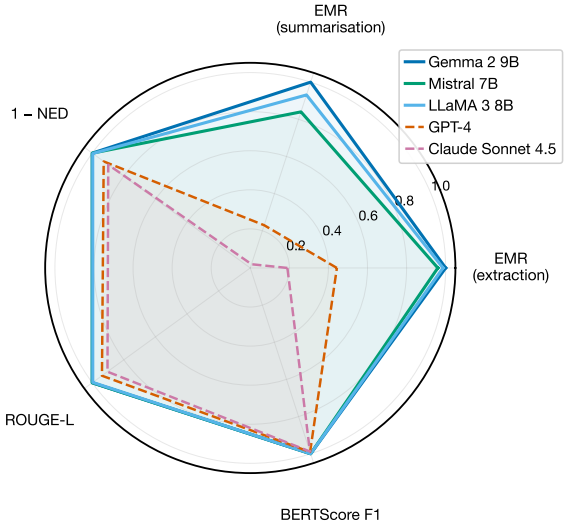

Fig. 3 API non-determinism is textual, not semantic. Three-level reproducibility profiles under greedy decoding. Local models (solid lines) occupy the outer region across all metrics. API models (dashed lines) show pronounced deficits in EMR and NED while maintaining BERTScore F1 > 0.97 (hidden non-determinism). Axes: EMR (extraction), EMR (summarization), 1−NED, ROUGE-L, BERTScore F1.

Prior work documented LLM non-determinism in isolated settings: inconsistent NLP benchmark outputs[27], failure of "deterministic" API settings[28], and non-deterministic code generation[29]. Our study advances beyond these observations by systematically comparing local versus API deployments across five providers, extending measurements to multi-turn and RAG workflows, and providing a causal attribution framework that distinguishes infrastructure complexity from cloud deployment.

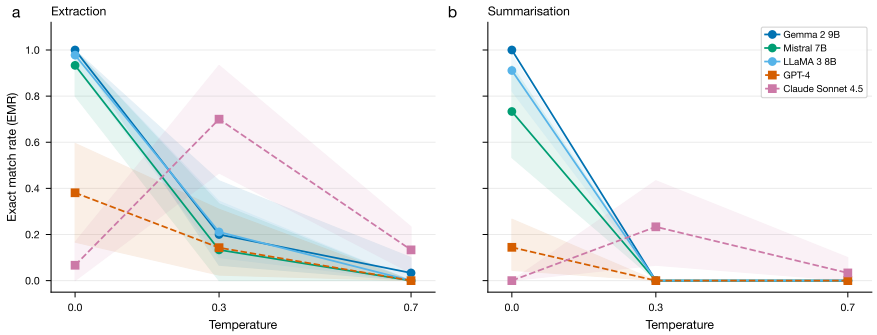

Fig. 4 Temperature paradox: greedy decoding is not greedy for API models. EMR versus temperature for five models (three local, two API). **a,** Extraction. **b,** Summarisation. Local models (solid lines) show the expected monotonic decline from near-perfect EMR at $t = 0$. API models (dashed lines) start from an already-low baseline. Claude Sonnet 4.5 (red dashed) exhibits a non-monotonic anomaly: extraction EMR *increases* from 0.067 at $t = 0$ to 0.700 at $t = 0.3$.

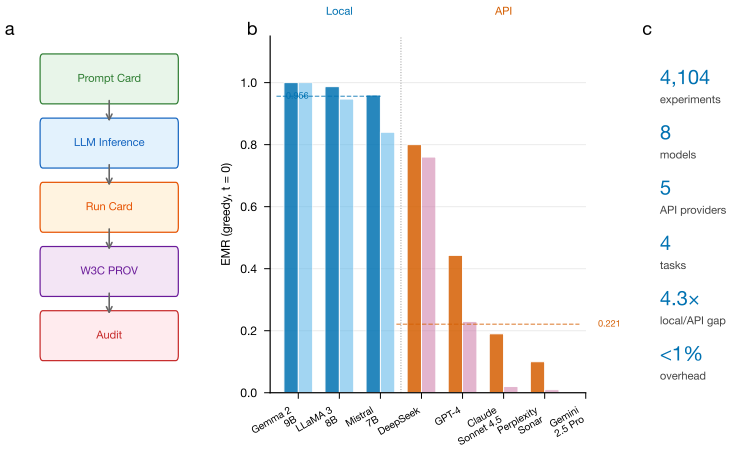

Fig. 5 Study overview and provenance protocol. Left: The provenance pipeline, from Prompt Card creation through Run Card logging and W3C PROV graph generation. **Centre:** EMR under greedy decoding for eight model deployments, illustrating the local–API reproducibility gap (local models green, EMR ≥ 0.840; API models red, EMR ≤ 0.800). **Right:** Key statistics: 4,104 experiments, 9 deployments, 7 execution environments, 4 tasks, <1% protocol overhead.

The Together AI quasi-isolation result provides a key mechanistic insight: because the same model architecture achieves near-local reproducibility via a different cloud API, the non-determinism observed in GPT-4, Claude, and Gemini is attributable to production serving complexity rather than cloud deployment itself. This suggests that reproducibility is a design choice that providers could address through deterministic execution modes or transparent documentation, rather than an inherent limitation. Documenting temperature zero as "deterministic" without disclosing infrastructure-level variation creates a false sense of reproducibility that may be more damaging than openly acknowledged stochasticity.

These findings also have regulatory implications. The EU AI Act[47] requires traceability for high-risk AI systems, and the NIST AI RMF[48] emphasizes transparency; hidden non-determinism directly undermines both. Ethical guidelines for LLM use in academic writing[20] similarly assume that documented configurations produce predictable outputs, an assumption our results show to be unwarranted.

Although the semantic preservation we observe (BERTScore F1 > 0.97) suggests that output quality is not corrupted, our field-level analysis shows that even "semantically equivalent" outputs differ in conclusion-relevant fields; the 78% non-match rate under "deterministic" settings represents a serious obstacle for regulatory submissions, systematic reviews, and automated pipelines.

The provenance protocol we provide offers a practical path forward. Grounded in W3C PROV[6] and adding less than 1% overhead (Extended Data Table 6), it creates auditable records linking every output to its generation context. Its key property is *differential diagnosis*: when all hashes match except the output hash, the divergence is automatically attributed to the generation process, a result confirmed across all 4,104 runs (Supplementary Section S4). This implements what recent editorials have advocated[7] and extends Model Cards[17] and Datasheets[18] to the inference layer. We propose a minimum reporting standard (model identity, parameters, reproducibility metrics, deployment mode, and output hashes) that our protocol automates at negligible cost (Supplementary Information).

Our study has several limitations. It covers eight models across four tasks but excludes code generation, mathematical reasoning, and creative writing. Input data comprises 30 English-language AI/ML abstracts; other languages and domains may show amplified effects. GPT-4 experiments used the `gpt-4-0613` snapshot; newer versions may differ, though the infrastructure-level mechanisms we identify are general. Multi-turn and RAG experiments include only two API providers; extending to others would strengthen generality. Sample sizes (10–30 abstracts per model) are well-powered for the large observed effects ($d > 1.6$) but may miss subtler phenomena.

Conclusions

We have shown that the "deterministic" settings offered by major LLM API providers do not deliver determinism in practice, creating a hidden reproducibility gap that affects any study relying on API-served models. The gap spans five independent providers, persists across multi-turn and retrieval-augmented generation workflows, and remains invisible to users because semantic equivalence masks textual divergence. By running the same open-weight model locally and via a cloud API, we isolate the cause: production infrastructure complexity, not cloud deployment itself, drives the non-determinism. Our lightweight provenance protocol, grounded in W3C PROV and requiring less than 1% overhead, offers a practical mechanism for making this variation visible, auditable, and attributable. Together with the minimum reporting standard we propose, it provides an actionable path toward reproducible LLM-based research.

These findings raise a broader reliability concern. Large language models are no longer confined to text generation; they are increasingly embedded in clinical decision support, environmental monitoring, engineering design, automated data collection and

analysis, and other domains where outputs directly inform decisions with consequences in the physical world. In such settings, the inability to reproduce or verify a model's output is not merely an academic inconvenience but a threat to the trustworthiness of the systems that depend on it. Our results demonstrate that, without explicit provenance tracking, the reliability of any LLM-based pipeline built on API inference cannot be assumed; it must be measured, documented, and continuously monitored. As the integration of AI into science and society accelerates, establishing the infrastructure for verifiable and reproducible model outputs is not a technical nicety but a scientific and societal necessity.

Methods

Protocol design

The provenance protocol addresses the question: what is the minimum metadata needed per generative AI run to enable reproducibility assessment, auditing, and provenance tracking? The FAIR principles[49] provide a foundation for data stewardship, yet no standard exists for documenting the full context of generative AI outputs. Existing experiment-tracking tools such as MLflow[50] were designed for training pipelines and numerical metrics, not for inference-time text-generation provenance. Data leakage analysis across 17 ML fields[25] underscores the urgency of structured documentation. It comprises four components.

Prompt Cards are versioned documentation artifacts capturing design rationale and metadata for prompt templates, including SHA-256 hash, version, task category, assumptions, limitations, and target models. The concept extends Model Cards[17] and Datasheets for Datasets[18] to the prompt layer.

Run Cards capture the complete execution context of a single generative AI run through 24 core fields organized into five groups: identification (run ID, prompt hash, prompt text), model context (name, version, weights hash), parameters (temperature, seed, decoding strategy, parameters hash), input/output (text and SHA-256 hashes), and execution metadata (environment fingerprint, timestamps, duration, logging overhead). For API models, optional extension fields capture provider-specific metadata: API request ID, response headers, resolved model version, and a `seed_status` field distinguishing between seeds that were "sent", "logged-only" (recorded for protocol parity but not transmitted, as with Claude), or "not-supported".

W3C PROV integration. Each experimental group is translated into a PROV-JSON document[6] expressing generation provenance as a directed graph of entities (Prompt, InputText, ModelVersion, InferenceParameters, Output), activities (Run-Generation), and agents (Researcher, SystemExecutor). The formal semantics of PROV enable automated traversal and comparison, identifying the exact factor that differs between non-identical outputs without custom parsing.

Reproducibility checklist. A 15-item checklist organized into four categories (Prompt Documentation, Model and Environment, Execution and Output, Provenance) enables self-assessment.

We formally define the protocol as a tuple $\mathcal{P} = (PC, RC, G, CL)$ and prove an *audit completeness* property: for 10 defined audit questions, every question is answerable if

and only if all field groups are populated. An ablation analysis confirms *minimality*: removing any field group renders at least one question unanswerable (Extended Data Table 6).

Models and infrastructure

We evaluate nine model deployments across three paradigms.

Local models (Ollama v0.15.5, Apple M4, 24 GB unified memory, macOS 14.6, Python 3.14.3): LLaMA 3 8B[30] (Q4_0), Mistral 7B[31] (Q4_0), Gemma 2 9B[32] (Q4_0). Weights hashes recorded via the Ollama API.

API-served models: GPT-4 (`gpt-4-0613`)[33] via OpenAI API; Claude Sonnet 4.5 (`claude-sonnet-4-5-20250929`)[34] via Anthropic API (urllib, no SDK); Gemini 2.5 Pro (`gemini-2.5-pro-preview-05-06`)[35] via Google AI Studio REST API (urllib); DeepSeek Chat via OpenAI-compatible API; Perplexity Sonar via Perplexity API.

Quasi-isolation probe: LLaMA 3 8B via Together AI (INT4, same architecture as local, `meta-llama/Llama-3-8b-chat-hf`).

Tasks

Four tasks spanning the output-structure spectrum: **Task 1, Scientific summarization:** produce a three-sentence summary of a scientific abstract. **Task 2, Structured extraction:** extract five fields (objective, method, key_result, model_or_system, benchmark) as JSON. **Task 3, Multi-turn refinement:** three-turn dialogue (extract, receive feedback, refine). **Task 4, RAG extraction:** structured extraction with a prepended retrieved context passage.

Input data

Thirty widely cited AI/ML abstracts (including Transformer, BERT, GPT-3, T5, and Chain-of-Thought[36–40]), varying in length (74–227 words) and technical complexity.

Experimental conditions

Five conditions systematically vary reproducibility factors: **C1** (fixed seed, greedy): $t = 0$, seed $= 42$, 5 repetitions; **C2** (variable seeds, greedy): $t = 0$, seeds $= \{42, 123, 456, 789, 1024\}$, 5 repetitions; **C3** (temperature sweep): $t \in \{0.0, 0.3, 0.7\}$, 3 repetitions each. Tasks 1–2 under all conditions for five models with full coverage; Tasks 3–4 under C1 for local models, Claude, and Gemini. DeepSeek and Perplexity: C1 only on Tasks 1–2. Together AI: C1 and C2 on Tasks 1–2. Grand total: 4,104 logged runs across 9 deployments and 7 execution environments.

For API models, the seed parameter is advisory (OpenAI[12]), absent (Anthropic), or empirically insufficient (Gemini); greedy decoding does not guarantee determinism[29].

Metrics

Exact Match Rate (EMR): fraction of all $\binom{n}{2}$ output pairs within a group that are character-identical. **Normalised Edit Distance (NED):** Levenshtein distance[42] normalized by the longer string. **ROUGE-L F1:** longest common subsequence overlap[41]. **BERTScore F1:** embedding-based semantic similarity[43]. For structured extraction, we additionally compute JSON validity rate, schema compliance, and field-level accuracy. Protocol overhead: logging time (ms), storage (KB), and overhead ratio (%).

Statistical analysis

All EMR values accompanied by 95% bootstrap confidence intervals (10,000 resamples, percentile method, per-abstract EMR)[44]. Primary comparisons: Wilcoxon signed-rank test (non-parametric) and paired t-test (parametric), with Holm–Bonferroni correction across 68 hypothesis tests. Effect sizes: Cohen's d (parametric) and Cliff's delta[45] (non-parametric). Power analysis confirms >0.95 for all primary comparisons at the observed effect sizes ($d > 1.6$)[46]. Balanced 10-abstract subsample analysis confirms robustness (local EMR = 0.953 vs. API EMR = 0.304, 3.1×).

Sources of non-determinism in distributed inference

Six well-documented mechanisms can independently produce non-deterministic outputs under greedy decoding in distributed GPU inference: (1) non-associative floating-point arithmetic[15]; (2) mixed-precision accumulation in BF16/FP16[11]; (3) tensor parallelism and all-reduce non-determinism[8]; (4) FlashAttention kernel non-determinism[16]; (5) dynamic batching and continuous request scheduling[10]; (6) speculative decoding[9]. Our single-GPU local deployment eliminates mechanisms (3)–(6) and GGML Q4 integer arithmetic mitigates (2), explaining near-perfect local reproducibility as a predicted consequence.

Protocol overhead

The protocol adds <1% overhead across all models profiled: mean logging time 21–30 ms versus inference latency of 4–182 s. Storage: ∼4 KB per run (16 MB total for 4,104 runs). The overhead is consistent across local and API deployment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability. All 4,104 run records (JSON), PROV-JSON provenance documents, Run Cards, Prompt Cards, input data (30 abstracts with DOIs), and generated figures are publicly available at https://github.com/Roverlucas/genai-reproducibility-protocol under CC-BY 4.0 license. Source data are provided with this paper.

Code availability. The reference implementation of the provenance protocol, all analysis scripts, and figure-generation code are publicly available at https://github.com/Roverlucas/genai-reproducibility-protocol under MIT License.

Supplementary information. The Supplementary Information (single PDF) contains: (S1) Full prompts for all four tasks; (S2) All 30 abstracts with DOIs; (S3) Retrieved contexts for RAG experiments; (S4) API payload documentation for all nine deployments; (S5) Protocol comparison table (our protocol vs. MLflow, W&B, DVC, OpenAI Evals, LangSmith); (S6) Ablation study: protocol minimality verification (10 audit questions $\times$ 8 field groups); (S7) Chat-format control experiment (200 runs); (S8) 15-item reproducibility checklist; (S9) Statistical test results (full Holm–Bonferroni table, 68 tests); (S10) Environment and provenance transparency documentation.

Acknowledgements. This work was supported by UTFPR—Universidade Tecnológica Federal do Paraná.

Declarations

- **Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
- **Author contributions:** L.R. conceived the study, designed the protocol, developed the software, conducted all experiments and analyses, and wrote the manuscript. E.T.B. contributed to experimental design and data analysis. A.T.d.A. contributed to methodology design and statistical analysis. Y.d.S.T. supervised the research, contributed to methodology design, and reviewed the manuscript.
- **Competing interests:** The authors declare no competing interests.
- **Correspondence:** Correspondence and requests for materials should be addressed to Lucas Rover (lucasrover@utfpr.edu.br).

Extended Data

Extended Data Table 1. Full three-level reproducibility assessment (EMR, NED, ROUGE-L, BERTScore F1) for all models under greedy decoding.

Extended Data Table 2. API versus local summary statistics with Cliff's delta effect sizes and bootstrap confidence intervals on the EMR ratio.

Extended Data Table 3. Temperature sweep: EMR at $t \in \{0.0, 0.3, 0.7\}$ for five models.

Extended Data Table 4. Balanced 10-abstract subsample robustness analysis.

Extended Data Table 5. Warm-up analysis: cold-start effect characterization.

Extended Data Table 6. Protocol overhead and ablation matrix (protocol minimality verification).

Extended Data Table 7. Conclusion divergence analysis: GPT-4 field-level differences.

Extended Data Fig. 1. Normalised Edit Distance comparison across all models and tasks.

Extended Data Fig. 2. Run Card schema and W3C PROV graph example for a single experimental group.

References

[1] Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).

[2] Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).

[3] Thakkar, N. *et al.* A large-scale randomized study of large language model feedback in peer review. *Nat. Mach. Intell.* **8**, online (2026).

[4] Xin, H., Kitchin, J. R. & Kulik, H. J. Towards agentic science for advancing scientific discovery. *Nat. Mach. Intell.* **7**, 1373–1375 (2025).

[5] Multi-agent AI systems need transparency [editorial]. *Nat. Mach. Intell.* **8**, 1 (2026).

[6] Moreau, L. & Missier, P. PROV-DM: The PROV Data Model. W3C Recommendation (2013); https://www.w3.org/TR/prov-dm/.

[7] What is in your LLM-based framework? [editorial]. *Nat. Mach. Intell.* **6**, 845 (2024).

[8] Shoeybi, M. *et al.* Megatron-LM: training multi-billion parameter language models using model parallelism. Preprint at https://arxiv.org/abs/1909.08053 (2019).

[9] Leviathan, Y., Kalman, M. & Matias, Y. Fast inference from transformers via speculative decoding. In *Proc. ICML* Vol. 202, 19274–19286 (PMLR, 2023).

[10] Kwon, W. *et al.* Efficient memory management for large language model serving with PagedAttention. In *Proc. 29th ACM SOSP* (2023).

[11] Yuan, J. *et al.* Understanding and mitigating numerical sources of nondeterminism in LLM inference. In *Advances in NeurIPS* Vol. 38 (2025).

[12] OpenAI. API Reference: Create Chat Completion—Seed Parameter. https://platform.openai.com/docs/api-reference/chat/create (2024).

[13] Kumar, A. *et al.* When large language models are reliable for judging empathic communication. *Nat. Mach. Intell.* **8**, 173–185 (2026).

[14] Serapio-García, G. *et al.* A psychometric framework for evaluating and shaping personality traits in large language models. *Nat. Mach. Intell.* **7**, 1954–1968 (2025).

[15] Higham, N. J. *Accuracy and Stability of Numerical Algorithms* 2nd edn (SIAM, 2002).

[16] Dao, T. *et al.* FlashAttention: fast and memory-efficient exact attention with IO-awareness. In *Advances in NeurIPS* Vol. 35 (2022).

[17] Mitchell, M. *et al.* Model cards for model reporting. In *Proc. FAccT* 220–229 (ACM, 2019).

[18] Gebru, T. *et al.* Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).

[19] Gundersen, O. E., Helmert, M. & Hoos, H. H. Improving reproducibility in AI research: four mechanisms adopted by JAIR. *J. Artif. Intell. Res.* **81**, 1019–1041 (2024).

[20] Porsdam Mann, S. *et al.* Guidelines for ethical use and acknowledgement of large language models in academic writing. *Nat. Mach. Intell.* **6**, 1272–1274 (2024).

[21] Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).

[22] Hutson, M. Artificial intelligence faces reproducibility crisis. *Science* **359**, 725–726 (2018).

[23] Gundersen, O. E. & Kjensmo, S. State of the art: reproducibility in artificial intelligence. *Proc. AAAI* **32**, 1644–1651 (2018).

[24] Ball, P. Is AI leading to a reproducibility crisis in science? *Nature* **624**, 22–25 (2023).

[25] Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**, 100804 (2023).

[26] Birhane, A. *et al.* Science in the age of large language models. *Nat. Rev. Phys.* **5**, 277–280 (2023).

[27] Chen, Y. *et al.* On the reproducibility of ChatGPT in NLP tasks. Preprint at https://arxiv.org/abs/2304.02554 (2023).

[28] Atil, B. *et al.* Non-determinism of "deterministic" LLM settings. Preprint at https://arxiv.org/abs/2408.04667 (2024).

[29] Ouyang, S. *et al.* An empirical study of the non-determinism of ChatGPT in code generation. *ACM Trans. Softw. Eng. Methodol.* **34**, 1–28 (2024).

[30] Grattafiori, A. *et al.* The LLaMA 3 herd of models. Preprint at https://arxiv.org/abs/2407.21783 (2024).

[31] Jiang, A. Q. *et al.* Mistral 7B. Preprint at https://arxiv.org/abs/2310.06825 (2023).

[32] Gemma Team *et al.* Gemma 2: improving open language models at a practical size. Preprint at https://arxiv.org/abs/2408.00118 (2024).

[33] Achiam, J. *et al.* GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).

[34] Anthropic. The Claude Model Family. https://www.anthropic.com/claude (2024).

[35] Reid, M. *et al.* Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. Preprint at https://arxiv.org/abs/2403.05530 (2024).

[36] Vaswani, A. *et al.* Attention is all you need. In *Advances in NeurIPS* Vol. 30 (2017).

[37] Devlin, J. *et al.* BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL* 4171–4186 (ACL, 2019).

[38] Brown, T. *et al.* Language models are few-shot learners. In *Advances in NeurIPS* Vol. 33, 1877–1901 (2020).

[39] Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**, 1–67 (2020).

[40] Wei, J. *et al.* Chain-of-thought prompting elicits reasoning in large language models. In *Advances in NeurIPS* Vol. 35, 24824–24837 (2022).

[41] Lin, C.-Y. ROUGE: a package for automatic evaluation of summaries. In *Proc. ACL Workshop on Text Summarization* 74–81 (2004).

[42] Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **10**, 707–710 (1966).

[43] Zhang, T. *et al.* BERTScore: evaluating text generation with BERT. In *Proc. ICLR* (2020).

[44] Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman & Hall/CRC, 1993).

[45] Romano, J. *et al.* Appropriate statistics for ordinal level data. In *Annual Meeting of the Florida Association of Institutional Research* (2006).

[46] Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Erlbaum, 1988).

[47] European Parliament. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act) (2024).

[48] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce (2023).

[49] Wilkinson, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

[50] Zaharia, M. *et al.* Accelerating the machine learning lifecycle with MLflow. *IEEE Data Eng. Bull.* **41**, 39–45 (2018).