

# Logging, Versioning, and Provenance in Generative AI Studies: A Protocol for Auditability and Scientific Reproducibility

LUCAS ROVER\*, UTFPR – Universidade Tecnológica Federal do Paraná, Brazil

YARA DE SOUZA TADANO, UTFPR – Universidade Tecnológica Federal do Paraná, Brazil

**Background:** Generative AI models produce non-deterministic outputs that vary across runs, even under nominally identical configurations. This variability threatens the reproducibility of studies that rely on large language model (LLM) outputs, yet most existing experiment-tracking tools were not designed for the specific challenges of text-generation workflows.

**Objectives:** We propose a lightweight, open-standard protocol for logging, versioning, and provenance tracking of generative AI experiments. The protocol introduces two novel documentation artifacts—Prompt Cards and Run Cards—and adopts the W3C PROV data model to create auditable, machine-readable provenance graphs linking every output to its full generation context.

**Methods:** We formalize the protocol and evaluate it empirically through 330 controlled experiments. These experiments employ two models—LLaMA 3 8B (locally deployed) and GPT-4 (cloud API)—on two NLP tasks (scientific summarization and structured extraction) across five experimental conditions that systematically vary the seed, temperature, and decoding strategy. We measure output variability using Exact Match Rate, Normalized Edit Distance, and ROUGE-L, and quantify the protocol’s own overhead in terms of time and storage.

**Results:** Under greedy decoding ( $t=0$ ), LLaMA 3 achieves perfect reproducibility on extraction (EMR = 1.000) and near-perfect on summarization (EMR = 0.840). In stark contrast, GPT-4 under identical greedy settings achieves only EMR = 0.520 for extraction and EMR = 0.200 for summarization, revealing substantial server-side non-determinism that is invisible without systematic logging. Increasing temperature to 0.7 eliminates exact matches for both models. The protocol adds a mean overhead of 33.56 ms per run (0.69% of inference time) and 4.17 KB per run record, totaling 4.87 MB for all 330 runs.

**Conclusions:** Our results demonstrate that (1) local inference is substantially more reproducible than API-based inference even under nominally identical parameters, (2) structured output tasks are inherently more reproducible than open-ended generation, (3) temperature is the dominant *user-controllable* factor affecting variability, and (4) comprehensive provenance logging can be achieved with negligible overhead. The protocol, reference implementation, and all experimental data are publicly available.

Additional Key Words and Phrases: reproducibility, generative AI, provenance, large language models, experiment tracking, W3C PROV

**JAIR Associate Editor:** Insert JAIR AE Name

## JAIR Reference Format:

Lucas Rover and Yara de Souza Tadano. 2026. Logging, Versioning, and Provenance in Generative AI Studies: A Protocol for Auditability and Scientific Reproducibility. *Journal of Artificial Intelligence Research* (2026), 18 pages. doi: [10.1613/jair.1.xxxxx](https://doi.org/10.1613/jair.1.xxxxx)

\*Corresponding Author.

Authors' Contact Information: Lucas Rover, ORCID: [0000-0001-6641-9224](https://orcid.org/0000-0001-6641-9224), lucasrover@utfpr.edu.br, UTFPR – Universidade Tecnológica Federal do Paraná, Programa de Pós-Graduação em Engenharia Mecânica, Curitiba, Paraná, Brazil; Yara de Souza Tadano, ORCID: [0000-0002-3975-3419](https://orcid.org/0000-0002-3975-3419), yaratadano@utfpr.edu.br, UTFPR – Universidade Tecnológica Federal do Paraná, Programa de Pós-Graduação em Engenharia Mecânica, Curitiba, Paraná, Brazil.



This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2026 Copyright held by the owner/author(s).

doi: [10.1613/jair.1.xxxxx](https://doi.org/10.1613/jair.1.xxxxx)

## 1 Introduction

2 The rapid adoption of large language models (LLMs) in scientific research has introduced a fundamental challenge:  
 3 how to ensure that studies relying on generative AI outputs are reproducible, auditable, and scientifically rigorous.  
 4 Unlike traditional computational experiments, in which deterministic algorithms produce identical results given  
 5 identical inputs, LLMs exhibit inherent variability in their outputs due to stochastic sampling, floating-point  
 6 non-determinism, and opaque model-versioning practices (Y. Chen et al. 2023; Zhu et al. 2023).

7 This reproducibility challenge is not merely theoretical. Baker (2016) reported that over 70% of researchers  
 8 have failed to reproduce another scientist’s experiment, a crisis that extends to AI research (Gundersen and  
 9 Kjensmo 2018; Hutson 2018). For generative AI specifically, the problem is compounded by several factors  
 10 unique to text-generation workflows: (1) the same prompt can yield semantically similar yet textually distinct  
 11 outputs across runs; (2) API-based models may undergo silent updates that alter behavior; (3) temperature and  
 12 sampling parameters create a high-dimensional space of possible outputs; and (4) no established standard exists  
 13 for documenting the full context needed to understand, audit, or reproduce a generative output.

14 Existing experiment-tracking tools such as MLflow (Zaharia et al. 2018), Weights & Biases (Biewald 2020),  
 15 and DVC (Miao et al. 2023) were designed primarily for training pipelines and numerical metrics. Although  
 16 valuable for their intended purposes, these tools lack features critical for generative AI studies: structured prompt  
 17 versioning, cryptographic output hashing for tamper detection, provenance graphs linking outputs to their full  
 18 generation context, and environment fingerprinting specific to inference-time conditions.

19 In this paper, we make three contributions:

- 20 (1) **A lightweight protocol** for logging, versioning, and provenance tracking of generative AI experiments.  
   The protocol introduces *Prompt Cards* and *Run Cards* as structured documentation artifacts, and adopts  
   the W3C PROV data model (Moreau and Missier 2013) for machine-readable provenance graphs.
- 21 (2) **An empirical evaluation** of both the protocol’s effectiveness and the reproducibility characteristics of  
   LLM outputs. Through 330 controlled experiments with LLaMA 3 8B (local) and GPT-4 (API) across two  
   tasks and five conditions, we quantify output variability using three complementary metrics and measure  
   the protocol’s overhead. Our results reveal a striking reproducibility gap between local and API-based  
   inference that is invisible without systematic logging.
- 22 (3) **A reference implementation** in Python that demonstrates the protocol’s practical applicability, together  
   with all experimental data, to facilitate adoption and independent verification.

23 The remainder of this paper is organized as follows. Section 2 reviews related work on reproducibility in AI and  
 24 experiment tracking. Section 3 formalizes the protocol design. Section 4 describes the experimental methodology.  
 25 Section 5 presents the empirical results. Section 6 discusses findings, limitations, and practical implications.  
 26 Section 7 concludes with directions for future work.

## 36 2 Related Work

### 37 2.1 Reproducibility in AI Research

38 The reproducibility crisis in AI has been documented extensively. Gundersen and Kjensmo (2018) surveyed 400  
 39 AI papers and found that only 6% provided sufficient information for full reproducibility. Pineau et al. (2021)  
 40 reported on the NeurIPS 2019 Reproducibility Program, which introduced reproducibility checklists and found  
 41 significant gaps between reported and actual reproducibility. Gundersen, Gil, et al. (2018) identified three levels  
 42 of reproducibility in AI—method, data, and experiment—and argued that all three are necessary for scientific  
 43 progress.

44 For generative AI specifically, Y. Chen et al. (2023) demonstrated that ChatGPT’s outputs on NLP benchmarks  
 45 exhibit non-trivial variability across identical queries, even with temperature set to zero. Zhu et al. (2023)  
 46

48 Table 1. Comparison of our protocol with existing reproducibility tools and frameworks for GenAI experiments. Checkmarks  
 49 ( $\checkmark$ ) indicate full support; tildes (~) indicate partial support; dashes (–) indicate no support.

Feature	Ours	MLflow	W&B	DVC	OpenAI Eval	LangSmith
Prompt versioning (Prompt Card)	$\checkmark$	–	~	–	~	~
Run-level provenance (W3C PROV)	$\checkmark$	–	–	–	–	–
Cryptographic output hashing	$\checkmark$	–	–	$\checkmark$	–	–
Seed & param logging	$\checkmark$	$\checkmark$	$\checkmark$	–	$\checkmark$	$\checkmark$
Environment fingerprinting	$\checkmark$	~	~	~	–	–
Model weights hashing	$\checkmark$	–	~	$\checkmark$	–	–
Overhead <1% of inference	$\checkmark$	~	~	N/A	N/A	~
Designed for GenAI text output	$\checkmark$	–	–	–	$\checkmark$	$\checkmark$
Open standard (PROV-JSON)	$\checkmark$	–	–	–	–	–
Local-first (no cloud dependency)	$\checkmark$	$\checkmark$	–	$\checkmark$	–	–

62  
 63  
 64 showed that reproducibility degrades further when tasks involve subjective judgment, such as social computing  
 65 annotations.

## 66 2.2 Experiment Tracking Tools

67 Several tools exist for tracking machine learning experiments, although none was designed specifically for  
 68 generative AI text-output workflows:

69 **MLflow** (Zaharia et al. 2018) provides experiment tracking, model packaging, and deployment. It logs parameters,  
 70 metrics, and artifacts, but focuses on training pipelines and numerical outcomes rather than text-generation  
 71 provenance.

72 **Weights & Biases** (Biewald 2020) offers experiment tracking with visualization dashboards. It supports prompt  
 73 logging but lacks structured prompt versioning, cryptographic output hashing, and provenance graph generation.

74 **DVC** (Miao et al. 2023) provides data versioning through git-like operations. While effective for dataset  
 75 management, it does not address run-level provenance or prompt documentation.

76 **OpenAI Eval** (OpenAI 2023) is a framework for evaluating LLM outputs against benchmarks. It provides  
 77 structured evaluation but is tightly coupled to OpenAI’s ecosystem and does not generate interoperable provenance  
 78 records.

79 **LangSmith** (LangChain 2023) offers tracing and evaluation for LLM applications. It captures detailed execution  
 80 traces but uses a proprietary format and requires cloud connectivity.

81 Table 1 provides a systematic feature-by-feature comparison of our protocol with these tools, highlighting the  
 82 gaps that motivate our work.

## 83 2.3 Provenance in Scientific Computing

84 Data provenance—the lineage of data through transformations—has a rich history in database systems and  
 85 scientific workflows (Herschel et al. 2017). The W3C PROV family of specifications (Moreau and Missier 2013)  
 86 provides a standardized data model for representing provenance as directed acyclic graphs of *entities*, *activities*,  
 87 and *agents*. Samuel and König-Ries (2022) applied provenance tracking to computational biology workflows,  
 88 demonstrating its value for reproducibility. However, to our knowledge, no prior work has applied W3C PROV  
 89 specifically to generative AI experiment workflows, in which the challenge involves not only tracking data  
 90 lineage but also capturing the stochastic generation context that determines output variability.

### 95 3 Protocol Design

96 Our protocol addresses the question: *What is the minimum set of metadata that must be captured for each generative*  
 97 *AI run to enable auditing, reproducibility assessment, and provenance tracking?* We address this question through  
 98 four complementary components.

#### 100 3.1 Scope and Design Principles

101 The protocol is designed around three principles:

- 103 (1) **Completeness:** Every factor that can influence a generative output must be captured—prompt text, model  
 104 identity and version, inference parameters, environment state, and timestamps.
- 105 (2) **Negligible overhead:** The logging process must not materially affect the experiment. We target <1%  
 106 overhead relative to inference time.
- 107 (3) **Interoperability:** All artifacts are stored in open, machine-readable formats (JSON, PROV-JSON) to  
 108 enable tool integration and long-term preservation.

#### 110 3.2 Prompt Cards

111 A *Prompt Card* is a versioned documentation artifact that captures the design rationale and metadata for a prompt  
 112 template used in experiments. Each Prompt Card contains:

- 114 • `prompt_id`: Unique identifier
- 115 • `prompt_hash`: SHA-256 hash of the prompt text, enabling tamper detection
- 116 • `version`: Semantic version number
- 117 • `task_category`: Classification of the task (e.g., summarization, extraction)
- 118 • `objective`: Natural-language description of what the prompt is designed to achieve
- 119 • `assumptions`: Explicit assumptions about inputs and expected behavior
- 120 • `limitations`: Known limitations or failure modes
- 121 • `target_models`: Models for which the prompt was designed and tested
- 122 • `expected_output_format`: Description of the expected output structure
- 123 • `interaction_regime`: Single-turn, multi-turn, or chain-of-thought
- 124 • `change_log`: History of modifications

125 Prompt Cards serve two purposes: they document design intent (supporting human understanding) and they  
 126 provide a citable, hashable reference for automated provenance tracking. The concept draws inspiration from  
 127 Model Cards ([Mitchell et al. 2019](#)) and Datasheets for Datasets ([Gebru et al. 2021](#)), extending the structured-  
 128 documentation paradigm to the prompt layer of the generative AI pipeline.

#### 130 3.3 Run Cards

131 A *Run Card* captures the complete execution context of a single generative AI run. Each Run Card records 23  
 132 fields organized into five groups:

- 134 (1) **Identification:** `run_id`, `task_id`, `task_category`, `prompt_card_ref`
- 135 (2) **Model context:** `model_name`, `model_version`, `weights_hash`, `model_source`
- 136 (3) **Parameters:** `inference_params` (temperature, top\_p, top\_k, max\_tokens, seed, decoding\_strategy),  
 137 `params_hash`
- 138 (4) **Input/Output:** `input_text`, `input_hash`, `output_text`, `output_hash`, `output_metrics`
- 139 (5) **Execution metadata:** `environment` (OS, architecture, Python version, hostname), `environment_hash`,  
 140 `code_commit`, `timestamps`, `execution_duration_ms`, `logging_overhead_ms`, `storage_kb`

<sup>142</sup> The separation of logging overhead from execution time is deliberate: it allows researchers to verify that the  
<sup>143</sup> protocol itself does not confound experimental measurements.

<sup>144</sup>

### <sup>145</sup> 3.4 W3C PROV Integration

<sup>146</sup> Each Run Card is automatically translated into a W3C PROV-JSON document ([Moreau and Missier 2013](#)) that  
<sup>147</sup> expresses the generation provenance as a directed graph. The mapping defines:

<sup>148</sup>

- **Entities:** Prompt, InputText, ModelVersion, InferenceParameters, Output, ExecutionMetadata
- **Activities:** RunGeneration (the inference execution)
- **Agents:** Researcher, SystemExecutor (the execution environment)

<sup>152</sup> PROV relations capture the causal structure:

- used: RunGeneration used Prompt, InputText, ModelVersion, InferenceParameters
- wasGeneratedBy: Output wasGeneratedBy RunGeneration
- wasAssociatedWith: RunGeneration wasAssociatedWith Researcher, SystemExecutor
- wasAttributedTo: Output wasAttributedTo Researcher
- wasDerivedFrom: Output wasDerivedFrom InputText

<sup>159</sup> This standardized representation enables automated reasoning about experiment provenance, including  
<sup>160</sup> detecting when two runs share identical configurations and identifying the specific factors that differ between  
<sup>161</sup> non-identical outputs.

<sup>162</sup>

### <sup>163</sup> 3.5 Reproducibility Checklist

<sup>164</sup> We provide a 15-item checklist organized into four categories—Prompt Documentation, Model and Environment,  
<sup>165</sup> Execution and Output, and Provenance—that researchers can use to self-assess the reproducibility of their  
<sup>166</sup> generative AI studies. The complete checklist is provided in Appendix A.

<sup>167</sup>

## <sup>168</sup> 4 Experimental Setup

<sup>169</sup> We designed a controlled experiment to simultaneously evaluate (a) the reproducibility characteristics of LLM  
<sup>170</sup> outputs under varying conditions and (b) the overhead imposed by our logging protocol.

<sup>171</sup>

### <sup>172</sup> 4.1 Models and Infrastructure

<sup>173</sup> We evaluate two models representing fundamentally different deployment paradigms:

<sup>174</sup> **LLaMA 3 8B** ([Grattafiori et al. 2024](#)): A locally deployed open-weight model served through Ollama ([Ollama 2024](#)) on an Apple M4 system with 24 GB unified memory running macOS 14.6. Local deployment provides complete control over the execution environment, eliminating confounding factors such as network latency, server-side batching, and silent model updates.

<sup>175</sup> **GPT-4** ([Achiam et al. 2023](#)): A cloud-based proprietary model accessed via the OpenAI API with controlled  
<sup>176</sup> seed parameters. Although we requested `model="gpt-4"`, the API returned `gpt-4-0613` as the resolved model  
<sup>177</sup> version in all 140 runs, which we recorded in the `model_id_returned` field of each run record. This represents  
<sup>178</sup> the typical deployment scenario where researchers have limited control over the inference environment. The  
<sup>179</sup> API introduces additional sources of variability: load balancing, server-side batching, potential model-version  
<sup>180</sup> updates, and floating-point non-determinism across different hardware.

<sup>181</sup>

### <sup>182</sup> 4.2 Tasks

<sup>183</sup> We evaluate two tasks that represent complementary points on the output-structure spectrum:

<sup>184</sup>

Table 2. Experimental design: conditions, parameters, and expected outcomes.

Cond.	Description	Temp.	Seed	Reps	Expected Outcome
C1	Fixed seed, greedy	0.0	42 (fixed)	5	Deterministic output
C2	Variable seeds, greedy	0.0	5 different	5	Near-deterministic
$C3_{t=0.0}$	Temp. baseline	0.0	per-rep	3	Deterministic
$C3_{t=0.3}$	Low temperature	0.3	per-rep	3	Low variability
$C3_{t=0.7}$	High temperature	0.7	per-rep	3	High variability

2 tasks = 10 groups per condition. Total: 330 logged runs (190 LLaMA 3 + 140 GPT-4).

**Task 1: Scientific Summarization.** Given a scientific abstract, produce a concise summary in exactly three sentences covering the main contribution, methodology, and key quantitative result. This is an open-ended generation task in which the model has considerable freedom in word choice and phrasing.

**Task 2: Structured Extraction.** Given a scientific abstract, extract five fields (objective, method, key\_result, model\_or\_system, benchmark) into a JSON object. This is a constrained generation task in which the output format is fixed and the model must select, rather than generate, content.

### 4.3 Input Data

We use five widely-cited scientific abstracts from landmark NLP papers: Vaswani et al. (2017) (Transformer), Devlin et al. (2019) (BERT), Brown et al. (2020) (GPT-3), Raffel et al. (2020) (T5), and Wei et al. (2022) (Chain-of-Thought). These abstracts vary in length (120–192 words), technical complexity, and the number of quantitative results reported, thereby providing diversity in the generation challenge.

### 4.4 Experimental Conditions

We define five conditions (Table 2) that systematically vary the factors hypothesized to affect reproducibility:

**C1 (Fixed seed, greedy decoding):** Temperature = 0, seed = 42 for all 5 repetitions. This represents the maximum-control condition and should yield deterministic outputs.

**C2 (Variable seeds, greedy decoding):** Temperature = 0, seeds = {42, 123, 456, 789, 1024}. This condition tests whether seed variation affects outputs when greedy decoding is used.

**C3 (Temperature sweep):** Three sub-conditions at  $t \in \{0.0, 0.3, 0.7\}$  with 3 repetitions each, using different seeds per repetition. This condition characterizes how temperature affects output variability.

For LLaMA 3, each task × abstract combination is evaluated under conditions C1 (5 runs), C2 (5 runs), and C3 (9 runs = 3 temperatures × 3 reps), yielding 19 runs per pair, or  $19 \times 5 \times 2 = 190$  runs. For GPT-4, C1 is omitted (seed control is less meaningful for API models), yielding C2 (5 runs) and C3 (9 runs) per pair, or  $14 \times 5 \times 2 = 140$  runs. **Total: 330 runs.**

### 4.5 Metrics

We measure output variability using three complementary metrics computed over all pairwise comparisons within each condition group:

**Exact Match Rate (EMR):** The fraction of output pairs that are character-for-character identical. EMR = 1.0 indicates perfect reproducibility; EMR = 0.0 indicates that no two outputs match exactly.

**Normalized Edit Distance (NED):** The Levenshtein edit distance (Levenshtein 1966) between each pair, normalized by the length of the longer string. NED = 0.0 indicates identical outputs; higher values indicate greater textual divergence.

236 Table 3. Output variability across experimental conditions for LLaMA 3 8B (local) and GPT-4 (API). Mean over 5 abstracts.  
 237 EMR = Exact Match Rate, NED = Normalized Edit Distance, ROUGE-L = word-level LCS F1.

239	Model	Task	Condition	EMR↑	NED↓	ROUGE-L↑
240	LLaMA 3 8B	Summarization	C1 (fixed seed, $t=0$ )	0.840	0.0148	0.9823
241			C2 (var. seeds, $t=0$ )	0.840	0.0148	0.9823
242			C3 ( $t=0.0$ )	0.733	0.0247	0.9706
243			C3 ( $t=0.3$ )	0.000	0.2289	0.7820
244			C3 ( $t=0.7$ )	0.000	0.4323	0.5550
245		Extraction	C1 (fixed seed, $t=0$ )	<b>1.000</b>	<b>0.0000</b>	<b>1.0000</b>
246			C2 (var. seeds, $t=0$ )	<b>1.000</b>	<b>0.0000</b>	<b>1.0000</b>
247			C3 ( $t=0.0$ )	<b>1.000</b>	<b>0.0000</b>	<b>1.0000</b>
248			C3 ( $t=0.3$ )	0.133	0.1883	0.8458
249			C3 ( $t=0.7$ )	0.000	0.3031	0.7447
250	GPT-4 (API)	Summarization	C2 (var. seeds, $t=0$ )	0.200	0.0718	0.9295
251			C3 ( $t=0.0$ )	0.000	0.0778	0.9248
252			C3 ( $t=0.3$ )	0.000	0.1721	0.8052
253			C3 ( $t=0.7$ )	0.000	0.3598	0.6143
254		Extraction	C2 (var. seeds, $t=0$ )	0.520	0.0343	0.9748
255			C3 ( $t=0.0$ )	0.333	0.0257	0.9770
256			C3 ( $t=0.3$ )	0.400	0.0679	0.9413
257			C3 ( $t=0.7$ )	0.000	0.1648	0.8557

259  
 260 **ROUGE-L F1:** The F1 score based on the longest common subsequence at the word level (Lin 2004). This  
 261 captures semantic similarity even when surface forms differ. ROUGE-L = 1.0 indicates identical word sequences.  
 262 For protocol overhead, we measure:

- 263
- 264 • **Logging time:** Wall-clock time spent on hashing, metadata collection, and file I/O, measured separately  
 265 from inference time.
  - 266 • **Storage:** Size of each run record (JSON) and total storage for all protocol artifacts.
  - 267 • **Overhead ratio:** Logging time as a percentage of total execution time.

## 269 5 Results

### 271 5.1 Output Variability

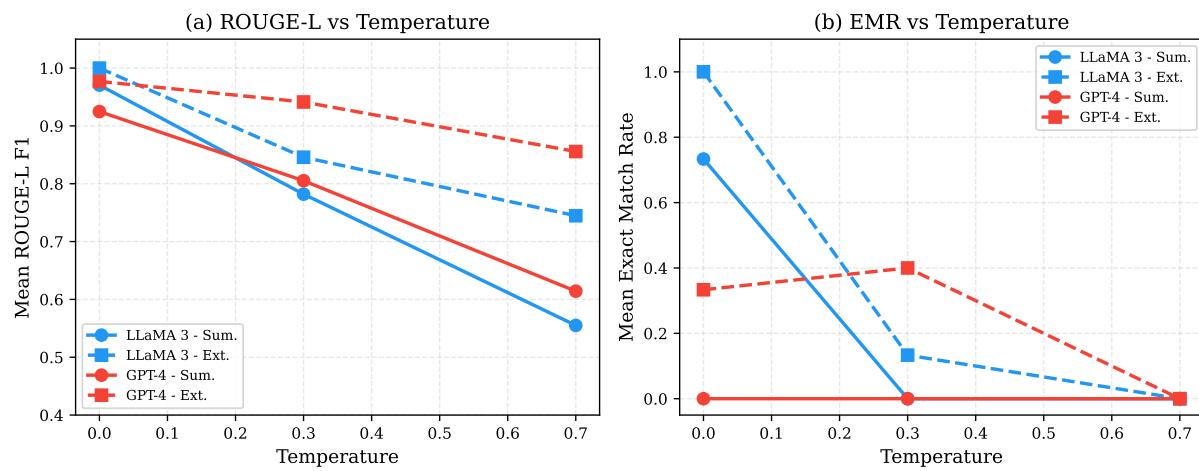
272 Table 3 presents the main variability results for both models, aggregated across all five abstracts.

273  
 274 **5.1.1 LLaMA 3 8B (Local Inference).** **Finding 1: Structured extraction achieves perfect reproducibility**  
 275 **under greedy decoding.** With  $t = 0$ , extraction produces EMR = 1.000 and NED = 0.0000 across all conditions  
 276 (C1, C2, C3 $_{t=0.0}$ ), meaning every output is character-for-character identical. Summarization achieves an EMR of  
 277 0.840 with NED = 0.0148, indicating near-perfect but not complete reproducibility.

278 **Finding 2: Seed variation has no effect under greedy decoding.** Conditions C1 and C2 produce identical  
 279 results despite using different seeds. With  $t = 0$ , the model always selects the highest-probability token, making  
 280 the seed irrelevant. This finding confirms that greedy decoding provides reliably deterministic inference with  
 281 locally deployed models.

283 Table 4. Reproducibility comparison: LLaMA 3 8B (local) vs. GPT-4 (API) under greedy decoding ( $t=0$ ). GPT-4 shows markedly  
 284 lower reproducibility due to server-side non-determinism ( $p < 0.05$  for EMR; see text for paired  $t$ -tests).

Task	Metric	LLaMA 3 8B	GPT-4
Summarization	EMR	0.840	0.200
	NED	0.0148	0.0718
	ROUGE-L	0.9823	0.9295
Extraction	EMR	1.000	0.520
	NED	0.0000	0.0343
	ROUGE-L	1.0000	0.9748



310 Fig. 1. Effect of temperature on output variability for both models. (a) ROUGE-L F1 decreases monotonically with temperature.  
 311 (b) Exact Match Rate: LLaMA 3 starts from near-perfect reproducibility at  $t = 0$ , whereas GPT-4 starts from a lower baseline;  
 312 however, both degrade at comparable rates with increasing temperature.

313  
 314  
 315 **5.1.2 GPT-4 (API Inference). Finding 3: API-based inference is substantially less reproducible than**  
 316 **local inference, even under greedy decoding.** This is the most striking result of our study. Under greedy  
 317 decoding ( $t = 0$ ) with controlled seeds, GPT-4 achieves only  $\text{EMR} = 0.200$  for summarization and  $\text{EMR} = 0.520$  for  
 318 extraction—compared to LLaMA’s 0.840 and 1.000, respectively, under the same C2 condition.

319 Table 4 highlights this reproducibility gap directly.

320 This gap is not due to parameter differences: both models use  $t = 0$  with the same seed. The variability  
 321 must originate from server-side factors that are invisible to the researcher: hardware-level floating-point non-  
 322 determinism across different GPU types in the serving cluster, request-batching and scheduling effects, and  
 323 potential silent model updates during the experimental window. *Without systematic logging, this non-determinism*  
 324 *would be entirely invisible.*

325  
 326 **5.1.3 Temperature Effects Across Models. Finding 4: Temperature is the dominant user-controllable factor**  
 327 **affecting variability.** Figure 1 shows the relationship between temperature and output variability for both  
 328 models.

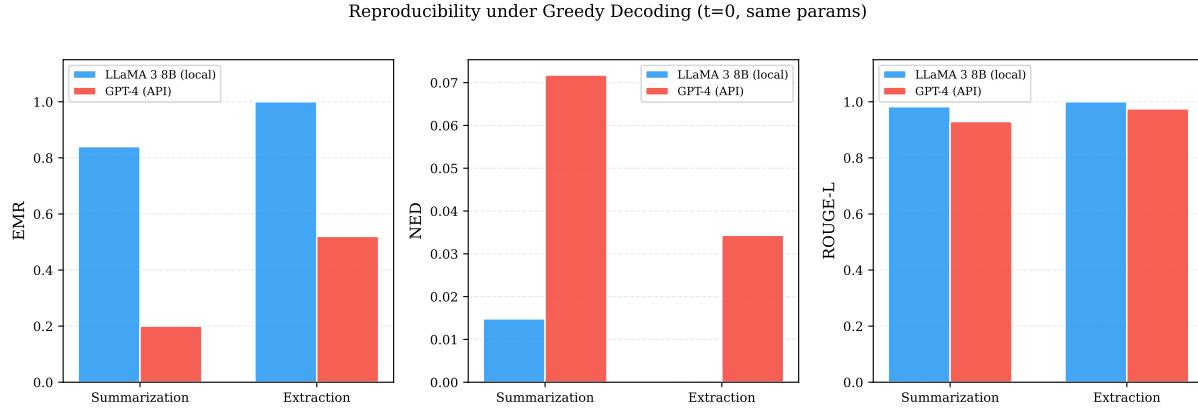


Fig. 2. Reproducibility under greedy decoding ( $t = 0$ ): LLaMA 3 8B (local) vs. GPT-4 (API). LLaMA 3 achieves near-perfect to perfect reproducibility, while GPT-4 shows measurable variability across all metrics, particularly for summarization.

For LLaMA 3, increasing temperature from 0 to 0.7 reduces ROUGE-L from 0.971 to 0.555 (summarization) and from 1.000 to 0.745 (extraction). For GPT-4, the same increase reduces ROUGE-L from 0.925 to 0.614 (summarization) and from 0.977 to 0.856 (extraction). The *relative* rate of degradation is comparable, but GPT-4 starts from a lower baseline owing to its inherent server-side non-determinism.

## 5.2 Cross-Model Comparison

Figure 2 provides a direct visual comparison of the two models under greedy decoding.

Figure 3 presents a comprehensive heatmap of EMR across all model-task-condition combinations.

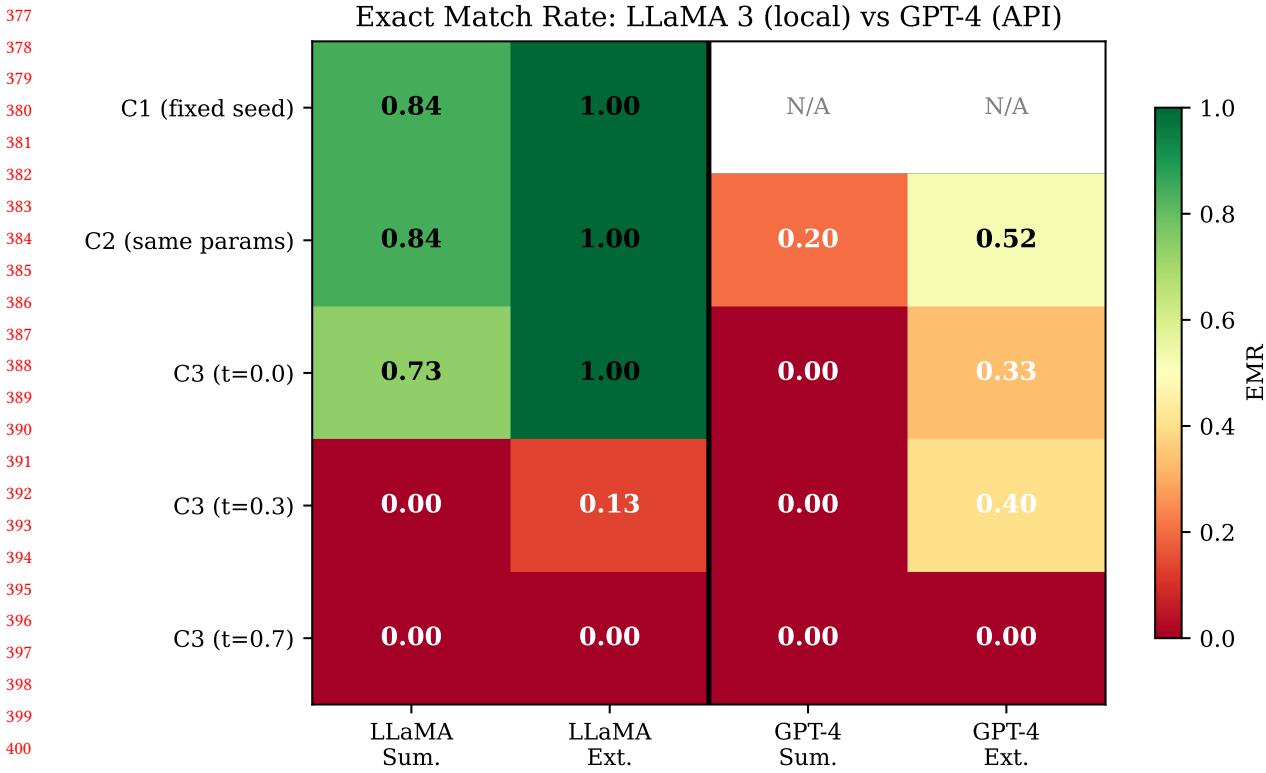
To quantify the reproducibility gap between local and API-based inference, we performed paired  $t$ -tests on per-abstract EMR values under condition C2 (greedy decoding,  $t=0$ ). For summarization, the difference is statistically significant:  $t(4) = 4.000$ ,  $p = 0.016$ , Cohen's  $d = 1.789$  (LLaMA 3 mean EMR = 0.840, 95% CI [0.568, 1.112]; GPT-4 mean EMR = 0.200, 95% CI [0.024, 0.376]). For extraction, the gap is also significant:  $t(4) = 3.639$ ,  $p = 0.022$ , Cohen's  $d = 1.627$  (LLaMA 3 EMR = 1.000; GPT-4 EMR = 0.520, 95% CI [0.154, 0.886]). Both effect sizes are very large ( $d > 1.2$ ), confirming that the reproducibility difference is not only statistically significant but practically meaningful. We note that no multiple-comparison correction was applied, as these were a small set of pre-specified comparisons ( $k = 6$ ); however, neither EMR  $p$ -value ( $p = 0.016$  and  $p = 0.022$ ) survives Bonferroni correction at the per-family threshold  $\alpha = 0.05/6 \approx 0.008$ , underscoring the limited statistical power inherent to  $n = 5$  abstracts ( $df = 4$ ). Notably, ROUGE-L differences did not reach significance ( $p > 0.05$ ), suggesting that while outputs differ at the exact-string level, their semantic content remains relatively stable.

## 5.3 Protocol Overhead

Table 5 presents the protocol's overhead metrics across all 330 runs.

The protocol adds a mean overhead of **33.56 ms** per run, representing **0.69%** of the mean inference time. This is well within our target of <1%. The overhead is dominated by SHA-256 hashing and environment metadata collection; JSON serialization and file I/O contribute minimally.

Storage overhead is similarly modest: each run record occupies approximately 4.17 KB, and the complete set of 330 run logs, 331 provenance documents, and 330 Run Cards totals 4.87 MB—less than a single high-resolution image.



402 Fig. 3. Heatmap of Exact Match Rate across all experimental conditions. The left columns (LLaMA 3) show high EMR under  
403 greedy decoding, while the right columns (GPT-4) show lower EMR even at  $t = 0$ . The vertical black line separates the two  
404 models.

405  
406 Figure 4 shows the overhead distribution broken down by model.  
407

## 408 6 Discussion

### 409 6.1 Implications for Reproducibility Practice

410 Our results yield several actionable recommendations for researchers conducting generative AI experiments:

411 **Use greedy decoding with local models for maximum reproducibility.** Under  $t = 0$  with LLaMA 3 (local),  
412 extraction achieved perfect reproducibility and summarization reached 84% EMR. This configuration should be  
413 the default for any study in which output consistency is critical.

414 **Be aware of API non-determinism.** Our most consequential finding is that GPT-4, even with  $t = 0$  and a  
415 fixed seed, produces substantially variable outputs (EMR = 0.200 for summarization). Researchers using API-based  
416 models should *never assume reproducibility* without verification, and should report multiple runs with variability  
417 metrics.

418 **Prefer structured output formats when possible.** The extraction task's consistently higher reproducibility  
419 across both models demonstrates that output-format constraints directly improve reproducibility. Researchers  
420 should consider whether their tasks can be reformulated as structured extraction rather than open-ended  
421 generation.

Table 5. Protocol overhead: logging time and storage costs for 330 runs (190 LLaMA 3 + 140 GPT-4).

Metric	Value	Unit
<i>Logging time overhead</i>		
Mean per run	33.56 ± 5.68	ms
Min / Max	12.85 / 51.20	ms
Total (330 runs)	11074	ms
Mean overhead ratio	0.694%	of inference time
Max overhead ratio	1.621%	of inference time
<i>Storage overhead</i>		
Run logs (330 files)	1382	KB
PROV documents (331 files)	1736	KB
Run Cards (330 files)	454	KB
Total output	4.87	MB

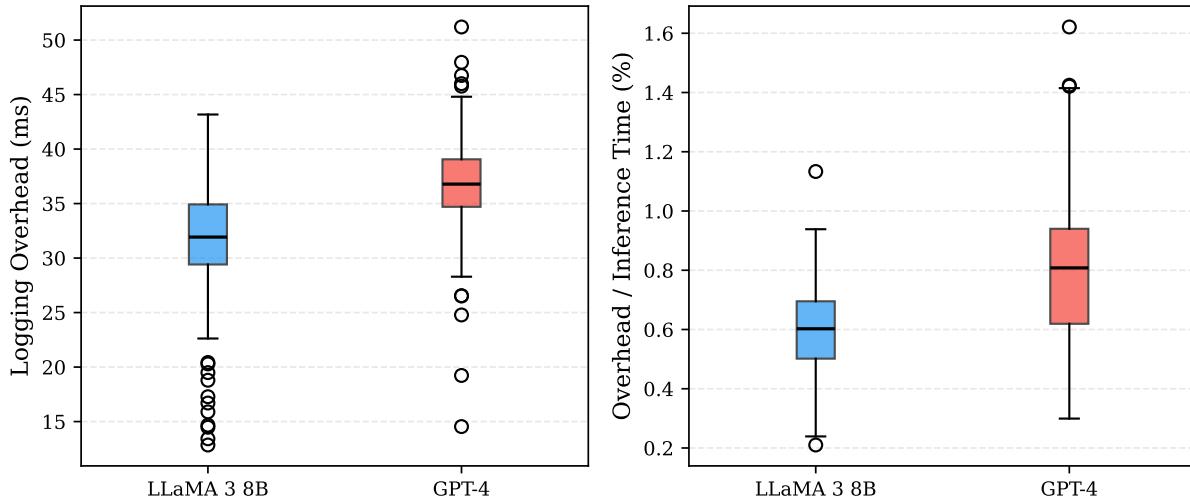


Fig. 4. Distribution of protocol overhead by model. Left: Absolute logging time (ms). Right: Overhead as a percentage of inference time. Overhead is comparable between local (LLaMA 3) and API (GPT-4) inference, consistently below 1.7%.

**Include warm-up runs for local models.** The per-abstract analysis revealed that the first inference call after model loading may differ from subsequent calls owing to cache initialization effects. Discarding the first run is a straightforward practice that improves measured reproducibility.

**Log comprehensively; the cost is negligible.** At 0.69% overhead and 4.17 KB per run, there is no practical reason not to apply comprehensive logging. The cost of not logging—namely, the inability to detect the kind of API non-determinism documented herein—far exceeds the protocol’s minimal requirements.

## 471 6.2 Local vs. API Inference: A Reproducibility Gap

472 The most significant finding of this study is the reproducibility gap between local and API-based inference. Under  
 473 nominally identical greedy decoding conditions, LLaMA 3 (local) achieves EMR = 1.000 for extraction while  
 474 GPT-4 (API) achieves only 0.520. For summarization, the gap is 0.840 vs. 0.200.

475 This gap has profound implications for the scientific use of API-based LLMs. *Without systematic logging, a*  
 476 *researcher using GPT-4 would have no way of knowing that their “deterministic” experiment produces different outputs*  
 477 *across runs.* The variability is not due to temperature or seed—it originates entirely from opaque server-side  
 478 factors. Our protocol makes this hidden non-determinism visible, measurable, and documentable.

## 480 6.3 Task-Dependent Reproducibility

481 The difference between summarization and extraction reproducibility under identical conditions—observed  
 482 consistently across both models—is, to our knowledge, the first empirical quantification of how task structure  
 483 affects LLM output reproducibility. This finding suggests a spectrum ranging from highly constrained tasks (struc-  
 484 tured extraction, classification) to open-ended tasks (summarization, dialogue), with the degree of output-space  
 485 constraint serving as a primary determinant. Notably, even GPT-4’s extraction task (EMR = 0.520) substantially  
 486 outperforms its summarization task (EMR = 0.200), confirming that this effect is not specific to any single model.  
 487

## 488 6.4 The Role of Provenance

489 The W3C PROV graphs generated by our protocol serve multiple purposes beyond simple audit trails:

- 490 (1) **Automated comparison:** By comparing PROV graphs of two runs, one can automatically identify which  
   491 factors differed (e.g., same prompt and model but different temperatures), enabling systematic diagnosis  
   492 of non-reproducibility.
- 493 (2) **Lineage tracking:** When outputs are used as inputs to downstream processes (e.g., summarization  
   494 outputs fed into a meta-analysis), the provenance chain can be extended to trace any final result back to  
   495 its full generation context.
- 496 (3) **Compliance:** For regulated domains (healthcare, legal, finance), PROV documents provide the formal  
   497 evidence trail required by audit standards.

## 500 6.5 Limitations

501 **Two models.** Our evaluation covers LLaMA 3 8B (local) and GPT-4 (API), representing two important deployment  
 502 paradigms. However, other models (e.g., Claude, Gemini, Mixtral, and smaller or larger LLaMA variants) may  
 503 exhibit different reproducibility characteristics. Future work should extend the evaluation to a broader model  
 504 suite.

505 **Two tasks.** While summarization and extraction represent distinct points on the output-structure spectrum,  
 506 they do not cover the full range of generative AI applications (e.g., dialogue, code generation, reasoning chains).  
 507 A broader task suite would strengthen the generalizability of our findings.

508 **English-only, narrow domain.** Our input data consists of five English scientific abstracts, all drawn from  
 509 high-profile NLP/ML papers. This selection may not be representative of scientific writing more broadly (e.g.,  
 510 biomedical, social science, or humanities abstracts), and the reproducibility characteristics we observe may differ  
 511 for other languages, domains, or document types.

512 **No multi-turn evaluation.** All experiments use single-turn interactions. Multi-turn dialogues introduce  
 513 additional variability through conversation history, which our current protocol logs but our experiments do not  
 514 evaluate.

**518 Privacy considerations.** The protocol’s environment metadata includes the machine hostname, which may  
**519** reveal institutional information in sensitive contexts. Deployments in privacy-sensitive settings should consider  
**520** anonymizing or omitting this field.

**521 Computational cost.** The total computational cost of our 330 experiments was modest: approximately 0.5  
**522** GPU-hours on a consumer laptop (Apple M4, 24 GB) for 190 LLaMA 3 runs, plus 140 API calls to GPT-4. We  
**523** estimate the carbon footprint to be negligible given this scale. Nevertheless, larger-scale adoption of the protocol  
**524** across thousands of runs would not materially increase energy consumption, as the logging overhead (34 ms per  
**525** run) is dominated by the inference itself.

526

## 527 6.6 Practical Costs and Adoption

**528** One concern with any new protocol is whether the adoption burden is justified. We address this concretely:

529

- 530 • Implementation effort:** Our reference implementation adds approximately 600 lines of Python (the  
 protocol core) to an existing workflow. Integration requires 3–5 function calls per run.
- 531 • Runtime cost:** 34 ms per run, negligible compared to inference times of seconds to minutes for typical  
 LLM calls.
- 532 • Storage cost:** 4 KB per run. Even at scale (10,000 runs), total storage is approximately 40 MB—less than a  
 single model checkpoint.
- 533 • Learning curve:** The protocol uses standard JSON and W3C PROV, requiring no specialized knowledge  
 beyond basic Python.

534

**535** Against these modest costs, the protocol provides complete audit trails, automated provenance graphs, tamper-  
**536** detectable outputs via cryptographic hashing, and structured metadata that enable systematic reproducibility  
**537** analysis.

538

## 539 7 Conclusion

540

**541** We presented a lightweight protocol for logging, versioning, and provenance tracking of generative AI experiments,  
**542** introducing Prompt Cards and Run Cards as novel documentation artifacts and adopting the W3C PROV data  
**543** model for machine-readable provenance graphs. Through 330 controlled experiments with LLaMA 3 8B (local)  
**544** and GPT-4 (API) on two NLP tasks, we demonstrated four key findings:

545

- 546 (1) Local inference is substantially more reproducible than API-based inference.** Under identical  
 greedy decoding settings, LLaMA 3 achieves EMR = 1.000 for extraction while GPT-4 achieves only 0.520,  
 revealing substantial server-side non-determinism that is invisible without systematic logging.
- 547 (2) Task structure is a primary determinant of reproducibility.** Structured extraction consistently  
 outperforms open-ended summarization across both models, with the JSON format constraint reducing  
 the model’s output space.
- 548 (3) Temperature is the dominant user-controllable factor.** Increasing from  $t = 0$  to  $t = 0.7$  reduces  
 ROUGE-L from 0.971 to 0.555 (LLaMA summarization) and from 0.977 to 0.856 (GPT-4 extraction), while  
 seed variation has no measurable effect under greedy decoding for local models.
- 549 (4) Comprehensive provenance logging adds negligible overhead:** 0.69% of inference time and 4.17 KB  
 per run, thereby removing any practical argument against systematic documentation.

550

**551** Future work will (i) expand the model suite to include Claude, Gemini, and open-weight models of varying  
**552** sizes; (ii) extend the task coverage to dialogue, code generation, and multi-turn interactions; and (iii) develop  
**553** automated reproducibility scoring based on provenance graph analysis.

554

**555** The reference implementation, all 330 run records, provenance documents, and analysis scripts are publicly  
**556** available to support adoption and independent verification.

557

## 565 Acknowledgments

566 This work was supported by UTFPR – Universidade Tecnológica Federal do Paraná. The experiments were  
 567 conducted using locally deployed open-weight models to ensure full reproducibility of the computational envi-  
 568 ronment.

## 570 Data Availability Statement

571 The reference implementation, all 330 run records (JSON), 331 PROV-JSON provenance documents, 330 Run  
 572 Cards, Prompt Cards, input data, analysis scripts, and generated figures are publicly available at:

573 <https://github.com/Roverlucas/genai-reproducibility-protocol>

574 The repository includes instructions for reproducing all experiments and regenerating all tables and figures from  
 575 the raw data.

## 577 Author Contributions

578 Following the CRediT (Contributor Roles Taxonomy) framework: **Lucas Rover**: Conceptualization, Methodol-  
 579 ogy, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Writing –  
 580 Review & Editing, Visualization, Project Administration. **Yara de Souza Tadano**: Supervision, Conceptualization,  
 581 Methodology, Writing – Review & Editing, Project Administration.

## 583 Conflict of Interest

584 The authors declare no conflicts of interest. This research was conducted independently at UTFPR with no  
 585 external funding from commercial AI providers. The use of OpenAI’s GPT-4 API was for research evaluation  
 586 purposes only and does not constitute an endorsement.

## 588 References

- 589 J. Achiam et al. 2023. *GPT-4 Technical Report*. (2023). arXiv: [2303.08774 \(cs.CL\)](https://arxiv.org/abs/2303.08774).
- 590 M. Baker. 2016. “1,500 Scientists Lift the Lid on Reproducibility.” *Nature*, 533, 7604, 452–454.
- 591 L. Biewald. 2020. *Experiment Tracking with Weights and Biases*. <https://www.wandb.com/>. (2020).
- 592 T. Brown et al.. 2020. “Language Models are Few-Shot Learners.” In: *Advances in Neural Information Processing Systems*. Vol. 33, 1877–1901.
- 593 Y. Chen, J. Li, X. Liu, and Y. Li. 2023. *On the Reproducibility of ChatGPT in NLP Tasks*. (2023). arXiv: [2304.02554 \(cs.CL\)](https://arxiv.org/abs/2304.02554).
- 594 J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 4171–4186.
- 595 T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM*, 64, 12, 86–92.
- 596 A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, et al.. 2024. *The LLaMA 3 Herd of Models*. (2024). arXiv: [2407.21783 \(cs.AI\)](https://arxiv.org/abs/2407.21783).
- 597 O. E. Gundersen, Y. Gil, and D. W. Aha. 2018. “On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications.” *AI Magazine*, 39, 3, 56–68.
- 598 O. E. Gundersen and S. Kjensmo. 2018. “State of the Art: Reproducibility in Artificial Intelligence.” *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, 1.
- 599 M. Herschel, R. Diestelkämper, and H. Ben Lahmar. 2017. “A Survey on Provenance: What for? What form? What from?” *The VLDB Journal*, 26, 6, 881–906.
- 600 M. Hutson. 2018. “Artificial Intelligence Faces Reproducibility Crisis.” *Science*, 359, 6377, 725–726.
- 601 LangChain. 2023. *LangSmith: A Platform for Building Production-Grade LLM Applications*. (2023). <https://smith.langchain.com/>.
- 602 V. I. Levenshtein. 1966. “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals.” *Soviet Physics Doklady*, 10, 8, 707–710.
- 603 C.-Y. Lin. 2004. “ROUGE: A Package for Automatic Evaluation of Summaries.” In: *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81.
- 604 J. Miao, J. Guo, J. Dougherty, and R. Dougherty. 2023. “DVC: Data Version Control for Machine Learning Pipelines.” *Software: Practice and Experience*.

611

- 612 M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. 2019. “Model Cards for Model  
613 Reporting.” In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 220–229.
- 614 L. Moreau and P. Missier. 2013. *PROV-DM: The PROV Data Model*. W3C Recommendation. World Wide Web Consortium. <https://www.w3.org/TR/prov-dm/>.
- 615 Ollama. 2024. *Ollama: Run Large Language Models Locally*. <https://ollama.com/>. (2024).
- 616 OpenAI. 2023. *OpenAI Eval: A Framework for Evaluating LLMs*. (2023). <https://github.com/openai/evals>.
- 617 J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and H. Larochelle. 2021. “Improving Reproducibility  
618 in Machine Learning Research: A Report from the NeurIPS 2019 Reproducibility Program.” *Journal of Machine Learning Research*, 22, 164,  
619 1–20.
- 620 C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. “Exploring the Limits of Transfer Learning  
621 with a Unified Text-to-Text Transformer.” *Journal of Machine Learning Research*, 21, 140, 1–67.
- 622 S. Samuel and B. König-Ries. 2022. “A Provenance-based Semantic Approach to Support Understandability, Reproducibility, and Reuse of  
623 Scientific Experiments.” *Journal of Biomedical Semantics*, 13, 1, 1–30.
- 624 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. “Attention is All You Need.” In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc.
- 625 J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. 2022. “Chain-of-Thought Prompting Elicits  
626 Reasoning in Large Language Models.” In: *Advances in Neural Information Processing Systems*. Vol. 35, 24824–24837.
- 627 M. Zaharia et al. 2018. “Accelerating the Machine Learning Lifecycle with MLflow.” *IEEE Data Engineering Bulletin*, 41, 4, 39–45.
- 628 Y. Zhu, P. Zhang, E. Haq, P. Hui, and G. Buchanan. 2023. *Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing  
Tasks*. (2023). arXiv: [2304.10145 \(cs.CL\)](https://arxiv.org/abs/2304.10145).

## 630 A Reproducibility Checklist

631 The following checklist is designed for self-assessment of reproducibility in generative AI studies. Each item  
632 maps to a specific field or artifact in our protocol.

### 634 Prompt Documentation

- 635 (1) Is the exact prompt text recorded and versioned? [Prompt Card: prompt\_text, prompt\_hash]
- (2) Are design assumptions and limitations documented? [Prompt Card: assumptions, limitations]
- (3) Is the expected output format specified? [Prompt Card: expected\_output\_format]
- (4) Is the interaction regime documented (single/multi-turn)? [Prompt Card: interaction\_regime]

### 640 Model and Environment

- 641 (5) Is the model name and version recorded? [Run Card: model\_name, model\_version]
- (6) Are model weights hashed for identity verification? [Run Card: weights\_hash]
- (7) Is the execution environment fingerprinted? [Run Card: environment, environment\_hash]
- (8) Is the source code version recorded? [Run Card: code\_commit]

### 647 Execution and Output

- 648 (9) Are all inference parameters logged? [Run Card: inference\_params]
- (10) Is the random seed recorded? [Run Card: inference\_params.seed]
- (11) Is the output cryptographically hashed? [Run Card: output\_hash]
- (12) Are execution timestamps recorded? [Run Card: timestamp\_start, timestamp\_end]
- (13) Is logging overhead measured separately? [Run Card: logging\_overhead\_ms]

### 654 Provenance

- 655 (14) Is a provenance graph generated per run? [PROV-JSON document]
- (15) Are provenance documents in an interoperable format? [W3C PROV standard]

## 659 B Run Card Schema

660 The complete Run Card schema, with data types and descriptions:

661

662 Listing 1. Run Card JSON schema (simplified).

```

663 {
664     "run_id": "string (unique identifier)",
665     "task_id": "string (task identifier)",
666     "task_category": "string (e.g., summarization)",
667     "prompt_hash": "string (SHA-256 of prompt)",
668     "prompt_text": "string (full prompt text)",
669     "input_text": "string (input to the model)",
670     "input_hash": "string (SHA-256 of input)",
671     "model_name": "string (e.g., llama3:8b)",
672     "model_version": "string (e.g., 8.0B)",
673     "weights_hash": "string (SHA-256 of weights)",
674     "model_source": "string (e.g., ollama-local)",
675     "inference_params": {
676         "temperature": "float",
677         "top_p": "float",
678         "top_k": "integer",
679         "max_tokens": "integer",
680         "seed": "integer|null",
681         "decoding_strategy": "string"
682     },
683     "params_hash": "string (SHA-256 of params)",
684     "environment": {
685         "os": "string",
686         "os_version": "string",
687         "architecture": "string",
688         "python_version": "string",
689         "hostname": "string",
690         "timestamp": "ISO 8601 datetime"
691     },
692     "environment_hash": "string (SHA-256)",
693     "code_commit": "string (git commit hash)",
694     "researcher_id": "string",
695     "affiliation": "string",
696     "timestamp_start": "ISO 8601 datetime",
697     "timestamp_end": "ISO 8601 datetime",
698     "output_text": "string (model output)",
699     "output_hash": "string (SHA-256 of output)",
700     "output_metrics": "object (task-specific)",
701     "execution_duration_ms": "float",
702     "logging_overhead_ms": "float",
703     "storage_kb": "float",
704     "system_logs": "string (raw system info)",
705     "errors": "array of strings"
706 }

```

**706 C Example PROV-JSON Document**

**707** An abbreviated example of a PROV-JSON document generated for a single summarization run:

**708**

**709 Listing 2. Abbreviated PROV-JSON for a summarization run.**

```

710 {
711   "prefix": {
712     "genai": "https://genai-prov.org/ns#",
713     "prov": "http://www.w3.org/ns/prov#"
714   },
715   "entity": {
716     "genai:prompt_c9644358": {
717       "prov:type": "genai:Prompt",
718       "genai:hash": "c9644358805b...",
719       "genai:task_category": "summarization"
720     },
721     "genai:model_llama3_8b": {
722       "prov:type": "genai:ModelVersion",
723       "genai:name": "llama3:8b",
724       "genai:source": "ollama-local"
725     },
726     "genai:output_590d0835": {
727       "prov:type": "genai:Output",
728       "genai:hash": "590d08359e7d..."
729     }
730   },
731   "activity": {
732     "genai:run_llama3_8b_sum_001_C1_rep0": {
733       "prov:type": "genai:RunGeneration",
734       "prov:startTime": "2026-02-07T21:54:34Z",
735       "prov:endTime": "2026-02-07T21:54:40Z"
736     }
737   },
738   "wasGeneratedBy": {
739     "_:wGB1": {
740       "prov:entity": "genai:output_590d0835",
741       "prov:activity": "genai:run_llama3_8b_..."
742     }
743   },
744   "used": {
745     "_:u1": {
746       "prov:activity": "genai:run_llama3_...",
747       "prov:entity": "genai:prompt_c9644358"
748     }
749   },
750   "agent": {
751     "genai:researcher_lucas_rover": {
752       "prov:type": "prov:Person",
753       "genai:affiliation": "UTFPR"
754     }

```

```
753    46  },
754    47  "wasAssociatedWith": {
755    48    "_:wAW1": {
756    49      "prov:activity": "genai:run_llama3_...",
757    50      "prov:agent": "genai:researcher_..."
758    51    }
759    52  }
760  }
```

761 Received February 2026  
762

763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799