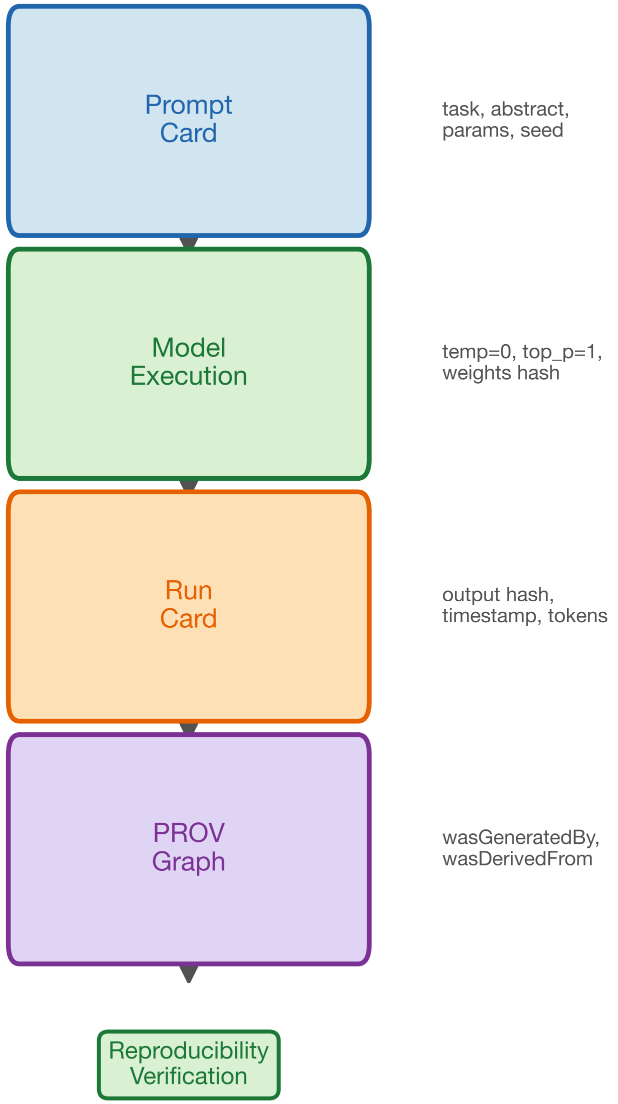
Hidden Non-Determinism in LLM APIs

A Lightweight Provenance Protocol for Reproducible Generative AI Research

Protocol Pipeline



Prompt Card — task, abstract, params, seed

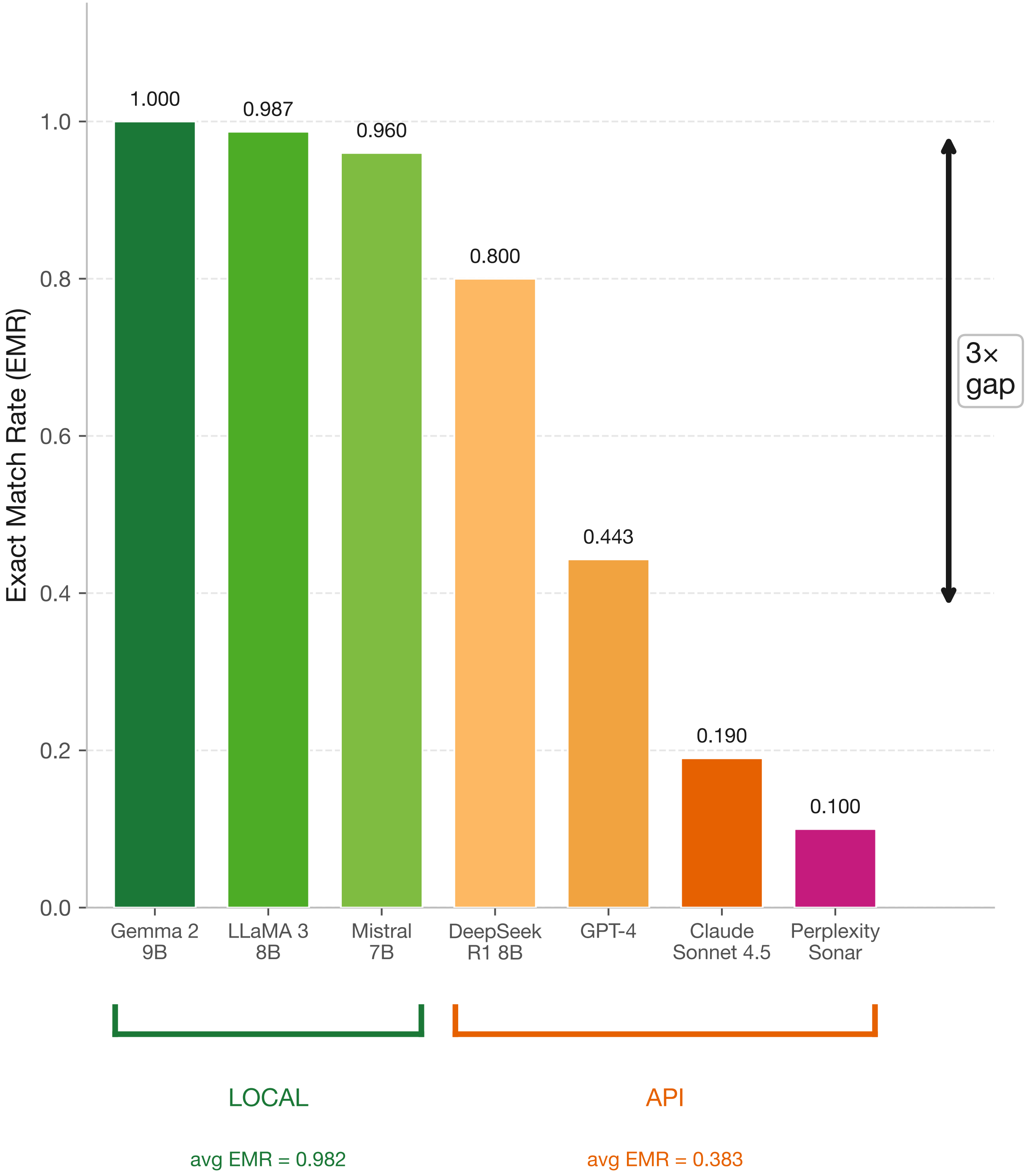
Model Execution — temp=0, top_p=1, weights hash

Run Card — output hash, timestamp, tokens

PROV Graph — wasGeneratedBy, wasDerivedFrom

Reproducibility Verification

Greedy-Decoding Reproducibility (Extraction Task)



Key Results

3,804 total runs

7 LLM models

4 tasks

4 API providers

<1% overhead

Local models reproduce near-perfectly; API outputs vary across identical calls.

github.com/Roverlucas/genai-reproducibility-protocol