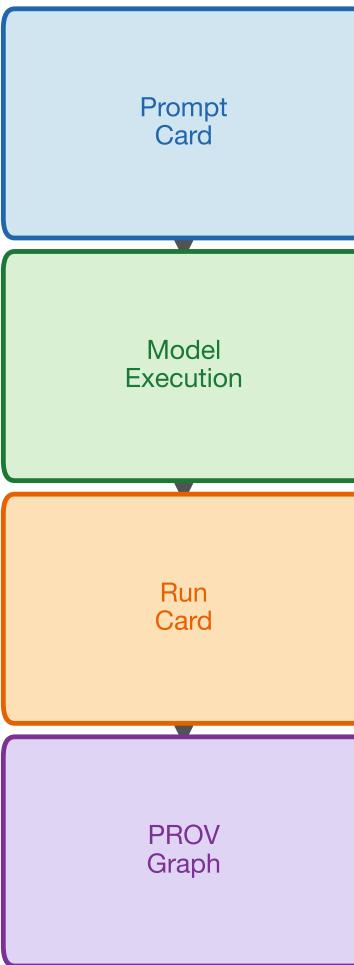


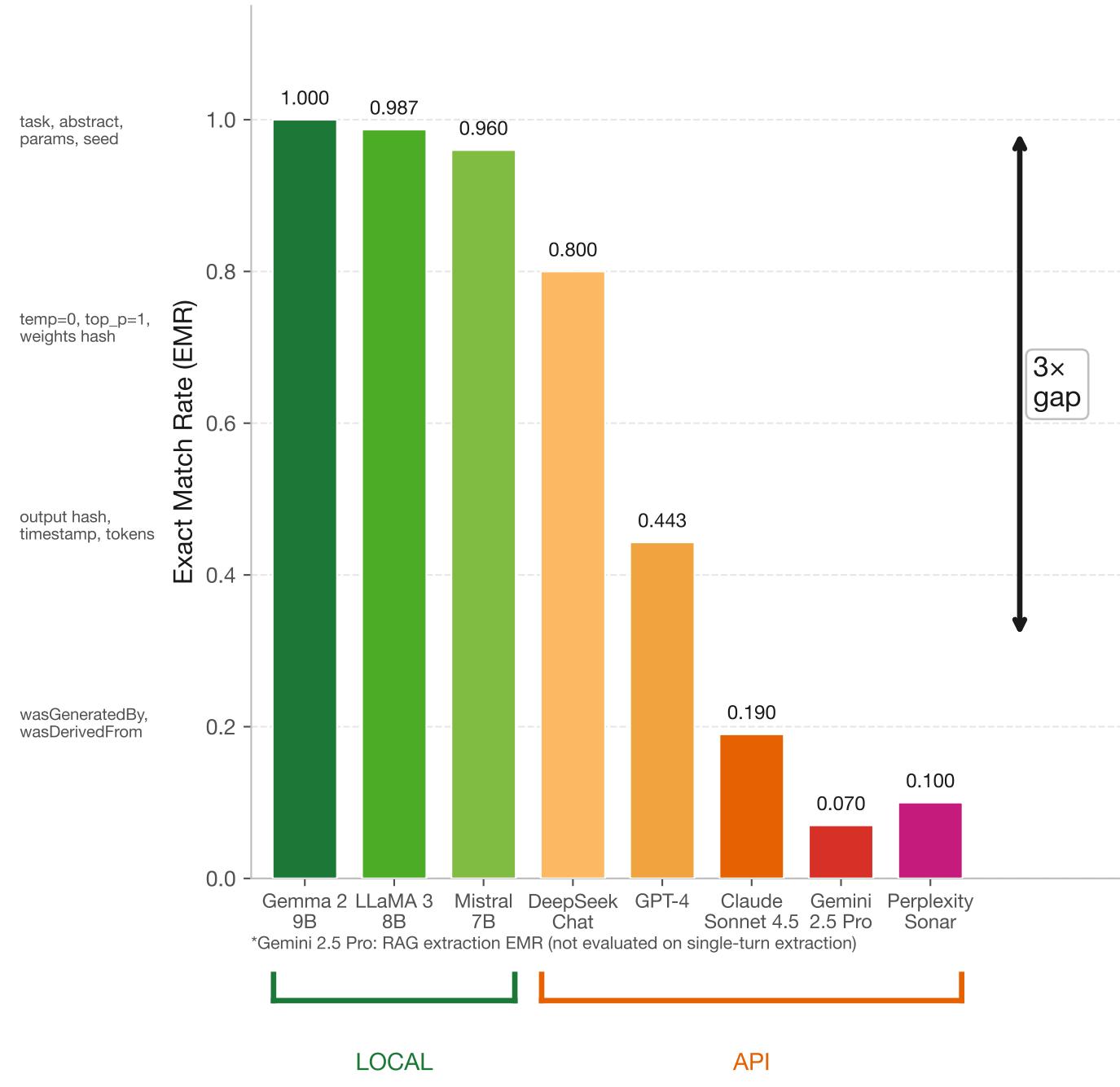
Hidden Non-Determinism in LLM APIs

A Lightweight Provenance Protocol for Reproducible Generative AI Research

Protocol Pipeline



Greedy-Decoding Reproducibility (Extraction Task*)



Key Results

3,904

total runs

8

LLM models

4

tasks

5

API providers

<1%

overhead

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