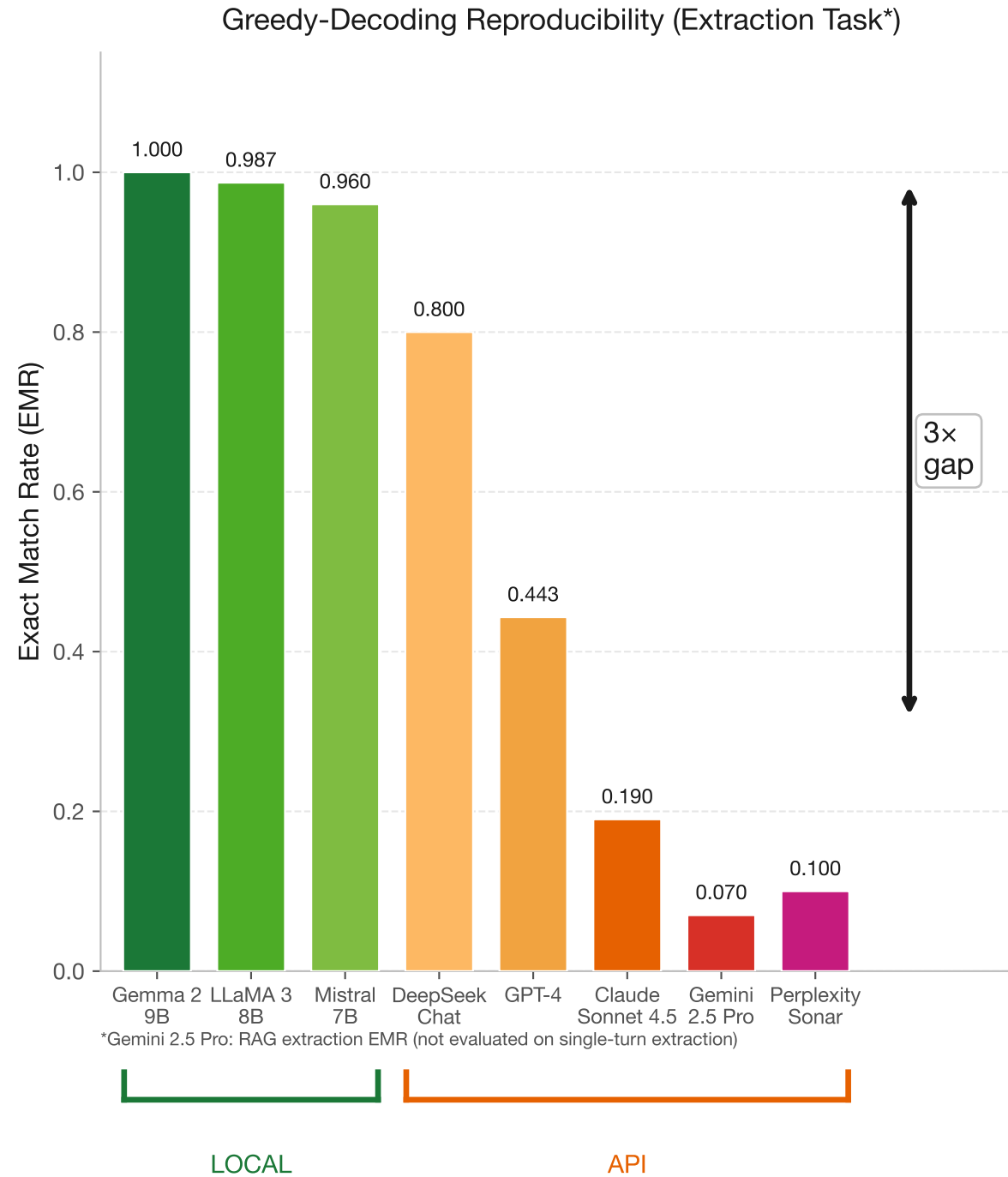
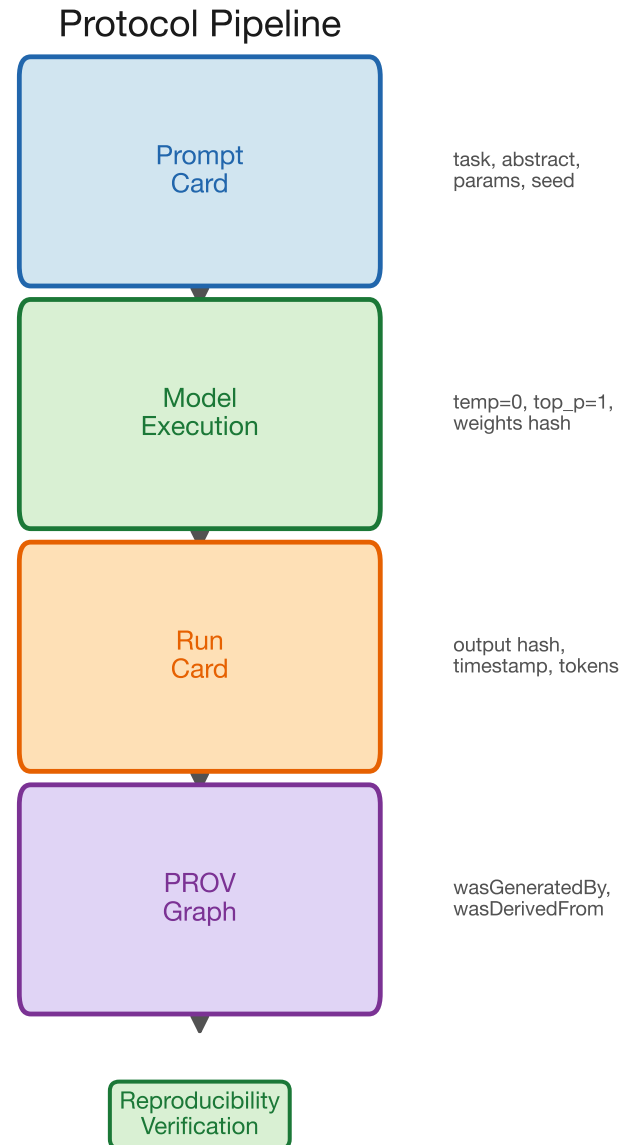


Hidden Non-Determinism in LLM APIs

A Lightweight Provenance Protocol for Reproducible Generative AI Research



*Gemini 2.5 Pro: RAG extraction EMR (not evaluated on single-turn extraction)

avg EMR = 0.982

avg EMR = 0.321

github.com/Roverlucas/genai-reproducibility-protocol

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.