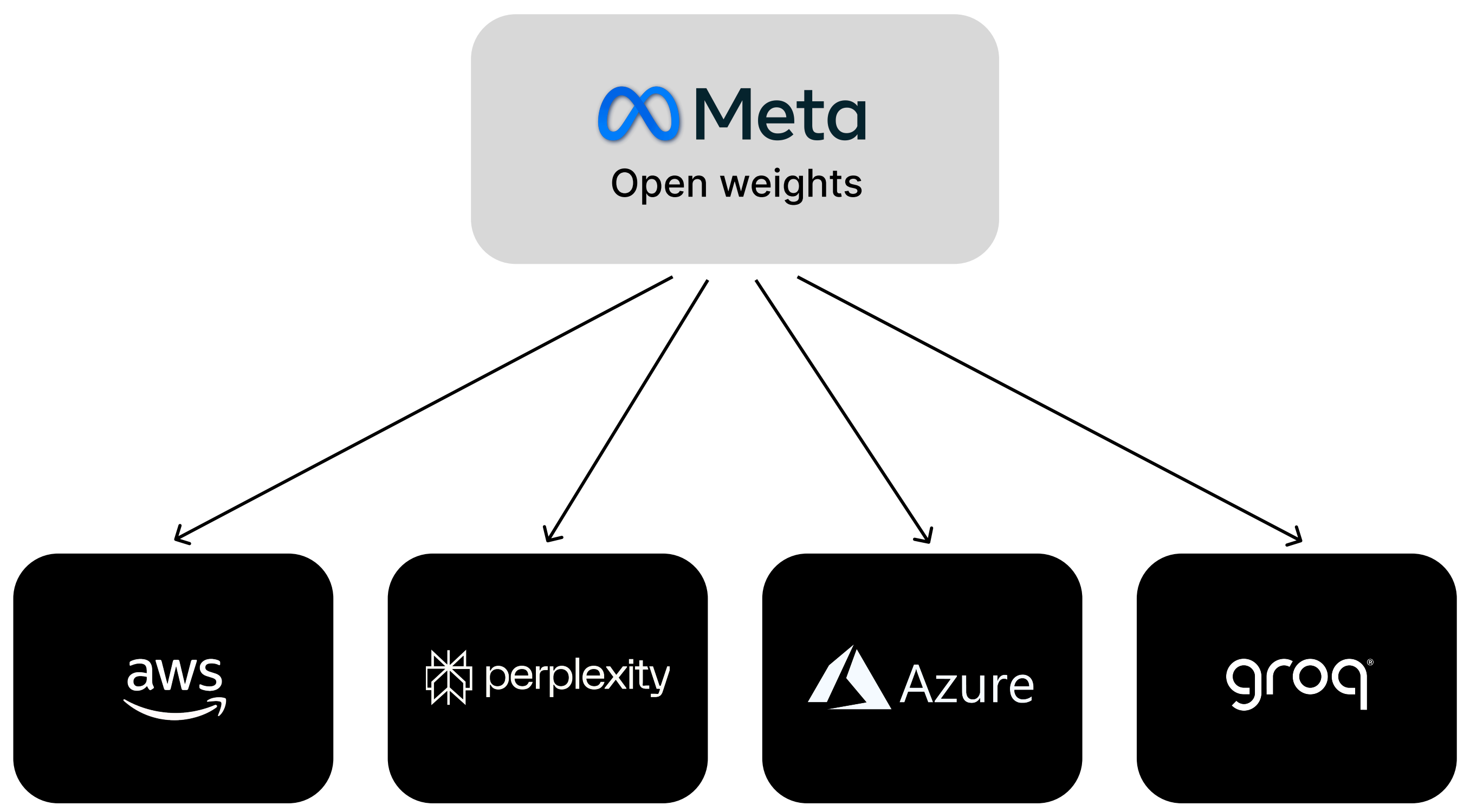
Model Equality Testing: Which model is this API serving?

Irena Gao, Percy Liang, Carlos Guestrin | Stanford University

`pip install model-equality-testing` to audit APIs for your own tasks.

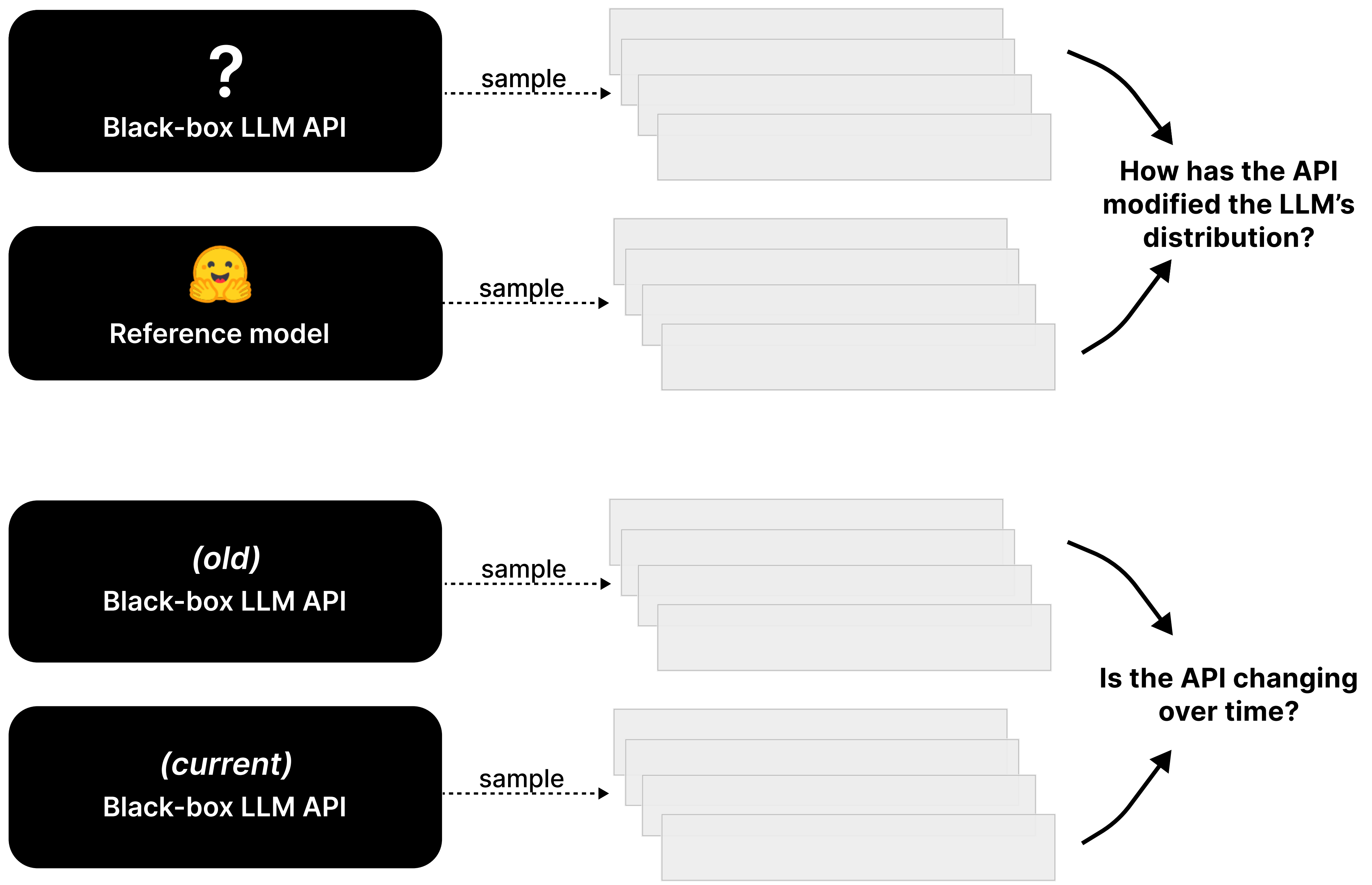
Paper / Data / Code

Many black-box inference APIs exist for the same language model, but this raises some questions...



Black-box inference API providers

...we answer some of these questions using Model Equality Testing!



? Black-box LLM API

Reference model

sample

sample

How has the API modified the LLM's distribution?

(old) Black-box LLM API

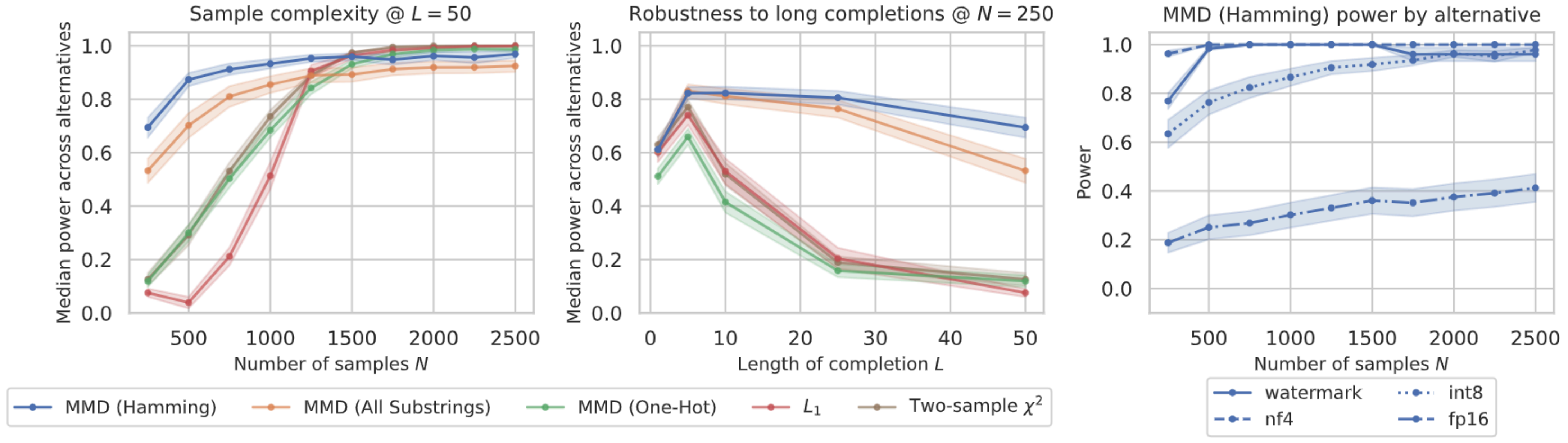
(current) Black-box LLM API

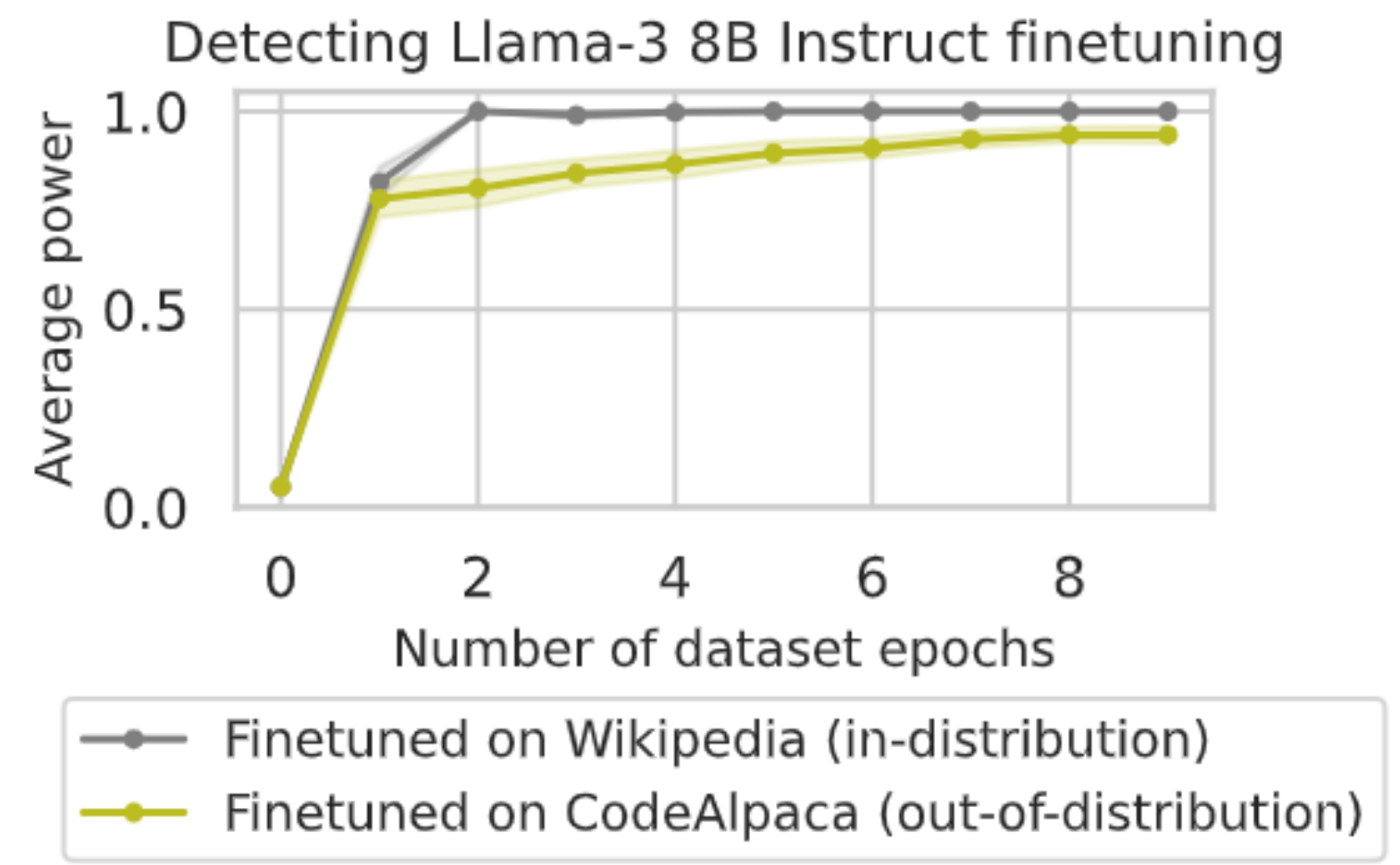
sample

sample

Is the API changing over time?

Experiments





(Top) Tests using the Hamming string kernel achieve a median of **77.4%** power against a range of model distortions, including quantization and watermarking, using an average of just 10 samples per prompt for a language modeling task defined over 25 prompts. This is more sample-efficient and robust to completion length than other two-sample tests.

(Left) The MMD (Hamming) test can also detect when a model has been fine-tuned.

Problem Statement: Model Equality Testing

We formalize these auditing questions as a two-sample hypothesis testing problem.

1. Sample N (prompt, completion) pairs from APIs P and Q, given a user-set prompt distribution π.

$$\mathcal{D}_P = \{z^{(i)} := (x^{(i)}, y^{(i)}) \mid x^{(i)} \sim \pi, y^{(i)} \sim P(\cdot \mid x^{(i)})\}_{i=1}^N$$
$$\mathcal{D}_Q = \{z^{(i)} := (x^{(i)}, y^{(i)}) \mid x^{(i)} \sim \pi, y^{(i)} \sim Q(\cdot \mid x^{(i)})\}_{i=1}^N$$

2. Use these samples to test between $P = Q$ (null hypothesis) and $P \neq Q$ at level α.

Method: Kernel Tests

We experiment with two-sample kernel tests based on the empirical Maximum Mean Discrepancy (MMD) between samples, and we empirically find that a simple string kernel related to the Hamming distance between samples captures how language models differ.

$$\widehat{\mathrm{MMD}}(\mathcal{D}_P, \mathcal{D}_Q) := \frac{1}{N(N_1)} \left[\sum_{z,z' \in \mathcal{D}_P} k(z,z') + \sum_{z,z' \in \mathcal{D}_Q} k(z,z') \right] - \frac{2}{N^2} \sum_{z \in \mathcal{D}_P, z' \in \mathcal{D}_Q} k(z,z') \quad [1]$$

where the proposed Hamming kernel is defined as $k(z, z') = \mathbf{1}\{x = x'\} \left(\sum_{i=1}^{L} \mathbf{1}\{y_i = y'_i\} \right)$.

[1] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. The Journal of Machine Learning Research, 13(1), 723-773.

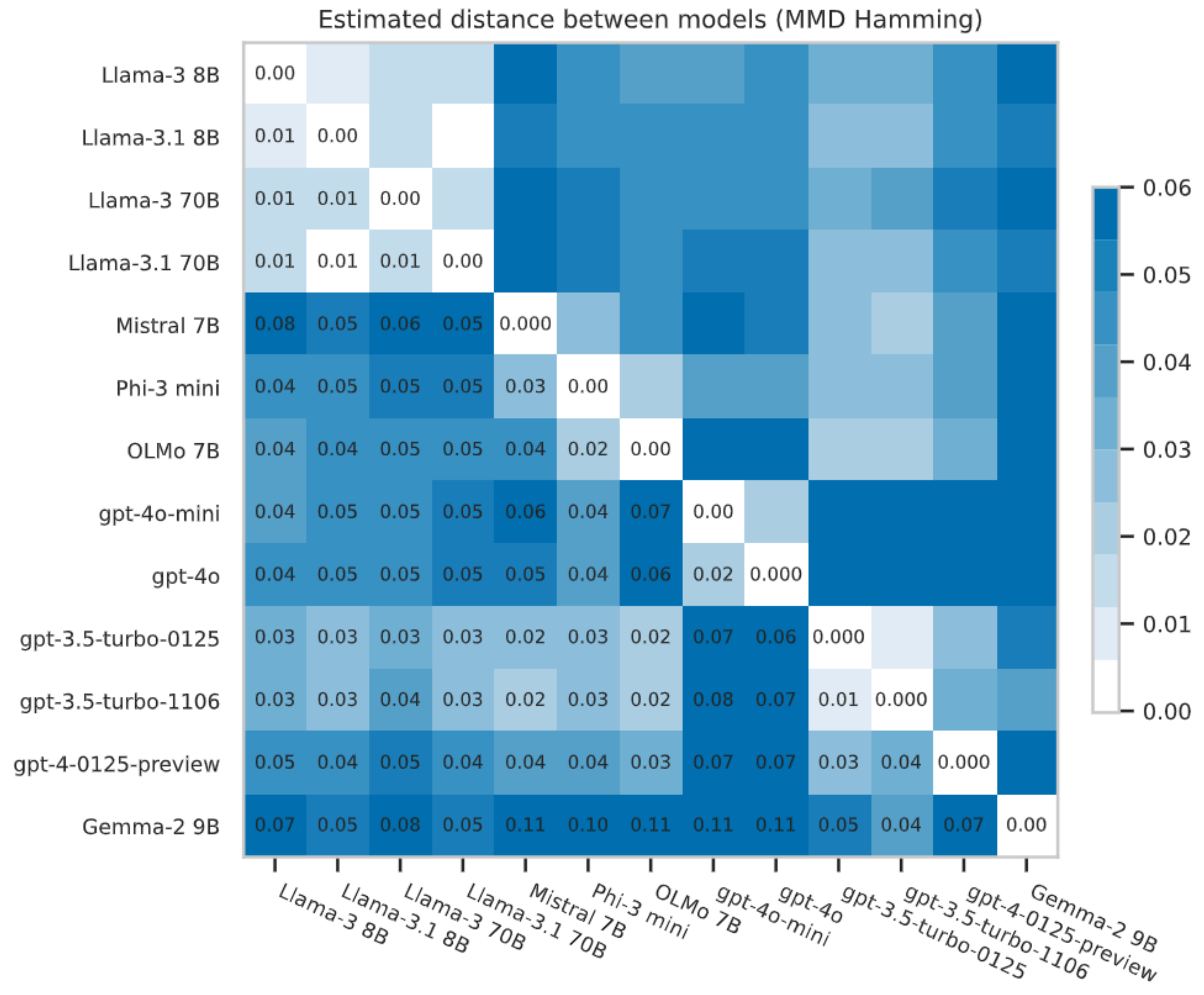
Model Equality Testing

	Llama-3		Llama-3.1	
API Provider	8B	70B	8B	70B
anyscale	✓	—	—	—
aws	✗	✗	✗	✗
Azure	✓	✓	✓	✓
deepinfra	✓	✓	✓	✓
Fireworks AI	✓	✓	✗	✓
groq	✓	✓	✗	✗
perplexity	✗	✗	✗	✗
replicate	✓	✓	—	—
together.ai	✓	✓	✓	✓

* Results for endpoints between July – August 2024; APIs are evolving, and results may have changed.

(Right) The test statistic can estimate the distance between different model completion distributions, which shows that models within the same family are more similar to each other than models of the same size.

(Left) Our test found that 11 out of 31 endpoints from Summer 2024 served different distributions than Llama reference weights released by Meta.


Estimated distance between models (MMD Hamming)
Llama-3 8B: 0.00
Llama-3.1 8B: 0.01 0.00
Llama-3 70B: 0.01 0.01 0.00
Llama-3.1 70B: 0.01 0.01 0.01 0.00
Mistral 7B: 0.08 0.05 0.06 0.05 0.000
Phi-3 mini: 0.04 0.05 0.05 0.05 0.03 0.00
OLMo 7B: 0.04 0.04 0.05 0.05 0.04 0.02 0.00
gpt-4o-mini: 0.04 0.05 0.05 0.05 0.06 0.04 0.07 0.00
gpt-4o: 0.04 0.05 0.05 0.05 0.05 0.04 0.06 0.02 0.000
gpt-3.5-turbo-0125: 0.03 0.03 0.03 0.03 0.02 0.03 0.02 0.07 0.06 0.000
gpt-3.5-turbo-1106: 0.03 0.03 0.04 0.03 0.02 0.03 0.02 0.08 0.07 0.01 0.000
gpt-4-0125-preview: 0.05 0.04 0.05 0.04 0.04 0.04 0.03 0.07 0.07 0.03 0.04 0.000
Gemma-2 9B: 0.07 0.05 0.08 0.05 0.11 0.10 0.11 0.11 0.11 0.05 0.04 0.07 0.00