Statistical Hypothesis Testing for Auditing Robustness in Language Models

UNIVERSITY OF CAMBRIDGE

Paulius Rauba[1] Qiyao Wei [1] Mihaela van der Schaar [1]

[1]University of Cambridge

Motivation & Importance

Motivation. LLMs exhibit stochastic behavior. As a result, two identical queries can differ purely by chance due to *sampling noise*.

In this paper, we ask how we can disentangle meaningful model changes from sampling variability. We can then use this for auditing:

- **Audit mechanism I**. Evaluating whether LLMs exhibit different behavior under arbitrary changes.
- **Audit mechanism II**. Understanding whether two LLMs exhibit the same behavior in the same setting.

Main takeaways

- We introduce a **general-purpose statistical hypothesis testing procedure** to test LLM behavior under different input or model perturbations.
- **Our procedure is model-agnostic** and provides effect sizes and interpretable p-values for any input perturbation and any model change with minimal assumptions.
- This can be used for understanding whether LLMs (or any LLM-based systems) are reliable in **high-stakes environments**.

Looking at LLMs via frequentist hypothesis testing

We make the insight that we can look at LLM outputs via the lens of frequentist hypothesis testing.

$$H_0 : \mathcal{D}_x = \mathcal{D}_{x'} \quad \text{(The perturbation has no effect)} \qquad (1)$$
$$H_1 : \mathcal{D}_x \neq \mathcal{D}_{x'} \quad \text{(The perturbation has an effect)} \qquad (2)$$

Distribution-based perturbation analysis: an overview of the procedure

Distribution-based perturbation analysis proceeds in four steps: response sampling, distribution construction, distributional comparison, and statistical inference.

I. Response Sampling. Draw k independent outputs from the original prompt and k from the perturbed prompt

$$\hat{\mathcal{D}}_x = \{y_i\}_{i=1}^k, \qquad \hat{\mathcal{D}}_{x'} = \{y_i'\}_{i=1}^k,$$

where $y_i \overset{i.i.d.}{\sim} \mathcal{S}(x)$ and $y_i' \overset{i.i.d.}{\sim} \mathcal{S}(x')$ with $x' := \Delta_x(x)$. Define the pooled vector $Z = (z_1, \dots, z_{2k})$ with

$$z_i = y_i \quad (1 \le i \le k), \qquad z_{k+i} = y_i' \quad (1 \le i \le k).$$

II. Distribution construction. Using a similarity function $s : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, build

$$P_0 = \{\, s(y_i, y_j) \,:\, 1 \le i < j \le k\},$$
$$P_1 = \{\, s(y_i, y_j') \,:\, 1 \le i, j \le k\}.$$

III. Distributional comparison. Measure the discrepancy between P_0 and P_1 with any non-negative functional

$$\omega : \mathcal{P} \times \mathcal{P} \longrightarrow \mathbb{R}_{\ge 0}, \qquad T_{\text{obs}} = \omega(P_0, P_1).$$

IV. Statistical inference. Formulate the hypotheses

$$H_0 : \mathcal{S}(x) = \mathcal{S}(x'), \qquad H_1 : \mathcal{S}(x) \neq \mathcal{S}(x').$$

We can evaluate this hypothesis via a simple permutation procedure that uses the pooled vector Z.

Objective. If $\hat{p}$ is small, this suggests that T_{obs} is unusually large relative to its null distribution. The value T_{obs} itself serves as the effect-size estimate, whereas the permutation test provides frequentist p-values

Example 1: Measuring true positive and false positive rates

Setup. We can evaluate LLMs' true positive/false positive rates by altering conditions which should / shouldn't affect their outputs. We can then compute p-values and FPR/TPR as we vary α.

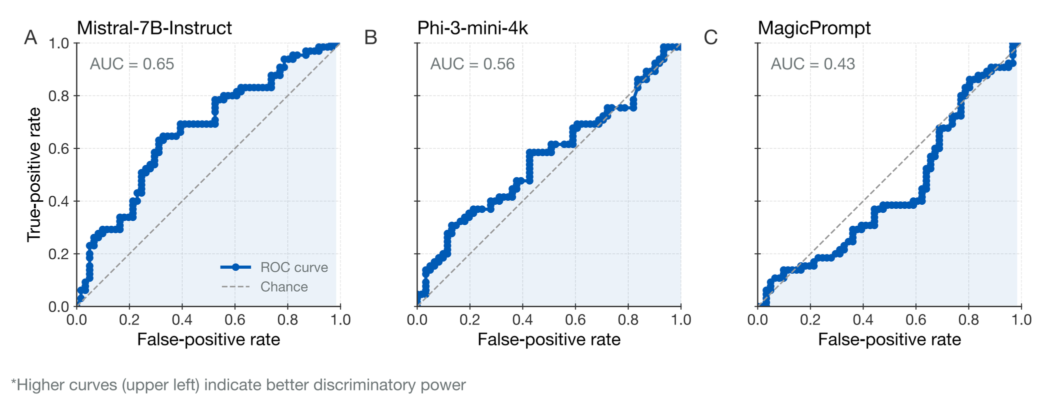

Figure 1. **TPR/FPR trade-offs for three language models.** Panels A–C show ROC curves (TPR vs. FPR) and AUC scores (varying $\alpha \in [0, 1]$); higher AUC indicates better detection of true perturbations and resistance to irrelevant changes.

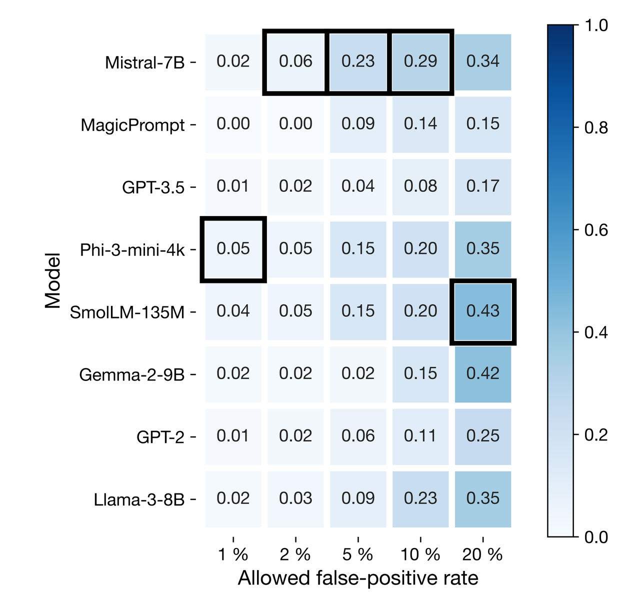

Figure 2. **TPR by selected FPR for multiple models.**

Example 2: Measuring alignment with a reference language model

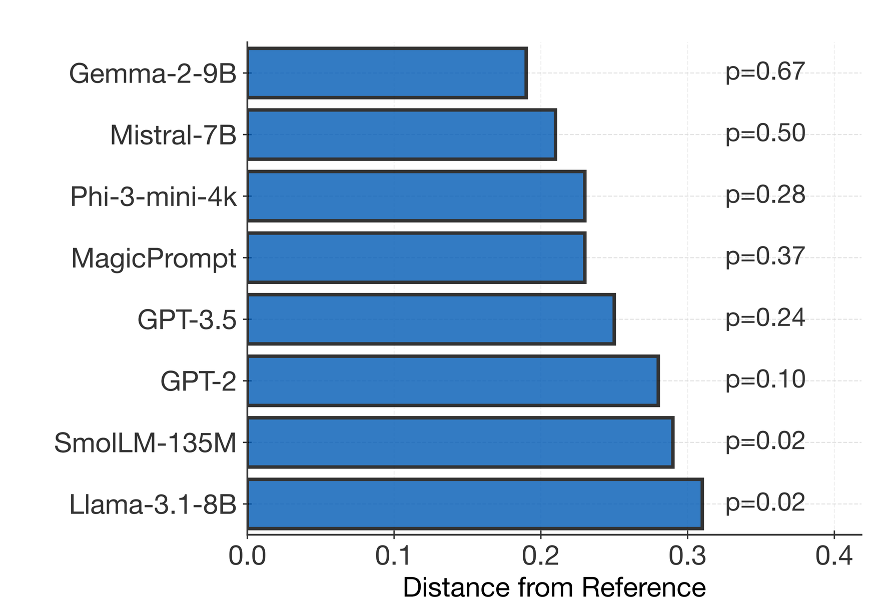

Figure 3. Distance of responses from a reference language model.

Takeaway. (Fig. 3) shows how to quantify inter-model alignment with respect to a reference language model.

The Big Picture. We establish a general-purpose procedure to ensure reliable LLM behavior for task-specific problems for any black-box LLM with minimal assumptions.