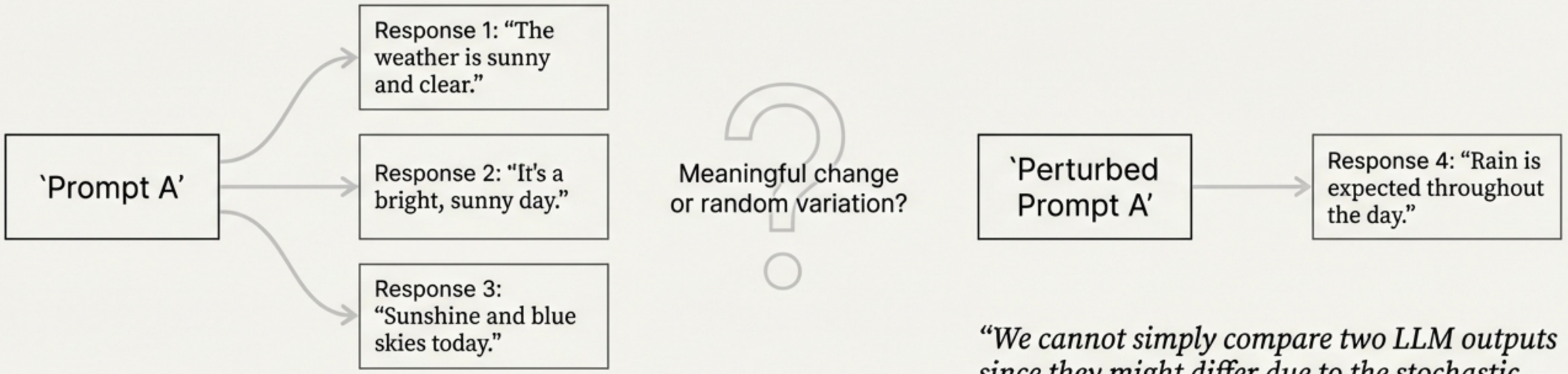
Beyond the Prompt: A Statistical Approach to Robust LLM Auditing

An exploration of Distribution-Based Perturbation Analysis (DBPA) and a new proposal for semantic stability, S-DBPA.

The Auditing Challenge: How Do We Measure Real Change in a Stochastic System?

LLMs are **not deterministic.** The same input can produce different outputs due to sampling parameters (e.g., temperature).

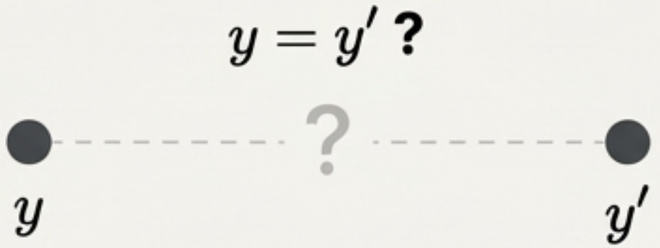
When we change an input prompt, is the change in output a meaningful behavioral shift, or just random noise?



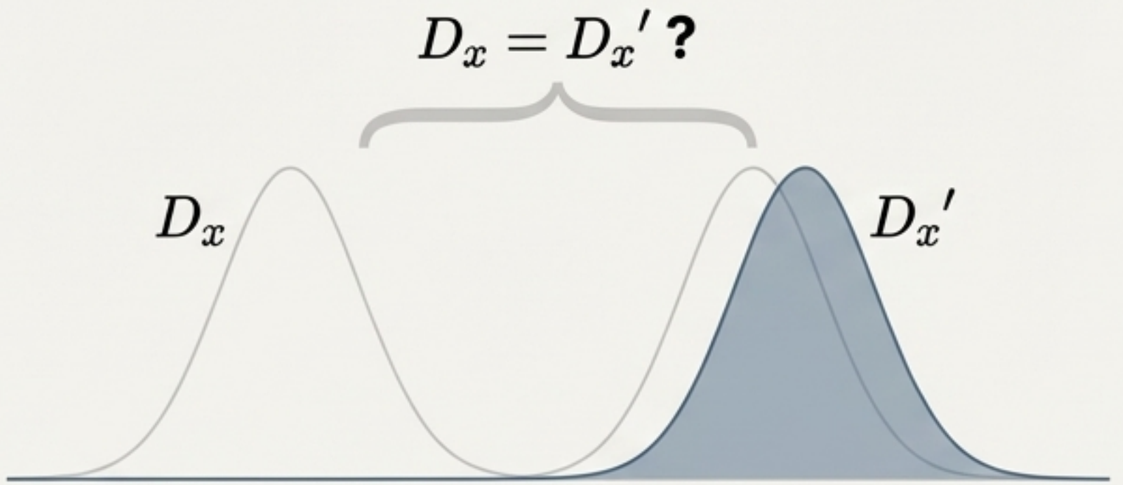
"We cannot simply compare two LLM outputs since they might differ due to the stochastic nature of the system." — Rauba et al.

DBPA's Insight: Reframe Auditing as a Frequentist Hypothesis Test

Insufficient



DBPA's Approach

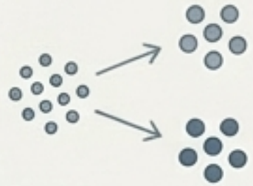


Null Hypothesis (H_0): $D_x = D_x'$ The perturbation has no statistically significant effect on the model's response distribution.

Alternative Hypothesis (H_1): $D_x \neq D_x'$ The perturbation causes a measurable shift in the response distribution.

This distributional approach captures the full stochastic behavior of the model, allowing for robust statistical inference.

The DBPA Procedure: A Four-Step Statistical Workflow

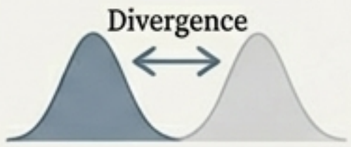


1. Response Sampling

Use Monte Carlo sampling to generate k responses from the original prompt ($\hat{D}_x$) and k from the perturbed prompt ($\hat{D}_x{}'$).

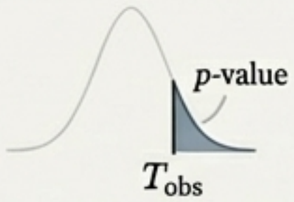
2. Distribution Construction

Create two distributions of *pairwise similarities* (P_0 and P_1) using an embedding model and cosine similarity to capture intrinsic and between-group variability.



3. Distributional Comparison

Measure the discrepancy between P_0 and P_1 using a metric like Jensen-Shannon Divergence (JSD). This yields the observed test statistic (T_{obs}).



4. Statistical Inference

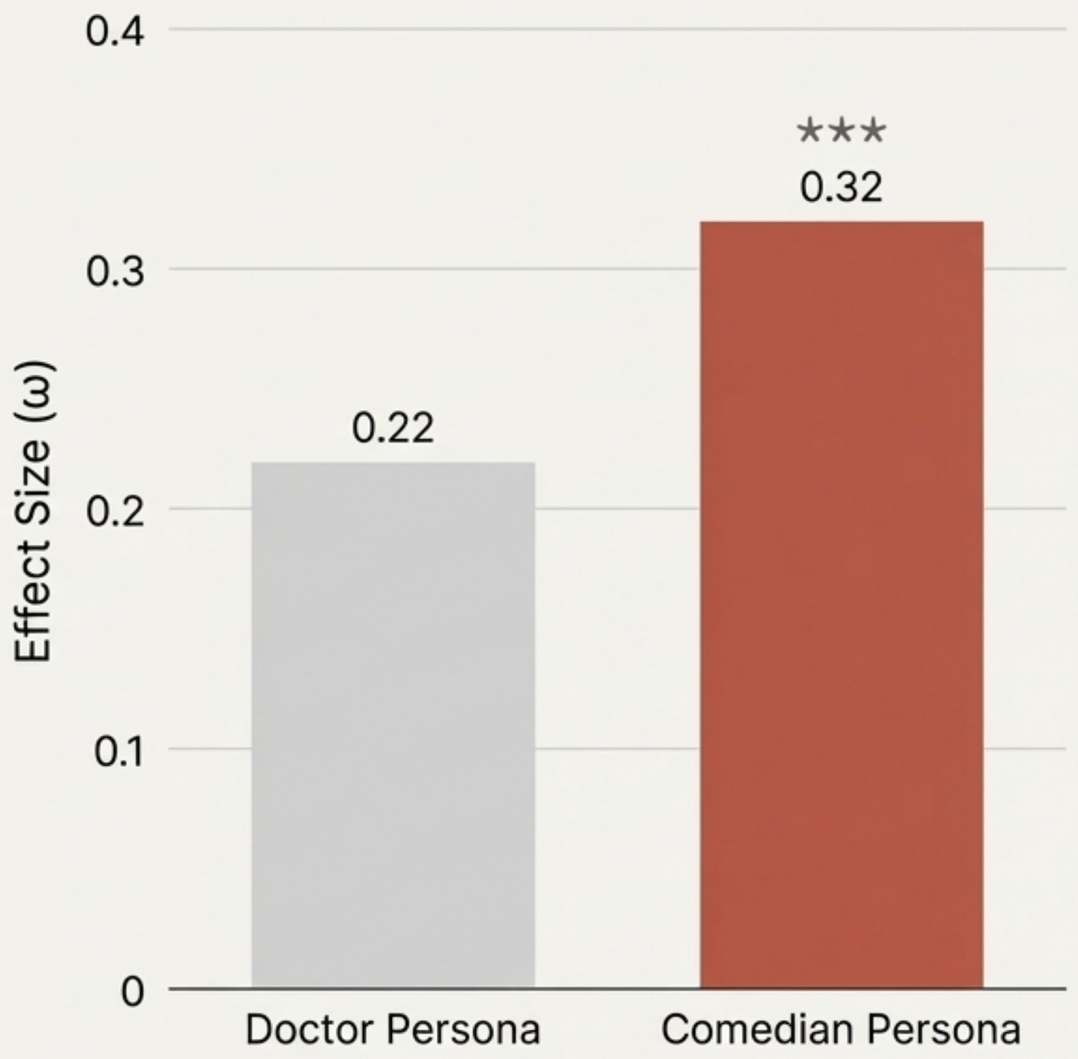
Use a permutation test on the pooled responses to calculate a p-value, determining if T_{obs} is statistically significant.

DBPA in Action: Quantifying the Impact of Personas

Testing the effect of prepending role-playing instructions (‘Act as a {role}’) to a healthcare question on GPT-4.

Key Findings

- **Medical Personas** (‘Doctor’, ‘Nurse’): Low effect size, not statistically significant. The model’s core medical knowledge is stable.
- **Non-Medical Personas** (‘Comedian’, ‘NeurIPS Reviewer’): High effect size, statistically significant ($p < 0.001$ and $p < 0.01$ respectively). The model’s behavior shifts meaningfully.



DBPA successfully quantifies and validates intuitive expectations about model behavior.

But a Hidden Flaw Undermines Its Robustness

The Problem:

The DBPA framework is highly sensitive to superficial, lexical variations in prompts.

The Critical Insight:

The Critical Insight: A robust auditing tool should not be fooled by simple rephrasing. The semantic intent is what matters.

An audit should yield the same conclusion for:

```
"Act as a doctor."
```

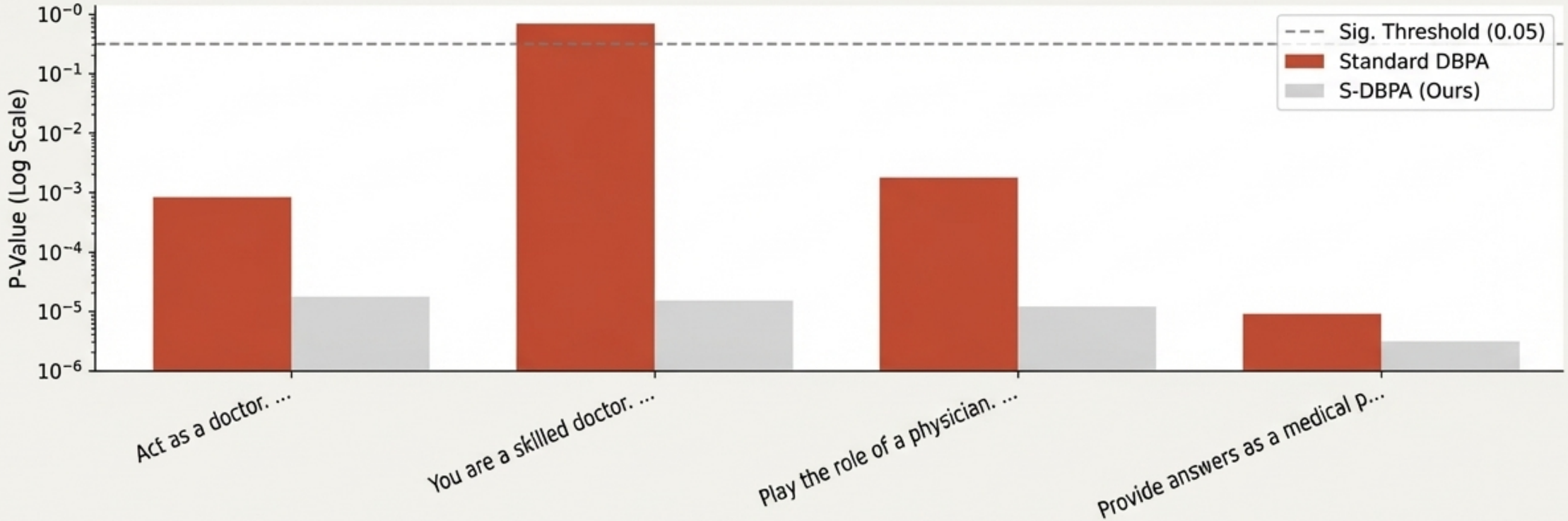
```
"You are a skilled doctor."
```

```
"Play the role of a physician."
```

“A prompt (‘Act as a doctor’) and its semantic equivalent (‘You are a doctor’) often yield statistically distinguishable response distributions under standard testing, leading to inconsistent auditing conclusions.” — Cohen-Eliya

The Evidence: Trivial Wording Changes Cause Wildly Fluctuating P-Values

We ran Standard DBPA on four semantically identical "Doctor" prompts. A robust method should produce consistently significant (or non-significant) p-values.

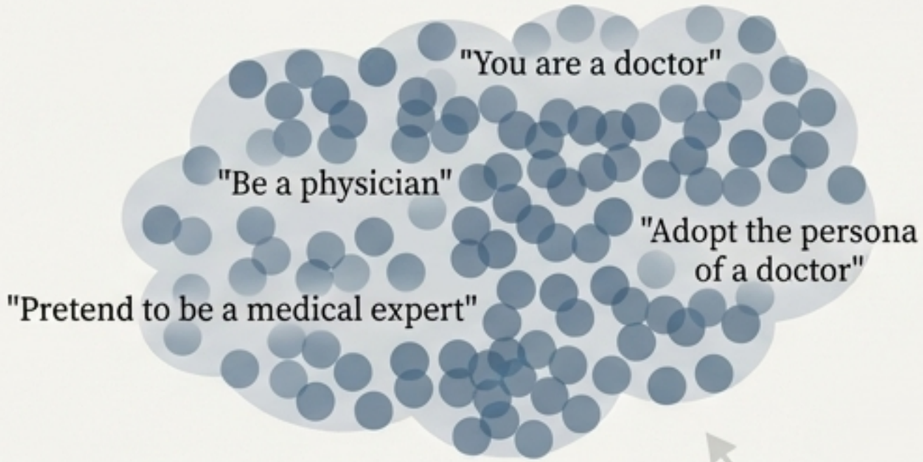


Standard DBPA's results are unstable. The p-value for "You are a skilled doctor" is 0.6230 (not significant), while **"Provide answers as a medical professional" is 0.0000** (highly significant). The audit's conclusion depends entirely on the phrasing.

The Solution: Audit the *Concept*, Not Just the Prompt

Introducing **S-DBPA (Semantic DBPA)**: A new methodology that builds on DBPA to achieve semantic robustness.

Standard DBPA: Tests a single point in "intent space" ("Act as a doctor")



S-DBPA: Audits a **"Semantic Neighborhood"** (Concept: Doctor)

The Goal: To marginalize out the noise from specific phrasing and measure the true effect of the semantic intent.

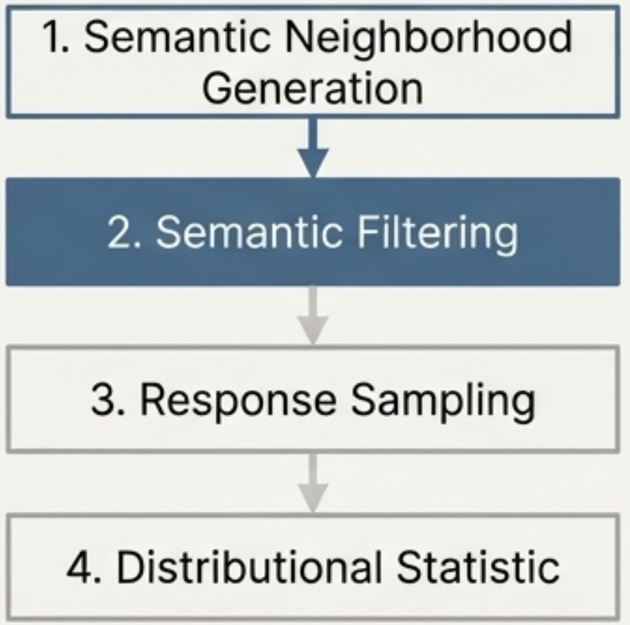
The S-DBPA Method: Controlled Semantic Sampling

Semantic Neighborhood Generation: Use a paraphrasing LLM to generate a large set of candidate variations for the base prompt.

Semantic Filtering: Apply a strict cosine similarity filter (e.g., $\tau > 0.95$) using an embedding model to retain only true paraphrases and discard semantic drift.

Response Sampling: Sample responses from the *entire filtered set* of prompts, not just one.

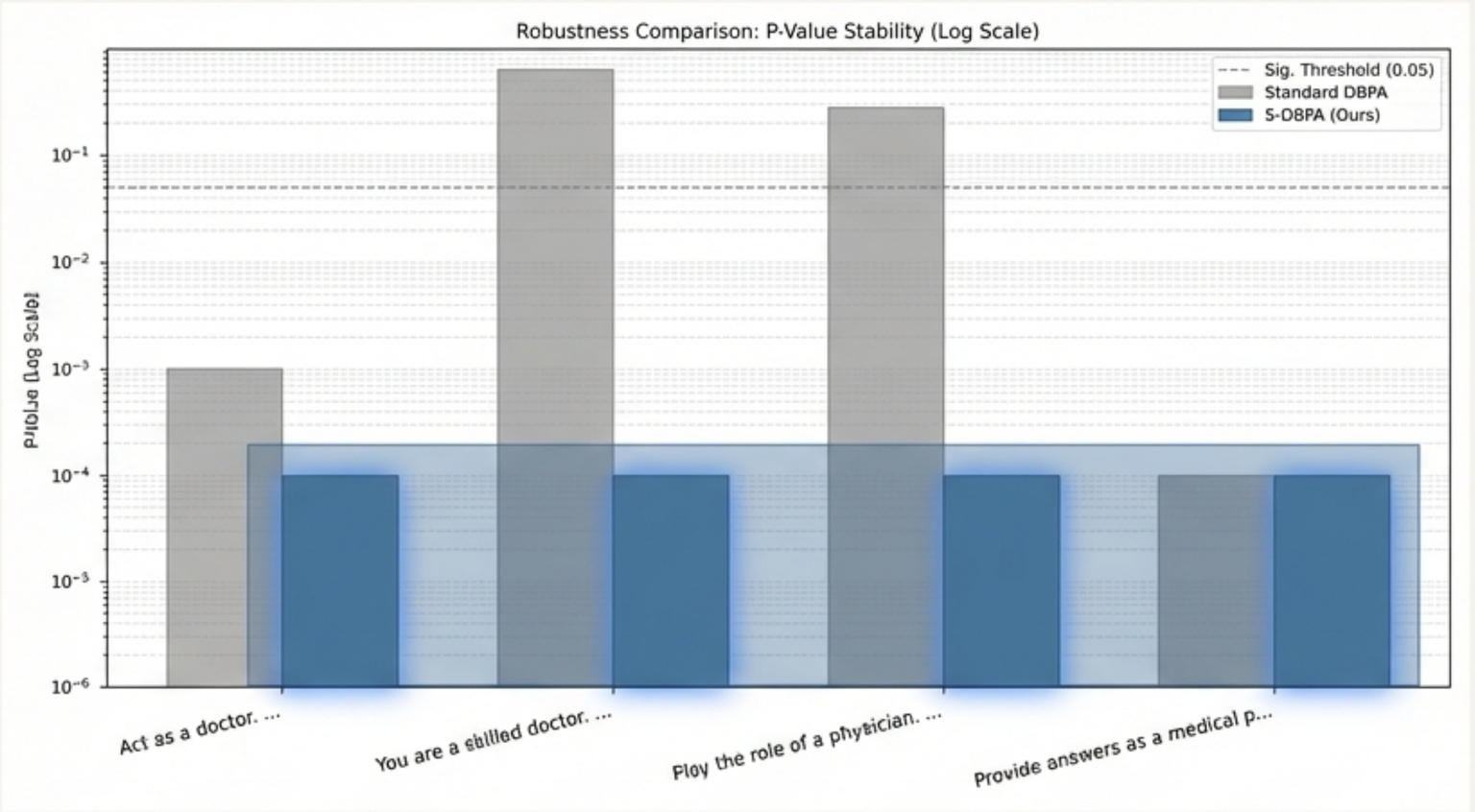
Distributional Statistic: Compute the JSD between the aggregated neighborhood response distribution and the reference distribution.



This process effectively performs a Monte Carlo integration over the local semantic manifold, yielding a stable estimate of the concept's impact.

The Resolution: S-DBPA Achieves a Stable, Consistent Signal

We ran S-DBPA on the same four ‘Doctor’ prompt variations.

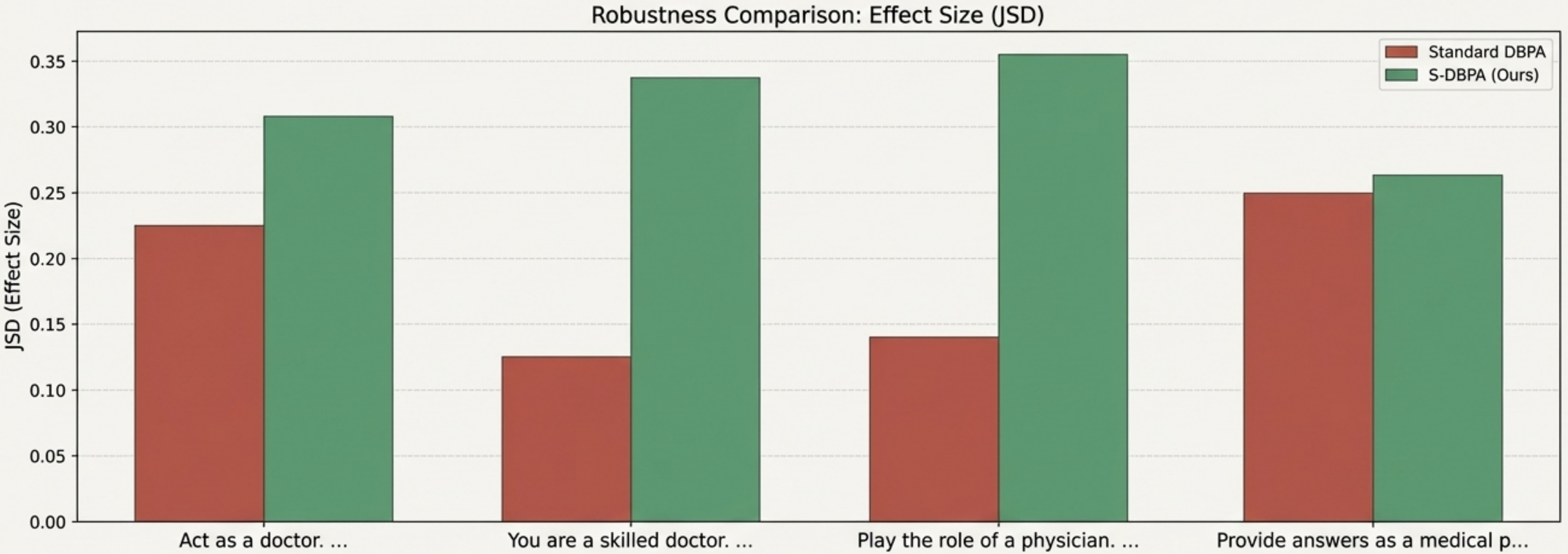


S-DBPA (blue bars) delivers consistently low p-values (**< 0.001**) across all variations. It correctly and reliably identifies that the ‘Doctor’ persona has a significant effect, regardless of the specific phrasing used.

S-DBPA successfully smooths out the noise introduced by lexical choices, providing a truly robust audit.

S-DBPA Also Stabilizes the Measured Effect Size

As seen in the orange bars, the measured JSD (effect size) for Standard DBPA is also highly variable, suggesting the persona's effect is weak or strong depending on phrasing.



The green bars show that S-DBPA provides a more stable and consistently higher estimate of the effect size. It captures the true, robust impact of the 'Doctor' concept on the model's output distribution.

Theoretical Grounding I: Statistical Validity

Is the permutation test still valid when we're sampling from a set of prompts instead of one?

Theorem 1 (Semantic Exchangeability)

Under the null hypothesis (H_0), where the persona has no effect, the responses generated from the semantic neighborhood are exchangeable with the reference responses.

Proof Sketch

If H_0 is true, the persona instructions are ignored. Therefore, responses from all prompts in the semantic neighborhood ($p' \in P_S$) and the neutral prompt are all i.i.d. samples from the same underlying distribution.

Because the pooled responses are exchangeable, the permutation p-value is exact and statistically valid.

Theoretical Grounding II: The Source of Robustness

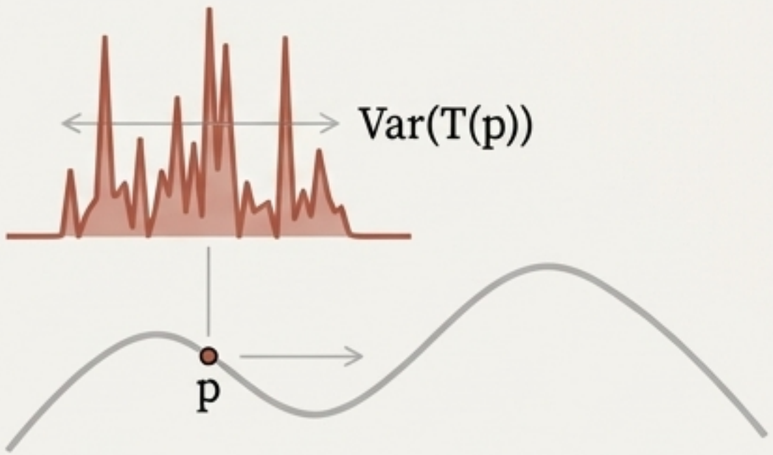
Why does sampling from a neighborhood produce a more stable estimate?

The Law of Large Numbers on the Semantic Manifold

Standard DBPA

Standard DBPA produces an estimate of the effect T(p) for a single, specific prompt p.

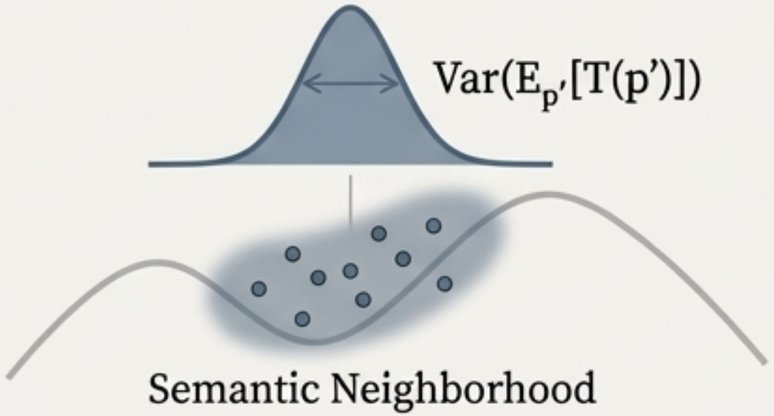
This estimate has high variance.



S-DBPA

S-DBPA estimates the *expected effect* over the entire semantic neighborhood: $E_{p'}[T(p')]$.

By the Law of Large Numbers, as the sample size N_S increases, the variance of our estimator decreases: $Var(E_{p'}[T(p')]) \to 0$.



S-DBPA provides a more robust audit metric because it is a principled estimate of the average effect over the concept, rather than a noisy measurement at a single point.

From Brittle Prompts to Robust Concepts

1. **Problem:** Auditing stochastic LLMs requires statistical rigor.
2. **Initial Solution (DBPA):** A powerful hypothesis testing framework.
3. **Hidden Flaw:** This framework is fragile, with results depending on superficial wording.
4. **Improved Solution (S-DBPA):** By auditing 'Semantic Neighborhoods,' we achieve stability and robustness.

For reliable and meaningful LLM evaluation, we must move beyond testing individual prompts to auditing the underlying semantic intents. S-DBPA provides a statistically-grounded methodology to do exactly that.