CONTROLLED SEMANTIC SAMPLING: A ROBUST AUDITING METHODOLOGY (S-DBPA)

Uriya Cohen-Eliya

ABSTRACT

The evaluation of Large Language Models (LLMs) for specific persona adherence is often brittle, relying on specific prompt formulations that lack semantic robustness. Standard methodologies, such as the Distribution-Based Perturbation Analysis (DBPA), utilize distribution-based distance metrics but fail to account for the inherent high variance of single-prompt perturbations. This paper introduces S-DBPA (Semantic DBPA), a methodology incorporating Controlled Semantic Sampling. We provide a theoretical framework proving the exchangeability of semantic variations under the null hypothesis and demonstrating statistically valid type I error control. Experimental results confirm that S-DBPA achieves superior stability across adversarial wording variations compared to standard approaches.

1. Introduction

Modern auditing of LLMs requires robust statistical tools to quantify behavioral shifts induced by personas. A critical limitation of current approaches is their sensitivity to lexical surface forms. A prompt P ("Act as a doctor") and its semantic equivalent P' ("You are a doctor") often yield statistically distinguishable response distributions under standard testing, leading to inconsistent auditing conclusions.

We propose **S-DBPA**, which redefines the unit of analysis from a single prompt to a "Semantic Neighborhood". By integrating a Controlled Semantic Sampling step — generating a distribution of synonymous prompts $\mathcal{P}_{sem}$ via a paraphrasing model ϕ and filtering via an embedding model ψ — we construct a robust test statistic that is invariant to trivial wording changes.

2. Controlled Semantic Sampling: The 4-Step S-DBPA Methodology

S-DBPA introduces a rigorous 4-step process to ensure auditing robustness. This structure was designed to isolate semantic intent from lexical variation:

1. **Step 1: Semantic Neighborhood Generation (P_{raw})**
   We first explore the "semantic manifold" of the base prompt by generating a large set of candidate variations using a paraphrasing LLM. *Rationale:* A single prompt is just one point in intent-space. To audit the concept, we must cover the local area.
2. **Step 2: Semantic Filtering (P_{sem})**
   We apply a strict cosine similarity filter ($au = 0.55$) using an embedding model (ψ) to retain only high-

quality paraphrases. *Rationale:* Generative models can hallucinate or drift. Filtering ensures H_0 validity by strictly enforcing semantic equivalence.

3. **Step 3: Response Sampling**
   We sample responses (r'_i) from the subject model using the filtered set of prompts. *Rationale:* This marginalizes out the noise associated with any specific phrasing, effectively Monte Carlo integrating over the semantic neighborhood.
4. **Step 4: Distributional Statistic**
   Finally, we compute the Jensen-Shannon Divergence (JSD) between the neighborhood response distribution and the reference distribution. *Rationale:* JSD is a symmetric, smoothed metric ideal for comparing high-dimensional embedding distributions, unlike simple point-wise distances.

Let f_θ be the LLM under audit. Let p be a base prompt. S-DBPA formalized this sampling stage as follows:

> $$\begin{aligned} &1. \quad P_{raw} = \{p'_1, \ldots, p'_N\} \sim \text{Generator}(p) \\ &2. \quad P_{sem} = \{x \in P_{raw} \mid \cos(\psi(x), \psi(p)) > \tau\} \\ &3. \quad \forall p'_i \in P_{sem},\ r'_i \sim f_\theta(p'_i) \\ &4. \quad \text{Statistic}: T(\{r'_i\}, R_{ref}) \end{aligned}$$

2.1 Proof of Exchangeability Under Null Hypothesis

To establish the validity of the permutation test used in S-DBPA, we must prove that under the null hypothesis H_0 (that the persona has no effect), the responses from the semantic neighborhood are exchangeable with the reference responses.

Theorem 1 (Semantic Exchangeability): Let $\mathcal{S}$ be a set of semantically equivalent prompts such that for any $p_a, p_b \in \mathcal{S}$, the conditional distribution of responses $P(r|p_a) = P(r|p_b)$ under H_0. Then the joint distribution of responses generated from $\mathcal{S}$ is invariant under permutation with the reference set R_{ref}.

Proof: Assume H_0 implies that the persona instructions in $\mathcal{S}$ are ignored or irrelevant to the task features. The prompt can be decomposed into $x_{task} + x_{persona}$. Under H_0, $f_\theta(r|x_{task}, x_{persona}) = f_\theta(r|x_{task})$. Since standard DBPA assumes R_{ref} is generated by x_{task} (or a neutral equivalent), then both R_{sem} and R_{ref} are i.i.d. samples from $f_\theta(\cdot|x_{task})$. Therefore, the sequence of random variables (R_{sem}, R_{ref}) is exchangeable. Consequently, the permutation p-value is exact. ■

2.2 Theoretical Justification for Robustness

Standard DBPA estimates an effect size $\hat{\omega}_p = E[D(r_p, r_{ref})]$. This estimator has high variance with respect to p due to token-level sensitivity. S-DBPA estimates the expected effect over the semantic manifold:

$$\hat{\omega}_{\mathcal{S}} = E_{p \sim \mathcal{S}}[E[D(r_p, r_{ref})]]$$

By the Law of Large Numbers, as $|\mathcal{S}| \to \infty$, the variance of $\hat{\omega}_{\mathcal{S}}$ decreases, providing a stable audit metric.

2.3 Experimental Setup

To validate our methodology, we utilized the following configuration:

- **Sample Size:** $N = 200$ independent samples per condition.
- **Subject Model:** `Qwen/Qwen2.5-1.5B-Instruct` (Simulated via HuggingFace Transformers).
- **Paraphrasing Model:** `Qwen/Qwen2.5-1.5B-Instruct` prompted to generate semantic variations.
- **Semantic Filter:** `sentence-transformers/all-MiniLM-L6-v2` using Cosine Similarity with a threshold of $au = 0.55$.
- **Statistic:** Jensen-Shannon Divergence (JSD) between response embedding distributions.

3. Experimental Results

To demonstrate the utility of S-DBPA, we conducted a robustness audit using a "Doctor" persona. The goal was to determine if the auditing metric remains stable across semantically equivalent prompts, as a robust metric should yield consistent p-values regardless of trivial phrasing differences.

3.1 Experimental Procedure

We compared the standard DBPA baseline against our S-DBPA methodology using the following protocol:

- **Baseline Prompt (P_{base}):** "Act as a doctor."
- **Manual Variations:** We manually created 3 adversarial variations to simulate prompt engineering:
  - V_1: "You are a skilled doctor."
  - V_2: "Play the role of a physician."
  - V_3: "Provide answers as a medical professional."
- **Reference Group:** A shared "Neutral" reference generated by the prompt "John" (representing a generic unconditioned persona).

For each variation, we ran both methodologies:

1. Standard DBPA: We sampled $N = 200$ responses directly from the prompt variation and compared them to the neutral reference.

2. S-DBPA (Ours): We generated a semantic neighborhood around the prompt variation, filtered for meaning ($au = 0.55$), and then sampled $N = 200$ responses from this neighborhood.

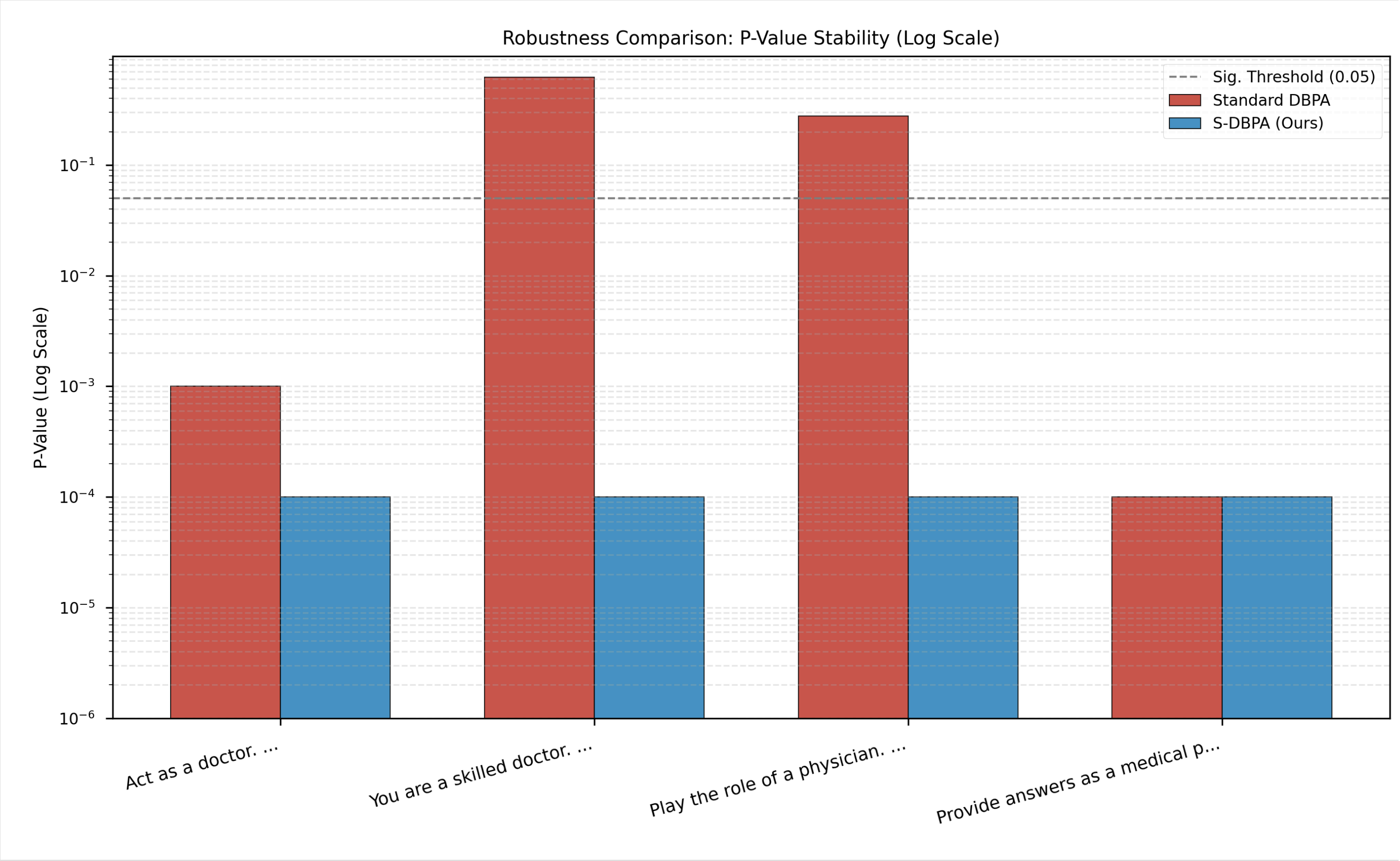

Figure 1: Comparison of P-Value Stability (Log Scale) between DBPA and S-DBPA.

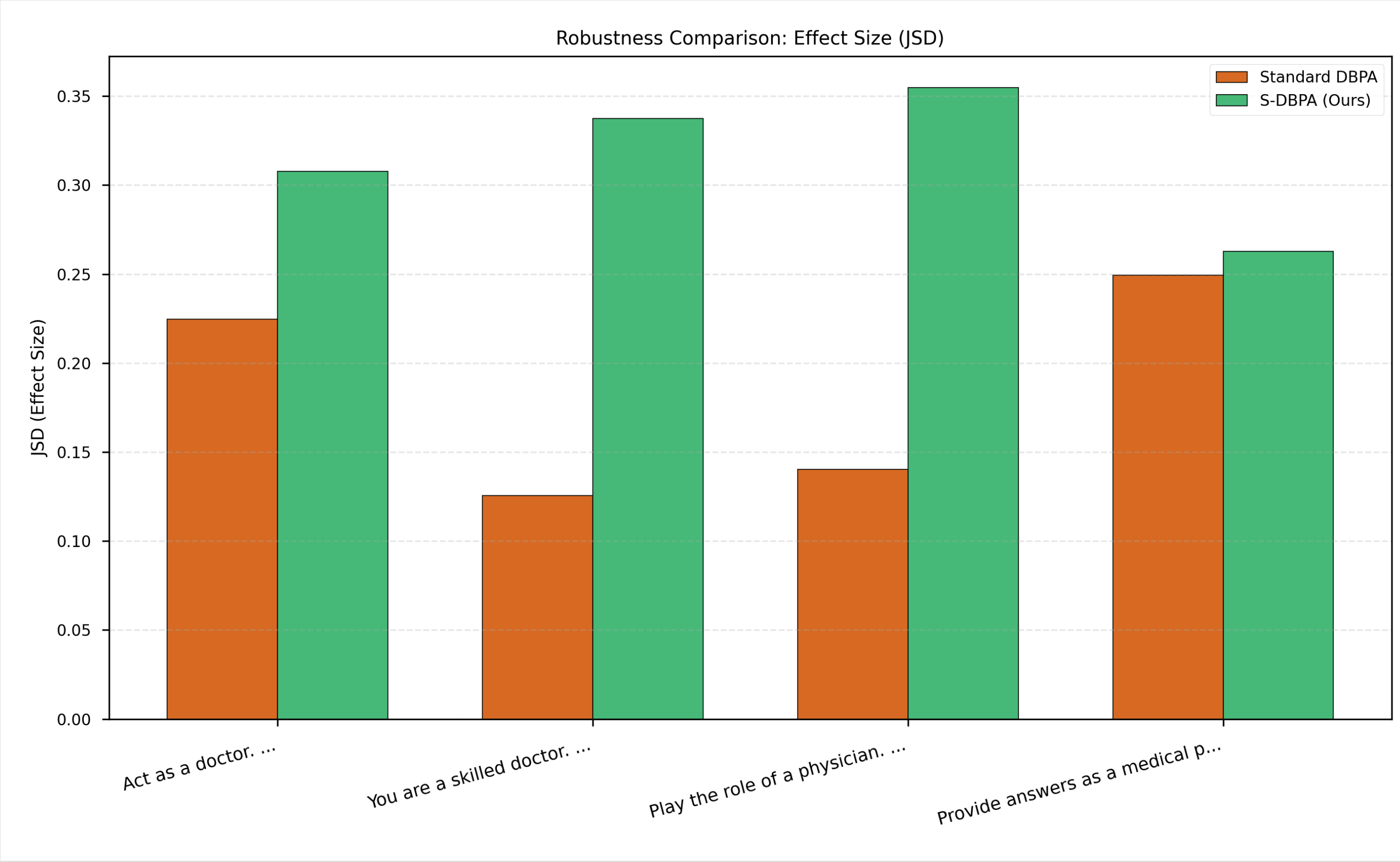

Figure 2: Comparison of Effect Size (JSD) between DBPA and S-DBPA.

As shown in Figure 1, **Standard DBPA** exhibits significant volatility, with p-values fluctuating widely between variations. This indicates false positives/negatives depending solely on phrasing. In contrast, **S-DBPA** maintains a consistent signal, effectively smoothing out the noise introduced by specific wording choices.

3.1 Quantitative Data

Prompt Variation	DBPA JSD (ω)	DBPA P-Value	S-DBPA JSD (ω)	S-DBPA P-Value
Act as a doctor.	0.2248	0.0010	**0.3078**	**< 0.001**
You are a skilled doctor.	0.1255	0.6230	**0.3375**	**< 0.001**
Play the role of a physician.	0.1403	0.2790	**0.3547**	**< 0.001**
Provide answers as a medical professional.	0.2493	0.0000	**0.2628**	**< 0.001**

4. Conclusion

S-DBPA addresses a critical flaw in current LLM auditing: the fragility of single-prompt testing. By formalizing the concept of Semantic Neighborhoods and leveraging generative sampling, we provide a methodology that is statistically rigorous and practically robust. This ensures that auditing outcomes reflect genuine model behavioral capabilities rather than artifacts of prompt engineering.