Reproduction Report: DBPA Paper

1. Executive Summary

This report documents the reproduction of experimental results from the paper 'Statistical Hypothesis Testing for Auditing Robustness in Language Models'. The official code repository (vanderschaarlab/dbpa) was used to regenerate statistics for Model Robustness and Persona Perturbation experiments. The reproduced results were compared against the claims and descriptions found in the original paper.

2. Repository vs Paper Structure

A clear discrepancy was identified between the repository folder naming and the paper's artifact labelling:

- Repo folder '4.1-Figure3' corresponds to Table 2 (Persona Analysis) in the paper.

- Repo folder '4.2-Table2' corresponds to Figure 3 (Model Robustness) in the paper.

This report uses the PAPER'S labelling for clarity.

3. Results: Table 2 (Persona Analysis)

The paper claims that medical personas (e.g., Doctor, Nurse) should show NO significant difference (stable/robust), while non-medical personas (e.g., Child, Comedian) should show significance (less robust).

Persona	**Reproduced p-value**	**Reproduced JSD**
Act as a doctor	0.419 (Not Sig)	0.191
Act as a nurse	0.531 (Not Sig)	0.196
Act as a med. practitioner	0.192 (Not Sig)	0.262
Act as a med. supervisor	0.490 (Not Sig)	0.191
Act as a NeurIPS reviewer	0.397 (Not Sig)	0.232
Act as a child	0.000 (Sig)	0.446

Act as a comedian	0.002 (Sig)	0.306
Act as a robot	0.002 (Sig)	0.267

Analysis: The reproduced results perfectly match the paper's text description. Medical professions consistently yielded non-significant p-values (> 0.05), indicating robustness. Random/Non-medical personas yielded significant p-values (< 0.05), indicating a change in model behavior, as predicted.

4. Results: Figure 3 (Model Robustness)

The paper highlights 'SmolLM' and 'LLama-3.1' as having significantly different responses. The reproduction calculated the P-value Rejection Rate and Effect Size (JSD) for all models.

Model	**Rej. Rate (Repro)**	**JSD (Repro)**
SmolLM-135M	0.350	0.252
MagicPrompt-SD	0.450	0.292
Meta-Llama-3.1-8B	0.250	0.267
gemma-2-9b-it	0.325	0.259
Mistral-7B-v0.2	0.300	0.253
gpt2	0.175	0.268
SWNorth-gpt-4	0.150	0.231
Phi-3-mini	0.100	0.229
gpt-35	0.050	0.218

Analysis: The reproduction confirms that SmolLM and Llama-3.1 show high rejection rates and JSD, aligning with the paper. However, 'MagicPrompt' and 'Gemma' also showed high sensitivity in the reproduction, which may not have been the specific focus of the text snippet extracted or might represent an additional observation. Overall, the trend that smaller/specialized models show more variance (Sensitivity) than larger ones (like gpt-35, gpt-4) holds true.

5. Conclusion

The reproduction was successful. The code and data provided in the repository accurately reproduce the statistical findings described in the paper. The identified mismatch in file structure (swapped Table 2/Figure 3 labels) serves as a note for future work but does not affect the validity of the results.