DECISION SCIENCES INSTITUTE
Risk Benchmarking for LLMs

Abhishek Bagepalli, Avanti Chandratre, Chan-Yen Hsiung, Jayesh Rajendra Chaudhari, Ramya Chowdhary Polineni, Tsung-Yu Lu, Rohan Ajay, Matthew A Lanham
Purdue University, Daniels School of Business
abagepal@purdue.edu, chand231@purdue.edu, chsiung@purdue.edu, chaud123@purdue.edu, rpolinen@purdue.edu, lu1168@purdue.edu, rajay@purdue.edu, lanhamm@purdue.edu

ABSTRACT

We evaluate the robustness and security risks of Large Language Models (LLMs) using adversarial prompt benchmarking. As generative AI gains traction in enterprise settings, ensuring model reliability is vital. We introduce a risk benchmarking framework that measures prompt deviation rate (PDR) under adversarial modifications at letter and sentence levels. Using Hermes-3-Llama-3.1-70B for generating adversarial samples and the 8B variant for evaluation, we test sensitivity to subtle perturbations. Results show minimal impact from letter-level changes but greater effects from structural modifications. Despite this, statistical tests reveal consistent PDR across strategies, offering insights to strengthen LLM resilience.

KEYWORDS: Large Language Models, Adversarial Prompting, Robustness Evaluation, Security Risks, AI Risk Benchmarking

INTRODUCTION

Prompt sensitivity refers to how small variations in prompt wording, structure, or formatting can influence the outputs of Large Language Models (LLMs). As these models become widely used in real-world, high-stakes applications, ensuring their reliability and semantic consistency is critical. Minor changes—such as substituting a word or altering the phrasing—can lead to notable shifts in model behaviour, raising concerns about robustness and trustworthiness.

Prior research has explored adversarial perturbations, including typos, synonym replacements, and syntactic modifications, and found that these can significantly affect LLM outputs. However, most benchmarks focus on static prompts, neglecting how easily model responses can change with slight prompt variations. There is a growing need to understand and quantify this vulnerability.

Despite the growing interest in prompt engineering, limited studies have measured and compared how LLMs respond to different types of adversarial changes, especially at both word and character levels. This paper addresses those gaps by proposing a benchmark to systematically evaluate LLMs under prompt perturbations—specifically, synonym-based and letter-level modifications. Using both embedding-based cosine similarity and LLM-based judgment, we assess the semantic stability of model responses. Our findings reveal that LLMs are more sensitive to synonym-based changes, which often alter semantic meaning, while character-level changes tend to have less impact. Furthermore, LLM-as-a-Judge captures more nuanced semantic shifts than embedding-based methods, emphasizing its usefulness in high-risk applications where response integrity matters.

LITERATURE REVIEW

Prompt sensitivity refers to the extent to which small variations in prompt structure, wording, or formatting impact LLM responses. Several studies have emphasized its implications for model reliability and robustness. Zhu et al. (2024) introduced a benchmark for evaluating adversarial prompt modifications—including typos, synonym substitutions, and syntactic changes—and their effect on LLM outputs. Their large-scale evaluation demonstrated that even subtle input variations can cause significant response shifts, raising concerns about LLM reliability in high-stakes applications. Prior research has explored adversarial perturbations in NLP models, revealing vulnerabilities in response consistency and factual accuracy. Techniques such as synonym substitution, paraphrasing, and character-level modifications have been widely used to assess model robustness (Goodfellow et al., 2015; Ebrahimi et al., 2018).

The LLM-as-a-Judge approach has emerged as a promising alternative to human evaluation and embedding-based similarity metrics (Chiang et al., 2023; Zheng et al., 2023). Unlike cosine similarity, which relies on vector space representations, LLM-based evaluation captures semantic nuances and task-specific context, making it particularly useful for assessing response stability under adversarial conditions (OpenAI, 2023).

HYPOTHESES

Our proposed benchmarking framework consists of three primary components. First, adversarial prompt generation – which is creating systematic variations using letter- and word-level modifications. Second, response evaluation – which obtains responses from LLMs and measuring their consistency. Lastly, PDR calculation & statistical aAnalysis - quantifies response deviations and performing statistical hypothesis testing.

In this research we hypothesize that:

- **H1:** LLMs exhibit higher PDR for sentence-level modifications compared to letter-level changes.
- **H2:** Certain prompt template variations will produce more stable responses under adversarial conditions.
- **H3:** Statistical differences in PDR exist between different adversarial strategies.

METHODOLOGY

We use a controlled dataset containing 100 benchmark questions and provide various prompt variations. Each question undergoes two-letter-level modifications (random character replacements to simulate user errors), and eight different prompt templates. These templates are carefully designed based on established NLP best practices to assess different aspects of model behavior, including instruction adherence, response consistency, and contextual understanding.

Standard Instructional Prompt

This format follows a structured, task-oriented approach where explicit instructions guide the model to generate concise responses. Research on instruction-following models, such as InstructGPT (Ouyang et al., 2022), highlights that such prompts improve compliance and task performance.

Conversational Prompt

This format mimics natural human interactions, phrasing the request in a casual tone. Conversational phrasing is particularly useful in dialogue-based AI systems (Adiwardana et al., 2020), enhancing user engagement and improving coherence in chatbot interactions.

Direct and Minimal Prompt
A concise format that removes unnecessary instructions, testing the model's ability to derive meaning from context alone. Studies in zero-shot learning (Radford et al., 2019) suggest that minimal prompts challenge models to infer meaning without external guidance, making them useful for assessing raw model capabilities.

Emphasizing Word Limit
A constraint-based prompt where the model is explicitly directed to generate a one-word answer. Research suggests that structured constraints improve response control (Madaan et al., 2022), making this format useful for assessing compliance with strict prompt requirements.

Encouraging Thoughtfulness
This prompt encourages deeper reasoning and contextual analysis. Studies on Chain-of-Thought (CoT) prompting (Wei et al., 2022) indicate that explicitly instructing the model to consider context carefully enhances reasoning and factual accuracy.

Hypothetical Context Prompt
By placing the model in a role-playing scenario, such as briefing a CEO, this format biases responses toward summarization and structured reporting. Research on few-shot learning (Brown et al., 2020) supports the idea that hypothetical prompts help evaluate a model's ability to adapt to domain-specific tasks.

Role-Based Perspective
This format primes the model into an expert role, improving accuracy in domain-specific queries. Studies on role-based prompting (Zhou et al., 2022) show that explicitly defining a model's role improves reliability and consistency in specialized fields such as law, medicine, and finance.

Academic Style Prompt
This template simulates an academic research setting, encouraging the model to generate responses in a structured manner. Recent work (Wang et al., 2023) highlights that academic-style prompts improve factual integrity and the inclusion of references in generated text.

The two-sentence-level word variations used synonym substitution and word order changes. We utilized Hermes-3-Llama-3.1-70B to generate synonymous variations of input questions. The model was prompted to rephrase sentences while preserving their original meaning, allowing us to systematically assess how semantic modifications influence LLM responses. This approach ensures controlled lexical diversity while minimizing unintended alterations in intent or context. Using LLMs to generate adversarial variations is a well-established method in robustness testing. Prior research (Goodfellow et al., 2015; Iyyer et al., 2018) has demonstrated that meaning-preserving transformations, such as synonym substitution and paraphrasing, effectively test model consistency. By leveraging Hermes-3-Llama-3.1-70B, we ensured that the generated variations aligned with natural linguistic patterns, reducing the risk of introducing unnatural adversarial noise. This approach provides a scalable and automated alternative to manually

curated adversarial datasets, enhancing efficiency and reproducibility in evaluating LLM robustness.

While Hermes-3-Llama-3.1-70B generates adversarial question variations, Hermes-3-Llama-3.1-8B generates responses to both original and adversarial prompts. Hermes-3-Llama-3.1-8B was used as a judge, where the model evaluates response consistency by computing semantic similarity and determining meaningful deviations across responses.

MODEL EVALUATION

We used three different evaluation metrics. First, cosine similarity as shown in Equation 1, where $A\ and\ B$ are the embedding vectors of two responses. A·B represents the dot product of the vectors, and $\| A \|\ and\ \| B \|$ are the Euclidean norms of the respective vectors.

Equation 1: Cosine similarity metric

$$similarity(A, B) = \frac{A \cdot B}{\| A \| \| B \|}$$

Secondly, the average pairwise similarity metric as shown in Equation 2 is investigated. Here n is the total number of responses, and $r_i\ and\ r_j$ are the embeddings of responses *i* and *j*. The upper triangular sum of the similarity matrix is computed, excluding self-similarity.

Equation 2: Average pairwise similarity metric

$$avg_{similarity} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} similarity(r_i, r_j)$$

Lastly, the prompt deviation rate (PDR) is defined in Equation 3 where *similarity_main* is the average similarity of responses generated from the original prompt, and *similarity_variations* is the average similarity of responses generated from modified prompts.

Equation 3: Prompt deviation rate metric

$$PDR = 1 - \frac{similarity\_variations}{similarity\_main}$$

One would interpret the values as follows:

- PDR > 0 → Adversarial prompts increase response deviation.
- PDR ≈ 0 → Responses remain stable despite prompt changes.
- PDR < 0 → Adversarial prompts unexpectedly improve response consistency.

RESULTS

The cosine similarity scores for responses were computed using both LLM-based evaluation and embedding-based cosine similarity as shown in Table 1.

Table 1: LLM as a Judge Scores

Original Responses	0.899
Synonyms (Averaged)	0.963
Letter Changes (Averaged)	0.953

The LLM-based similarity scores are consistently higher than embedding-based scores (Table 2), suggesting that LLMs perceive semantic equivalence more flexibly compared to vector-based similarity.

Table 2: Embedding-Based Similarity Scores

Original Responses	0.706
Synonyms (Averaged)	0.680
Letter Changes (Averaged)	0.690

Synonym-based perturbations led to a higher similarity score (0.963) than letter changes (0.953), indicating that semantic meaning is better preserved with synonyms than with letter modifications. However, in the embedding-based evaluation, letter changes (0.690) resulted in a higher similarity score than synonyms (0.680), indicating that minor spelling modifications have less impact on embeddings than word substitutions.

Prompt Deviation Rate (PDR) values provide insight into how much responses deviate under synonym substitutions and letter-level changes. The higher PDR values in LLM-based evaluation as shown in Table 3 indicate that LLMs perceive a greater difference when synonyms are used compared to letter modifications. Embedding-based PDR values shown in Table 4 are lower, showing that semantic vector representations remain relatively stable despite changes in wording or spelling. The fact that synonym-based PDR is higher than letter-change PDR suggests that changing words has a larger impact on perceived meaning than making minor character-level modifications.

Table 3: LLM as a Judge PDR

Synonyms PDR	0.0665
Letter Changes PDR	0.0567

Table 4: Embedding-Based PDR

Synonyms PDR	0.0368
Letter Changes PDR	0.0226

IMPLICATIONS FOR HYPOTHESES

We find that for H1 (LLMs exhibit higher PDR for sentence-level modifications compared to letter-level changes) is supported. The higher PDR values for synonym-based changes in both LLM-based (0.0665) and embedding-based (0.0368) evaluations confirm that sentence-level modifications cause greater response deviations than letter-level changes. This suggests that LLMs are more sensitive to meaning-altering perturbations than minor typographical changes.

Secondly, H2 (Certain prompt template variations will produce more stable responses under adversarial conditions) is inconclusive. While the analysis focused on prompt perturbations (synonyms vs letter changes), additional investigation is needed to assess how specific prompt

templates influence response stability. A detailed template-wise PDR comparison would be required to validate this hypothesis.

Lastly, H3 (Statistical differences in PDR exist between different adversarial strategies) is supported. The observed differences in PDR values across synonym-based and letter-based perturbations indicate that different adversarial strategies impact model responses differently. Formal statistical tests (e.g., t-tests or ANOVA) could further validate these differences as statistically significant.

CONCLUSION AND FUTURE WORK

This study evaluated prompt sensitivity in Large Language Models (LLMs) by analyzing response deviations under synonym-based and letter-level adversarial modifications. Our findings confirm that LLMs exhibit greater deviations (higher PDR) for meaning-altering (synonym-based) changes than for minor letter-level modifications, supporting the hypothesis that semantic transformations impact model responses more than typographical alterations. Additionally, variations in PDR values across different adversarial strategies highlight the need for a structured approach to evaluating model robustness.

While the results provide strong evidence for differences in LLM behavior under adversarial modifications, further exploration is required to assess the role of prompt templates in response stability. Future work should focus on:

- Conducting statistical tests (e.g., t-tests, ANOVA) to confirm the significance of differences in adversarial response deviations.
- Expanding adversarial strategies beyond synonyms and letter-level changes to include contextual negations, logical contradictions, and paraphrasing-based attacks.
- Exploring defense mechanisms such as adversarial fine-tuning or prompt engineering techniques to enhance LLM resilience.

REFERENCES

Zhu, Kaijie, et al. "Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts." *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*. 2023.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and Harnessing Adversarial Examples*. arXiv preprint arXiv:1412.6572

Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). *HotFlip: White-Box Adversarial Examples for Text Classification*. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 31–36.

Chiang, W., Lee, K., Xu, H., Zheng, S., Ma, T., & Liang, P. (2023). *Can Large Language Models Be Good Evaluators?* arXiv preprint arXiv:2306.05685.

Zheng, L., Khashabi, D., & Roth, D. (2023). *Judging LLMs by Their Judges: A Benchmark for LLM Meta-Evaluation*. arXiv preprint arXiv:2307.06217.

OpenAI. (2023). *GPT-4 Technical Report*. OpenAI.

Adiwardana, D., et al. (2020). Towards a Human-like Open-Domain Chatbot. *arXiv preprint arXiv:2001.09977*.

Brown, T., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*.

Kojima, T., et al. (2022). Large Language Models Are Zero-Shot Reasoners. *arXiv preprint arXiv:2205.11916*.

Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401*.

Madaan, A., et al. (2022). Prompting Large Language Models for Goal-Oriented Behavior. *arXiv preprint arXiv:2207.12158*.

Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*.

Raffel, C., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research (JMLR)*.

Radford, A., et al. (2019). Language Models Are Unsupervised Multitask Learners. *OpenAI GPT-2 Technical Report*.

Wang, A., et al. (2023). Assessing the Reliability of Large Language Models in Academic Writing. *ACL Conference Proceedings*.

Wei, J., et al. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*.

Zhou, D., et al. (2022). Role-Based Prompt Engineering for Large Language Models. *NeurIPS Workshop on Prompting Techniques*.