# Security and Privacy of Machine Learning, 2025 Critique G3: Adversarial Attack on LLMs – (1) Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment (2) DA$^3$: A Distribution-Aware Adversarial Attack against Language Models

Shih-Yu Lai
*National Taiwan University*
Taipei, Taiwan
akinesia112@gmail.com

## Critique:(1) Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment

### Summary

This paper investigates the adversarial robustness of Large Language Models (LLMs) when used as zero-shot evaluators ("LLM-as-a-judge"). The authors propose simple yet effective universal adversarial attacks, where short token sequences are appended to candidate texts to inflate quality scores. They introduce a surrogate-based attack method, learning adversarial phrases on FlanT5-xl and transferring them to stronger judge models (Llama2-7B, Mistral-7B, GPT-3.5). The results show that absolute scoring is highly vulnerable (with four-token phrases consistently forcing maximum scores), while comparative assessment is more robust. Perplexity-based detection is proposed as an initial defense, though adaptive attacks could bypass it. The study raises concerns about deploying LLM judges in high-stakes applications such as benchmarking and academic grading.

### Strengths

- **Novel contribution:** The first systematic study of adversarial robustness in zero-shot LLM assessment, filling an important gap left by prior evaluation work [1], [2].
- **Clear methodology:** The greedy search algorithm for generating universal adversarial phrases is simple and reproducible.
- **Transferability analysis:** Demonstrates that adversarial phrases learned on a small surrogate model generalize to larger LLMs, highlighting real-world feasibility.
- **Comparative vs. absolute evaluation:** The contrast between paradigms provides an actionable insight: robustness is higher for comparative setups but at greater computational cost.
- **Responsible framing:** The authors explicitly address risks, ethics, and licensing, ensuring transparency.

### Weaknesses

- **Narrow attack design:** Only concatenative attacks with greedy search are studied. Other adversarial paradigms, such as paraphrase-based or optimization-driven methods [3], are excluded.
- **Simplistic defense:** Perplexity-based detection, while promising [4], is unlikely to withstand adaptive attackers.
- **Dataset limitations:** The experiments focus on SummEval [5] and TopicalChat, leaving out broader domains like machine translation, legal writing, or scientific text.
- **Over-reliance on zero-shot:** Few-shot or fine-tuned evaluators, which may exhibit greater robustness, are not explored.
- **Limited interpretability:** Although some adversarial phrases are interpretable (e.g., "outstandingly"), a systematic linguistic analysis is missing.

### Potential Improvements and Extensions

- Explore **richer adversarial strategies** beyond concatenation, including paraphrasing, synonym substitution, or reinforcement learning-based prompt optimization.
- Conduct **cross-domain evaluations** on additional tasks (e.g., factual QA, translation, long-form essay scoring).
- Investigate **few-shot and fine-tuned assessment** systems to test whether they are inherently more robust.
- Develop **adaptive defenses**, such as adversarial training, ensemble comparative scoring, or robust prompting strategies.

- Provide a deeper **linguistic analysis** of adversarial phrases, bridging adversarial robustness with interpretability.

### Questions for the Authors

1) How would the attacks perform on evaluation tasks with longer or more technical texts (e.g., scientific abstracts, legal writing)?
2) Could adversarial phrases be detected using attribution methods (e.g., gradient-based saliency) rather than perplexity?
3) Would combining absolute and comparative assessment (hybrid scoring) balance efficiency and robustness?
4) Are there systematic linguistic features of effective attack phrases (e.g., sentiment, fluency markers)?
5) How do these attacks interact with evaluation systems trained via reinforcement learning from human feedback (RLHF)?

## Critique: (2) DA$^3$: A Distribution-Aware Adversarial Attack against Language Models

### Summary

The paper proposes **DA3 (Distribution-Aware Adversarial Attack)**, a new adversarial attack framework targeting language models (LMs). The key insight is that adversarial examples generated by prior methods (e.g., BERT-Attack) exhibit **distribution shifts** from original data in terms of *Maximum Softmax Probability (MSP)* and *Mahalanobis Distance (MD)*, which makes them easily detectable by out-of-distribution (OOD) detection methods. To address this, the authors design a **Data Alignment Loss (DAL)** that aligns adversarial examples with original examples across MSP and MD. They also propose a novel metric, **Non-detectable Attack Success Rate (NASR)**, which incorporates both attack success and detectability. Experiments across four NLP tasks (SST-2, CoLA, RTE, MRPC) show that DA3 achieves strong attack success, better resistance to detection, and transferability to black-box LLMs (LLAMA2-7B). Human evaluation further confirms that DA3 generates natural and semantically preserved adversarial text.

### Strengths

- **Novel perspective:** The paper highlights distribution shifts (via MSP and MD) as a key weakness of existing attacks. This is an insightful contribution rarely emphasized in prior work.
- **Methodological innovation:** The proposed DAL is simple yet effective, balancing adversarial success with detectability resistance.
- **New evaluation metric:** NASR provides a more realistic measure of adversarial quality, penalizing attacks that are trivially detected.
- **Comprehensive experiments:** Evaluations on both white-box (BERT, RoBERTa) and black-box (LLAMA2-7B) models, with ablations and human studies, strengthen the empirical claims.
- **Transferability:** Demonstrating effectiveness against large LLMs underscores DA3's practical significance.

### Weaknesses

- **Limited scope of distribution metrics:** The method only considers MSP and MD. Other OOD detection methods (energy scores [6], deep k-NN [7]) may also expose vulnerabilities.
- **Dataset dependence:** The MD distribution shift is not consistent across datasets (e.g., weak on MRPC). This questions whether DAL always improves robustness.
- **Attack diversity:** The experiments are restricted to classification tasks. Sequence generation (e.g., summarization, translation) may present different challenges.
- **Trade-offs:** The paper notes tension between increasing adversarial confidence (MSP alignment) and preserving closeness to training distribution (MD alignment), but does not fully resolve this optimization conflict.
- **Ethical discussion is minimal:** Stronger analysis of misuse risks would be valuable, given DA3 produces harder-to-detect adversarial content.

### Potential Improvements and Extensions

- Extend DAL to incorporate other distributional measures such as **energy-based scores** [6] or **deep nearest-neighbor distances** [7].
- Apply DA3 to **generative LLM tasks** (dialogue, summarization) to test robustness beyond classification.
- Investigate **adaptive defenders** that retrain on DA3 examples; assess whether DA3 maintains effectiveness under adversarial training.
- Explore **lightweight alternatives** to DAL that reduce computational cost, making DA3 more practical for large-scale attacks/defenses.

### Questions

- How would DA3 perform against more advanced OOD detectors beyond MSP and MD?
- Could DAL overfit to specific datasets, limiting cross-domain generalization?
- What is the computational overhead of fine-tuning with DAL compared to standard attacks?
- Could defenders exploit the trade-off between MSP and MD alignment to detect DA3 examples?
- How does DA3 interact with recent **defensive training methods** (e.g., adversarial fine-tuning in NLP [8])?

## References


[1] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *arXiv preprint arXiv:2306.05685*, 2023.

[2] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 2511–2522.

[3] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2014.
[5] A. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," in *Transactions of the Association for Computational Linguistics*, vol. 9, 2021, pp. 391–409.
[6] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 21 464–21 475.
[7] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *International Conference on Machine Learning (ICML)*, 2022, pp. 20 827–20 840.
[8] J. Y. Yoo and Y. Qi, "Towards improving adversarial training of nlp models," in *Findings of the Association for Computational Linguistics: EMNLP*, 2021, pp. 945–956.