View Review

Paper ID
3878

Paper Title
Incremental Exploits: Efficient Jailbreak on Large Language Models with Multi-round Interactions

Track Name
Main Track

REVIEW QUESTIONS

1. Reviewer's confidence
Very good

2. Relevance to IJCNN
Very good

3. Technical quality
Good

4. Novelty
Good

5. Quality of presentation
Good

6. Award quality?
No

7. Suggested type of presentation
Poster

9. Overall recommendation
Borderline

9. Comments to Authors
The paper provides a clear and comprehensive approach to addressing multi-round attacks on large language models. The experimental results show that MIEJ outperforms existing methods in terms of both success rate and query efficiency. The method's scalability and cross-model applicability are particularly noteworthy. It would be beneficial to explore further how these findings can be used to improve safety mechanisms in real-world LLM deployments.

10. Was Authors' anonymity ensured? (If No, please explain in confidential comments to TPC)
Yes

11. Confidential comments to Technical Program Committee

This paper introduces a highly efficient and effective multi-round jailbreak attack that reveals significant vulnerabilities in large language models. Its methodology is robust and addresses an emerging security concern in AI.

12. All reviewers must accept and agree to follow IEEE Policies. By submitting your review, you acknowledge that you have read and agree to IEEE's Privacy Policy (https://www.ieee.org/security-privacy.html).

Agreement accepted