View Review

Paper ID

2279

Paper Title

AttentionDefense: Leveraging System Prompt Attention for Explainable Defense Against Novel Jailbreaks

Track Name

Main Track

REVIEW QUESTIONS

1. Reviewer’s confidence

Very good

2. Relevance to IJCNN

Very good

3. Technical quality

Very good

4. Novelty

Excellent

5. Quality of presentation

Good

6. Award quality?

No

7. Suggested type of presentation

Oral

9. Overall recommendation

Accept

9. Comments to Authors

The paper presents a highly innovative approach to detecting jailbreaks in language models, and the results are impressive. The use of system-prompt attention is an elegant solution that balances explainability, computational efficiency, and robustness. A suggestion for future work could involve exploring how the system-prompt attention mechanism could be further generalized to other forms of adversarial input beyond jailbreaks. Additionally, while the results for novel jailbreaks are strong, more practical deployment scenarios or case studies could help better communicate the impact of your approach in real-world systems.

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

Yes

11. Confidential comments to Technical Program Committee

The proposed approach is highly relevant and brings substantial novelty to the field of AI model security. The paper's ability to handle both known and novel jailbreaks effectively, using computationally light models, sets it apart from existing solutions. There are no major concerns, and it stands out as a solid contribution to the field. Further discussions could be focused on the practical integration of this method into production environments.

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