

 **Microsoft CMT** <email@msr-cmt.org>

to me ▾

9:36 PM (10 minutes ago)



Hello,

The following submission has been edited.

Track Name: Special Session 2: Machine Learning for Natural Language Processing

Paper ID: 684

Paper Title: Detecting Prompt Injection Attacks Against Application Using Classifiers

**Abstract:**

Prompt injection attacks are a growing concern in various domains, as they can compromise the security and stability of critical systems, from power grids to large-scale web applications. In this research, we propose a comprehensive approach for detecting and mitigating prompt injection attacks against web applications using predefined system prompts— by highly curating a prompt injection dataset and training it on classifiers like Long Short-Term Memory (LSTM) networks to FNN and machine learning models such as Random Forest Classifier and Naive Bayes. We build upon a pre-existing dataset from HuggingFace, named HackAPrompt-PlaygroundSubmissions. Our proposed solution aims to improve the detection and mitigation of prompt injection attacks, ensuring the security and stability of the targeted applications and systems.

Created on: Tue, 05 Sep 2023 15:31:54 GMT

Last Modified: Tue, 05 Sep 2023 15:36:17 GMT

**Authors:**

- [safwan.shaheen@g.bracu.ac.bd](mailto:safwan.shaheen@g.bracu.ac.bd) (Primary)
- [gw.refatul.islam@g.bracu.ac.bd](mailto:gw.refatul.islam@g.bracu.ac.bd)
- [mohammad.rafid.hamid@g.bracu.ac.bd](mailto:mohammad.rafid.hamid@g.bracu.ac.bd)
- [md.abrar.faiyaz.khan@g.bracu.ac.bd](mailto:md.abrar.faiyaz.khan@g.bracu.ac.bd)
- [md.omar.faruk@g.bracu.ac.bd](mailto:md.omar.faruk@g.bracu.ac.bd)
- [yaseen.nur.taz@g.bracu.ac.bd](mailto:yaseen.nur.taz@g.bracu.ac.bd)

Secondary Subject Areas: Not Entered

Submission Files: Detecting Mitigating Prompt Injection Attacks Against Applications Using Classifiers.pdf (2 Mb, Tue, 05 Sep 2023 15:31:37 GMT)

Submission Questions Response: Not Entered

Thanks,

CMT team.

Download the CMT app to access submissions and reviews on the move and receive notifications:

<https://apps.apple.com/us/app/conference-management-toolkit/id1532488001><https://play.google.com/store/apps/details?id=com.microsoft.research.cmt>

To stop receiving conference emails, you can check the 'Do not send me conference email' box from your User Profile.

Microsoft respects your privacy. To learn more, please read our [Privacy Statement](#).

Microsoft Corporation

One Microsoft Way

Redmond, WA 98052

 Reply Forward