CS839
Presentation Proposal

Group Members -

- **Atharv Prajod Padmalayam**
- **Brennen Hill**
- **Sujay Sharma**
- **Surendra Parla**
- **Venkata Abhijeeth Balabhadruni**

Selected Paper - TrojText: Test-time Invisible Textual Trojan Insertion

Much emphasis has been placed on alignment, accuracy, and ensuring model robustness in the landscape of foundation models. However, there has been less focus on understanding potential pathways for model misalignment, especially in the context of textual Trojan attacks, where models can be subtly influenced to produce malicious or unexpected outputs without obvious triggers. This paper introduces TrojText, a novel approach for implementing efficient, invisible Trojan attacks on NLP models without the need for large training data. It presents an innovative algorithm, Representation-Logit Trojan Insertion (RLI), along with techniques like accumulated gradient ranking (AGR) and Trojan Weights Pruning (TWP), to minimize attack overhead and tuning requirements. We chose this paper because it offers an opportunity to explore vulnerabilities in language models that are difficult to detect and defend against, helping us gain a deeper understanding of adversarial NLP and misalignment risks within foundation models.

This paper aligns closely with several topics covered in CS 839, such as model robustness, security, and future areas in foundation models. It extends discussions from the *Security, Privacy, Toxicity + Future Areas* lecture by examining how NLP models can be exploited through subtle, invisible triggers, posing unique challenges for defense mechanisms in foundation models. This presentation will provide insights into the vulnerabilities of foundation models, reinforcing the importance of robustness and security as we continue to depend on foundation models and try to improve them.