Week 2 Workshop

FIT5230 Malicious AI

Adversarial Machine Learning I

Activity 1

1. **EXPLORING THE ADVERSARIAL CAPABILITIES OF LARGE LANGUAGE MODELS**

   Struppek, L., Le, M. H., Hintersdorf, D., & Kersting, K. (2024). Exploring the Adversarial Capabilities of Large Language Models. arXiv preprint arXiv:2402.09132.

   https://arxiv.org/pdf/2402.09132

2. **CODEATTACK: CODE-BASED ADVERSARIAL ATTACKS FOR PRE-TRAINED PROGRAMMING LANGUAGE MODELS**

   Jha, A., & Reddy, C. K. (2023, June). Codeattack: Code-based adversarial attacks for pre-trained programming language models. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 12, pp. 14892-14900).

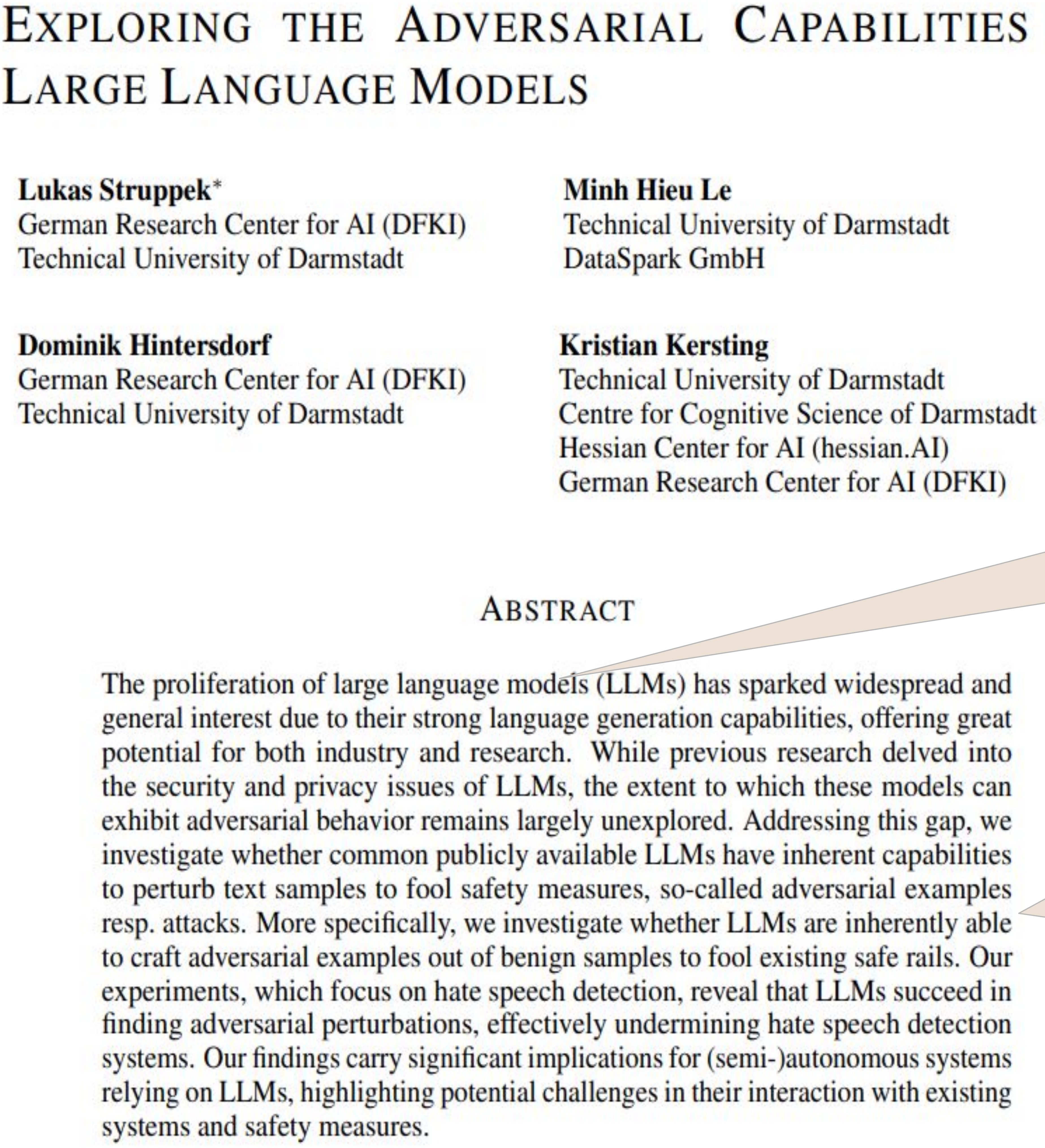
   https://doi.org/10.1609/aaai.v37i12.26739

3. **ONE PROMPT WORD IS ENOUGH TO BOOST ADVERSARIAL ROBUSTNESS FOR PRE-TRAINED VISION-LANGUAGE MODELS**

   Li, L., Guan, H., Qiu, J., & Spratling, M. (2024). One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 24408-24419).

   https://openaccess.thecvf.com/content/CVPR2024/papers/Li_One_Prompt_Word_is_Enough_to_Boost_Adversarial_Robustness_for_CVPR_2024_paper.pdf

Dark



Exploring the Adversarial Capabilities of Large Language Models

Lukas Struppek*
German Research Center for AI (DFKI)
Technical University of Darmstadt

Minh Hieu Le
Technical University of Darmstadt
DataSpark GmbH

Dominik Hintersdorf
German Research Center for AI (DFKI)
Technical University of Darmstadt

Kristian Kersting
Technical University of Darmstadt
Centre for Cognitive Science of Darmstadt
Hessian Center for AI (hessian.AI)
German Research Center for AI (DFKI)

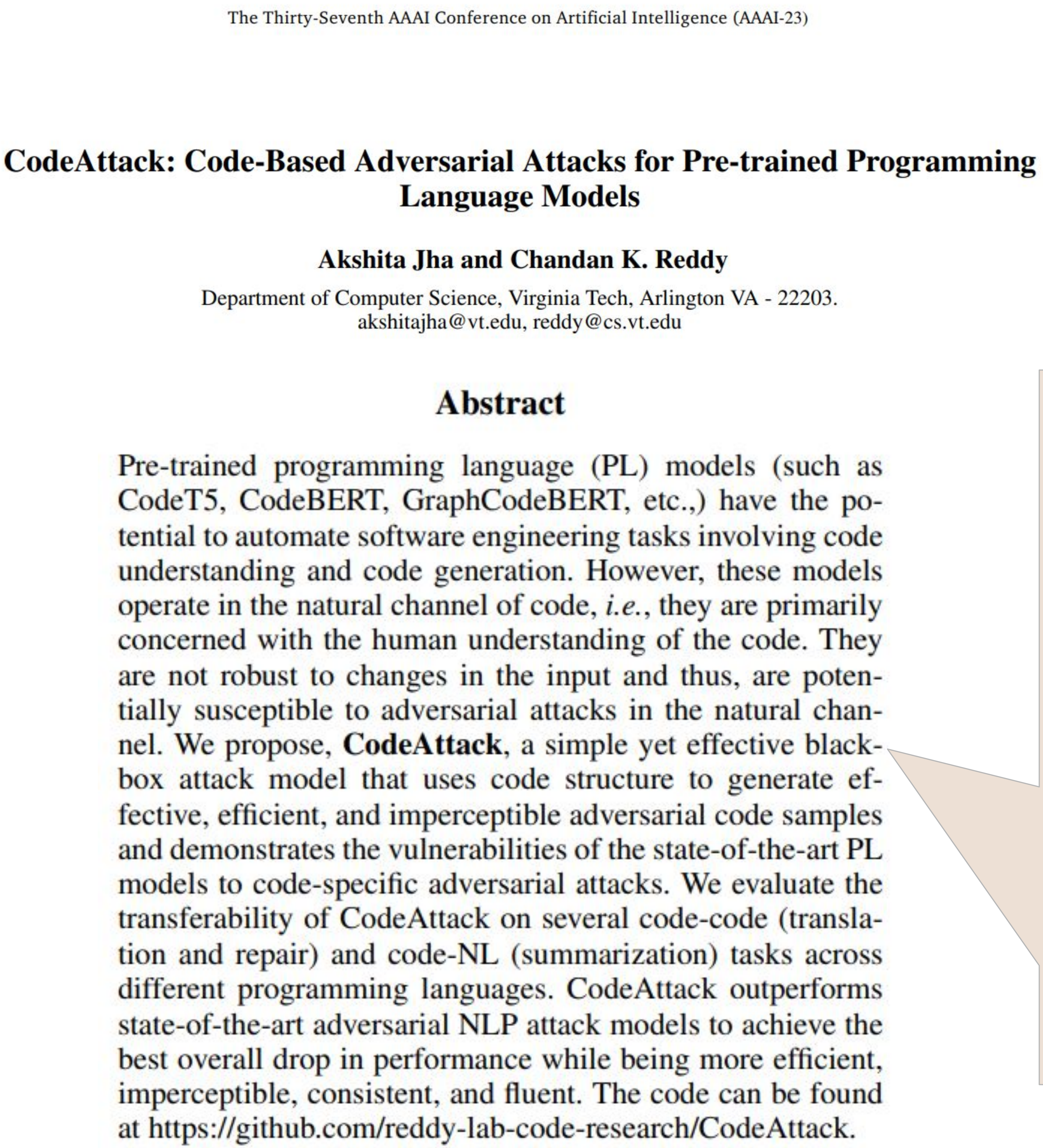
Abstract

The proliferation of large language models (LLMs) has sparked widespread and general interest due to their strong language generation capabilities, offering great potential for both industry and research. While previous research delved into the security and privacy issues of LLMs, the extent to which these models can exhibit adversarial behavior remains largely unexplored. Addressing this gap, we investigate whether common publicly available LLMs have inherent capabilities to perturb text samples to fool safety measures, so-called adversarial examples resp. attacks. More specifically, we investigate whether LLMs are inherently able to craft adversarial examples out of benign samples to fool existing safe rails. Our experiments, which focus on hate speech detection, reveal that LLMs succeed in finding adversarial perturbations, effectively undermining hate speech detection systems. Our findings carry significant implications for (semi-)autonomous systems relying on LLMs, highlighting potential challenges in their interaction with existing systems and safety measures.

A large language model (LLM) is a type of artificial intelligence (AI) program that can recognize and generate text, among other tasks, e.g., ChatGPT.

Aim: To check if the existing LLMs are able to craft adversarial examples from benign samples.

Dark



The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

CodeAttack: Code-Based Adversarial Attacks for Pre-trained Programming Language Models

Akshita Jha and Chandan K. Reddy

Department of Computer Science, Virginia Tech, Arlington VA - 22203.
akshitajha@vt.edu, reddy@cs.vt.edu

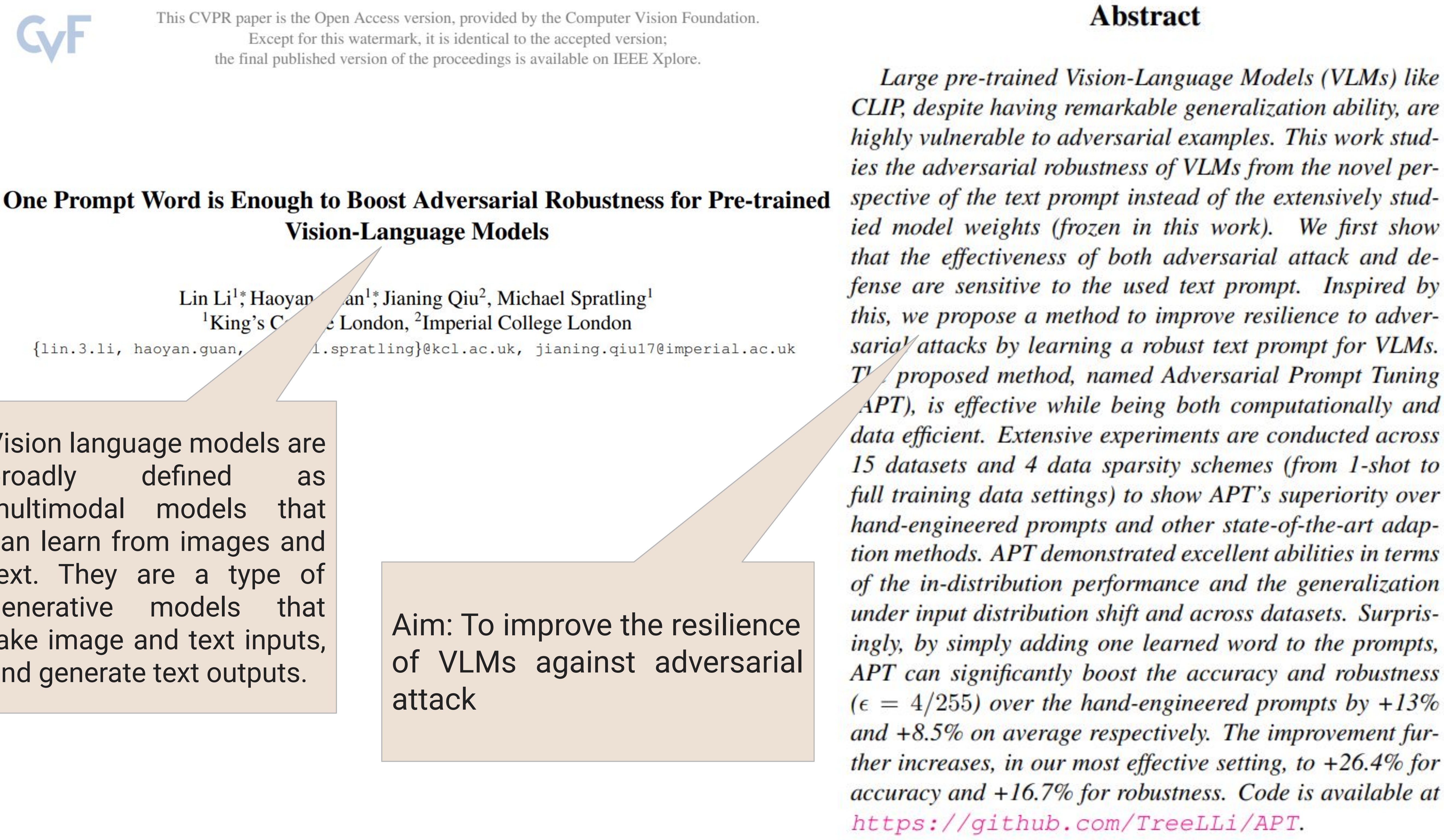
Abstract

Pre-trained programming language (PL) models (such as CodeT5, CodeBERT, GraphCodeBERT, etc.,) have the potential to automate software engineering tasks involving code understanding and code generation. However, these models operate in the natural channel of code, *i.e.*, they are primarily concerned with the human understanding of the code. They are not robust to changes in the input and thus, are potentially susceptible to adversarial attacks in the natural channel. We propose, **CodeAttack**, a simple yet effective black-box attack model that uses code structure to generate effective, efficient, and imperceptible adversarial code samples and demonstrates the vulnerabilities of the state-of-the-art PL models to code-specific adversarial attacks. We evaluate the transferability of CodeAttack on several code-code (translation and repair) and code-NL (summarization) tasks across different programming languages. CodeAttack outperforms state-of-the-art adversarial NLP attack models to achieve the best overall drop in performance while being more efficient, imperceptible, consistent, and fluent. The code can be found at https://github.com/reddy-lab-code-research/CodeAttack.

Aim:

- To explore the vulnerabilities of pre-trained programming language (PL) models to adversarial attacks using CodeAttack, specifically focusing on code-based attacks that exploit the structure of the code to generate adversarial examples.
- To evaluate the transferability of CodeAttack across different programming language.

Light



One Prompt Word is Enough to Boost Adversarial Robustness for Pre-trained Vision-Language Models

Lin Li[1]*, Haoyan Guan[1]*, Jianing Qiu[2], Michael Spratling[1]
[1]King's College London, [2]Imperial College London
{lin.3.li, haoyan.guan, michael.spratling}@kcl.ac.uk, jianing.qiu17@imperial.ac.uk

Abstract

Large pre-trained Vision-Language Models (VLMs) like CLIP, despite having remarkable generalization ability, are highly vulnerable to adversarial examples. This work studies the adversarial robustness of VLMs from the novel perspective of the text prompt instead of the extensively studied model weights (frozen in this work). We first show that the effectiveness of both adversarial attack and defense are sensitive to the used text prompt. Inspired by this, we propose a method to improve resilience to adversarial attacks by learning a robust text prompt for VLMs. The proposed method, named Adversarial Prompt Tuning (APT), is effective while being both computationally and data efficient. Extensive experiments are conducted across 15 datasets and 4 data sparsity schemes (from 1-shot to full training data settings) to show APT's superiority over hand-engineered prompts and other state-of-the-art adaption methods. APT demonstrated excellent abilities in terms of the in-distribution performance and the generalization under input distribution shift and across datasets. Surprisingly, by simply adding one learned word to the prompts, APT can significantly boost the accuracy and robustness ($\epsilon = 4/255$) over the hand-engineered prompts by +13% and +8.5% on average respectively. The improvement further increases, in our most effective setting, to +26.4% for accuracy and +16.7% for robustness. Code is available at https://github.com/TreeLLi/APT.

Vision language models are broadly defined as multimodal models that can learn from images and text. They are a type of generative models that take image and text inputs, and generate text outputs.

Aim: To improve the resilience of VLMs against adversarial attack

Activity 2

What you need to do: (30 mins)

1. Choose **1 paper** from the list.
2. **Summarize** the key points of the paper here: https://shorturl.at/FPSyM
3. Teams will be randomly selected to **present** their ideas.