October 25, 2023

Large Language Models, Security, and Privacy

Description

This course will cover advanced topics in Larger Language Models (LLM) for security, security and privacy issues of LLM, and harmful misuses of LLMs.

Grading

- Weekly reading responses: 11%
- Paper presentation: 9%
- Midterm exam: 30%
- Midterm project report: 20%
- Final project (1-3 students): 30%

If you have not officially registered but you attend the class, please email me and let me know.

Office Hour

Instructor Yizheng: Tuesday 2pm to 3pm, IRB 5224 (First week: 1pm to 2pm on Thursday)
Email: yzchen@umd.edu
Homepage: https://surrealyz.github.io/
TA Yanjun: Thursday 2pm to 3pm, IRB 4214

Themes

- Large Language Model Families
- Code Generation Large Language Models
- LLM for Security Analysts and Developers
- LLM for Vulnerability Detection
- LLM for Binary Analysis
- LLM for Network Security
- Robustness Evaluation of LLM
- Poisoning LLMs
- Prompt Injection of LLM
- Privacy of LLM
- Watermarking of LLM
- Adversarial Usages of LLMs

Tentative Paper List

- Can Language Models Help in System Security? Investigating Log Anomaly Detection using BERT https://aclanthology.org/2022.alta-1.19/
- Do Users Write More Insecure Code with AI Assistants? (ACM CCS'23) https://arxiv.org/abs/2211.03622
- Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions (IEEE Security & Privacy 2021) https://arxiv.org/abs/2108.09293
- Pop Quiz! Can a Large Language Model Help With Reverse Engineering? https://arxiv.org/abs/2202.01142
- Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants https://www.usenix.org/system/files/sec23fall-prepub-353-sandoval.pdf
- SecurityEval dataset: mining vulnerability examples to evaluate machine learning-based code generation techniques https://dl.acm.org/doi/abs/10.1145/3549035.3561184
- DiverseVul: A New Vulnerable Source Code Dataset for Deep Learning Based Vulnerability Detection https://arxiv.org/abs/2304.00409
- Controlling Large Language Models to Generate Secure and Vulnerable Code https://arxiv.org/abs/2302.05319
- Palmtree: Learning an assembly language model for instruction embedding https://arxiv.org/abs/2103.03809
- Trex: Learning Execution Semantics from Micro-Traces for Binary Similarity https://arxiv.org/abs/2012.08680
- CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation https://arxiv.org/abs/2109.00859
- NatGen: Generative pre-training by "Naturalizing" source code https://arxiv.org/abs/2206.07585
- VulRepair: A T5-Based Automated Software Vulnerability Repair https://dl.acm.org/doi/abs/10.1145/3540250.3549098
- Fixing Hardware Security Bugs with Large Language Models https://arxiv.org/abs/2302.01215
- Extracting Training Data from Large Language Models https://arxiv.org/abs/2012.07805
- Just fine-tune twice: Selective differential privacy for large language models https://arxiv.org/abs/2204.07667
- More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models https://arxiv.org/abs/2302.12173

- MGTBench: Benchmarking Machine-Generated Text Detection https://arxiv.org/abs/2303.14822
- Can AI-Generated Text be Reliably Detected? https://arxiv.org/abs/2303.11156
- A Watermark for Large Language Models https://arxiv.org/abs/2301.10226
- On the Possibilities of AI-Generated Text Detection https://arxiv.org/abs/2304.04736
- Adversarial Prompting for Black Box Foundation Models https://arxiv.org/abs/2302.04237
- ReCode: Robustness Evaluation of Code Generation Models https://arxiv.org/abs/2212.10264
- TrojanPuzzle: Covertly Poisoning Code-Suggestion Models https://arxiv.org/abs/2301.02344
- Provably Confidential Language Modeling https://aclanthology.org/2022.naacl-main.69.pdf
- How Should Pre-Trained Language Models Be Fine-Tuned Towards Adversarial Robustness? https://proceedings.neurips.cc/paper/2021/hash/22b1f2e0983160db6f7bb9f62f4dbb39-Abstract.html
- Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution https://arxiv.org/abs/2202.10054#
- Evading watermark based detection of AI generated content https://arxiv.org/abs/2305.03807
- https://arxiv.org/abs/2305.06212 Privacy preserving prompt tuning
- Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants https://arxiv.org/abs/2208.09727
- More on the class website

Other resources:

LLaMA: Open and Efficient Foundation Language Models https://arxiv.org/abs/2302.13971

ChatLLaMA https://github.com/juncongmoo/chatllama

Cybercriminals starting to use ChatGPT https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/

https://www.cyberark.com/resources/threat-research-blog/chatting-our-way-into-creating-a-polymorphic-malware?utm_source=substack&utm_medium=email

https://hackernoon.com/how-i-solved-the-passman-ctf-challenge-with-gpt-4?utm_source=substack&utm_medium=email

https://gizmodo.com/gpt4-open-ai-chatbot-task-rabbit-chatgpt-1850227471?utm_source=substack&utm_medium=email

https://micahflee.com/2023/04/capturing-the-flag-with-gpt-4/?utm_source=substack&utm_medium=email

Products

https://blog.virustotal.com/2023/04/introducing-virustotal-code-insight.html

https://cloud.google.com/blog/products/identity-security/rsa-introducing-ai-powered-investigation-chronicle-security-operations

https://cloud.google.com/security/ai

https://cloud.google.com/blog/products/identity-security/rsa-google-cloud-security-ai-workbench-generative-ai

---

Reading Responses

- Due every Tuesday before the class
- Respond to 2 papers each week, write ~ 1 page: choose 1 paper from each of 2 topics in the week. I am looking for nontrivial response.
- We accept reasonable extension request *before the deadline*.
- If you missed the deadline not for medical absences, I would accept reading response for 3 papers, due Thursday before the class by email.

Late Policy

In addition to what is stated above in the “reading responses”, we accept pre-arranged reasonable deadline extension for project reports if requested *before the deadline*.

Presentation

- What is the problem the paper is trying to solve?
- What are the related works?
- What is the technique?
- Why is this paper doing it better?
- Does the new method makes sense?
- How are the results?
- Has the problem been solved? Is there nothing else left to do?
- How does it inspire your class project (or not)?

25 min presentation + discussions

Mid-term project report

- Define the problem
- Write a related work section
- Propose the method
- What is at least one experiment you have run (the more experiment the better), and what have you learned from it.
  - Alternatively, what is one theoretical result you have obtained, and what have you learned from it.
- Plan the tasks for the remaining semester

Example Project Topics

- Attacker use ChatGPT to write malware, spearphishing emails, underground economy
- Fine tune LLMs for security tasks
  - Log analysis
  - Vulnerable source code
  - Variable renaming
  - Binary analysis
- Study effect of source code suggestion on users
  - Note on IRB approval
- Study how LLM helps threat analysts
- Evade watermarking detection
- Poisoning attacks on LLM
- …