Jiazhao Li

Ph.D. Candidate at the School of Information, University of Michigan
3442 North Quad, 105 S. State St., Ann Arbor, MI, USA, 48109-1285
(E) jiazhaol@umich.edu (C) (734) 604-1596 (W) https://jiazhaoli.github.io/

RESEARCH INTERESTS

Generative AI
Cybersecurity of LLMs (attack & defense)
Health Informatics

EDUCATION

University of Michigan, Ann Arbor, US Aug 2020 -- Aug 2024
Ph.D. candidate in School of Information
Advisor: V.G. Vinod Vydiswaran
Committee: Chaowei Xiao, Paramveer Dhillon, Dallas Card, Liyue Shen

M.S. in Electrical Computer Engineering (Computer Vision Track) Aug 2017 -- Apr 2019

Nankai University, Tianjin, China Sep 2013 -- Jun 2017
B.E. in the Electrical Engineering

WORKING EXPERIENCE

Yahoo Research, Content team Apr 2023 -- Aug 2023
ML Research Intern

PUBLICATIONS

ChatGPT as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger
Jiazhao Li, Yijin Yang, Zhuofeng Wu, V. G. Vydiswaran, Chaowei Xiao
In Proceedings of NAACL 2024. (pdf)

Defending against Insertion-based Textual Backdoor Attacks via Attribution
Jiazhao Li, Zhuofeng Wu, Wei Ping, Chaowei Xiao, VG Vydiswaran
In Proceedings of Findings ACL 2023. (pdf)

PharmMT: A neural machine translation approach to simplify prescription directions
Jiazhao Li, Corey Lester, Xinyan Zhao, Yuting Ding, Yun Jiang, VG Vydiswaran
In Proceedings of Findings of EMNLP 2020. (pdf)

Re-ranking biomedical literature for precision medicine with pre-trained neural models
Jiazhao Li, Adharsh Murali, Qiaozhu Mei, V.G. Vinod Vydiswaran
In Proceedings of ICHI 2020. (pdf)

PREPRINT

Mitigating Fine-tuning Jailbreak Attack with Backdoor Enhanced Alignment
Jiongxiao Wang, Jiazhao Li, Yiquan Li, Xiangyu Qi, Muhao Chen, Junjie Hu, Yixuan Li, Bo Li, Chaowei Xiao
arXiv preprint arXiv: 2402.14968 (In submission to ACL 2024) (pdf)

JOURNALS

Accelerating Theme Analysis on clinical tele-visit narrows via Active Learning
Jiazhao Li, VG Vydiswaran (Under Review of JAMIA)

Performance evaluation of a prescription medication image classification model: an observational cohort.
Corey A. Lester, Jiazhao Li, Yuting Ding, Brigid Rowell, Jessie 'Xi' Yang, Raed Al Kontar
NPJ Digit. Med (2021) (IF=15.2) (pdf)

PROJECT EXPERIENCE

School of Information, University of Michigan Aug 2020 – Present

Mitigating Fine-tuning Jailbreak Attack with Backdoor Enhanced Alignment

- We proposed a Backdoor Enhanced Safety Alignment method by integrating a secret prompt, acting as a `backdoor trigger', that is prefixed to safety examples against Fine-tuning-based Jailbreak Attack.
- Comprehensive experiments demonstrate that through by adding as few as 11 prefixed safety examples, the maliciously fine-tuned LLMs will achieve similar safety performance as the original aligned models.
- Our method is also proved to be effective in a more practical setting where the fine-tuning data consists of both FJAttack examples and the fine-tuning task data.
- The drop of instruction-following ability is shown to be mitigated on MT-Bench benchmark.

ChatGPT as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger

- Propose a stealthy, input-dependent backdoor attack method to mislead textual classifiers (BERT, LLaMA2-7B) utilizing paraphrasing models (ChatGPT, mBART, BART) as LM-based triggers.
- The generated backdoor embedded examples are less noticeable for human cognition with diverse evaluation metrics including Perplexity, CoLA score, grammar error, semantic invariance and human evaluations.
- ChatGPTAttack is easily accessible and avoids detected by both GPT-detection and defense methods from literature.

Defending against Insertion-based Textual Backdoor Attacks via Attribution

- Build a defense framework against backdoor attacks on text classifier (pre-training and post-training).
- Apply a poisoned sample detector ELECTRA to identify poisoned samples.
- Identify triggers by calculating the attribution score of tokens using Partial LRP (trigger word contributes most to mislabeling)
- Achieve SOTA performance, an average accuracy of 79.97% (56.59 up) and 48.34% (3.99 up) on 4 benchmarks against pre-training attack and post-training attack respectively.
- This work was presented as a poster in **ACL'23**.

PharmMT: A Neural Machine Translation Approach to Simplify Prescription Directions.

- Built Seq-to-Seq Text Simplification Model between parallel prescription and pharmacy directions corpus using OpenNMT framework.
- Archive 60.27 BLEU score against pharmacists' reference and 94.3% of the simplified directions could be used as-is or with minimal changes evaluated by pharmacists.
- This work was presented in **EMNLP'20.**

Yahoo Research, Content Team Apr 2023 – Aug 2023

Robustness of LLM against backdoor attack during Instruction Fine-tunning under label space shift.

- Applied backdoor attack during instruction fine-tuning and task-specific fine-tuning against LLaMA-2 7B via Parameter-Efficient Fine-Tuning, LoRA.
- Backdoor Attack can be transferred between datasets under same label space of LLMs
- Backdoor Attack can be transferred between datasets under different textual but the same semantic labels of LLMs.

SKILLS

Programming Languages: Python, C, C++, Verilog, Pascal
Frameworks & Tools: PyTorch, Fairseq, LaTeX, Vim, Git

SERVICE

Journal Reviewer

Frontiers in Big Data, section Cybersecurity and Privacy.

Conference Reviewer

EMNLP 2020, 21, 22, 23, ACL 2023, EACL 2023

ACL Rolling Reviewer

February 2024 Cycle, October 2023 Cycle, June 2023 Cycle

TEACHING

Graduate Student Instructor

SI 630 Natural Language Processing (Winter 2024)
LHS 712 Natural Language Processing for Health (Winter 2023)