Research on Bias and Security & Privacy Limitations of LLMs at IUB &WPI

Present by: Rui Zhu

INDIANA UNIVERSITY BLOOMINGTON
and
WORCESTER POLYTECHNIC INSTITUTE

Our Team

Xiaofeng Wang (IUB)
xw7@indiana.edu
https://homes.luddy.indiana.edu/xw7

Haixu Tang (IUB)
hatang@indiana.edu
https://homes.luddy.indiana.edu/hatang

Xiaozhong Liu (WPI)
xliu14@wpi.edu
https://www.wpi.edu/people/faculty/xliu14

IARPA TROJAI Project
(2020-2024)
- **Multiple times top-2 in the leaderboard**

Trojan Detection Challenge
(2022)
- **Champion on both tracks**

Publication(LLM related)
(2022-2023)
- **12 publications**

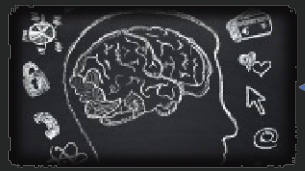
LLM Bias & Personalization

HTML understanding

GEM: Gestalt Enhanced Markup Language Model for Web Understanding via Render Tree (EMNLP 23)

LLM-Enhanced Software Patch Localization (USENIX 24)

Vulnerability detection



Personalized Memorization

Memory-Augmented LLM Personalization with Short- and Long-Term Memory Coordination (AAAI 24)

LLM

LLM-domain model collaborations

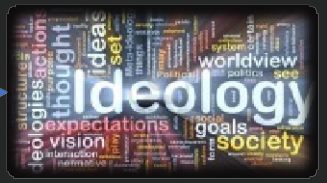
Precedent- Enhanced Legal Judgment Prediction with LLM and Domain-Model Collaboration (EMNLP 23)

LLM Attack Method

Order-Disorder: Imitation Adversarial Attacks for Black-box (CCS 22, Best Paper Nomination)

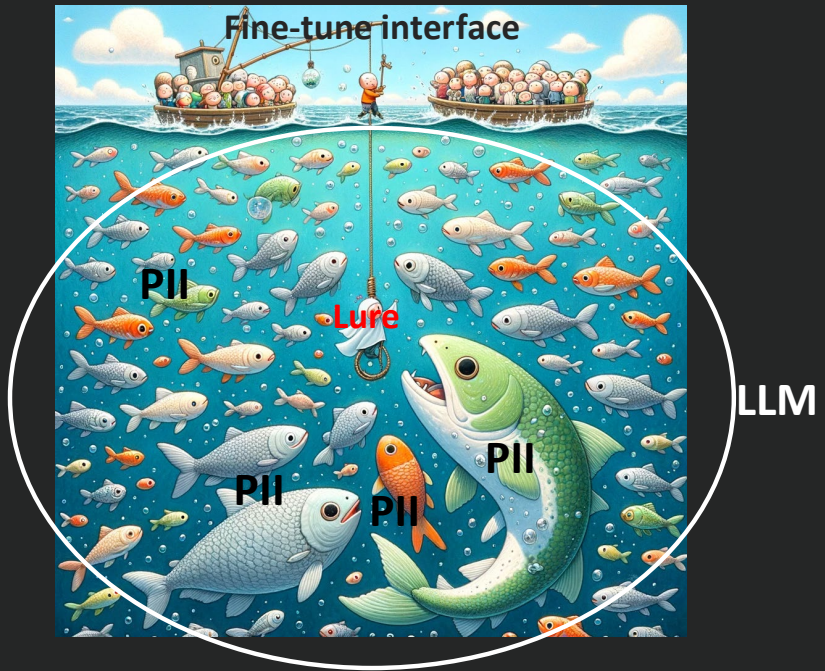
MAWSEO: Adversarial Wiki Search Poisoning for Illicit Online Promotion (S&P 23)

LLM Ideologization

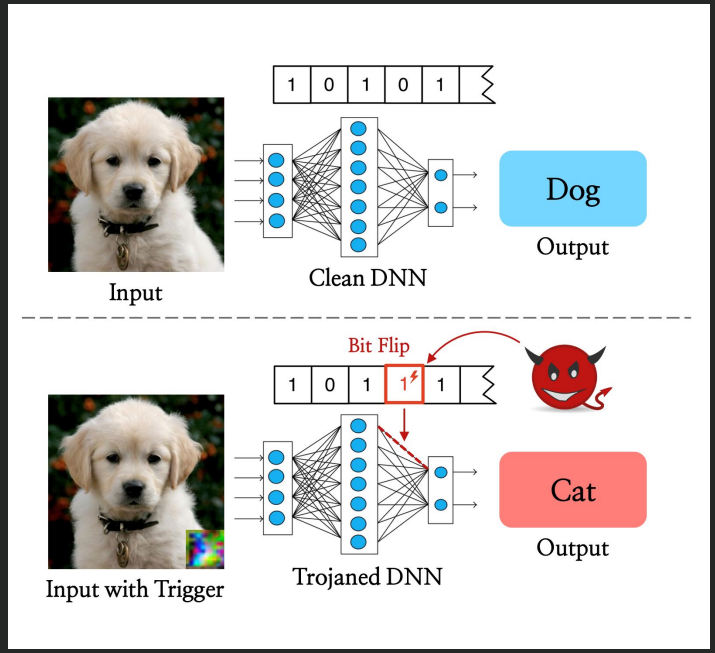


Large Language Model Soft Ideologization via AI-Self-Consciousness

LLM Security & Privacy



The Janus Interface: How Fine-Tuning in Large Language Models Models Amplifies the Privacy Risks



Tossing in the Dark: Practical Bit-Flipping on Gray-box Deep Neural Networks for Runtime Trojan Injection (USENIX 2024)

Thanks!

We are looking for partner.

If you have any questions, please contact:

Xiaofeng Wang: xw7@indiana.edu
Haixu Tang: hatang@indiana.edu
Xiaozhong Liu: xliu14@wpi.edu