Yang Ouyang

Garner, NC 27529, United States

☎ (984) 325-2686 ✉ youyang7@ncsu.edu LinkedIn GitHub Website

Education Experience

North Carolina State University **Aug. 2024 – Present**

Doctor of Philosophy in Electrical and Computer Engineering *Raleigh, U.S.A*

- **Advisor: Jung-Eun Kim**
- **Research Interests:** Trustworthy&Efficient ML; Interpretable LLMs (Chain-of-Thought);

Duke University **Aug. 2022 – May 2024**

Master of Engineering in Electrical and Computer Engineering *Durham, U.S.A*

- GPA: 3.83 / 4.0
- **Teaching Assistant** of ECE 551K: Programming, Data Structures, and Algorithms in C++

Shenzhen University **Sep. 2018 – July 2022**

Bachelor of Engineering in Computer Science and Technology *Shenzhen, China*

- GPA: 3.75 / 4.0
- Honors/Awards: Two times winner of The Second Award of Studying Star in 2020 & 2021 (Ranked in 4 & 6); Outstanding Graduate of the Year 2022

Selected Publication

- **[In Proceedings of NAACL 2025] Yang Ouyang**, Hengrui Gu, Shuhang Lin, Wenyue Hua, Jie Peng, Bhavya Kailkhura, Meijun Gao, Tianlong Chen, Kaixiong Zhou. "Layer-AdvPatcher: Layer-Level Self-Exposure and Patch for Jailbreak Defense"
- [**In Proceedings of ICLR 2025**] Jingyang Zhang, Jingwei Sun, Eric Yeats, **Yang Ouyang**, Martin Kuo, Jianyi Zhang, Hao Frank Yang, Hai Li. "Min-K%++: Improved Baseline for Detecting Pre-Training Data of LLMs"

Project Experience

Fact-Enhanced CoT: Steering for Factual Reasoning **Aug. 2025 – Present**

Targeting ICML 2026 *Raleigh, U.S.A*

- Constructed style-controlled **positive/negative first-step CoT** pairs on GSM8K to learn **factual subspaces**; performed **Activation Steering** for online control during the first step of reasoning.
- Observed up to **20% improvement** on factuality metrics with minimal accuracy degradation

Layer-AdvPatcher: Layer-Level Self-Exposure and Patch for Jailbreak Defense **Aug. 2024 – Oct. 2024**

In Proceedings of NAACL 2025 *Raleigh, U.S.A*

- Developed the Layer-AdvPatcher framework to defend against jailbreak attacks in LLMs, including a three-step pipeline for defense: i) toxic layer identification, ii) adversarial augmentation, and iii) localized toxic layer editing.
- Achieved a **25% reduction** in Attack Success Rate using our method across models including Mistral-7B and Llama2-7B compared to modification-based defense methods.

Internship Experience

Trip.com Group Ltd | *Java, Spring Framework* **May 2023 – Aug. 2023**

Back End Developer Intern, Flight Ticket Department *Shanghai, China*

- Contributed to the optimization of MegaSearch which serves as an aggregation and cache layer for Trip's international ticket responses using **Java**.
- Optimized the response size to fit AWS's smaller bandwidth while saving some storage costs. Reduced the **Protobuf response size by 50%** in total using a variety of methods.
- Compared a variety of serialization and deserialization means using **JMH**: including the latest open source Fury, Kryo, and ultimately found that Protobuf is the most efficient serialization, but Kryo in the serialization of the size of a small advantage.

Amazon Web Services | *Java, K8s* **July 2022 – Oct. 2022**

Back End Developer Intern, DeepJavaLibrary Department *Mountain View, U.S.A (remote)*

- Integrated DeepJavaLibrary Model Server with KServe by developing 3 robust HTTP APIs in **Java** for KServe's inference engine, supporting DJL-Serving health checks, model information retrieval, and inference result processing with request data, each passing unit tests and providing clear response codes.
- Deployed containerized DJL-Serving on KServe using **yaml** files to configure ports and parameters, facilitating model deployment and testing within the KServe framework.

Technical Skills

- Programming Languages: Python, Java, C++, Javascript
- Deep Learning Frameworks: PyTorch, Huggingface