Yang Ouyang

Raleigh, NC 27606, United States
(984) 325-2686 youyang7@ncsu.edu LinkedIn GitHub

Education Experience

North Carolina State University **Aug. 2024 – Present**
Doctor of Philosophy in Electrical and Computer Engineering *Raleigh, U.S.A*

- **Advisor: Kaixiong Zhou**
- **Research Interests:** AI for biology (Interpretable RNA), Trustworthy Large Language Model

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 Submission] **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 submission to NAACL 2025, meta-review 4.0 out of 5.0
- [In Submission] Shuhang Lin, Wenyue Hua, Lingyao Li, **Yang Ouyang**, Jie Peng, Bhavya Kailkhura, Kaixiong Zhou, Tianlong Chen. "Skeletonic Speculative Decoding", in submission to ACL Rolling Review - October 2024
- [In Submission] 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", in submission to ICLR 2025

Project Experience

Layer-AdvPatcher: Layer-Level Self-Exposure and Patch for Jailbreak Defense **Aug. 2024 – Oct. 2024**
Submittd to ACL Rolling Review - October 2024 *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.

Skeletonic Speculative Decoding **May. 2024 – Oct. 2024**
Submittd to ACL Rolling Review - October 2024 *Raleigh, U.S.A*

- Introduced Skeletonic Speculative Decoding (SSD), a divide-and-conquer approach to enhance speculative decoding efficiency by decomposing complex queries into manageable sub-questions, improving the acceptance rate and enabling parallel processing.
- Achieved up to a 2.8x speedup over standard speculative decoding methods across various model configurations by leveraging parallelization and increased acceptance rates for simplified sub-questions.

Min-K%++: Improved Baseline for Detecting Pre-Training Data of LLMs **Jan. 2024 – Apr. 2024**
Submitted to ICLR 2025 *Durham, U.S.A*

- Proposed a theoretically motivated methodology, Min-K%++, for pre-training data detection in LLMs, leveraging local maxima of the modeled distribution to identify training data effectively.
- Achieved new state-of-the-art performance, surpassing existing methods by 6.2% to 10.5% in AUROC on the WikiMIA benchmark and performing competitively on the challenging MIMIR benchmark.

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 *San Jose, 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.

Tencent Music Entertainment Group | ***Javascript, Vue*** **May 2021 – Sep. 2021**

Front End Developer Intern, Security Center *Shenzhen, China*

- Applied **Vue2.0** framework based on JavaScript to develop the inner front-end of content audit security platform.
- Built and maintained middle ground management system.
- Developed search, collection, and recently used functions for the middle ground management system.
- Utilised Least Recently Used (LRU) to design a cache that was able to clear the cache efficiently.
- Configured Webpack to optimize the local development and deployment increased the packaging speed by 75% and decreased the packaging size by 10%.

Technical Skills

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