HONGXIANG ZHANG

Lafayette, IN, USA ◇ (+1) 530-564-2580 ◇ hxxzhang@gmail.com ◇ LinkedIn ◇ Homepage ◇ Google Scholar

EDUCATION BACKGROUND

Purdue University Augest 2025 - Present
Visiting Scholar

University of California, Davis (GPA: 3.9/4.0) Sep 2022 - June 2025
M.Eng of Science, Computer Science

Australian National University Jul 2020 - Sep 2022
Bachelor of Engineering (Honours), Computer Science

Shandong University Sep 2018 - Jun 2022
Bachelor of Engineering, Computer Science

ACADEMIC EXPERIENCE

Multi-agent System alignment Ongoing
Hongxiang Zhang, *Yuan Tian, Tianyi Zhang*

· Designing an alignment framework for **multi-agent systems** to improve inter-agent consistency and task coordination.
· Focusing on **inter-agent communication**, we make attention steering, allowing the agent to maintain focus throughout the generation.

Self-Anchor: Large Language Model Reasoning via Step-by-step Attention Alignment
Hongxiang Zhang, *Yuan Tian, Tianyi Zhang* *[arXiv]*

· Proposed Self-Anchor, a reasoning-time alignment method that leverages the inherent structure of reasoning to steer LLM attention, allowing the model to maintain focus throughout generation.
· Self-Anchor significantly reduces the performance gap between **"non-reasoning"** models and specialized reasoning models, with the potential to enable most LLMs to tackle complex reasoning tasks without retraining.

Active Layer-Contrastive Decoding Reduces Hallucination in Large Language Model Generation **EMNLP 2025 Main**
Hongxiang Zhang, *Hao Chen, Muhao Chen, Tianyi Zhang* *[arXiv] [Project][Code]*

· Proposed Active Layer-Contrastive Decoding (ActLCD), a decoding algorithm that **actively contrasts model layers** to suppress hallucination.
· By casting decoding as a sequential decision-making problem, ActLCD employs a **reinforcement learning** policy guided by a reward-aware classifier to optimize factuality **beyond the token level**.
· Achieved SOTA hallucination reduction across five benchmarks.

SteerDiff: Steering towards Safe Text-To-Image Diffusion Model
Hongxiang Zhang, *Yifeng He, Hao Chen* *Under review [arXiv]*

· Developed **SteerDiff**, a lightweight plug-in module that filters unsafe or inappropriate concepts in diffusion model prompts via latent-space steering.
· Proposed a semantic-preserving projection mechanism between text encoder and UNet, enabling safety control without retraining.
· Demonstrated scalability to concept removal and fairness alignment tasks with minimal fine-tuning.

LLAMAFUZZ: Large Language Model Enhanced Greybox Fuzzing
Hongxiang Zhang, *Yuyang Rong, Yifeng He, Hao Chen* *Under review [arXiv]*

- Introduced **LLAMAFUZZ**, a hybrid fuzzing framework that leverages LLM-driven mutation to learn structured input formats for both binary and text-based targets.
- Integrated an adaptive seed-selection policy that balances exploration and structure preservation, achieving significant coverage improvements on real-world benchmarks.
- Provided interpretability analyses showing how language-based mutation heuristics enhance fuzzing efficiency and semantic validity.

INDUSTRY EXPERIENCE

Software and Operation Engineer Intern Volkswagen, Beijing, China Feb 2022 - Jul 2022

- Achieved data masking and auto-populated to **JIRA** log in Python.
- Intuitively analyzed and provided services and finance data to CIO with **Tableau**, helping the management level adjust the company service strategy agilely.

Quality Assurance Intern Didi Global, Beijing, China Dec 2020 - Feb 2021

- Developed distributed Java invoice services to enable the interaction between server and end-user devices that served more than 1 million users per day.
- Achieved unit testing automation by using **JUnit** and the test case in **Redis**, reduced over 15% of the testing engineer's workload.

TECHNICAL SKILLS

Programming Languages	Python, Java, C/C++, JavaScript HTML, CSS
Software and skills	MetaGPT, Autogen, MySQL, Numpy, Pytorch, Numpy, Pandas, Azure

TEACHING

ECS 153 Computer Security - UC Davis	2024 Spring
ECS 036A Programming & Problem Solving - UC Davis	2024 Winter
ECS 036C Data Structures, Algorithms, & Programming - UC Davis	2023 Fall
ECS 140A Programming Language - UC Davis	2022 Winter
ENGN 4528 Computer Vision - Australian National University	2021

ACADEMIC SERVICES

Conference reviewer
2025 IEEE 43rd International Conference on Consumer Electronics (ICCE 2025)
2025 ACL Rolling Review (ARR)