

BANG AN

bangann@umd.edu

<https://bangann.github.io/>

EDUCATION

University of Maryland, College Park Ph.D. in Computer Science. Advisor: Furong Huang	2020 - 2025 (<i>Expected</i>)
Tsinghua University, China M.S. in Automation, School of Information Science and Technology	2013 - 2016
Northeastern University, China B.S. in Automation, School of Information Science and Engineering	2009 - 2013

RESEARCH INTEREST

My research interests focus on developing **Responsible AI** systems with an emphasis on three key areas:

Safety of Large Language Models (LLMs): including automatic red-teaming methods for identifying safety vulnerabilities [2] and false refusals [1], detecting AI-generated content [4], watermarking and copyright issues [5], improving the test time alignment, and recently the safety of AI agents.

Robustness in Generative AI: including the robustness of invisible image watermarks [3], improving the reasoning ability of Vision Language Models (VLMs) [7], and robustness of machine unlearning.

Distribution Shift: including spurious correlations in LLMs [6], maintaining the fairness under distribution shifts [9], understanding multi-head attention mechanism [11], and improving OOD generalization [8, 10].

SELECTED PUBLICATIONS

Please visit my [Google Scholar](#) for the complete list. * denotes equal contribution

- Automatic Pseudo-Harmful Prompt Generation for Evaluating False Refusals in Large Language Models
B. An*, S. Zhu*, R. Zhang, MA. Panaitescu-Liess, Y. Xu, F. Huang. [COLM](#), 2024
[Website](#)
- AutoDAN: Automatic and Interpretable Adversarial Attacks on Large Language Models
S. Zhu, R. Zhang, **B. An**, G. Wu, J. Barrow, Z. Wang, F. Huang, A. Nenkova, T. Sun. [COLM](#), 2024
[Media Coverage](#)
- WAVES: Benchmarking the Robustness of Image Watermarks
B. An*, M. Ding*, T. Rabbani*, A. Agrawal, C. Deng, Y. Xu, S. Zhu, A. Mohamed, Y. Wen, T. Goldstein, F. Huang. [ICML](#), 2024
[Website](#)
[NeurIPS'24 Competition](#)
- Position: On the Possibilities of AI-Generated Text Detection
S. Chakraborty*, AS. Bedi*, S. Zhu, **B. An**, D. Manocha, F. Huang. [ICML](#), 2024
[Media Coverage](#)
- Can Watermarking Large Language Models Prevent Copyrighted Text Generation and Hide Training Data?
MA. Panaitescu-Liess, Z. Che, **B. An**, Y. Xu, P. Pathmanathan, S. Chakraborty, S. Zhu, T. Goldstein, F. Huang. [Arxiv](#), 2024
- Explore Spurious Correlations at the Concept Level in Language Models for Text Classification
Y. Zhou, P. Xu, X. Liu, **B. An**, W. Ai, F. Huang. [ACL](#), 2024
- PerceptionCLIP: Zero-shot Visual Classification by Inferring and Conditioning on Contexts
B. An*, S. Zhu*, MA. Panaitescu-Liess, CK. Mummadi, F. Huang. [ICLR](#), 2024
- Learning Unforeseen Robustness from Out-of-distribution Data Using Equivariant Domain Translator
S. Zhu, **B. An**, F. Huang, S. Hong. [ICML](#), 2023

9. Transferring Fairness under Distribution Shifts via Fair Consistency Regularization
B. An, Z. Che, M. Ding, F. Huang. NeurIPS, 2022
10. Understanding the Generalization Benefit of Model Invariance from a Data Perspective
S. Zhu*, B. An*, F. Huang. NeurIPS, 2021
11. Repulsive Attention: Rethinking Multi-head Attention as Bayesian Inference
B. An, J. Lyu, Z. Wang, C. Li, C. Hu, F. Tan, R. Zhang, Y. Hu, C. Chen. EMNLP, 2020

EMPLOYMENT

- Bloomberg** May 2024 - Present
Research Intern, CTO Support Team & AI Safety Team, Mentor: [Mark Dredze](#) New York, NY
- **Safety of RAG LLMs.** Investigated how and why RAG impacts model safety. Explored **red-teaming** methods for RAG. Adapted and **accelerated** gradient-based optimization methods to **long-context** input.
 - Consulting on **red-teaming** strategies for an internal LLM assistant with **tool-use** capabilities.
- Capital One** Jun 2023 - Aug 2023
Research Intern, Applied AI Research Team, Mentor: [Sam Sharpe](#) McLean, VA
- **Interpret the Representation Space of Language Embedding Models.** Applied a contrastive interpretation method to an internal foundation model to assist regulation.
- Google** Jun 2022 - Aug 2022
Research Intern, Google Brain (now Google DeepMind), Mentor: [Zhe Zhao](#) Mountain View, CA
- **Distill Pre-trained Knowledge to Downstream Models.** Proposed an interactive communication method.
- Microsoft Research Asia** Sep 2020 - Dec 2020
Research Intern, System Intelligence Team, Mentor: [Xueting Han](#) Beijing, China
- **Transfer Learning on Graph Neural Networks.** Proposed an adaptation method by introducing auxiliary tasks to mitigate the gap between **self-supervised** training and downstream tasks.
- State University of New York at Buffalo** Jul 2019 - May 2020
Visiting Researcher, Machine Learning Lab, Mentor: [Changyou Chen](#) Buffalo, NY
- Rethinking **Multi-head Attention** as **Bayesian** Inference. Investigated the attention collapse problem from a Bayesian view and proposed a technique to diversify multi-head attention.
- IBM Research - China** Aug 2018 - Jun 2019
Research Scientist, NLP Team, Manager: [Zhong Su](#) Beijing, China
- Applied research on semantic analyses. Built a semantic compliance advisor for unstructured documents.

HONORS & AWARDS

- | | |
|---|------------------|
| COLM 2024 DEI Travel Scholarship | 2024 |
| Outstanding Graduate Assistant Award of the University of Maryland (top 2%) | 2023 |
| NeurIPS 2022 Travel Award | 2022 |
| First Prize of National Mathematics Competition in Liaoning Province, China | 2013 |
| National Undergraduate Scholarship (top 1%), China | 2010, 2011, 2012 |

SERVICES

- Reviewer:** ICML 2022, NeurIPS 2022, ICML2023, NeurIPS2023, ICLR2024, NAACL2024.
- Organizer:** NeurIPS 2024 Competition, Erasing the Invisible: A Stress-Test Challenge for Image Watermarks.