BANG AN

bangan@umd.edu https://bangann.github.io/

EDUCATION

University of Maryland, College Park *2020 - 2025 (Expected)*
Ph.D. in Computer Science. Advisor: Furong Huang

Tsinghua University, China *2013 - 2016*
M.S. in Automation, School of Information Science and Technology

Northeastern University, China *2009 - 2013*
B.S. in Automation, School of Information Science and Engineering

RESEARCH INTEREST

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

Safety Alignment of LLMs: including automatic red-teaming for safety [1,4] and false refusals [3], detecting AI-generated content [6], watermarking and copyright issues [7], controllable decoding for test time alignment[2], post-training for alignment, safety of AI Agents, and understanding alignment through interpretability.

Robustness in Generative AI: including the robustness of invisible image watermarks [5], enhancing the reasoning ability of Vision Language Models (VLMs) [9], and robustness of AI agents against accidental misuse.

Distribution Shift: including spurious correlations in LLMs [8], maintaining the fairness under distribution shifts [11], diversity multi-head attention [13], and improving OOD generalization [10, 12].

SELECTED PUBLICATIONS

Please visit my Google Scholar for the complete list. * denotes equal contribution

1. RAG LLMs are Not Safer: A Safety Analysis of Retrieval-Augmented Generation for Large Language Models
B. An, S. Zhang, M. Dredze — Under Review, 2024

2. GenARM: Reward Guided Generation with Autoregressive Reward Model for Test-time Alignment
Y. Xu, UM. Sehwag, A. Koppel, S. Zhu, **B. An**, F. Huang, S. Ganesh — Arxiv, 2024

3. Automatic Pseudo-Harmful Prompt Generation for Evaluating False Refusals in Large Language Models — COLM, 2024 *Website*
B. An*, S. Zhu*, R. Zhang, MA. Panaitescu-Liess, Y. Xu, F. Huang.

4. AutoDAN: Automatic and Interpretable Adversarial Attacks on Large Language Models — COLM, 2024
S. Zhu, R. Zhang, **B. An**, G. Wu, J. Barrow, Z. Wang, F. Huang, A. Nenkova, T. Sun. *Media Coverage*

5. WAVES: Benchmarking the Robustness of Image Watermarks — ICML, 2024
B. An*, M. Ding*, T. Rabbani*, A. Agrawal, C. Deng, Y. Xu, S. Zhu, A. Mohamed, *Website*
Y. Wen, T. Goldstein, F. Huang. *NeurIPS'24 Competition*

6. Position: On the Possibilities of AI-Generated Text Detection — ICML, 2024
S. Chakraborty*, AS. Bedi*, S. Zhu, **B. An**, D. Manocha, F. Huang. *Media Coverage*

7. 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

8. 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

9. PerceptionCLIP: Zero-shot Visual Classification by Inferring and Conditioning on Contexts
B. An*, S. Zhu*, MA. Panaitescu-Liess, CK. Mummadi, F. Huang. ICLR, 2024

10. Learning Unforeseen Robustness from Out-of-distribution Data Using Equivariant Domain Translator
S. Zhu, **B. An**, F. Huang, S. Hong. ICML, 2023

11. Transferring Fairness under Distribution Shifts via Fair Consistency Regularization
B. An, Z. Che, M. Ding, F. Huang. NeurIPS, 2022

12. Understanding the Generalization Benefit of Model Invariance from a Data Perspective
S. Zhu*, **B. An***, F. Huang. NeurIPS, 2021

13. 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.

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
Student Researcher, 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**.

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 (full-time), NLP Team, Manager: Zhong Su Beijing, China

- Applied research on semantic analyses. Built a semantic compliance advisor for unstructured documents.

China Transport Information Center Co., Ltd Jul 2016 - Jul 2018
Machine Learning Engineer (full-time) Beijing, China

SKILLS

Python, Pytorch, JAX, Linux, AWS, LLMs, VLMs, RAG, Watermarks

HONORS & AWARDS

Outstanding Graduate Assistant Award of the University of Maryland (top 2%) 2023
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, NAACL2025.
Organizer: NeurIPS'24 Competition, Erasing the Invisible: A Stress-Test Challenge for Image Watermarks.