Huaizhi Ge

New York, NY
hg2590@columbia.edu — +1 (917) 826-2280 — Personal Website

RESEARCH INTERESTS

My research interests center on advancing the trustworthiness, robustness, and capabilities of AI systems while deepening our understanding of their core mechanisms through interpretability techniques.

EDUCATION

Columbia University, New York, NY — Sep 2021 — Dec 2022
Master of Science in Data Science — Cumulative GPA: 4.04/4.00

Shanghai Jiao Tong University, Shanghai, China — Sep 2017 — Jun 2021
Bachelor of Science in Aeronautics and Astronautics Engineering
Graduated with Distinction (top 5%): College Graduate Excellence Award of Shanghai, 2021

RESEARCH EXPERIENCE

Understanding LLM Backdoor Attacks Through Model-Generated Explanations — Rutgers University
Research Assistant, Advised by Prof. Ruixiang Tang — Jun 2024 — Present

- Injected backdoor functions into LLaMA 3-8B and LLaMA 2-13B models via QLoRA finetuning, achieving over 95% attack success rate while maintaining the models' core performance.
- Generated explanations for poisoned inputs of backdoored models and compared them to those of clean inputs, using techniques such as Semantic Textual Similarity and Jaccard Similarity to identify differences in explanation quality.
- Conducted in-depth analysis of backdoored model mechanics through logit lens and prediction trajetory visualization, revealing distinct patterns in the attention dynamics of generated explanation.
- Offered insights into how backdoors influence model behavior: when dealing with poisoned inputs, the backdoored model tends to look at the newly generated token rather than the context.

Circuits Analysis in Language Models through Knowledge Editing — Stevens Institute of Technology
Research Assistant, Advised by Prof. Zining Zhu — Apr 2024 — Sep 2024

- Developed a novel circuit-aware method for editing circuits in GPT-2 Base, revealing that circuits containing related knowledge are more resistant to knowledge editing.
- Extracted circuits of 5 different sizes, finding that optimal circuits containing related knowledge often encompass 5%–50% of the model's parameters.
- Analyzed the proportion of each module in the extracted circuits, uncovering that layer normalization modules account for up to 60% of circuit parameters, offering new insights into language model interpretability.
- Extracted circuits from 8 different datasets and evaluated overlaps among these circuits.

Evaluating Knowledge Editing Methods Efficacy on Editing Perplexing Knowledge — Vector Institute
Research Assistant, Advised by Prof. Zining Zhu — Sep 2023 — Sep 2024

- Evaluated effectiveness of knowledge editing techniques for modifying perplexing knowledge in large language models (GPT-2 L, GPT-2 XL, GPT-J), utilizing methods such as Rank-One Model Editing (ROME).
- Coined the term 'Perplexingness' to capture the finding that pre-edit probabilities and hierarchical relationships significantly influence the efficacy of knowledge editing.
- Developed a novel dataset, HierarchyData, to study the impact of knowledge hierarchy on editing outcomes, discovering that more abstract concepts (hypernyms) are more challenging to modify.

Statistical Analysis on Medical Data with Visualization — Columbia University Irving Medical Center
Data Scientist, Advised by Prof. Jose Gutierrez — July 2024 — Present

- Processed large medical datasets to ensure data reliability and conducted comprehensive exploratory data analysis.
- Performed data exploration, survival analysis, and visualization for a research paper examining the relationship between various types of arterial stenosis and treatment outcomes in stroke patients.
- Led a project and authored a paper that applied data imputation methods and diverse clustering algorithms, discovering insights from the cohorts.
- Collaborated with medical professionals to interpret analytical results, directly contributing to clinical research efforts.

PUBLICATIONS

When Backdoors Speak: Understanding LLM Backdoor Attacks Through Model-Generated Explanations

- **Huaizhi Ge**, Yiming Li, Qifan Wang, Yongfeng Zhang, and Ruixiang Tang. Under Review at *NAACL 2025.*

What Do the Circuits Mean? A Knowledge Edit View

- **Huaizhi Ge**, Frank Rudzicz, and Zining Zhu. Under Review at *ARR*.

How Well Can Knowledge Edit Methods Edit Perplexing Knowledge?

- **Huaizhi Ge**, Frank Rudzicz, Zining Zhu. 2024. arXiv preprint.

Stall Angle of Attack Prediction of Ridge Ice on Airfoil Using Deep Neural Networks

- Dinghao Yu, **Huaizhi Ge**, Zhirong Han, Bin Zhang, Fuxing Wang. Journal of Aeronautics, Astronautics and Aviation Volume 55 Issue 1 Vol.55 No.1 (2023/03) Pp. 29-38.

Impact of Concomitant Extracranial Atherosclerosis on Treatment Outcomes in Intracranial Arterial Stenosis: A Post hoc analysis of the SAMMPRIS Trial

- Edgar R. Lopez-Navarro, **Huaizhi Ge**, Jose Gutierrez. Work in Progress.

Cluster Analysis of Stroke Risk Factors in the Northern Manhattan Study

- **Huaizhi Ge**, Jose Gutierrez. Work in Progress.

PROJECTS

Detection of Dialogue Act New York, NY
Accenture Sep 2022 — Dec 2022

- Led a team of 4 in leveraging more than 1000 conversation transcripts for dialogue act detection, involving preprocessing, tokenization, and extraction of 21 learnable features.
- Leveraged distillBERT and roBERTa-base, for generating sentiment and emotion features.
- Experimented with various machine learning models such as Random Forest, XGBoost, LightGBM, and CatBoost, while optimizing performance through cross-validation and hyperparameter tuning.
- Delivered model and results with comprehensive documentation of architecture decisions and feature importance analysis to stakeholders, enabling seamless deployment and future improvements.

Multi-Task Learning Model Performance Evaluation New York, NY
Columbia University Jan 2022 — May 2022

- Developed comprehensive evaluation framework for multi-task learning (MTL) architectures to identify the most effective model for specific real-world datasets.
- Trained multiple MTL structures on the consolidated dataset in TensorFlow, including MMOE and CTRCVR (ESMM) models, and evaluated their performance in predicting multiple related targets simultaneously.
- Demonstrated superiority of MTL models in AUC scores on different datasets through rigorous evaluation and analysis, while identifying key challenges: task interference in competing objectives and the need for specialized optimization strategies when handling unstable, antagonistic subtasks.

Drug Recommendation System Based on LDA and Sentiment Analysis New York, NY
Columbia University Oct 2021 — Dec 2021

- Developed a system showing promising results in recommending drugs to users based on their past reviews and predicted ratings, suggesting potential use in personalized medicine.
- Used Word2Vec for understanding context in reviews and CNN in TensorFlow for rating prediction.
- Designed and implemented a drug recommendation system, embedding users and items using LDA over PCA for a more accurate dimensional reduction and understanding of user-item relationships.
- Generated a ranking matrix based on predicted ratings from sentiment analysis, successfully linking user preferences to corresponding drug recommendations through case studies.

Aircraft Icing Analysis based on Machine Learning Shanghai, China
Shanghai Jiao Tong University Nov 2020 — Jun 2021

- Collaborated with 3 teams on solving aerodynamic problems such as the impact of icing, utilizing Machine Learning as a new method. Successfully reduced prediction time from 2 days to 1 second, and published a paper in the Journal of Aeronautics, Astronautics and Aviation.

SKILLS

- **Programming:** Python, Java, C++, SQL, R, SPSS, SAS, MATLAB, Spark, HTML, CSS, JavaScript, LaTeX
- **Python Packages:** PyTorch, TensorFlow, Scikit-Learn, NumPy, Pandas, SciPy, NLTK, Matplotlib, Seaborn, PySpark
- **Developer Tools:** Linux, Vim, Git, Docker, GCP, AWS, Spring Boot, MySQL, Spark, REST api, DevOps Cycle