Benhao Huang

huskydogewoof@gmail.com

https://huskydoge.github.io/ | Google Scholar

EDUCATION

School of Electronics Information and Electrical Engineering, Shanghai Jiao Tong University Sep 2021 – Current

B.ENG. in Computer Science and Technology (IEEE Honor Class), GPA: 93.14/100.00, 4.08/4.30, (Rank 5/129)

Selected Courses:

- **Computer Science:** Design and Analysis of Algorithms (A+), Computer Networks (A+), Operating System (A+), Programming Languages and Compilers (A+), Natural Language Processing (A+), Database System Technology (A+), Computer Vision (A+), Principles and Methods of Program Design (A+), Program Design Practice (A+), Introduction to Data Science (A+)
- **Mathematics:** Mathematical Analysis (A+), Linear and Convex Optimization (A+), Information Theory (A+), Complex Analysis (A+), Probability and Statistics (A+)
- **Additional Academic Pursuit:** Engaged in a dual degree program in Mathematics and Applied Mathematics. Coursework includes: Complex Analysis, Abstract Algebra, Linear Algebra II

TOEFL: 107, S24, R27, L29, W27, **GRE**: V157, Q170, AW4.0

RESEARCH INTEREST

- LLM Interpretability
- LLM Agent, World Model, LLM Reasoning
- LLM Alignment, Long Context Modeling

PUBLICATIONS

- DCA-Bench: A Benchmark for Dataset Curation Agents [paper | code]
  Benhao Huang, Yingzhuo Yu, Jin Huang, Xingjian Zhang, Jiaqi Ma. **(In Submission)**
- Seeing is not always believing: The Space of Harmless Perturbations [paper]
  Lu Chen, Shaofeng Li, *Benhao Huang*, Fan Yang, Zheng Li, Jie Li, Yuan Luo. **(In Submission)**
- Defining and Extracting Generalizable Interaction Primitives from DNNs. [paper | code]
  Lu Chen, Siyu Lou, *Benhao Huang*, Quanshi Zhang. **ICLR 2024.**

RESEARCH PROJECTS

Pandora: Towards General World Model with Natural Language Actions and Video States

Work in Process (co-leading) *Advisor: Prof. Zhiting Hu* Jun 2024 – Present

DCA-Bench: A Benchmark for Dataset Curation Agents

NIPS 2024 Under Review (1st author) *Advisor: Prof. Jiaqi Ma* [paper | code] Jan 2023 – Present

- Identified a novel task for LLM agents, detecting dataset quality issues for the purpose of automating AI training data curation, and developed the first benchmark for this task.
- Developed an LLM-based automatic evaluator, which is shown to be reliable and robust to self-preference or length bias through experiments.
- I led and fully engaged in this project, including surveying, code implementing, experiments and writings.

Defining and Extracting Generalizable Interaction Primitives from DNNs

ICLR 2024 *Advisor: Prof. Quanshi Zhang* [paper | code] Sep 2023 – Jan 2024

- Given different DNNs trained for the same task, developed a method to extract their shared interactions.
- By conducting contrast experiments, we showed that the extracted interactions can better reflect common knowledge shared by different DNNs.
- I implemented the main experiment codes and engaged in algorithm design.

RESEARCH EXPERIENCES

- Research Intern, MAITRIX Lab, University of California San Diego. Apr 2024 – Present

Worked on World Model video generation, benchmark development, and LLM reasoning. *Advisor: Prof. Zhiting Hu*

- Research Intern, Alignment Team, Moonshot AI Mar 2024 – Jun 2024

Explored prompt priorities alignment of LLM *Advisor: Flood Sung, Yanan Zheng*

- Research Intern, TRAIS Lab, University of Illinois Urbana-Champaign Nov 2023 – Present

Constructed a LLM Agent benchmark for dataset issues detection *Advisor: Prof. Jiaqi Ma*

- Research Intern, XAI Lab, Shanghai Jiao Tong University — Apr 2023 – Jan 2024

Worked on extracting common knowledge of different LLMs *Advisor: Prof. Quanshi Zhang*

AWARDS

- Rui Yuan-Hong Shan scholarship (Top 2%), SJTU — 2022 - 2023
- Shao Qiu scholarship (Top 4%), SJTU — 2021 - 2022
- Meritorious Winner of Mathematical Contest In Modeling — 2022

OTHERS

- Student Mentor of CS2612 Programming Languages and Compilers — 2023 - 2024
- Student Mentor of CS2601 Convex Optimization — 2023 - 2024
- Volunteer of Shanghai Marathon — 2022 - 2024
- Member of the Outreach Department, SJTU Spark Program Student Associations — 2021 - 2022