Hongxin Zhang

University of Massachusetts, Amherst
Email: hongxinzhang@umass.edu
Tel: 4132759957
Homepage: https://icefoxzhx.github.io/

EDUCATION

University of Massachusetts Amherst, Amherst, MA, USA — Sep. 2023 - Present
Ph.D. in Computer Science — *Advisor: Chuang Gan*

Shanghai Jiao Tong University, Shanghai, China — Sep. 2019 - June 2023
B.E. in Computer Science — *ACM Honors Class*

PUBLICATIONS

1. **Hongxin Zhang***, Weihua Du*, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, Chuang Gan. "Building Cooperative Embodied Agents Modularly with Large Language Models" *International Conference on Learning Representations (ICLR)*, 2024.
2. Qinhong Zhou, Sunli Chen, Yisong Wang, Haozhe Xu, Weihua Du, **Hongxin Zhang**, Yilun Du, Joshua B. Tenenbaum, Chuang Gan. "HAZARD Challenge: Embodied Decision Making in Dynamically Changing Environments" *International Conference on Learning Representations (ICLR)*, 2024.
3. Zhiqing Sun, Yikang Shen, **Hongxin Zhang**, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, Chuang Gan. "SALMON: Self-Alignment with Principle-Following Reward Models" *International Conference on Learning Representations (ICLR)*, 2024.
4. Zhiqing Sun, Yikang Shen, Qinhong Zhou, **Hongxin Zhang**, Zhenfang Chen, David Cox, Yiming Yang, Chuang Gan. "Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision" *Neural Information Processing Systems (NeurIPS)*, 2023.
5. Omar Shaikh, **Hongxin Zhang**, William Held, Michael Bernstein, Diyi Yang. "On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning" *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.
6. Bolin Lai*, **Hongxin Zhang***, Miao Liu*, Aryan Pariani*, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M. Rehg, Diyi Yang. "Werewolf Among Us: A Multimodal Dataset for Modeling Persuasion Behaviors in Social Deduction Games" *Findings of the Association for Computational Linguistics: ACL*, 2023.
7. Albert Lu*, **Hongxin Zhang***, Yanzhe Zhang, Xuezhi Wang, Diyi Yang. "Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints" *Findings of the Association for Computational Linguistics: EACL*, 2023.
8. **Hongxin Zhang**, Yanzhe Zhang, Ruiyi Zhang, Diyi Yang. "Robustness of Demonstration-based Learning Under Limited Data Scenario" *Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

RESEARCH EXPERIENCE

University of Massachusetts Amherst — Amherst, Massachusetts
Graduate Research Assistant, advised by Prof. Chuang Gan — Sep. 2023 - Present

- **Building Cooperative Embodied Agents Modularly with Large Language Models**
  Hongxin Zhang*, Weihua Du*, Jiaming Shan, Qinhong Zhou, Yilun Du, Joshua B. Tenenbaum, Tianmin Shu, Chuang Gan
  ICLR 2024 Poster
  - Conducted the first systematic study on LLMs' capacity for planning and communication in embodied multi-agent cooperation.
  - Introduced a novel framework that utilizes LLMs to build cooperative embodied agents, surpassing strong planning-based methods.
  - Conducted a user study to evaluate the possibility of achieving effective and trustworthy human-AI cooperation using LLMs.
- **HAZARD Challenge: Embodied Decision Making in Dynamically Changing Environments**
  Qinhong Zhou, Sunli Chen, Yisong Wang, Haozhe Xu, Weihua Du, **Hongxin Zhang**, Yilun Du, Joshua B. Tenenbaum, Chuang Gan
  ICLR 2024 Poster
  - Proposed a new simulated embodied benchmark, called HAZARD, specifically designed to assess the decision-making abilities of embodied agents in dynamic situations.
  - Developed an LLM-based agent and performed an in-depth analysis of its promise and challenge in solving these challenging tasks.
- **SALMON: Self-Alignment with Principle-Following Reward Models**
  Zhiqing Sun, Yikang Shen, **Hongxin Zhang**, Qinhong Zhou, Zhenfang Chen, David Cox, Yiming Yang, Chuang Gan
  ICLR 2024 Poster
  - Proposed a new RLAIF paradigm for self-aligning language models from scratch, using only a small set of human-defined principles as guidance.

  - Open-sourced Large Language Model Dromedary-2, which is trained with the SALMON paradigm on the LLaMA-2-70b base language model, with Principle-Driven Self-Alignment as the Supervised Fine-Tuning (SFT) strategy to initialize the policy model.

- **Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision**
  Zhiqing Sun, Yikang Shen, Qinhong Zhou, **Hongxin Zhang**, Zhenfang Chen, David Cox, Yiming Yang, Chuang Gan
  Neurips 2023 Spotlight
  - Proposed a novel approach called `SELF-ALIGN`, which combines principle-driven reasoning and the generative power of LLMs for the self-alignment of the AI agents with minimal human supervision.
  - Released Dromedary, an open-source self-aligned language model trained with minimal human supervision.

Stanford University Stanford, California

Visiting Student Researcher, advised by Prof. Diyi Yang Sep. 2022 - Jan. 2023

- **On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning**
  Omar Shaikh, **Hongxin Zhang**, William Held, Michael Bernstein, Diyi Yang **ACL 2023**
  - Performed a controlled evaluation of zero-shot CoT across two sensitive domains: harmful questions & stereotype benchmarks.
  - Found that using zero-shot CoT reasoning in a prompt can significantly increase a model's likelihood of producing undesirable output.
- **Werewolf Among Us: A Multimodal Dataset for Modeling Persuasion Behaviors in Social Deduction Games**
  Bolin Lai*, **Hongxin Zhang***, Miao Liu*, Aryan Pariani*, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James M. Rehg, Diyi Yang. **ACL 2023**
  - Presented the first multimodal dataset for persuasion modeling which is collected in naturalistic social game scenarios with intensive in-person conversations with multiple players.
  - Conducted comprehensive experiments to show the importance of context and visual signals for persuasion strategy prediction.
  - Provided additional experiment results to investigate model generalization on the persuasion modeling task, and discuss how persuasion strategy influences the game voting outcome.
- **Bounding the Capabilities of Large Language Models in Open Text Generation with Prompt Constraints**
  Albert Lu*, **Hongxin Zhang***, Yanzhe Zhang, Xuezhi Wang, Diyi Yang. **EACL 2023**
  - Provided a taxonomy of prompts containing stylistic or structural constraints to facilitate finer-grained analyses of open text generation.
  - Conducted a systematic experiment using proposed taxonomy by creating 288 different prompts and evaluating 3000+ generated outputs to analyze the capabilities and limitations of current LLMs on open-ended text generation.
  - Analyzed in-context mitigation strategies to improve model performance.

Georgia Institute of Technology Atlanta, Georgia

Visiting Student Researcher, advised by Prof. Diyi Yang Aug. 2021 - Aug. 2022

- **Robustness of Demonstration-based Learning Under Limited Data Scenario**
  Hongxin Zhang, Yanzhe Zhang, Ruiyi Zhang, Diyi Yang. **EMNLP 2022**
  - Investigated the robustness of demonstration-based learning by designing pathological demonstrations.
  - Experiments show demonstrations composed of random tokens still make the model a better few-shot learner; the length of random demonstrations and the relevance of random tokens are the main factors affecting the performance; demonstrations increase the confidence of model predictions on captured superficial patterns.

SELECTED PROJECTS

Mx* Compiler Java

- Developed a compiler that compiles C-and-Java-like language (Mx*) to Assembly Language
- Designed LLVM-like IR
- Implemented optimizations based on data-flow and control-flow analysis

RISC-V CPU Verilog

- Designed a RISC-V CPU that supports RV32I Base Integer Instruction Set V2.0 (2.1 2.6) with the following property:
  - 5-stage pipelined
  - 1KB iCache, direct mapped
  - Branch Prediction using 2-bit saturating counter BHT with Branch Target Buffer (Size is 128*4 Byte)
  - Running on FPGA with 200MHz
  - Data forwarding supported

TEACHING EXPERIENCE

Teaching Assistant: Great Ideas in Computer Science — Fall 2020
Teaching Assistant: Data Structure — Spring 2021
Student Instructor: Principle and Practice of Computer Algorithms — Summer 2021
Teaching Assistant: Mathematical Logic — Fall 2021
Teaching Assistant: Machine Learning — Spring 2022

SKILLS AND INTERESTS

Programming: C++ / Python / Java / Go / Verilog
Language: Mandarin (native), English (TOEFL 109/120), Latin (Beginner)
Interests: Literature, Movie, Billiards

HONORS AND AWARDS

Shanghai Excellent Graduate (Awarded for overall performance in undergraduate career) — 2023
Zhiyuan Outstanding Student Scholarship (Highest award for undergraduate in SJTU) — 2023
Foresight-Sequoia Scholarship (**5** winners at Zhiyuan College) — 2022
Academic Excellence Scholarship — 2020, 2021, 2022
35th China's National Olympiad in Informatics (NOI) Silver Medal — 2018