Zhaoyi (Joey) Hou

website: joeyhou.github.io email: joey.hou@pitt.edu

Summary I am a Ph.D. researcher focused on the foundations of generative AI, model behavior, and large-scale evaluation methodologies. My work develops computational frameworks for understanding and improving LLM reasoning, creativity, and tool-augmented workflows, with an emphasis on reliability, failure analysis, and inference-time behavior. I have experience across the full research lifecycledesigning datasets, building novel evaluation pipelines, fine-tuning and analyzing large models, and validating hypotheses through rigorous experimentation.

Education

University of Pittsburgh 2023 - Present
Computer Science, Doctor of Philosophy, Graduate Student Researcher
Research Advisor: Dr. Xiang Lorraine Li

University of Pennsylvania 2021 - 2023
Data Science, Master of Science in Engineering
Research Advisor: Dr. Chris Callison-Burch

University of California, San Diego 2017 - 2021
Data Science, Bachelor of Science Cum Laude
Research Advisor: Dr. Jingbo Shang

Work Experience

Amazon Web Service May 2025 - Aug 2025
Applied Scientist Intern (AWS Q Console)
- Designed a generalized evaluation framework for identifying fundamental failure modes in large models, including tool misuse, hallucinated actions, and breakdowns in multi-step reasoning.
- Conducted research on agentic system reliability, analyzing how LLMs behave under complex, multi-step workflows and diverse prompting conditions.
- Developed and tested behavioral diagnostics for large models, contributing insights relevant to trustworthiness, inference-time behavior, and scalable evaluation of emerging AI systems.
- Collaborated with scientists and engineers to integrate findings into broader AI oversight and robustness pipelines.

United Imaging Intelligence May 2023 - Aug 2023
LLM Research Intern
-Fine-tuned and evaluated domain-specific large language models, studying how training data, adaptation methods, and retrieval augmentation influence model generalization and robustness.
- Built and analyzed pipelines for knowledge-intensive reasoning and high-stakes QA tasks, investigating model reliability under distribution shift.
- Explored strategies for improving data quality, grounding, and controlled generation, contributing to broader questions in scalable training and trustworthy model deployment.
- Collaborated closely with domain experts, emphasizing a rigorous, human-centered approach to model behavior analysis and evaluation.

Publications

[11] **Zhaoyi Joey Hou**, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, Daniel Khashabi, Xiang Lorraine Li. *CreativityPrism: A Holistic Benchmark for LLM Creativity* [In Submission]

[10] **Zhaoyi Joey Hou**, Adriana Kovashka, Xiang Lorraine Li. *Leveraging Large Models for Evaluating Novel Content: A Case Study on Advertisement Creativity* [EMNLP2025]

[9] **Zhaoyi Joey Hou***, Tanya Shourya*, Yingfan Wang, Shamik Roy, Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah. *Multi-Faceted Evaluation of Tool-Augmented Dialogue Systems* [In Submission]

[8] **Zhaoyi Joey Hou**, Alejandro Ciuba, Xiang Lorraine Li. *Improve LLM-based Automatic Essay Scoring with Linguistic Features* [Innovation and Responsibility in AI-Supported Education (Spotlight Paper) - AAAI2025]

[7] **Zhaoyi Hou**, Li Zhang, Chris Callison-Burch. *Choice-75: A Dataset on Decision Branching in Script Learning* [LREC-COLING2024]

[6] Tianyi Zhang*, Li Zhang*, **Zhaoyi Hou**, Ziyu Wang, Yuling Gu, Peter Clark, Chris Callison-Burch, Niket Tandon. *PROC2PDDL: Open-Domain Planning Representations from Texts* [2nd Natural Language Reasoning and Structured Explanations Workshop - ACL2024]

[5] Alyssa Hwang*, Bryan Li*, **Zhaoyi Hou***, Dan Roth. *Large Language Models as Sous Chefs: Revising Recipes with GPT-3*

[4] Tianyi Zhang*, Isaac Tham*, **Zhaoyi Hou***, Jiaxuan Ren, Liyang Zhou, Hainiu Xu, Li Zhang, Lara J. Martin, Rotem Dror, Sha Li, Heng Ji, Martha Palmer, Susan Brown, Reece Suchocki, and Chris Callison-Burch. *Human-in-the-Loop Schema Induction* [ACL2023]

[3] Xiaochen Kev Gao, **Zhaoyi Hou**, Yifei Ning, Jingbo Shang, Vish Krishnan. *Towards Comprehensive Patent Approval Predictions: Beyond Traditional Document Classification* [ACL2022]

[2] Caitlin A. Stamatis, Jonah Meyerhoff, Tingting Liu, **Zhaoyi Hou**, Garrick Sherman, Brenda L. Curtis, Lyle H. Ungar, David C. Mohr. *The Association of Language Style Matching in Text Messages with Symptoms of Affective Psychopathologies* [Procedia Computer Science]

[1] Artemis Panagopoulou, Manni Arora, ...(6 more), **Zhaoyi Hou**, Alyssa Hwang, Lara Martin, Sherry Shi, Chris Callison-Burch, Mark Yatskar. *QuakerBot: A Household Dialog System Powered by Large Language Models* [Alexa Prize TaskBot Challenge Proceedings]

(*Equal contribution)

Projects

Amazon Alexa TaskBot Competition Nov 2021 - Apr 2022
Information Retrieval
- Implemented the document retrieval module for the Alexa TaskBot competition;
- Improved the retrieval success rate by 25% and advanced to the final list.

Awards

Best Problem-Solution 2024
Annual Doctoral Guild Poster Slam
School of Computing and Information at University of Pittsburgh

HDSI Undergraduate Scholarship 2019
Halıcıoğlu Data Science Institute
University of California, San Diego

Service

Reviewer
Association of Computational Linguistics Rolling Review (ARR) 2024, 2025
Journal of Educational Data Mining 2025

Teaching

University of Pittsburgh
Teaching Assistant 2024, 2025
- CS1675 (Introduction to Machine Learning)
- CS1503 (Mathematical Foundation for Machine Learning)
- CS1671 (Human Language Technologies)

Penn Engineering Online
Course Development Assistant 2022, 2023
- CIS5300 (Computational Linguistics)

UC San Diego
Student Tutor 2018, 2019, 2021
- CSE151A (Intro to Machine Learning)
- DSC20 (Intro to Data Structure)