Keyu He

(213) 713-2973 | keyuhe@cmu.edu | keyu-he.github.io | www.linkedin.com/in/keyuhe/

Education

Carnegie Mellon University, Pittsburgh, PA **Aug. 2025 – May 2027**
Master of Science in Intelligent Information Systems

University of Southern California, Los Angeles, CA **Aug. 2021 – May 2025**
Bachelor of Science in Computer Science **GPA: 3.98/4.00**
Bachelor of Arts in Applied and Computational Mathematics
Minor in Artificial Intelligence Applications

Research Interests

Explainable NLP Systems, Socially Aware AI, Human-AI Interactions, etc.

Publications

1. **Keyu He**, Tejas Srinivasan, Brihi Joshi, Xiang Ren, Jesse Thomason, Swabha Swayamdipta. *Believing without Seeing: Quality Scores for Contextualizing Vision-Language Model Explanations*. Submitted to NeurIPS 2025. Under Review.
2. Brihi Joshi*, **Keyu He***, Sahana Ramnath, Sadra Sabouri, Kaitlyn Zhou, Souti Chattopadhyay, Swabha Swayamdipta, Xiang Ren. *ELI-Why: Evaluating the Pedagogical Utility of LLM Explanations*. Findings of ACL 2025. 10.48550/arXiv.2506.14200
3. Huihan Li*, Arnav Goel*, **Keyu He**, and Xiang Ren. *Attributing Culture-Conditioned Generations to Pretraining Corpora*. ICLR 2025. 10.48550/arXiv.2412.20760.
(* Indicates equal contribution)

Research Experience

University of Southern California, Viterbi School of Engineering
Research fellow in Center for Undergraduate Research in Viterbi Engineering (CURVE) Jan. 2024 – May 2025

- First-authored *Believing without Seeing: Quality Scores for Contextualizing Vision–Language Model Explanations* (hosted in DILL Lab); under review at NeurIPS 2025.
- Designed visual-faithfulness and contrastiveness metrics; on A-OKVQA/VizWiz they are better-calibrated to correctness; showing the scores improved humans' correctness judgments by 11.1% and cut false belief by 15.4%.
- Co-first-authored *ELI-Why*—a 13.4K "Why"-question benchmark for evaluating LLM pedagogy (hosted in INK lab); **accepted to ACL Findings 2025**.
- Led study design & execution (tasks/rubrics/prompts), built the human-study web app and statistical analyses; showed GPT-4 under-personalizes to learner profiles and keeps grade-level complexity nearly constant despite prompting.
- Secured **$6,750** across four CURVE terms (Spring 2024, Summer 2024, Fall 2024, Spring 2025).

University of Southern California, Viterbi School of Engineering
Research contributor in INK lab Aug. 2024 – Oct. 2024

- Contributed to *MEMOed*—a framework attributing culture-conditioned generations to pretraining memorization (applied across 110 cultures); **accepted to ICLR 2025**.
- Implemented the survey infrastructure: integrated Google APIs to host culture-specific forms and built a routing service that directs participants to the correct survey, enabling scalable data collection.

Major Awards

- **Silver Medal**, Kaggle Competition, Ranked 75/2175 (Top 3.4%) on the global leaderboard, LLM-Prompt-Recovery Project, 2024
- **USC Academic Achievement Award**, awarded in 4 terms (F22, S23, S24, F24); total support ≈**$24,000**
- **4th Place**, USC Integral Bee Competition, 2022
- **1st Prize**, International Linguistics Olympiad (Senior Level), Individual Open Round, China, 2021
- **1st Prize**, International Linguistics Olympiad (Senior Level), Team Open Round, China, 2021

Projects

LLM Prompt Recovery Project — USC Mar. 2024 – Apr. 2024

- Developed a system to recover user prompts given original text and modified text generated by Gemma.
- Fine-tuned the Mixtral model using custom metrics, achieving a score of 0.65 with sentence-T5-base and sharpened cosine similarity (exponent = 3).
- Awarded a silver medal in the Kaggle competition for outstanding performance (ranked 75/2175, top 3.4%).
- See the final fine-tuned model here: Mixtral-8x7b Instruct Finetuned.

AI-Based Career Advisor — USC Nov. 2024 – Dec. 2024

- Developed an AI advisor to assist users in planning career paths based on skills and interests, leveraging datasets such as the JobSkills Dataset (1.3M entries) and LinkedIn Jobs Dataset.
- Implemented a cosine similarity search on sentence embeddings for matching user skills with most-fit jobs and identifying skill gaps.
- Integrated Bing AI for real-time resource and job application link retrieval, enhancing usability with advanced support metrics for internal consistency verification.
- Enabled post-hoc evaluation using a T5 fine-tuned entailment verification model to validate skill-job relevance, ensuring reliable recommendations.

Enhancing Debugging Skills of LLMs with Prompt Engineering — USC Aug. 2023 – Nov. 2023

- Improved the debugging capabilities of LLMs using innovative prompt engineering techniques.
- Conducted experiments using various prompting strategies (Zero-Shot, Few-Shot, Chain of Thought) to enhance the efficiency of GPT models in debugging tasks.
- See the technical report here: https://swabhs.com/fall23-csci499-lm4nlp/assets/reports/KeyuHe_MaxLi_JosephLiu.pdf

Automated Hate Speech Detection in Social Media — USC Sep. 2023 – Dec. 2023

- Led the development of an advanced machine learning model for detecting hate speech on social media, employing a mix of techniques with a focus on BERT fine-tuning.
- Achieved a 94% accuracy rate in classification tasks, underlining the model's effectiveness in enhancing online safety and inclusivity through rigorous evaluation and optimization strategies.

Teaching Experience

University of Southern California (USC), Los Angeles, CA Sep. 2022 – May 2025

Teaching & Grading Assistant

- **Course Producer**: CSCI-102 (Fundamentals of Computation) and CSCI-360 (Introduction to Artificial Intelligence).
- **Grader**: Calculus sequence — MATH-117, MATH-126, MATH-129, MATH-226.
- Supported instruction via office hours and discussion help; coordinated grading rubrics and turnaround with instructor/grader teams.

Technical Skills

Programming: C++, C, Python, Java, MySQL, HTML, CSS, JS, x86-64 Assembly
Frameworks/Tools/Software: PyTorch, Pandas, NumPy, Git, AWS, LaTeX, Mathematica, Matlab
Areas of Expertise: Machine Learning, Natural Language Processing (NLP), Large Language Models (LLMs), Data Science / Data Engineering
Language: Mandarin (native), English (professional)