Keyu He

Los Angeles, CA — (213) 713-2973 — keyuhe@cmu.edu — keyu-he.github.io

Education

Carnegie Mellon University, Pittsburgh, PA **August 2025 – May 2027 (Expected)**
Master of Science in Intelligent Information Systems (MIIS)

University of Southern California, Los Angeles, CA **August 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
Dean's List Fall 2021 - Fall 2024
USC Center for Undergraduate Research in Viterbi Engineering (CURVE) Fellow
Member of Phi Kappa Phi Honor Society

Research Interests

Explainable NLP Systems, Interpretable Machine Learning, Cross-Disciplinary AI Research, etc.

Conference Publications and Working Papers

1. Huihan Li*, Arnav Goel*, **Keyu He**, and Xiang Ren. *Attributing Culture-Conditioned Generations to Pretraining Corpora*. ICLR 2025. 10.48550/arXiv.2412.20760.
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*. Submitted to ACL 2025. Under review.
3. **Keyu He**, Tejas Srinivasan, Brihi Joshi, Swabha Swayamdipta. *Beyond the Text: How Explanation Qualities Influence User Trust in Visual Language Models*. Under preparation. Aiming for NeurIPS 2025.

(* Indicates equal contribution)

Research Experience

Center for Undergraduate Research in Viterbi Engineering (CURVE) Fellow
— USC Viterbi School of Engineering, US Jan. 2024 - Feb. 2025

- Co-first authored a paper introducing ELI-WHY—a benchmark of 13.4K "Why" questions designed to evaluate the pedagogical capabilities of large language models in education.
- Demonstrated that, despite their widespread use, language models like GPT-4 struggle to tailor explanations to diverse learner needs and knowledge backgrounds.
- Complemented human evaluations with automated metrics, which revealed that GPT-4 explanations remained indistinguishable in grade-level complexity, underscoring that inference-time instructions alone are insufficient for producing high-utility, tailored explanations.
- Fellowship awarded multiple times (Spring 2024, Summer 2024, Fall 2024, and Spring 2025), with a total funding amount of **$6,750**.
- Paper **submitted to ACL 2025** and currently under review.

Research Contributor on Cultural Bias in Language Models
— USC Viterbi School of Engineering, US Aug. 2024 - Oct. 2024

- Collaborated with PhD student Huihan Li to develop the MEMOed framework, analyzing cultural biases in language models.
- Conducted a comprehensive survey across 110 cultures, gathering qualitative and quantitative data.
- Contributed to insights showing high-frequency cultures yield more memorized generations, while low-frequency cultures often produce none.
- Our work led to a paper **accepted by ICLR 2025**.

Research Contributor on Visual Language Models (VLM) Project
— USC Viterbi School of Engineering, US Jun. 2024 - Present

- Leading the "Beyond the Text: How Explanation Qualities Influence User Trust in Visual Language Models" project as first author under the guidance of Prof. Swabha Swayamdipta and PhD mentors Brihi Joshi and Tejas Srinivasan, focused on evaluating the faithfulness, relevance, and completeness of VLM rationales to help users assess model reliability.
- Conducting both automatic and human evaluations to identify limitations of current text-only metrics and explore new vision-specific metrics.
- Running human studies to examine how presenting explanation qualities can enhance user trust in the model.

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

Teaching and Grading Assistant — USC, US Sep. 2022 - Present
- Selected for multiple roles: **Course Producer** for CSCI-102 and CSCI-360, **Grader** for MATH-117, MATH-126, MATH-129 and MATH-226.
- Ensured a consistent approach to teaching and grading by regularly collaborating with faculty and fellow graders.

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**, Fall 2022, Spring 2023, Spring 2024, Fall 2024.
  - This award covered 11 units of tuition costs in total, amounting to approximately **$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

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
Language: Mandarin (native), English (professional)
Areas of Expertise: Machine Learning, Natural Language Processing (NLP), Large Language Models (LLMs), Data Science / Data Engineering