Keyu He

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

Education

University of Southern California, Los Angeles, CA **August 2021 – May 2025**
Double Major in Computer Science and Applied and Computational Mathematics **GPA: 3.97/4.00**
Minor in Artificial Intelligence Applications
Dean's List Fall 2021, Spring 2022, Fall 2022, Spring 2023, Fall 2023, Spring 2024
Member of Phi Kappa Phi Honor Society
USC Center for Undergraduate Research in Viterbi Engineering (CURVE) Fellow

Research Interests

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

Relevant Coursework

Language Models in Natural Language Processing (A)
Directed Research (A)
Introduction to Machine Learning (A)
Applications of Machine Learning (A)
Professional C++ (A)
Introduction to Artificial Intelligence (A)
Applied Neural Networks (A)
Introduction to Data Analysis (A)
Introduction to Computer Systems (A)
Introduction to Embedded Systems (A)
Introduction to Internetworking (A)
Data Structure and Object-Oriented Design (A)
Linear Algebra and Differential Equations (A)
Probability Theory (A)
Mathematical Statistics (A)
Numerical Methods (A)
Introduction to Algorithms and the Theory of Computing (A-)

Conference Publications and Working Papers

Huihan Li*, Arnav Goel*, **Keyu He**, and Xiang Ren. Attributing Culture-Conditioned Generations to Pretraining Corpora. *Submitted to ICLR 2025, under review.* Accepted to *SoCalNLP Symposium 2024*.

Brihi Joshi*, **Keyu He***, Kaitlyn Zhou, Sadra Sabouri Halestani, Souti Chattopadhyay, Swabha Swayamdipta, Xiang Ren. Assessing Language Models' Capability to Explain to Different Audiences. *Under preparation. Aiming for ACL 2025.*

Keyu He, Brihi Joshi*, Tejas Srinivasan*, Swabha Swayamdipta. Beyond the Text: How Explanation Qualities Influence User Trust in Visual Language Models. *Under preparation. Aiming for ACL 2025.*

(* Indicates equal contribution)

Research Experience

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

- Contributing to the "Assessing Language Models' Capability to Explain to Different Audiences" project under Prof. Xiang Ren, Prof. Swabha Swayamdipta and PhD mentor Brihi Joshi, focusing on the comparative study of mechanistic and teleological explanations of "Why" questions by humans and GPT-4.
- Conducted quantitative analysis using lexical and semantic metrics and qualitative assessment to evaluate text complexity and alignment with user roles in curiosity-driven scenarios.
- Engaged in interdisciplinary workshops on cognitive science and NLP to refine methodologies for assessing context-driven explanation types.

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

- Working on the "Beyond the Text: How Explanation Qualities Influence User Trust in Visual Language Models" project under Prof. Swabha Swayamdipta and PhD mentors Brihi Joshi and Tejas Srinivasan, focusing on developing tools to assess the faithfulness, relevance, and completeness of VLM rationales, aiming for helping users better judge whether to rely on explanations.
- Conducting both automatic and human evaluations to identify limitations of current text-only metrics and explore new vision-specific metrics.

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.

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.

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.

USC Study Room/Area Rating Web Application — USC Aug. 2022 – Dec. 2022
- Contributed to a team in the development of a web interface that allows students to rate and comment on study rooms located throughout the USC campus.
- Implemented sorting and filtering features, as well as a simplified reservation system, showcasing technical skills and contribution to the application's functionality.

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
- 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
Software: LaTeX, Mathematica, Matlab
Language: Mandarin (native), English (professional)