ZEKUN ZHAO

Department of Computer Science and Engineering
University of California, Santa Cruz
zzhao99@ucsc.edu
zekunzhao.github.io

BIO

My research focuses on large language model inference with **knowledge-based reasoning**. I am particularly interested in developing methods that enhance both the understanding and generation of natural language with an emphasis on **efficient and effective performance**.

EDUCATION

UNIVERSITY OF CALIFORNIA, SANTA CRUZ — Sep 2021 - Present
Ph.D. in Natural Language Processing
Advisors: Jeffrey Flanigan

UNIVERSITY OF CALIFORNIA, SANTA CRUZ — Sep 2018 – Mar 2021
Master of Science in Computer Science

UNIVERSITY OF CALIFORNIA, BERKELEY — Dec 2017 – Jun 2018
Exchange Student in the Department of Electrical Engineering and Computer Sciences

NANKAI UNIVERSITY, TIANJIN — Sep 2014 – Jun 2018
Bachelor of Engineering in Intelligent Science and Technology

PROJECT

FAST LLM INFERENCE WITH PARALLEL PROMPTING — Sep 2024 - Jun 2025

- Developing a novel inference method for Transformer Large Language Models(LLMs)
- Improving the inference efficiency without compromising generation quality
- Optimizing generation latency and throughput with fast parallel generation
- Reducing inference time on various popular datasets (SQUAD, QuAC, DROP) by over 70%

IMPLICIT ROLE RECOGNITION IN DOCUMENT — Sep 2024 - Jun 2025

- Designed a novel prompt method with the knowledge graph for document-level implicit role recognition
- Constructed the QA Paris with predicate-argument relations extracted from PropBank
- Implemented full-document semantic parsing by incorporating concept coreference and implicit role recognition
- Improving over existing state-of-the-art methods in semantic knowledge graph representation

ABSTRACT MEANING REPRESENTATION (AMR) PARSING — Sep 2023 – Jun 2024

- Designed a novel method for generating out-of-domain semantic representation AMR pairs
- Implemented a data pipeline with Automatic Keyword Extraction, Back Generation, and Pseudo-AMR Parsing
- Developed a quality estimation method based on the semantic similarity score
- Fine-tuned a language model for boosting the performance of AMR Paring in the out-of-domain scenario

PUBLICATIONS

- Jon Cai, Kristin Wright-Bettner, Zekun Zhao, Shafiuddin Rehan Ahmed, Abijith Trichur Ramachandran, Jeffrey Flanigan, Martha Palmer, and James Martin. 2025. LiDARR: Linking Document AMRs with Referents Resolvers. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 426–435, Vienna, Austria.