Chuangji Li

+1 (878) 999 6607 | chuangjl@andrew.cmu.edu | Linkedin | GitHub

EDUCATION

Carnegie Mellon University **Sept 2021 - Expected May 2025**
BS in Statistics and Machine Learning (GPA 3.67/4.00) *Dietrich / Computer Science*

WORK EXPERIENCE

Teaching Assistant **Carnegie Mellon University**
36225 Intro to Probability Theory *May 2023 - July 2023*

RELEVANT COURSEWORKS

Generative AI (10423), Machine Learning (10701), Deep Learning (11785), Advanced Natural Language Processing (11711), Computer Vision (16385), Deep Learning Systems (10714), Probabilistic Graphical Models (10708), Probability and Mathematical Statistics (36700)

TECHNICAL SKILLS

Programming Languages: Python, C, C++, R, Rust, JAVA
Libraries and Tools: PyTorch, Sklearn, Pandas, Numpy, Git, Docker, AWS EC2
DL Architectures CNN, RNN, Transformers, VAE, GAN, Stable Diffusion, CLIP, ViT

RESEARCH EXPERIENCE

Handwritten English Recognition System [Link] **South China University of Technology**
Research Assistant, Mentored by Lianwen Jin *June 2023 - Aug 2023*

- The goal is to build a handwritten **English recognition system** using segmentation based method for Chinese recognition. However, English calligraphy is difficult to segment.
- Generated **synthetic** data of handwritten English by randomly sampling from human written text and fonts.
- **Modified** the model framework including the dimensionality, architecture layout, and recognition settings.
- **Improved** the performance comparing to baseline by **2.5%** in accurate rate (AR) and **4%** in correct rate (CR)

PROJECTS

Retrieval Augmented Generation (RAG) Question-Answering System [GitHub Repo] **Carnegie Mellon University**
Leader, Mentored By Graham Neubig *Jan 2024 - March 2024*

- Builded a **Question-Answering system** which answers questions about CMU faculty, courses, history and events.
- Scrapped and annotated CMU raw data with 12,000 samples, extracted important information using **Llama2**
- **augmented** it using uniform templates, and segregated them for better search.
- Experimented with 6 different LLMs to **encode** each file, 4 different **retriever**, and 2 different **re-ranker**.
- Achieved **0.91 accuracy** and **0.82 recall**, statistically significantly out-performing base model, as evaluated by the significance test.

Dexcom Jira Ticket Clustering System **Carnegie Mellon University**
Mentored By Peter Freeman *Jan 2024 - May 2024*

- Jira Ticketing System is a database storing user response from Dexcom. The goal is to **cluster** 450,000 replicated message.
- Experimented with various **LLM** and trained **Auto-encoder** to encode strings to vectors, extracted keywords to **enhance** the representation
- Using **K-Nearest Neighbor** to conduct clustering, and visualizing using **t-SNE/PCA** algorithm.
- Achieved **silhoutte score** of **0.475**, improved by around **0.36** comparing to the baseline model.

Sui-GPT [Link] **Carnegie Mellon University**
Research Assistant, Mentored By Eason Chen *April 2024 - Present*

- Sui Move is a rust-based programming language for **smart contract**. The goal is to build a **code-generation** model which supports writing smart contract using sui move.
- Developed **automatic data collection pipeline** which automatically updates smart contracts, libraries and base codes and use ChatGPT-4o to annotate.
- Use **Abstract Syntax Tree** to generate module, use **RAG** to improve correctness and use multi-hop **compiler** to examine function and values
- Improved generation correctness by **Expected 10%**; Won Second Place in **Sui Overflow Hackathon** in Infrastructure and Tools