Chuyan Zhou

chyzhou24@gmail.com · Github · Blog

Education

ShanghaiTech University, Shanghai, China 2022.9 – 2026.7 (expected)

Bachelor of Engineering in Computer Science
Supervised by Dr. Kewei Tu
Overall GPA 3.86/4.0, ranked 3/162 (<2%), *Major GPA* 4.0/4.0

University of California, Berkeley, California, USA 2024.8 – 2025.1

GLOBE Program in Computer Science major, College of Engineering
Overall GPA 4.0/4.0

Research Experience

VDI Center ShanghaiTech University 2023.10 – Present

Undergraduate Researcher Supervised by Dr. Kewei Tu

FreqLLM: KV Cache Pruning Based on Token Frequency (In Progress) 2025.3 – Present

- **Investigated** and (WIP) **visualized** (with Bert-viz) the impact of corpus frequency statistics on **averaged unmasked attention scores** as the importance metric for KV cache pruning.
- (WIP) Proposing a novel **KV cache pruning method** based on the above investigation, with a focus on improving the efficiency of LLMs.

Pushdown Layers for Semantic Dependency Parsing (EMNLP '25) 2025.2 – 2025.5 (expected)

- Reconstructed the model architecture (and the training pipeline) of Pushdown Layers for Constituency Parsing on the codebase of **Transformer-XL** transferring the original model into an important comparable baseline and optimizing the efficiency of its multi-head attention mechanism.
- (WIP) Building an evaluation pipeline including **perplexity based on marginal word probabilities by beam search** and **SG Test** for the model.
- (WIP) Conducting experiments on the PTB 1987-89 WSJ Treebanks datasets for comparison.

Other Contributions 2023.12 – 2025.2

- (2023.12-2025.1, as a main contributor) Developed a binary classification and regression model predicting packaging efficiency leveraging ESM Encoders, for synthesis (impainting) of gene sequence insertions into AAV2 caspid proteins with the motif for gene therapy vectors.
- (2024.9-2025.2, as the primary contributor) Investigated the feasibility of a Jacobi decoding method in a continuous fashion. Further work e.g. parallel continuous CoT is under investigation.

Miscellaneous Projects

Reconstruction and Re-evaluation of SFCNN for PLA Scoring ShanghaiTech 2025.4 – Present
Project Leader Teamed Project

- (2025.4) Constructed a PyTorch implementation of the model architecture, training and evaluation of **SFCNN** (Scoring Function 3D Convolutional Neural Network) for **protein-ligand binding affinity prediction**.
- (2025.4-6 (expected)) Developed a **novel benchmark** for similar models for inference directly on the 3D structure of protein-ligand complexes, instead of that on decoupled structures of proteins and ligands.

LLM-powered Lecture Generation UC Berkeley 2024.8 – 2024.12
Primary Contributor Teamed Project

- (2024.8-9) Independently developed the backend framework of **the lecture generation pipeline** based on FastAPI as a deployable web service. The backend service involves asynchronous task running using multithreading, task managing by API powered by Redis databases, and a metadata system managing the generated data.

- (2024.9-10) Worked as a main developer for and **integrated** the respective model components, which run on the above backend framework.
- (2024.10-11) Developed an additional **LLM-powered QA agent** based on the built backend, which involves interacting with LLMs using a long context of generated lectures with a RAG system dynamically indexing the sources (e.g. textbooks) for LLM grounding.

Activities and Competitions

SI 140A Course (Probability and Statistics for EECS) ShanghaiTech 2025.2 – 2025.7
Head Teaching Assistant

- Held discussion classes in a weekly basis for 40 students, covering the course materials, assignments and providing additional insights.
- Mainly responsible for creating & grading assignments and exams.

Multi-step ASR Pipeline Bengali.AI Speech Recognition Challenge 2023.7 – 2023.10
Primary Contributor Teamed Competition

- Developed an automatic speech recognition (**ASR**) pipeline for a **low-resource language** (Bangla/Bengali) in a team of 2. With supervised finetuning from augmented low-resource speech datasets, the inference pipeline mainly consists of denoising → inference main ASR decoder model using finetuned Wav2Vec2 whose decoder probabilities are fused with those of an n-gram draft language model → post-processing by rearranging the punctuation from plain words using a BERT-based open source model.

Honors and Awards

- ***Merit Student***, ShanghaiTech University 2025
- ***Silver Medal*** (24-th place), Award on Bengali.AI Speech Recognition Challenge on Kaggle 2023.10

Skills and Abilities

- Programming Languages: Python > C/C++ == MATLAB == Go == CUDA
- Machine Learning Frameworks: PyTorch == Huggingface
- Platform: Linux & Windows; Also familiar with usage of schedulers such as Slurm.
- TOEFL: 101, achieved in 2023.8.26; GRE: 331/340 (Q 170/170, V 161/170), achieved in 2025.3.8.
- JLPT N2 170/180, achieved in 2024.7.
- Languages: English - Fluent, Japanese - Fluent, Mandarin - Native speaker