Zheng Zheng

Arlington, TX, USA, 76019

☎ 401-215-4595 ✉ zxz7934@mavs.uta.edu linkedin portfolio

Education

University of Texas at Arlington **Aug. 2024 – Present**

Ph.D. in Computer Science *Advisor: Prof. Junzhou Huang*

- Developed multimodal alignment framework for 7 modalities under missing-modality settings in pathology prediction.
- Analyzed and designed algorithms for multi-type gradient conflict in multi-task learning, including structural subspace conflict, batch-induced stochastic noise conflict, and dominant gradient imbalance.
- Designed and fine-tuned protein foundation models (PTM) and implemented structured knowledge injection mechanisms to enhance representation adaptation and transferability.

Brandeis University **Sep. 2021 – Dec. 2023**

Master of Science in Computer Science

Missouri University of Science and Technology **Sep. 2019 – May 2021**

Bachelor of Science in Geology and Geophysics

Publications

[1] **Zheng Zheng**, Y. Guo, X. Hu, Y. Miao, H. Ma, J. Gao, and J. Huang. Heterogeneous Aligned Fusion for Survival Prediction with Missing Modalities. In *Medical Imaging with Deep Learning (MIDL)*, 2026. (To appear). Paper

[2] **Zheng Zheng**, X. Ni, and P. Hong. Multiple Abstraction Level Retrieve Augment Generation. *arXiv:2501.16952*, 2025.Paper

[3] Pengtao Zhang, **Zheng Zheng**, and Junlin Zhang. FiBiNet++: Reducing model size by low rank feature interaction layer for CTR prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM' 23, pages 4425–4429, New York, NY, USA, 2023. Association for Computing Machinery. Paper

Research Experience

Multiple Abstraction Level Retrieval-Augmented Generation **Apr. 2024 - Apr.2025**

Visiting Research Scientist with Prof. Pengyu Hong *Waltham, MA*

- Architected a hierarchical Retrieval-Augmented Generation (RAG) system with multi-level document abstraction to support coarse-to-fine retrieval.
- Built a large-scale document ingestion and parsing pipeline (8,000+ PDFs) using **Requests**, **BeautifulSoup4**, **Selenium-wire**, and **Grobid**, converting raw documents into structured nodes across four abstraction levels via **LlamaIndex**.
- Designed semantic segmentation and embedding workflows with **Linq-Embed-Mistral**, and implemented MapReduce-based summarization using **Vicuna-13B** to compress long-context information.
- Constructed QAR (question-answer-reason) datasets using **GPT-4-mini** with rule-based and BERT-assisted node selection strategies to enhance supervision quality.
- Developed evaluation and benchmarking framework using **RAGAS** metrics, achieving a **25.73%** improvement in answer correctness over standard RAG baselines.

FiBiNet++ - Recommended System **2022 - 2023**

Research Fellow *Remote*

- Conducted research focused on optimizing the model structure of both **bi-linear** and **SENet** structures within Recommandation System model **FibiNet** based on **Tensorflow** framework
- Designed and implemented an innovative "Low Rank Layer" and also transformed the **element-wise Hadamard product** into an **inner product** method within the bi-linear interaction, leading to a significant reduction in the size of the FiBiNet model
- Enhanced the efficiency of the "squeeze" step in the SENet structure by strategically splitting data points into multiple groups, improving the model's ability to capture one-dimensional features
- Conducted rigorous benchmarking and performance evaluations of various models including **DNN**, **DeepFM**, **xDeepFM**, **DCN**, **AutoInt**, **DCN v2**, **FiBiNet**, and the newly developed **FiBiNet++**, achieving significant improvements in training and inference efficiencies, ranging from **37.50% to 81.03%** across the datasets
- Achieved a substantial reduction in model size for **FiBiNet**, decreasing it by **12x to 16x** across datasets, while also delivering remarkable performance improvements over state-of-the-art models

Teaching Experience

University of Texas at Arlington **Jan. 2026 - Present**
Teaching Assistant *Arlington, TX*

- CSE-1325 : Object-Oriented Programming (Java) - Spring 2026

Brandeis Computer Science Department **Sep. 2022 - May. 2023**
Teaching Assistant with Prof. Timothy J. Hickey *Waltham, MA*

- COSI-10A : Introduction to Problem Solving in Python - Fall 2022 & Summer 2023
- COSI-153A : Mobile Application Development - Summer 2023
- COSI-29A : Discrete Structures - Fall 2023

Projects

Hire Me Now | *Docker, AWS, MERN stack, Django, Nginx, Certbot, Stripe, RESTful, LLM model* **Mar. 2023**

- Designed and developed a full-stack website powered by **LLM** models aimed at assisting employees in their job search competition, utilizing **Django** framework, and **MERN** stack (MongoDB, Express.js, React, Node.js)
- Built **Express** and **Django** microservices with serverless **REST API**. The Django implemented with **PEP8** code style
- Architected a React frontend with **Redux**, **Google Authentication** and **Stripe** API to forbid 100% unauthorized access and enhance data status management
- Configured **Nginx** and **Certbot** with **SSL-certification** to facilitate communications between different services
- Containerized the entire system using **Docker** and developed **shell scripts** to automate installation and deployment
- Deployed the whole system on the **AWS EC2** service and incorporated a domain name using **GoDaddy** and **Route 53**

Multimodal QA system | *LLM, Transformer, CNN, Decoder* **Nov. 2024**

- Reproduced and analyzed the **multimodal vision-language model**, exploring its architecture and understanding the integration of visual and textual components for vision-language tasks.
- Fine-tuned the model on question answering benchmark **VQAv2** and **ScienceQA**, achieving state-of-the-art results across various benchmarks.
- Implemented and tested the multimodal pre-training process for PaliGemma, experimenting with different image resolutions (224px, 448px, and 896px) to optimize performance for high-resolution image tasks.

Predicting Effective Arguments | *RNN, GRU, LSTM, Bert, Pytorch, Matplotlib* **Mar. 2023**

- Preprocess datasets into Torchtext data iterators, including steps such as **lowercasing**, **removing stopwords**, performing **stemming** and **lemmatization**, and handling **out-of-vocabulary (OOV)** words.
- Designed and trained various neural network models, including **RNN**, **LSTM**, **GRU**, and **BERT**, to classify argumentative elements within essays written by U.S. students.
- Conducted an ablation study combined with a grid search to systematically evaluate model performance, achieving the lowest loss and highest accuracy across different architectures.

Technical Skills

Programming: Python (advanced), Java, SQL, Shell, Go
Deep Learning: PyTorch, TensorFlow, Distributed Training (DDP), Mixed Precision (AMP), Gradient Optimization, Multi-Task Learning, FlashAttention
LLM Engineering: Transformers (HuggingFace), Fine-tuning (LoRA/PEFT), RAG Systems, Embedding Models, Prompt Optimization, Evaluation Frameworks, Multimodal Systems
Data & Modeling: Large-scale data preprocessing, Feature Engineering, Vector Indexing, Clustering
Systems & Infrastructure: Docker, AWS (EC2), Linux, GPU Training, Experiment Tracking, Git