Huanmi TAN

huanmi.tan@gmail.com | (412)-860-1066 | LinkedIn | huanmit.github.io

Education

Carnegie Mellon University, School of Computer Science Pittsburgh, PA
Master of Software Engineering, Scalable Systems, GPA: 3.95/4.0 *Dec 2024*

Tongji University, School of Software Engineering Shanghai, China
Bachelor of Engineering in Software Engineering, GPA: 90/100 *July 2023*

Skills

Programming Languages: Python, C++/C, R, SQL, Swift, Java
Frameworks: PyTorch, HuggingFace, LangChain, Flask, Django, Django Rest Framework**,** Vue, React
Others: Git, Kubernetes, Slurm, Unix/Linux, LaTeX, GCP, AWS EC2, Azure, CI/CD, Docker, Jenkins, Wandb

Professional Experience

ByteDance Shanghai, China
PM Intern | Volcano Machine Translation, NLP, AI Lab *March 2023–Aug 2023*

- Architected and deployed a full-stack translation quality assessment platform integrating BLEU and COMET metrics, reducing manual evaluation time by 70%
- Spearheaded VolcTrans 2.4.0 plugin development, adding domain translation and personal glossary functions, coordinated a team of 6, reaching over **110,000** users across Chrome and Edge extension stores with 4.2/5 user rating
- Curated 100+ rounds of dialog for supervised fine-tuning based on selected user data for Douyin (Chinese TikTok) Xiao'an Caring bot, the first LLM product at ByteDance

Shanghai AI Laboratory Shanghai, China
Research Intern | Neuromorphic Computing, AI for Imaging Group *Nov 2022–Feb 2023*

- Contributed to enhance neural network robustness on memristor-based hardware—crucial for energy-efficient AI systems
- Conducted 50+ Bayesian optimization experiments, tuning network-wide dropout rates under real-world constraints to identify resilience factors and developing a dynamic training methodology that mitigated memristor weight drift
- Validated the approach on 10 architectures across varied Gaussian noise levels, consistently outperforming four baseline methods with reduced weight drifting and improved efficiency/accuracy

Selected Projects

Onnxpected: A CLI Tool for Unified ML Training & Evaluation Pipeline | CMU, Master Capstone Aug 2024 – Dec 2024

- Developed a standardized CLI integrating YOLOX, SuperGradients, NVIDIA TAO, and TensorFlow for training, evaluation, and ONNX export—serving a team of 10 ML engineers and streamlining workflows for our client partner
- Designed an extensible architecture accommodating new ML repositories with minimal integration overhead
- Integrated MLflow for experiment tracking, enabling real-time visualization and model comparisons via a web UI
- Managed multi-repo setup and dependencies with Docker to eliminate version conflicts and boost development efficiency

Self-Boost: Boosting LLMs with Iterative Self-Generated Data | CMU, 11711 Research Project *March 2024–June 2024*

- Pioneered Self-Boost, an iterative data augmentation framework for low-data fine-tuning, generating high-quality data pairs to expand the SFT dataset and boost model performance
- Engineered a fine-tuning pipeline to filter erroneous predictions, then self-generate and self-verify new examples—expanding the training set by 96–300% and enhancing model accuracy through continual feedback loops
- Conducted experiments and ablation studies using LoRA fine-tuned Llama3-8b-Instruct on GSM8K and TREC, boosting test accuracy by up to 21.6% in low-data scenarios

End-to-end RAG-based QA System | CMU, 11711 Coursework *Feb 2024–March 2024*

- Built an end-to-end RAG pipeline with LangChain, integrating a multi-source knowledge base—including 24 webpages and PDFs via Selenium and vector storage, and achieving a 2-second query response time
- Curated a dataset of 300 question–answer–context pairs and fine-tuned Llama2-13b with QLoRA, reaching a 0.56 F1 score with paraphrase-mpnet-base-v2 embeddings and highlighting opportunities for further QA enhancements

Individual Carbon Credits Assessment Platform | Citi Cup Fintech Contest [Backend, DRF] *Feb 2022–June 2022*

- Led an 8-person remote tech team to develop an AI-powered carbon credit assessment platform, enabling banks to evaluate individuals' daily behaviors for determining loan quotas
- Implemented 39 RESTful APIs with Django Rest Framework, integrating voice assistant and image search, enabling seamless multimodal user interactions
- Achieved top 5 ranking among ~150 teams in China, demonstrating the effectiveness and innovation of the solution