Xinyuan Tong

(323)791-9928 | xinyuantong0323@gmail.com | justintong0323.github.io

Education

University of Southern California — Los Angeles, CA, USA
Bachelor of Science in Computer Science, Exchange Student — *Aug 2024 – May 2025*

University of Edinburgh — Edinburgh, UK
Bachelor of Science in Computer Science — *Sep 2022 – Jun 2026*

- GPA: 3.9/4.0
- Highlighted Courses: Operating Systems, Introduction to Machine Learning, Algorithms and Data Structures, Computer Systems, Data Science, Software Engineering

Projects

ServerlessLLM | *Python, C, Ray, Docker, Git* — Jun 2024 – Present

- One of the main developers of ServerlessLLM, an open-source serving system designed for affordable multi-LLM deployment, optimizing for environments with limited GPU resources
- Implemented a distributed profiling component for ray workers
- Built and containerized the project using Docker to simplify deployment processes across various platforms
- Improved the auto-scaling component, enabling elastic scaling of model instances and efficient GPU multiplexing
- Developed a command-line interface and comprehensive tests to ensure reliability and ease of use

SER using Self-Supervised Learning and LLM | *Python, PyTorch, scikit-learn* — Sep 2024 – Dec 2024

- Developed a state-of-the-art Speech Emotion Recognition (SER) system by transitioning from traditional ML methods to fine-tuning self-supervised models.
- Fine-tuned the cross-lingually pre-trained model to achieve SOTA performance.
- Implemented extensive data augmentation and hyperparameter optimization techniques to enhance model robustness and generalization.

Virtual Memory & Cache Simulator | *C, Assembly, Git* — Sep 2023 – Dec 2023

- Engineered a C-based simulator that merges cache systems with virtual memory management, featuring TLB and Page Tables, to accurately simulate address translation from virtual to physical
- Introduced adjustable settings for cache sizes, TLB entries, and page replacement methods, offering the ability to mimic different computing environments. This adaptability is key for analyzing how system performance varies with configuration changes
- Established comprehensive error handling to detect, report, and resolve simulation issues, ensuring the simulator's reliability and producing precise outcomes while enhancing the user experience by minimizing disruptions

Experience

Undergraduate Research Assistant — Jun 2024 – Present
University of Edinburgh — *Edinburgh, UK*

- Collected and preprocessed over 100 research data samples from online sources, including data cleaning and classification using Python
- Annotated more than 20 data samples and utilized AI tools to establish a standardized data annotation workflow, enhancing efficiency by 30%

Teaching Assistant — Sep 2023 – Dec 2023
University of Edinburgh — *Edinburgh, UK*

- Guided over 50 students, simplified complex concepts for enhanced comprehension
- Adapted teaching strategies based on student feedback, improving the clarity and effectiveness of instruction
- Contributed to students' ability to understand and apply computation principles effectively, as evidenced by positive feedback and improved course performance metrics

Technical Skills

Languages: Python, Java, C/C++, Shell script, C#, Haskell, Assembly
Tools: Linux, kubernetes, Git, GitHub, Docker, Huggingface, GCP, LaTeX
Libraries: PyTorch, TensorFlow, Ray, scikit-learn, pandas, NumPy, Matplotlib