DAVID (JUNDA) SU

js202@rice.edu | (832) 341-3296 | davids048.github.io | linkedin.com/in/david-su-257124228

EDUCATION

Rice University **Houston, TX**

Bachelor of Science in Computer Science *Expected Graduation: May 2025*

- **GPA:** 3.98/4.00
- **Honors**: President's Honor Roll, Rice University (2022, 2023, 2024)

RESEARCH EXPERIENCE

Rice University **Houston, TX**

Research Assistant *Dec 2023 – Present*

- Proposed SpartanServe, a system designed for concurrent LLM serving using multiple structurally sparse adapters
- Developed a unified matrix multiplication operation and memory management technique that enables efficient batching
- Applied Triton kernels and CUDA graphs to further accelerate matrix multiplication in concurrent LLM serving
- Achieved 2.12x speedup over S-LoRA when serving 96 adapters using a single NVIDIA A100 GPU (40GB)
- Authored paper "In Defense of Structural Sparse Adapters for Concurrent LLM Serving," accepted to ES-FoMo'2024

Rice University **Houston, TX**

Research Assistant *Aug 2023 – Oct 2023*

- Contributed to the development of a CNN + BiLSTM model for arrhythmia classification using real-world ECG data
- Trained and benchmarked a ResNet18 model against the proposed model using the MIT-BIH arrhythmia database
- Demonstrated superior performance compared to existing baselines on proprietary dataset, achieving an average accuracy of 95% for binary classification and 88% for multi-label classification
- Co-authored a paper submitted to the Digital Health journal, currently under full consideration for publication

Baylor College of Medicine **Houston, TX**

Research Assistant *Aug 2022 – Dec 2022*

- Developed a sequence-sampling API for a whole-genome DNA methylation analysis software in a team of four
- Implemented a resampling algorithm using NumPy, improving selection efficiency of target DNA region by 2 times
- Visualized discovered DNA regions and created summaries of resampling results using Pandas and Matplotlib
- Used parallel programming on a Linux cluster server to improve API efficiency, allowing 20x data processing speedup

PUBLICATION & MANUSCRIPT

- **Junda Su**, Zirui Liu, Zeju Qiu, Weiyang Liu, Zhaozhuo Xu. "In Defense of Structural Sparse Adapters for Concurrent LLM Serving" *In submission to EMNLP'2024. Accepted in ES-FOMO at ICML'24)* [paper] [poster]
- Guangyao Zheng, Sunghan Lee, Jeonghwan Koh, Khushbu Pahwa, Haoran Li, Zicheng Xu, Haiming Sun, **Junda Su,** Sung Pil Cho, Sung Il Im, In cheol Jeong, Vladimir Braverman. "Hierarchical Deep Learning for Autonomous Multi-label Arrhythmia Detection and Classification on Real-world Wearable ECG Data" *In submission to Digital Health*

PROFESSIONAL EXPERIENCE

Tokio Marine HCC **Houston, TX**

Technology Advancement Program Intern *May 2023 – Aug 2023*

- Designing and developing quote submission and retrieval APIs for an insurance website, implementing RESTful architecture to ensure scalability and flexibility
- Employed AWS API Gateway for traffic scaling and Mongo DB, AWS, and PostgreSQL for data management
- Led daily standup meeting and biweekly sprint planning; represented the team in company-wide demo sessions
- Designed and wrote specific documentation to help developers quickly and effectively use our tools

TEACHING EXPERIENCE

Rice University **Houston, TX**

Teaching Assistant

- COMP 321: Introductions to Computer Systems *Jan 2024 – May 2024*
- COMP 382: Reasoning about Algorithms *Aug 2023 – Dec 2024*
- COMP 182: Algorithmic Thinking *Jan 2023 – May 2023*

PROJECT EXPERIENCE

LLM Finetuning Project **Houston, TX**

Team Member *Jan 2024 – May 2024*

- Evaluated Huggingface parameter-efficient fine-tuning methods for aligning LLMs such as Falcon, Gemma, and Phi-2.
- Investigated the impact of different 4-bit quantization schemes on fine-tuning LLMs for NLP tasks.
- Demonstrated that fine-tuning smaller LLMs (under 3 billion parameters) can achieve comparable performance to larger LLMs (around 7 billion parameters) such as Llama2-7B on domain-specific tasks.

NoSQL Document Database Project **Houston, TX**

Team Member *Aug 2023 – Oct 2023*

- Used Golang to create a network accessible NoSQL document database in a team of three
- Implemented RESTful web services to allow concurrent database queries, updates, and subscription
- Implemented robust data synchronization mechanisms, achieving strong reliability in a distributed system
- Utilized advanced database indexing and query optimization techniques to improve query response times by 30%

SKILLS

- **Programming Languages**: Python, Java, JavaScript, Golang, C, C++, C#
- **Tools**: PyTorch, Triton-lang, CUDA, Hugging Face, Git
- **Frameworks**: .Net, React, HTML, CSS, GraphQL, MongoDB, AWS, SQL
- **Skills**: Machine Learning Systems, Natural Language Processing, LLM, Deep Learning