

# JUNDA (DAVID) 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* Aug 2021 - 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* Aug 2024 - Present

- Proposed SpaLLM, a new fine-tuning paradigm on compressed LLM models using parameter-sharing algorithms
- Designed and executed comprehensive benchmarks alongside LLM-as-a-judge evaluations, testing SpaLLM across a diverse range of models, including LLaMA-2-7B, 13B, LLaMA-3-8B, and 70B, showcasing SpaLLM's adaptability
- Achieved favorable accuracy and up to 3x inference speedup than SOTA adapter-based compressive fine-tuning methods
- Authored paper "SpaLLM: Unified Compressive Adaptation of Large Language Models with Sketching"

**Rice University** Houston, TX  
*Research Assistant* Dec 2023 – Jun 2024

- Proposed SpartanServe, a system designed for fast concurrent LLM adapter serving using 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"

**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
- 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

- Tianyi Zhang\*, Junda Su\*, Oscar Wu, Zhaozhao Xu, Anshumali Shrivastava. "SpaLLM: Unified Compressive Adaptation of Large Language Models with Sketching" *In submission to ICLR'2025*
- Junda Su, Zirui Liu, Zeju Qiu, Weiyang Liu, Zhaozhao Xu. "In Defense of Structural Sparse Adapters for Concurrent LLM Serving" *Accepted in EMNLP'2024 findings. Presented 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" *Accepted in Digital Health*

## PROFESSIONAL EXPERIENCE

---

### Tokio Marine HCC

Technology Advancement Program Intern

Houston, TX

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

Teaching Assistant

Houston, TX

- COMP 318: Concurrent Program Design Aug 2024 – Present
- 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

Team Member

Houston, TX

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

Team Member

Houston, TX

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