Jiawei (Allen) Zhu

Mobile: (615) 974-6888 | **E-Mail:** jiaweizh@andrew.cmu.edu | **LinkedIn:** linkedin.com/in/allenjwzhu724/ | **Github:** AJWZhu

EDUCATION

Carnegie Mellon University *Expected Dec 2025*
Master of Science, Computer Systems (Information Networking), GPA: 3.65/4.0 – ML Systems & Infra *Pittsburgh, PA*

Vanderbilt University *Aug 2019 – May 2023*
Honors Bachelor of Science, Computer Science, minor in Applied Mathematics, GPA: 3.72/4.0 *Nashville, TN*

WORK EXPERIENCE

Meta (Facebook) *Expected May 2025 – Aug 2025*
Incoming Software Engineer Intern – ML Infrastructure *Menlo Park, CA*

- Will be interning at Meta during Summer 2025 as a Software Engineer Intern, expected project on ML Infra and Gen AI Infra.

Ericsson *Jun 2024 – Aug 2024*
Machine Learning Engineer Intern – ML Training Data *Beijing, China*

- **Root Cause Analysis and Anomaly Detection**: Increased false return identification accuracy by 9.1% through leveraging GBDT and XGBoost to build baseline for anomaly detection, collaborated with 10+ teams to reduce return rates for 5G Radio products.
- **LLM Parameter Optimization**: Achieved significant performance improvements in model deployment efficiency by 5% by building parameter-efficient fine-tuned open-source LLMs (7B/13B) using LoRA and QLoRA for optimization.
- **Big Data Pipeline and Feature Engineering**: Scaled feature engineering to handle TB-level data, streamlining production using PySpark for parsing, integrating with SageMaker and S3 to develop ETL pipelines to process 1M+ unstructured telemetry data.

SenseTime *Apr 2024 – Jun 2024*
Software Engineer Intern – LLM Application Development *Beijing, China*

- **LLM Quantization Optimization:** Deployed state-of-the-art large language models using ONNX and TensorRT, applying post-training quantization with SmoothQuant, achieving a 53% increase in model speed performance by enabling INT8 inference.
- **ML Model Deployment and Feature Development**: Deployed SenseTime foundational models in a high-concurrent distributed service environment, contributing over 30 PRs and merging 1500+ lines of code in C++ and Go.
- **Distributed System Scalability** : Utilized gRPC, multi-threading, and goroutines channel workflows to handle over 200,000+å daily requests and more than 100,000 active users, ensuring scalability and performance in a demanding production environment.

Vanderbilt University *May 2021 – Oct 2022*
Machine Learning Research Assistant – Deep Learning for Large Scale Image Processing *Nashville, TN*

- **Image Processing Algorithm Development**: Developed and fine-tuned image processing algorithms for TB-scale high-density imaging data, applying NN-Gaussian Processing and Otsu's method to improve accuracy on CircleNet by 9.89%.
- **Large Dataset Optimization and Performance Enhancement**: Implemented inference optimization through PyTorch and ONNX, achieving a 15% increase in processing speed for large-scale datasets, significantly enhancing model efficiency and scalability.

IFLYTEK *May 2020 – Jul 2020*
Software Engineer Intern – NLP Backend Platform Development *Hefei, China*

- **API Integration and Optimization**: Engineered integrations for voice recognition APIs using JSON-RPC in C++, improving client-server response time by 2 seconds for insurance sector clients in a cloud-based system.

PROJECTS

LlamaInfer: High-Performance Large Language Model Inference Engine – (HPC, ML System) *May 2024 – Present*

- **Inference Engine Development**: Architected a CUDA C++ accelerated LLaMA inference engine with custom memory management, achieving a processing speed of 60.34 tokens/s for LLaMA (1.1B) on NVIDIA RTX 3060 Laptop GPU.
- **Memory Optimization and Efficient Operators**: Developed critical operators including MatMul, LayerNorm, RMSNorm and attention mechanisms, implemented KV-Caching, reducing memory overhead and improving inference efficiency.

Needle Machine Learning Framework – (ML System) *Sep 2024 – Present*

- **ML Training/Inference Framework Development**: Developed a custom ML framework using Python and C++, featuring auto-differentiation via computational graph for efficient model training and optimization.
- **LLaMA2 Model Deployment with Optimized Inference**: Deployed the LLaMA2 language model within the framework, incorporating speculative decoding techniques to achieve faster and more optimized inference.

BERT-TensorRT Inference Optimization – (Inference Optimization) *Jul 2024 – Present*

- **Inference Optimization Documentation**: Authored a comprehensive guide on inference optimization with TensorRT, detailing techniques such as FP16/INT8 quantization and operator fusion to enhance model efficiency and reduce latency.

SKILLS

Languages: Python, C/C++, CUDA, SQL
Frameworks: PyTorch, TensorFlow
Platforms/Tools: AWS, GCP, Docker, Kubernetes, TensorRT, Hadoop, Spark, ONNX, Linux, Git
Courses: Machine Learning [Python], Generative AI [Python], Deep Learning Systems [Python, C++, CUDA], LLM Agents [Python], Cloud Infrastructure [AWS, GCP], Parallel Programming [C++, CUDA], Nonlinear Optimization [R], Computer Systems [C], Distributed Systems [Go], Computer Networks [C], Operating Systems [C], Big Data [Python], Data Structures [C++], Algorithms [C++]