Srinath Naik Ajmeera

ajmeerasrinath@gmail.com | +12133576173 | linkedin.com/in/srinath-naik-ajmeera-805784197

EDUCATION

Master's in Computer Science | GPA 4.00/4 2023 — University of California Los Angeles (UCLA), Los Angeles, CA

- Graduate Teaching Assistant for COM SCI 31, 32 – Introduction to Computer Science I, II (C++)
- Key Courses: Machine Learning Algorithms, Deep Learning, Distributed ML, Computational Robotics

Bachelor's in Computer Science & Engineering | GPA 8.26/10 2018 — Indian Institute of Technology Bombay (IITB), Mumbai, India

SKILLS

Frameworks: PyTorch, TensorFlow, OpenCV, Android, CUDA

Libraries: Python, Scikit, NLTK, Numpy, Pandas, Keras, R, Spark, AutoTokenizer

Languages: Python, C++, Java, Go, Typescript, C

AWS Tools: Lambda, Kubernetes, KMS, EC2, ECS, EKS, AutoScaling, Step Functions, SQS, SageMaker, ECR

Inference & Serving: Triton Inference Server, TensorRT-LLM, vLLM, ONNX, Custom Triton Backends, Constrained Decoding, KV-cache, Quantization, Tensor Parallelism

LLMs & ML: Tokenizer, LLMPerf, RoPE, Fine-tuning, Search Re-ranking, BERT, YOLO

EXPERIENCE

Moveworks

ML Engineer

June 2025 – Present | Mountain View, CA

- Designed and built an end-to-end high-performance inference stack for production LLMs and NLP models using **Triton Inference Server**, enabling scalable serving across translation, scoring, and completions workloads.
- Integrated backends including **ONNX, TensorRT-LLM, and vLLM**; wrote custom Triton backends for tokenization and constrained decoding, reducing latency by **90%** for translation models.
- Ran systematic server-side experiments on batch sizes and model configs to identify optimal serving configurations, improving throughput and reducing cost per query.
- Benchmarked standalone vLLM vs vLLM-behind-Triton; designed Triton integration into an in-house **LLM Gateway** (akin to OpenRouter/LiteLLM), enabling upstream apps to switch models with minimal code changes.
- Fine-tuned BERT-based classifiers on proprietary data for **search re-ranking**, improving relevance quality and ranking precision across search surfaces.

Amazon

Software Engineer – AI/ML

April 2023 – May 2025 | Seattle, WA

AWS Bedrock

- Built AWS Bedrock Custom Model Import infrastructure for hosting and inference of LLMs (Llama 2/3, Mistral, FLAN-T5 and custom models), supporting **300+ models** in production.
- Applied quantization, tensor parallelism, speculative decoding, and KV-cache streaming to improve latency by **1.8x** on long-context models.
- Designed a container patching pipeline to update active customer models with the latest vLLM engine changes, achieving **100% automation** and faster release cycles.
- Developed custom fine-tuning solutions for foundation models on user data with selectable hyperparameters; integrated guardrails and benchmarked in-house LLMs using LLMPerf.

Amazon One

- Developed core image processing software for Amazon One palm-recognition devices (**20,000+ devices globally**), handling cropping, validation, analysis, and formatting at 15 FPS.
- Built a 1P YOLO-v7–based barcode cropping model trained on custom data, reducing rejections by **~82%**.

UCLA – Visual Machines Group

Graduate Student Researcher

Spring 2022 | Los Angeles, CA

- Implemented a modular neural network for snow removal from images using dilated convolutions; trained on 20K images from Snow100K dataset achieving **82.3% test accuracy**.
- Devised a physics-guided neural network POC combining physics loss and regression loss for 2D projectile motion training.

Apple

Software Engineer – Identity Management Services

June 2018 – Feb 2020 | Hyderabad, India

- Managed the Registration, Access Management & Provisioning (RAMP) platform overseeing critical applications for access control across Apple's suite of applications.
- Collaborated with business teams to design and implement new features within RAMP, ensuring seamless UX and optimal functionality.