Srinath Naik Ajmeera

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

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

Inference & Serving: Triton Inference Server, vLLM, TensorRT-LLM, ONNX Runtime, Custom Triton Backends, KV-cache-aware Routing (llm-d / GAIE / EPP), Dynamic & Continuous Batching, Prefill/Decode Disaggregation, Constrained Decoding, Quantization (FP8/FP16/INT4), Speculative Decoding, Tensor Parallelism, Model-name Routing

Training & Fine-tuning: SFT, LoRA/PEFT, FSDP, Knowledge & Chain-of-Thought (CoT) Distillation, Rejection-Sampling Fine-tuning (RSFT/STaR), Listwise Ranking Losses (ListNet/ListMLE/KPO), Curriculum Learning, Mid-training, Reranker Fine-tuning

Agentic AI & Evaluation: Agentic Reasoning, Tool-use / Function Calling, Multi-turn Eval, Inspect AI, τ^2-bench, BFCL, BEIR/BRIGHT/MIRACL, LLM-as-judge, LLMPerf, Weights & Biases, MLflow

LLM Gateway & Platform: Provider-neutral LLM Gateway (OpenRouter/LiteLLM-style), Multi-provider Integration (Gemini, Azure OpenAI, GPT-5.x), Distributed Tracing & Metrics

Infra & MLOps: Kubernetes, Istio, Helm, AWS (SageMaker, Step Functions, Bedrock, EventBridge, ECR), Argo CD / Argo Rollouts, Prometheus / VictoriaMetrics, Bazel, Docker, Snowflake

Languages & Frameworks: Python, Go, C++, Java, PyTorch, TensorFlow, Hugging Face Transformers, CUDA (NIXL/UCX), Numpy, Pandas, SQL

Models: Qwen, Llama, Gemma, mBART, M2M100, FLAN-T5, BERT, YOLO, GPT-4o / GPT-5.x

EXPERIENCE

Moveworks (acquired by ServiceNow)

Senior Machine Learning Engineer June 2025 – Present | Mountain View, CA

- Architected an end-to-end high-performance inference stack for production LLMs and NLP models on **Triton Inference Server** (ONNX Runtime, TensorRT-LLM, vLLM backends) with custom Triton backends for tokenization, multilingual support, and constrained decoding — cutting translation latency **~90%** and scaling the legacy fleet to zero.
- Led design and phased production rollout of **KV-cache-aware inference routing** (llm-d / Gateway API Inference Extension) across self-hosted vLLM fleets with per-org canary and instant rollback — lifting prefix-cache hit rate **33%→62%**, throughput **~54%**, and tail latency up to **~10×** under load, validated on a 13.6M-request production analysis.
- Designed a reference-guided **chain-of-thought distillation** pipeline (rejection-sampling / STaR-style) and a top-K-truncated listwise ranking loss (KPO) to train reasoning-capable rerankers — reaching **83% data yield**, cutting prompt leakage 13%→0.4%, and gaining **+0.65 NDCG@10** on BEIR at the measured K=5 optimum.
- Driving training of an **open-weight core reasoning model** to match a proprietary frontier model on agentic reasoning at lower cost and latency (mid-training, RSFT, curriculum learning) — plus a reproducible AWS agentic-evaluation pipeline (Inspect AI, ephemeral SageMaker serving, Step Functions) benchmarking any checkpoint on **τ^2-bench and BFCL** with one-command cost/latency-parity evaluation.
- Integrated external providers (Google Gemini, Azure OpenAI) into the gateway and led a **zero-downtime cutover of GPT-5.x traffic** through an mTLS-secured managed proxy with a provider-agnostic error taxonomy — unlocking new model families and cutting long-context tail latency up to **~45%** vs. direct-cloud calls at ~5–10 ms added overhead.
- Drove serving-optimization experiments (dynamic batching, model/config sweeps; vLLM vs. vLLM-behind-Triton) exposing an ~8–10% throughput regression from the Triton wrapper, then architected a provider-neutral **LLM Gateway** (OpenRouter/LiteLLM-style) with model-name routing — migrating **100%** of callers off the legacy interface for zero-code model/provider switching.

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 **1.8×** on long-context models.
- Built a container-patching pipeline to roll active customer models onto the latest vLLM engine (**100% automation**); developed custom fine-tuning for foundation models with selectable hyperparameters, integrated guardrails, and benchmarked in-house LLMs with 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 barcode-cropping model reducing rejections by **~82%**.

Apple

Software Engineer – Identity Management Services June 2018 – Feb 2020 | Hyderabad, India

- Owned the Registration, Access Management & Provisioning (RAMP) platform for access control across Apple's application suite; designed and shipped new features with business teams.