VARAD VITTHAL MORE

+1 (623) 806-5756 • Sunnyvale, CA • varadmore1@gmail.com • linkedin.com/in/varad-more • github.com/varad-more • varadmore.me • Publications

SUMMARY

Software engineer with 3+ years building LLM and machine learning systems, backend services, and cloud infrastructure, now an Applied AI Engineer at Retell AI. Merged vLLM Metal contributor and author of Triton GPU kernels and LLM benchmarks. Co-founded a conversational AI company with 1,500+ daily users across 87 enterprise clients before its 2022 acquisition, then scaled the acquirer's AWS sync engine to 5,000+ concurrent executions at 99.9% uptime. M.S. CS at ASU (3.9 GPA), AWS Community Builder (2026).

TECHNICAL SKILLS

LLM Inference & ML Systems: PyTorch, CUDA, Triton, vLLM, SGLang, TGI, llama.cpp, NIXL, FlashAttention, paged attention, KV cache optimization, speculative decoding, quantization, GPU profiling (Nsight), MLX

Applied AI & Agents: LLM agents, function calling, prompt engineering, RAG, pgvector, vector search, LLM evaluation (TruLens), voice AI, ASR, TTS, conversational AI, NLU, OpenAI API, Anthropic API, Gemini, AWS Bedrock, LangChain

Languages & Backend: Python, C/C++, SQL, TypeScript, JavaScript, Java, Bash, FastAPI, Django, Flask, React, Next.js, Node.js, REST APIs, GraphQL, gRPC, WebSocket, SSE, OAuth 2.0, microservices, distributed systems, pytest

Cloud, Data & Observability: GCP, AWS (Lambda, EC2, S3, SQS, DynamoDB, Glue, Athena, SageMaker), serverless, Terraform, Docker, Kubernetes, Kafka, Redis, CI/CD, PostgreSQL, Spark, Prometheus, Grafana

EXPERIENCE

Applied AI Engineer: Retell AI, CA, USA Sep 2026 – Present

- Built a call replay simulator that measures ASR transcription accuracy on an enterprise insurance account's calls, which Retell leadership approved sharing with the customer.
- Wrote the Salesforce function calls a Retell voice agent makes during a call, testing each one in isolation before wiring it in.
- Prototyped IVR agents and diagnosed voice agent failures across **4+** enterprise accounts in insurance, healthcare, and higher education, turning recurring ones into scoped feature requests for the product team.

Founding Software Engineer: Sharp Data Analytics, Inc., CA, USA Sep 2022 – Jun 2024

- Scaled the sync engine to **5,000+** concurrent executions at **99.9%** uptime by splitting a monolith into serverless microservices on AWS with Terraform, SQS, and EC2 Auto Scaling.
- Built **7+** integrations across Twilio, Salesforce, Zapier, and the OpenAI API with rate limiting and exponential backoff, syncing **100K+** records a day into customer CRMs and AI workflows.
- Cut production defects **78%** with CI/CD on AWS CodeDeploy and CircleCI plus automated test suites, and reduced MTTR **42%** after adding Grafana dashboards on Prometheus with distributed tracing.

Co-Founder & Software Engineer: Botonomics Automations LLP (Acquired), Mumbai, India Feb 2021 – Sep 2022

- Ran engineering as one of two technical founders and shipped TeeDee, a conversational AI platform that reached **1,500+** daily active users across **87** enterprise clients, automated **82%** of support workflows, and cut client support costs **64%**.
- Cut cost per query **46%** with a hybrid NLU router that paired pgvector similarity matching with deterministic rules and sent simple intents down a fast path that skipped the LLM call at **<100ms** p95.
- Designed a FastAPI microservices backend with tenant isolation on GCP and AWS that held **99.8%** reliability.

OPEN SOURCE CONTRIBUTIONS

vLLM Metal Backend (MLX on Apple Silicon, 1.9K+ stars) github.com/vllm-project/vllm-metal (PR #550, #549 merged)

- Unblocked GPT-OSS 20B serving on Apple Silicon by wiring attention sinks through the Metal paged attention kernel and fixing a warp merge defect in the softmax denominator, guarded by a **389** line parity suite across fp32, fp16, and bf16.
- Fixed a circular import that kept a tokenizer patch from installing and leaked byte level artifacts into served Qwen text.

PROJECTS

KV Transfer Tax: Disaggregated Prefill/Decode LLM Serving (vLLM + NIXL) github.com/varad-more/kv-transfer-tax

- Built a disaggregation testbed on **4**× A10G with a routing proxy and goodput harness (**17** configs, **79** tests). Without NVLink, PCIe KV transfer near **7.4 GB/s** kept 1P3D from meeting the SLO, while **4** data parallel replicas sustained **0.91** req/s.

Triton GPU Kernel Lab for LLM Inference github.com/varad-more/triton-kernel-lab-for-llm-inference

- Wrote **8** Triton kernels (**58** tests): FlashAttention-2 matched SDPA at **61.8** TFLOP/s and RoPE ran **11.9**× faster than eager PyTorch.

Selfhosted Chat API: OpenAI and Anthropic Compatible LLM Gateway github.com/varad-more/selfhosted-chat-api

- Built a FastAPI gateway that serves OpenAI and Anthropic compatible APIs, including translated SSE streaming, over **7** open source runtimes such as vLLM, SGLang, and Ollama, with API key auth, token bucket rate limiting, Prometheus metrics, and CI.

Affordability Index (live): Salary Needed in Every US County github.com/varad-more/affordability-index

- Shipped a live tool that inverts a federal and state tax engine to find the gross salary each of **3,123** US counties needs for the MIT living wage, 70/20/10, or 50/30/20 budget, served as static files and covered by **619** tests.

EDUCATION

M.S. Computer Science | Arizona State University, Tempe, AZ | 3.9/4.0 GPA Aug 2024 – May 2026

Relevant coursework: Statistical Machine Learning, Artificial Intelligence, Data Mining, Cloud Computing, Data Processing at Scale

B.E. Information Technology | University of Mumbai, Mumbai, India | 3.5/4.0 GPA Jul 2017 – May 2021

2 IEEE papers at ICNTE 2021 · Smart India Hackathon Grand Finalist · eYantra Robotics finalist (IIT Bombay)