MUHAMMED SALIM K T

AI/ML Engineer · LLM, RAG & Speech Systems
mohdsalimkt@gmail.com | +91 62382 41231 | Ernakulam, Kerala, India
github.com/salimkt | linkedin.com/in/muhammed-salim-k-t-498125165
Public work: github.com/salimkt/small-model-agents — agent evaluation harness with CI regression gate; published MCP server
Open to remote | relocation within India | international with sponsorship | Fluent English, Malayalam (native), Hindi, Urdu; Arabic (proficient: reading, comprehension, writing)

SUMMARY

AI/ML engineer, 4 years 4 months, all of it on production LLM systems.

- **Built Echo, a production speech-AI service** transcribing voicemails and call recordings for a US hosted-telephony provider — NVIDIA Canary and Whisper on vLLM with engine failover, Redis priority queues and Celery workers on HA nodes, speaker diarization and LLM summaries.
- **Authored the multi-agent and tool-calling engine that replaced CrewAI** in a live enterprise platform, re-architected so a **12B model does the work the previous stack needed 32B for** — taking each customer deployment from a GPU cluster to a single node.
- **On-premise and air-gapped delivery into enterprise customers' own infrastructure**, through their security review and to HIPAA frameworks. **10,000+ queries/day**, RAG over **10M+ embeddings**, Kubernetes across eight environments including two GPU/DGX clusters.
- Speech (Whisper STT, Tacotron2 TTS, live captioning in production) and document AI (PaddleOCR-VL, Qwen-VL, LayoutLMv3) shipped into the same platform.

TECHNICAL SKILLS

Agentic Systems: Multi-agent orchestration engine design (author) · CrewAI · LangChain · ReAct loops with escalation · structured delegation · schema-validated tool calling · 73-toolset registry and manifest system · tiered model routing · SSE event streaming · Redis-backed admission control
LLM & RAG: RAG pipeline design (Docling + RapidOCR ingestion, chunking, hybrid keyword + semantic retrieval, re-ranking, evaluation harnesses) · Hugging Face Transformers · Sentence-Transformers · LangChain · LlamaIndex · model training and fine-tuning with Unsloth, LoRA / QLoRA and PEFT (incl. Arabic-language LLM fine-tuning) · model quantization · prompt engineering · benchmark and regression-gate design for LLM releases
Model Serving: vLLM · Ollama · on-premise and air-gapped deployment · engine failover and load-balanced inference tiers · batching, caching and autoscaling for inference · latency and cost profiling
Speech AI: NVIDIA Canary · Whisper (STT) on vLLM · speaker diarization (Silero VAD, SpeechBrain ECAPA-TDNN + clustering) · spoken language identification (SpeechBrain) · long-audio chunking · Tacotron2 (TTS) · streaming transcription and live captioning
Vision & Docs: Vision-language models in production (PaddleOCR-VL, Qwen-VL) · LayoutLMv3 fine-tuning · Docling · OCR pipelines (EasyOCR, RapidOCR, PP-OCRv4 ONNX)
Vector & Data: Cassandra Vector DB · PostgreSQL + pgvector · FAISS · Qdrant · Pinecone · MongoDB · Redis · ETL for training-data preparation
Backend: Python (FastAPI, Django, Flask) · PyTorch · Celery (priority queues, worker pools) · REST, WebSocket and event-driven services · RabbitMQ / Kafka · authn/z, rate limiting
Platform: Kubernetes (multi-environment manifests, Kustomize overlays, StatefulSets, PDBs, RBAC, cert-manager TLS, GPU/DGX node targets, in-cluster service networking) · Docker · GCP (Cloud Run, Vertex AI) · GitLab CI · Jenkins · Grafana & Prometheus · model versioning, drift monitoring, A/B testing
Frontend: React.js · React Native (custom turbo modules) · TypeScript

EXPERIENCE

AI/ML Engineer & Full-Stack Developer
Netstratum Technologies Pvt Ltd, Ernakulam, Kerala **June 2022 – Present**

SPEECH AI — ECHO (FLAGSHIP)

- **Built Echo, a production speech-AI service** transcribing voicemails and call recordings for a US hosted-telephony provider: NVIDIA Canary for accuracy-first voicemail and Whisper served on vLLM for call recordings, with automatic engine failover; Redis priority queues (0–99) and Celery workers across stateless HA nodes; FastAPI + Cassandra.
- Engineered the audio pipeline: hash-based duplicate suppression, 45-minute chunking of recordings up to 10 hours with absolute timestamps, automatic language ID, and speaker diarization (Silero VAD per channel for stereo; ECAPA-TDNN embeddings + clustering for mono), plus per-tenant transcript corrections.
- Added Gemma summaries and keywords on a separate Celery queue and LLM load balancer, so a slow LLM never blocks transcription; results delivered by webhook or polling.

LLM & RETRIEVAL SYSTEMS

- Built and shipped **BlueMesh**, an on-premise enterprise AI chat platform in Python/FastAPI, serving **10,000+ queries daily** at sub-2-second response times with no customer data leaving their own infrastructure.
- Designed its **RAG pipeline end to end** — Docling + RapidOCR document ingestion, chunking, Sentence-Transformers embeddings, a Cassandra vector store schema tuned for similarity search, and hybrid keyword-plus-semantic retrieval across **10M+ document embeddings** — improving answer accuracy **45%** over the bare LLM baseline.

- Reduced average inference latency to **1.8s** through model quantization, request batching and response caching, and held **99.7% uptime** using model failover and warm caching.
- Fine-tuned Hugging Face models (BERT, T5, GPT-family) for classification, NER and generation, and built an **LLM response-quality evaluation framework** used to gate releases.
- Implemented semantic search with FAISS and embedding models, cutting document retrieval time **70%**; added PostgreSQL **pgvector** for hybrid search.

AGENTIC SYSTEMS & TOOL CALLING

- Authored **bm-agent-engine**, BlueMesh's proprietary multi-agent and tool-calling engine, as a drop-in replacement for the platform's CrewAI-based orchestration — a versioned internal Python package that now serves all agent traffic in production.
- Removed the CrewAI and LangChain runtime dependency entirely and re-architected orchestration so a **12B model handles the work the previous stack needed a 32B model for**, escalating to a larger planner only for the genuinely hard step — cutting GPU cost and letting the platform run on a single node.
- Built the tool layer: a registry, manifest and catalogue system exposing **73 toolsets** — GitLab, SharePoint / MS Graph, AWS, Azure, Google BigQuery and Drive, Gmail, Jira, Asana, Zoom, Snowflake, Kafka, code execution and Playwright-driven web retrieval — with graceful degradation when an optional driver is absent.
- Implemented ReAct agents with tiered escalation, structured delegate/ask/finish delegation between agents, and a FAST / SMALL / LARGE / THINKING model router over a provider-agnostic layer (vLLM, Ollama, any OpenAI-compatible endpoint).
- Preserved a live **SSE streaming contract** exactly — agent-session, tool-usage and completion event taxonomy — so the existing UI and workflow orchestrator cut over with a single config change and no client rewrite.
- Added Redis-backed concurrency leases for admission control, internal-service auth gating, and hardened container builds under GitLab CI.

SPEECH & REAL-TIME AI

- Built the speech-to-text and text-to-speech stack using Whisper and Tacotron2, reaching **94% transcription accuracy**, and shipped live captioning into a production video-calling product at **92% accuracy**.
- Architected WebSocket and REST services (FastAPI, Node.js) supporting real-time communication for **50,000+ concurrent users**, with event-driven inference workers on RabbitMQ and Kafka.

VISION & DOCUMENT UNDERSTANDING

- Built and owned **BlueMesh Vision**, a FastAPI service serving vision-language models (PaddleOCR-VL, Qwen-VL) through Transformers with CUDA device handling and attention-implementation tuning for throughput — plus image analysis, generation, editing and avatar endpoints.
- Made documents fully searchable in RAG by wiring a pluggable OCR layer — EasyOCR, RapidOCR and PP-OCRv4 ONNX behind a single interface — into the retrieval pipeline, with configurable per-document figure understanding so charts and diagrams become retrievable text rather than dead pixels.
- Shipped on-device detection with MediaPipe Tasks and a quantized TFLite face detector for real-time background blur in the video product, holding mobile inference **under 30ms**.

PLATFORM, DEPLOYMENT & RELIABILITY

- Own the **Kubernetes deployment surface** for the AI platform across **eight environments** — development, test, demo, minikube and two GPU/DGX clusters — as the leading contributor to a Kustomize-layered manifest tree (config → data → apps → edge) covering the RAG service, agent engine, workflow orchestrator and search. Ship new services end to end: Deployment, Service, PVC, ConfigMap and Secret, cert-manager TLS, reverse-proxy edge routing and in-cluster service discovery.
- Also deploy with Docker on GCP Cloud Run with load-based autoscaling, and build GitLab CI pipelines for automated model testing, validation and release.
- Built monitoring dashboards tracking accuracy drift, latency and resource cost, and established model versioning and A/B testing practice for the AI platform.
- Designed **DebugTrail**, an ML-assisted log-analysis framework (Grafana LogQL, Isolation Forest and LSTM anomaly detection) over **5M+ log events daily** — cutting time to detection **60%** and false positives **70%** — and a streaming content-moderation pipeline (distilled BERT at 50ms over Kafka) that cut inappropriate content **85%**.

CUSTOMER & ENTERPRISE DELIVERY

- Built **OfficeSuite Live**, a video-conferencing product, for Windstream Enterprises as a named external client — and shipped **NCS**, an enterprise VCaaS platform, to **50,000+ downloads** across both app stores.
- Delivered the AI platform into enterprise customers' own infrastructure — **on-premise and air-gapped**, with no customer data leaving their environment — and took it through their security review: encryption at rest and in transit, RBAC, audit logging, and **HIPAA** compliance frameworks for regulated data.
- Ran multi-tenant rollouts with tenant-aware routing and dynamic service discovery, including customer-specific demo and test targets, so each deployment stayed isolated.

EDUCATION & CERTIFICATIONS

B.Tech, Computer Science & Engineering — KMEA Engineering College, Aluva, Kerala · 2018–2022 · CGPA 7.53/10

Certifications: Architecting with Google Compute Engine (Google Cloud / Coursera, 6 courses) · Fundamentals of Deep Learning (NVIDIA DLI) · Robotic Process Automation (UiPath)