Mohammed Usmani

+91 86899 08731 | mohammedusmani2005@gmail.com | mohammedusmani.com
Mumbai, India | linkedin.com/in/mohammed-usmani | github.com/mohammed-usmani

Summary

Applied AI engineer building production agentic, retrieval and realtime-voice systems and the backends that carry them (FastAPI, LangGraph, Qdrant/pgvector, OpenAI/Anthropic/Gemini, Deepgram), plus a Go control plane that deploys open-weight LLMs on GPUs.

Experience

AI Engineer Jan 2026 – Present
TechNsure IT Hub Pvt. Ltd. *Remote*

- Own the AI layer of **Garage Connect**, a live relationship-selling platform: social and messaging ingestion, identity resolution, a relationship model, signal detection, product opportunities and an agentic sales copilot.
- Measured **956 production recommendations** and found the relevance score couldn't separate good from bad matches; replaced it with a per-item LLM judge allowed to say nothing fits, cutting forced product pitches from **100% to 8%** and hedging from 83% to 0% on replayed cases.
- Built **EarnGPT**, a 5-iteration tool-calling copilot over a 7-layer context stack (relationship replica, memory slots, contextual retrieval with HyDE, document RAG) that skips raw RAG when the replica is fresh (estimated **~96K → ~2K tokens/turn**); the reranker returns IDs only, so it can't mislabel products.
- Designed and built **DeployMachine** (Go, PostgreSQL, Next.js), which deploys open-weight LLMs with vLLM on NVIDIA and AMD GPUs in the **customer's own cloud** (DigitalOcean; AWS via a cross-account role), with a **fit check** before any GPU is bought. It now serves **three models (27B FP8, 9B, embeddings) on one AMD MI325X** to our production sales-assistant backend. Enabling AMD's **AITER kernels** made the 27B decode **4.2x faster** on 7k-token prompts (24 to 102 tok/s).
- Mirrored the Garage catalog (9 item types) into Qdrant through a **transactional outbox** and HMAC-signed webhooks, skipping re-embedding when an item's text hash is unchanged, with an hourly reconcile and a daily orphan sweep as backstops.
- Shipped a live call copilot (one Deepgram stream per audio source, LangGraph), a crash-recoverable LiveKit note-taker, and turned **OpenClaw**, a single-user CLI agent runtime, into a multi-tenant agent service with per-user OAuth and an MCP server.

Back End Developer Jun 2025 – Sep 2025
Freelance *Remote*

- Built multi-tenant **NestJS/PostgreSQL** backends with RBAC, JWT auth and rate limiting, deployed in Docker behind Nginx, with **Redis caching** on high-traffic endpoints.

Projects

voiceAgent: AI Phone Screener | *Python, LiveKit, LangGraph, FastAPI, PostgreSQL* | GitHub

- Turns a job description into screening criteria a recruiter approves, calls candidates over **SIP with a LiveKit voice agent**, and scores each answer with the candidate's own words as evidence; rejections wait for a recruiter. Logs every turn-taking decision: median reply **2.44 s → 1.22 s** at the agent, stopped for **15/15 interruptions** on 10 test calls, and dropped a faster model after the log showed it returning empty replies.

Clipwright: Fine-Tuned 4B Model for Media Jobs | *PyTorch, Qwen3.5-4B, LoRA SFT, GRPO, llama.cpp* | GitHub

- A 4B model that drives ffmpeg, ImageMagick and SoX from plain English, trained with **LoRA SFT then GRPO**. Built the eval first: **154 tasks** graded by running its commands and checking the output files. **115 → 137/154**, above the 27B teacher's 119; runs on a 16 GB Mac.

Engram: Shared Memory for AI Assistants | *Python, FastAPI, MCP, PostgreSQL, pgvector, mem0* | GitHub

- Self-hosted memory shared by Claude Code, ChatGPT and other assistants over a **22-tool MCP server** with OAuth 2.1. One LLM call extracts typed events, facts and procedures; recall fuses 8 channels with rank fusion into a token-budgeted brief. Picked the extraction model by eval (**gpt-oss-120b, 126/126**).

Technical Skills

AI & Agents: Agentic tool-calling, RAG, LoRA fine-tuning (Unsloth/PEFT, distillation), PyTorch, Hugging Face (Transformers, sentence-transformers), NumPy, Pandas, vLLM, llama.cpp (GGUF), hybrid retrieval (RRF, HyDE, rerank), LangGraph, LangChain, MCP, OpenAI, Anthropic Claude, Gemini, Vertex AI, Qdrant, pgvector, Deepgram, LiveKit
Backend: Python (FastAPI), Go, TypeScript (Node.js, Express, NestJS), Celery, BullMQ, WebSockets, REST, GraphQL, SSE
Databases: PostgreSQL, MongoDB, Redis, Neo4j
DevOps: Docker, GitHub Actions, Nginx, AWS (EC2, S3, Lambda, RDS, CloudFormation), GCP, DigitalOcean, Linux, CI/CD
Frontend: React 18, Next.js 15, TailwindCSS

Education

Viva Institute of Technology Mumbai, India
B.E. Computer Science (AIML) | CGPA: 8.27/10 *Graduated May 2026*

Achievements

Top 17 Nationally, AI For Good Hackathon, Blend360 (Aug 2025); **3rd Prize, Codeathon**, Vidyavardhini College of Engineering (Jun 2024); **3rd Prize**, GNEC International Hackathon – Global NGO Executive Community (Aug 2024)