Muzammil Khan

P: +91 9515515529 | Hyderabad, India | E: muzxmmilkhxn@gmail.com | https://github.com/muzzlol | leetcode

Education

Chaitanya Bharathi Institute of Technology — Hyderabad, Telangana
Bachelor of Engineering in Computer Science, Minor in AIML [CGPA: 9.14] — *Nov 2022 — Apr 2026*

Technical Skills

- **Languages**: Python, TypeScript, Go, Java, Bash, SQL
- **Databases**: PostgreSQL, MySQL, MongoDB, Redis, Neo4j
- **Frameworks**: Next.js, React, Convex, Spring Boot, FastAPI, ExpressJS, Node.js.
- **AIML**: PyTorch, Weaviate, DSPy, LangChain, Transformers, Unsloth, vLLM, MCP, Vertex AI, Kubeflow, Pandas.
- **DevOps & Cloud Infra**: Docker, Kubernetes, AWS (Lambda, EC2, ECS), Cloudflare (Workers, R2, DO), Git, Traefik, Linux, Jest, pytest, Jenkins, GitHub Actions, Agile.
- **Core Competencies**: Machine Learning, RAG, Context Engineering, LLM Fine-Tuning (PEFT/LoRA/GRPO), API Design (REST/gRPC), Microservices, Data Engineering, CI/CD.

Work Experience

AI Engineering Intern — July 2025 — Oct 2025
Grass Labs PTE Ltd. — *Singapore (Remote)*

- Led technical evaluation of LLM memory systems (Mem0, Zep, Supermemory, MemOS…); benchmarked retrieval recall@k, latency (p50/p99), and semantic drift over multi-turn conversations — identified gaps in session persistence for our agent use cases.
- Architected a RAG enabled AI Agent with memory & websearch tooling — built API layer for memory CRUD and an observability UI for memory inspection, trace visualization, and debugging.
- Improved **token cache hits** by **80%**, **tool call accuracy** by **70%** through context engineering optimizations: prompt caching, constrained decoding via token logit masking, and DSPy prompt compilers on top of evals.
- Deployed embedding model on serverless GPU nodes using **vLLM**; built scalable vector store service on **AWS ECS**.

Projects

Spectra | live | A multiplayer game w/ a live viewing enviornment for drawing, coding and typing tests. *Personal Project*
Typescript, Cloudflare workers, DurableObjects, Convex, Tanstack start, React, Websockets, transformers

- Orchestrated real-time multiplayer sync via Cloudflare Durable Objects as single-source-of-truth with WebSocket fan-out; event-driven persistence flushing to DB on session end.
- Implemented CRDT-inspired conflict resolution with optimistic client updates and server reconciliation.
- Game lifecycle via FSM (lobby → active → voting → results) with validated transitions and timeout-based progression.

RuneBuddy | live | AI chat app w/ native memory system and observability features for parents. *Personal Project*
Typescript, Python, Convex, Tanstack start, React, R2, ai-sdk, MCP, RAG, Better-Auth, Redis, PostgreSQL

- Implemented full chat parity: multi-model support, tool-calling, image gen, voice input, file attachments, edit/regen, artifact previews, folders, shared threads.
- Schema-driven memory extraction against confined ontology (academics, interests, wellbeing) with amortized per-turn processing.
- Built resumable token streaming using KV queues in Redis.
- Hybrid memory storage: relational DB for structured facts, vector DB for episodic recall, Redis for hot-path lookups.

Nomodit | repo *Personal Project*
Python, Go, llama.cpp, PEFT, Transformers, GRPO

- TUI and inference server in go for custom fine-tuned SLM inference using the llama.cpp engine on nomodit-4b (fine-tune of gemma3-4b-it) achieving SOTA performance on BEA-19(language tasks bench) with >**80%** improvement over base model.

Certifications and Extracurricular Activities

- Contributed bug fixes and enhancements to open-source projects including Kubeflow, Hugging Face, Unsloth AI, OpenCode, Mem0, tanstack and MemOS.
- Led **front-runner** teams in 3 (inter)national-level hackathons.
- **Certifications** : MongoDB Python Dev Path | Google Cloud Computing - Fundamentals | Meta - FE Dev; Programming with JS Docker Mastery: with k8s and swarm