K Ashraf Ahmed

Software Engineer — AI Systems & Backend · Bengaluru, India
LinkedIn | GitHub | Portfolio

Email: ashrafahmed1232@gmail.com
Mobile: +91-9901916318

Experience

- **The Jobs Jungle — AI-native Applicant Tracking System** — Bengaluru, India
  Associate Software Engineer | Java 17, Spring Boot, MongoDB, OpenAI, React — *Aug 2026 - Present*
  - **AI Products**: Shipped **2 paid AI features** for recruiting agencies: a voice interviewer that writes questions and scores answers with bias guardrails, and an LLM pipeline that sources and ranks candidates.
  - **LLM Infrastructure**: Gave the team its first view of LLM spend with token and cost tracking across **8 LLM call sites**; added prompt caching where identical requests were being re-sent.
  - **Backend & Security**: Cut extra database lookups per application page from **40 to 2** and job-list payloads by **46%**; fixed **5 production security vulnerabilities**, including leaked OAuth tokens.

- **Open-Source Contributor — Zalando Skipper, CNCF etcd & CALL-E**
  12 merged PRs across 8 upstream repositories — *June 2026 - Present*
  - **Skipper**: Added **weighted round-robin load balancing** to Zalando's production HTTP proxy and fixed a bug that let streaming responses exceed concurrency limits. [Routing PR] [Concurrency PR]
  - **CALL-E**: Developed an AI phone-call agent and **MCP** server with crash-safe dispatch; raised confirmation accuracy from **87.7% to 94.6%** across **300 simulated trials**. [PR]

Projects

- **Kiln — LLM Inference Engine** | *C++, CUDA, Python, PyTorch* [Link]:
  - Built an LLM inference engine in C++ with custom **CUDA** kernels; serves SmolLM2-135M at **219–231 tokens/s** (single request) through an OpenAI-compatible API, validated against Hugging Face logits.
  - Sped up text generation **11.2**× over full recomputation with KV caching in a CPU benchmark, and built a paged KV cache that shares common prompt prefixes copy-on-write.
  - Made INT8 matrix multiplication **1.60**× **faster than FP32** in a GTX 1660 Ti microbenchmark by pairing cuBLAS integer math with a custom CUDA dequantization kernel.

- **Consensa — Distributed Vector Database** | *Go, Raft, HNSW, gRPC, TLA+* [Link]:
  - Architected a distributed vector database in Go with hand-written **Raft** consensus; checked linearizability through server crashes and network partitions, and model-checked quorum intersection and range ownership in **TLA+**.
  - Returned **98.6%** of the true top-10 matches in **156** μ**s** average local HNSW search, on 5,000 synthetic vectors checked against exact brute-force search.

- **Aegis — Predictive Observability Platform** | *Python, FastAPI, Kafka, Scikit-Learn, Prometheus* [Link]:
  - In simulated outages, found the root cause in **78 of 78** incidents and warned a median **193s** before failure on 18 held-out cases, using Kafka event correlation and gradient boosting.

- **ClearText API — Async ML Inference Platform** | *FastAPI, Redis, PostgreSQL, Celery* [Link]:
  - Moved BERT inference off the request path: the API returns a task ID while batched Celery workers run the model, with model-versioned Redis caching, backlog-based rejection and a replayable dead-letter queue.

Technical Skills

Languages	Java, Python, Go, C/C++, SQL, TypeScript, Bash
AI & ML	PyTorch, Hugging Face, CUDA, LLM serving, KV caching, INT8/INT4 quantization, LoRA fine-tuning, LLM APIs, model evaluation, vector search, MLOps
Infrastructure	Spring Boot, FastAPI, REST, gRPC, Kafka, Docker, AWS (Lambda), Linux, CI/CD, Prometheus, Grafana
Core	Data Structures and Algorithms, OOP, Distributed Systems, System Design, PostgreSQL, MongoDB, Redis

Publications & Awards

- **IEEE ASIACONF 2026 (accepted)**: "Lightweight Radar-Camera Fusion for Real-Time Drone Detection and Trajectory Prediction" — co-author; precision **0.91→0.94** at **8.3 FPS** (CPU).
- **IEEE C2I6 2025**: "Lightweight Real-Time Drone Detection and Trajectory Forecasting Using YOLOv8 and LSTM Networks" — co-author; **mAP 0.89** at ∼**9 FPS** (CPU).
- **Global Innovation Build Challenge V2 (2026)**: **3rd place (Bronze)**, Medical Tech & Finance track; in a historical backtest on Q3 2022 filings, Marked ranked Silicon Valley Bank **#1 of 4,713** US banks for run risk.

Education

- **New Horizon College of Engineering** — Bengaluru, India
  Bachelor of Engineering - Computer Science and Engineering; GPA: 7.71 — *Nov 2022 - July 2026*