

Mohammad Alhan Bellary

Generative AI & Agentic AI Engineer

Bengaluru, India | +91 8867634490 | alhanbellary123@gmail.com | linkedin.com/in/alhan-bellary | github.com/alhaannn | alhan.me

PROFESSIONAL SUMMARY

Agentic AI & LLM Systems Engineer pairing algorithmic depth (NeetCode 150) with production LangGraph, hybrid and vectorless RAG, and guardrailed FastAPI serving. Built multi-agent systems gated by Ragas/DeepEval evals and routed through LiteLLM. Targeting GenAI and Agentic AI roles.

TECHNICAL SKILLS

Agentic Frameworks: LangGraph (Stateful, Multi-Agent, Human-in-the-Loop), LangChain v1, LiteLLM
Vector DBs & Retrieval: Agentic RAG, Vectorless RAG, Hybrid Retrieval (Dense + BM25), ChromaDB, Qdrant
Security & Evals: PII Redaction, Prompt-Injection Defense, Ragas, DeepEval, Langfuse, LangSmith
Backend & Cloud: Python 3.13, FastAPI (SSE), Pydantic v2, Redis, PostgreSQL, Docker, Kubernetes, GitHub Actions

CORE PROJECTS

Enterprise Multi-Agent Research & Knowledge Assistant | *LangGraph, Qdrant, FastAPI, Ragas, LangSmith, Docker*

- Architected a cyclical LangGraph orchestrator (Planner, Retriever, Fact-Checker) that cut context hallucination by 34%, as measured by Ragas faithfulness and context-recall gates, by replacing single-pass RAG with verify-then-synthesize loops.
- Guardrailed the system with PII redaction (Aadhaar, PAN, phone) and prompt-injection blocking while sustaining sub-800ms time-to-first-token on FastAPI SSE endpoints, by pairing Qdrant + BM25 retrieval with a vectorless SQL fallback.

Self-Healing BI & Text-to-SQL Agent | *LangGraph, PostgreSQL, Plotly*

- Engineered a self-correcting LangGraph text-to-SQL agent that lifted first-pass query accuracy from 61% to 89% on multi-table joins, by feeding database tracebacks back to the generator for up to 3 retries.
- Orchestrated a five-node flow with human-in-the-loop interrupts gating destructive SQL, as validated by sandboxed execution.

Resilient Enterprise AI Gateway & Routing Engine | *FastAPI, LiteLLM, Redis, Pydantic v2, Docker Compose*

- Engineered an async LLM gateway sustaining 250+ concurrent agent requests at sub-50ms queue latency, by routing OpenAI, Groq, and open-source models through dynamic fallback.
- Cut vendor token spend by 28%, as measured by provider token usage, by implementing Redis semantic LRU caching, token-bucket rate limiting, and Pydantic v2 validation.

PROFESSIONAL EXPERIENCE & TRAINING

AI Technocrat Industry Training (Mandatory OJT, 120 Hours) | IIT DHARWAD | Dharwad | Sep 27 – Oct 22, 2027

- Containerized AI inference and agent microservices with Docker and scaled Kubernetes deployments with Horizontal Pod Autoscaling.
- Engineered Jenkins and GitHub Actions CI/CD pipelines with automated rollback triggered by anomalous performance telemetry.
- Configured Prometheus and Grafana dashboards tracking LLM latency, throughput, and memory.

Junior Software Developer (Internship), HeagI Technologies Pvt Ltd | Hubballi | Feb – May 2026

- Delivered 5 features across 2 production Django codebases, cutting feature turnaround by ~25%, by collaborating with senior developers.
- Resolved 15+ bugs, reducing debugging time by ~30%, by diagnosing defects with AI coding tools (Claude Code, Cursor).

EDUCATION & CERTIFICATIONS

AI Technocrat, NSQF Level 5.0 Professional Certification | IIT DHARWAD | Oct 22, 2026 – Oct 22, 2027

BCA, Information Technology, IBMR Business School | Hubballi | Sep 2023 – Jul 2026 | CGPA 7.66/10

Certifications: Oracle Certified Data Science Professional, Oracle, 2025 | AI Literacy, IBM, 2025 | Certified Product Manager (CPM), GeeksforGeeks

Awards: 1st Place, CheTech Hackathon (Chetan College), 2026 | Empathy Encryption Hackathon (UnsaidTalks), 2025