

# AL AMIN AHAMED

AI Engineer · RAG Pipeline Architect · Python / FastAPI / pgvector · LLM Agents

Dhaka, Bangladesh (UTC+6 · ~4hr CET overlap) · [mrabir.ahamed@gmail.com](mailto:mrabir.ahamed@gmail.com) · +880 1794 301713  
[linkedin.com/in/mralaminahamed](https://linkedin.com/in/mralaminahamed) · [github.com/mralaminahamed](https://github.com/mralaminahamed) · [alaminahamed.com](https://alaminahamed.com) (live RAG demo)

## PROFESSIONAL SUMMARY

---

Senior software engineer with 4+ years building production AI-powered systems — the last year focused exclusively on RAG pipelines, LLM agents, vector search, and multi-provider orchestration. Architect of JobPulse RAG (12-source job-discovery with grounded generation via Claude Sonnet 4.6), codebase-research-agent (ReAct tool-using agent exposed as a Claude Code MCP server), and the AI layer of EasyCommerce (GPT-4o + Gemini + Claude + DeepSeek with 35% conversion uplift in beta). Specialised in hybrid retrieval (pgvector + BM25 + RRF), agentic ReAct loops, prompt engineering, eval pipelines, and production-grade Python backends with FastAPI. Applied AI engineering — not ML research. Ships to production with measurable quality targets.

## TECHNICAL SKILLS

---

**AI / LLM:** OpenAI (GPT-4o, function calling, text-embedding-3-large) · Anthropic Claude (Sonnet 4.6, Haiku 4.5, tool use) · Gemini · DeepSeek · Ollama (codellama, deepseek-coder) · LangChain · RAG pipelines · agentic workflows (ReAct, planner-executor) · prompt engineering · evals · MCP servers · LLM cost optimisation · fallback routing

**Vector / Retrieval:** pgvector (primary; cosine / HNSW) · ChromaDB / FAISS (evaluated) · hybrid vector + BM25 · Reciprocal Rank Fusion · LLM reranking · sentence-transformers (BGE / E5) · semantic chunking · metadata filtering

**Python Backend:** Python 3.12 (advanced) · FastAPI · Pydantic v2 · SQLAlchemy 2.0 async · Alembic · Celery · httpx · REST + WebSocket + SSE · PostgreSQL 16 + pgvector · Redis 7

**Languages:** Python 3.12 · TypeScript · JavaScript / ES6+ · PHP 8.x · Go · SQL

**DevOps & Testing:** Docker · GitHub Actions CI/CD · Linux · pytest + asyncio + VCR · mypy strict · ruff · OpenTelemetry · Prometheus · structlog

**Frontend:** React 18 · TypeScript strict · Vite · Tailwind CSS · TanStack Query · Zustand · shadcn/ui

## SELECTED AI PROJECTS

---

*All AI projects are personal builds (private repos at [github.com/mralaminahamed](https://github.com/mralaminahamed); read access available on request for technical screening). Live RAG demo: [alaminahamed.com](https://alaminahamed.com)*

### JobPulse RAG — AI Job-Discovery Platform

*Python 3.12 · FastAPI · pgvector · Redis · Celery · OpenAI · Anthropic Claude · React 18*

- 12-source, 3-tier async ingestion pipeline: search engines (SerpAPI, JSearch for LinkedIn/Indeed, Bing), free job boards (RemoteOK, WWR, Remotive, HN), and ATS public APIs (Greenhouse, Lever, Ashby, Workable, SmartRecruiters) behind a unified async adapter protocol.
- Resume embedding pipeline: DOCX/PDF parsing → 512-token chunking with overlap → text-embedding-3-large vectorisation → pgvector cosine retrieval. All cover letter generation grounded in actual resume content via Claude Sonnet 4.6. Composite scoring: semantic + BM25 + salary + geo with configurable per-track weights.
- Engineered to measurable quality targets: ≥75% top-10 precision · ≤4s p95 latency · ≤\$15/month LLM cost · ≥95% ingestion success. React 18 dashboard with kanban, ATS gap analysis, and WebSocket progress tracking.

### codebase-research-agent — Tool-Using AI Agent / MCP Server

*Python 3.12 · FastAPI · tree-sitter · pgvector · BM25 · OpenAI · Anthropic · Ollama · MCP SDK*

- ReAct-style agent loop (think → act → observe) wrapping semantic search, AST navigation, symbol lookup, grep, and git blame as stateless async tools via OpenAI function calling + Anthropic tool use. Planner-executor fallback. Max iteration cap with fallback answer path.
- AST-aware hybrid retrieval: tree-sitter parsers for Python, PHP, TypeScript, JavaScript, and Go chunk code at function/class/method boundaries. Hybrid fusion: pgvector cosine + BM25 over identifiers/comments + symbol-graph traversal (callers, callees, references) via Reciprocal Rank Fusion → LLM rerank.

- Dual mode: cloud (OpenAI text-embedding-3-large + Claude Sonnet 4.6) or fully local (Ollama with deepseek-coder + nomic-embed-text) for proprietary codebases. Exposed as a Claude Code MCP server with streaming SSE transport and file/line citations in every response.

### EasyCommerce AI Layer — Multi-Provider LLM Integration

OpenAI GPT-4o · Gemini · Anthropic Claude · DeepSeek · Redis · FastAPI · Python 3.12

- Multi-provider AI content pipeline (product descriptions, DALL·E 3 image generation, personalised marketing copy) with cost-aware routing — small models for extraction, large for generation. Content-hash caching in Redis keyed on prompt + model + version.
- Smart product recommendation engine via vector similarity search and collaborative filtering: 35% conversion uplift in beta. Fraud detection via behavioural analysis + predictive scoring via REST API.
- Production reliability: Anthropic Claude + DeepSeek as fallback providers, LLM cost circuit breakers (abort + alert on per-request threshold breach), prompt versioning with changelog, structured evals in CI.

### CortexCMS — Multi-Tenant AI Document SaaS

Anthropic · OpenAI · Gemini · pgvector · Meilisearch · FastAPI

- 5 AI agents (classification/extraction, streaming drafting, RAG dual-similarity search, PII anonymisation, NL → search) across three LLM providers with controlled failover. Privacy-by-design two-tier RAG: tenant embeddings never leave the workspace; platform embeddings anonymised over 24h before promotion to shared index.

## UPWORK FIXED-SCOPE SERVICES

---

Available as fixed-price Upwork catalog items (all Dockerised, documented, ready to integrate):

• RAG pipeline MVP (single source, FastAPI + pgvector)	<b>\$499</b>	5 days
• Hybrid retrieval upgrade (BM25 + RRF reranking)	<b>\$349</b>	3 days
• Claude / OpenAI MCP server (up to 5 tools, streaming SSE)	<b>\$299</b>	3 days
• ReAct agent with tool use (up to 5 tools)	<b>\$449</b>	4 days
• Multi-provider LLM abstraction + fallback routing	<b>\$299</b>	3 days
• AI feature code review + architecture report	<b>\$199</b>	2 days

## PROFESSIONAL EXPERIENCE

---

### Senior Software Engineer · Codexpert Inc., Dhaka

Aug 2025 – Present

Core engineer of EasyCommerce — AI-powered e-commerce platform. Multi-provider AI content pipeline (GPT-4o, Gemini, Claude, DeepSeek), smart recommendations via vector similarity search, fraud detection, LLM cost circuit breakers, prompt versioning, structured evals. 35% conversion uplift in beta. Custom DB architecture delivering 3–5× performance vs WooCommerce.

### Software Engineer · weDevs, Dhaka

Feb 2024 – Jul 2025

Full-stack contributor to Dokan (60,000+ active businesses). React/TypeScript dashboards, REST API integrations, vendor dashboard features, commission logic, and WooCommerce API work.

### Software Engineer · Riseup Labs / EchoaSoft, Dhaka

Aug 2021 – Jan 2024

Custom WordPress plugins and themes; Gutenberg block development; published plugins on WordPress.org.

## EDUCATION & OPEN SOURCE

---

### Post Graduate Diploma in Information Technology (Computer Engineering) May 2023 – Apr 2024

Jahangirnagar University, Dhaka

WordPress Core Contributor (since Aug 2021) · Author of 8 published open-source plugins · Bengali Translation Editor for WooCommerce (7M+ installs) and Dokan (60,000+ installs) · WordCamp Dhaka 2025 Sponsor Team.