# RAKESH ALGOT

Location: Hyderabad, India
Email: rakeshalgot.career@gmail.com
Mobile: +91-8143130499
LinkedIn: linkedin.com/in/algot-rakesh

## APPLIED AI BACKEND ENGINEER | RAG & LLM SYSTEMS

Applied AI Backend Engineer with ~2 years of experience building RAG pipelines, vector search systems, and LLM-driven workflows using Python and FastAPI. Focused on designing reliable retrieval systems, improving search relevance with hybrid and reranking approaches, and developing scalable, secure backend APIs.

## CORE SKILLS

**Languages & Backend:** Python, FastAPI, REST APIs

**Applied AI Systems:** RAG Pipelines, Hybrid Retrieval (Dense + BM25), Cross-Encoder Reranking, Hierarchical Chunking

**LLM Engineering:** Structured Outputs, Prompt Design, Validation Pipelines, Hallucination Handling

**Vector & Data:** Qdrant, FAISS, MongoDB, Redis

**Infra & Systems:** Docker, Linux, Async Processing, Worker Pipelines, Caching Strategies

## PROFESSIONAL EXPERIENCE

### AI Backend Engineer

Yensi Solutions — Hyderabad | 2024 – Present

- Designed and built end-to-end RAG pipelines covering ingestion, recursive chunking, hybrid retrieval (dense + BM25), reranking, and response generation
- Enhanced retrieval accuracy using cross-encoder reranking and hierarchical parent-child chunking for richer context understanding
- Developed a fully local AI inference setup using SentenceTransformers and BM25, removing external API dependency and enabling cost-efficient processing
- Boosted system performance using Redis caching to reduce repeated computation and improve response latency
- Built scalable async worker-based pipelines to handle heavy ingestion and background processing efficiently
- Implemented structured logging and secure backend APIs with role-based access control for reliable and maintainable systems

## PROJECTS

### Citex — Citation-grade Context Retrieval System

**Tech**: FastAPI, Qdrant, Redis, MongoDB, Docker

- Built a multi-stage hybrid retrieval system with dense + sparse search and cross-encoder reranking
- Implemented recursive parent-child chunking (1500/500) to preserve semantic structure and improve retrieval quality
- Designed metadata-rich indexing (parent_id, file_hash, source tracking) for traceability and citation grounding
- Integrated caching layer using Redis to optimize repeated query performance
- Developed ingestion pipeline with duplicate detection and consistency guarantees

### AH Ledger — AI-Powered S106 Compliance System

**Tech:** FastAPI, MongoDB, Docker, LLM APIs (Claude/OpenRouter), AsyncIO

- Built a backend system to automate tracking of S106 obligations, conditions, and compliance workflows
- Designed an end-to-end AI pipeline (document ingestion → chunking → LLM extraction → validation → storage)
- Implemented validation and fallback mechanisms to handle unreliable AI outputs and ensure data quality
- Built asynchronous document processing and implemented document comparison with conflict detection

### AI Pujari — AI-Enabled Digital Platform

**Tech**: FastAPI, MongoDB, Keycloak, Docker ,Elevenlabs,Heygen

- Built AI-driven pipelines for audio and avatar generation with automated processing and background job handling
- Designed concurrent task execution with job tracking, logging, failure recovery, and secure access control

### Golf Addicts — Tournament & League Platform

**Tech**: FastAPI, MongoDB, Keycloak ,Docker

- Optimized leaderboard and tournament APIs for scalable performance using efficient queries, indexing, and caching
- Implemented background processing, bulk data ingestion, and secure endpoints with authentication and role-based access control

## EDUCATION

### Master of Computer Applications (MCA)

Nizam College — 2022