

YASH KHADSE

Hyderabad, Telangana, India | +91 9704640901 | khadseyash4@gmail.com
linkedin.com/in/yash-khadse-4yk | github.com/Yash-Khadse | yashkhadse.site

PROFESSIONAL SUMMARY

Programmer Analyst Trainee at Cognizant in the AI/ML domain and a B.Tech Computer Science & Engineering (Data Science) graduate (Jul. 2026), with experience across five engineering roles delivering Generative AI, full-stack, and cloud solutions. Proficient in Python, SQL, and LLM/RAG application development, with a track record of translating business problems into documented technical designs. Experienced in client-facing POC delivery, stakeholder communication, and handover documentation within agile teams using CI/CD, automated testing, and AI productivity tools such as GitHub Copilot and Claude.

TECHNICAL SKILLS

Programming Languages: Python, SQL, JavaScript, TypeScript, Java

Databases & Data: RDBMS concepts, SQL query optimization, schema design, MongoDB, PostgreSQL, vector databases (ChromaDB, Pinecone, pgvector), embeddings, Prisma

Generative AI & LLM: LangChain, LangGraph, CrewAI, RAG pipelines, agentic AI, LLM orchestration, prompt engineering, structured outputs, tool calling, OpenAI & Anthropic APIs, AI guardrails, RAGAS, DeepEval

AI/ML & Cloud: TensorFlow, Keras, Hugging Face Transformers, spaCy, OpenCV, NLP, Whisper/ElevenLabs (STT/TTS); AWS, Google Cloud Platform (Cloud Run), Firebase

Web, API & DevOps: FastAPI, REST APIs, Node.js, Express.js, Next.js, React, React Native (Expo), MERN stack, WebSockets, JWT/OAuth, Streamlit, Docker, Git, Turborepo monorepos, Redis/BullMQ job queues, CI/CD pipelines, automated regression testing, monitoring

Practices & Tools: Agile delivery, algorithms and data structures, technical design documentation, business requirement analysis, architecture and code reviews, GitHub Copilot, Claude, AI-assisted SQL generation

PROFESSIONAL EXPERIENCE

Programmer Analyst Trainee

Mar. 2026 – Present

Cognizant

Chennai, Tamil Nadu

- Build and support AI/ML solutions for enterprise clients in Python, developing data preprocessing, model training, and evaluation pipelines and exposing inference through FastAPI REST services
- Contribute to LLM and RAG-based delivery accelerators covering prompt engineering, retrieval tuning, and structured output validation; participate in architecture discussions, code reviews, and technical design documentation within an agile delivery team
- Write and optimize SQL queries for data extraction and feature preparation, and support CI/CD deployment, automated testing, and defect triage in line with client quality and delivery standards

AI Engineer

Dec. 2025 – Feb. 2026

NyraAI

Hyderabad, Telangana

- Built a production LLM voice assistant in Python using LangChain, agentic workflows, and RAG pipelines (ChromaDB), designing prompt templates, structured outputs, and tool-calling patterns that improved context accuracy by 35% and reduced hallucination rate by 40%
- Integrated OpenAI and Anthropic LLM APIs with FastAPI backends, implementing AI guardrails, monitoring, fallback degradation strategies, and output-quality metrics to meet production reliability requirements
- Translated ambiguous requirements into documented technical designs for agentic tool-orchestration workflows with multi-step intent resolution; maintained CI/CD pipelines with automated regression tests for AI output correctness, safety, and latency

Full Stack Developer

Sep. 2025 – Nov. 2025

Defence in Depth Solutions

Hyderabad, Telangana

- Developed AI-augmented enterprise applications (Next.js, React, Node.js, MongoDB) with LLM-powered search and form parsing, delivering short-cycle POCs and client-facing demos to tight timelines
- Analyzed business problems with client stakeholders and translated them into AI-native solution designs with clear API and data-flow specifications for both technical and non-technical audiences
- Architected secure RESTful APIs with JWT/OAuth, RBAC, and audit logging, applying stack-specific security and code-quality best practices; produced handover documentation enabling independent client operation

Technical Lead

Jul. 2024 – Aug. 2025

Atharva Data Science Community

Hyderabad, Telangana

- Led end-to-end delivery of 5+ MERN stack applications with AI/ML integrations, owning implementation planning, architecture reviews, and delivery timelines; improved MongoDB query and schema efficiency by 30% through data-model optimization
- Mentored a 6-member development team on architecture, code review discipline, and responsible AI practices

Web Developer Intern

Nov. 2023 – Feb. 2024

Brutanix

Hyderabad, Telangana

- Reduced API latency by 15% through database schema optimization and query tuning; developed 20+ reusable React/TypeScript UI components that improved front-end performance by 40%, and implemented JWT/OAuth authentication with security-first API design

PROJECTS

Social Brain — AI Knowledge Base | *TypeScript, Next.js 15, React Native, PostgreSQL/pgvector, Claude API* [social-brain-web.vercel.app](#)

- Built and deployed a cross-platform AI knowledge base (web and Android) from a single Turborepo monorepo, sharing TypeScript types, Zod validation schemas, and AI logic across a Next.js 15 app and an Expo React Native client to eliminate duplicated business logic
- Designed a four-stage asynchronous processing pipeline (extract, analyze, embed, complete) on BullMQ and Redis workers with exponential-backoff retry, summarizing, tagging, and vector-embedding saved content through the Claude API so no user request blocks on LLM inference
- Implemented hybrid retrieval over PostgreSQL with pgvector, merging HNSW-indexed cosine similarity with full-text search and re-ranking; powers a RAG assistant that grounds every response in the user's own indexed items with inline source citations

Multi-Agent Workflow Automation Platform | *Python, LangGraph, CrewAI, FastAPI, React, AWS*

- Architected a multi-agent system (LangGraph + CrewAI) with stateful memory, inter-agent tool handoffs, and human-in-the-loop approval gates, replacing manual document processing and cutting processing time by 70%
- Built a React observability dashboard with hallucination flagging and cost/latency telemetry; deployed on AWS with client handover documentation

Enterprise RAG Knowledge Assistant | *Python, LangChain, Pinecone, OpenAI, Next.js, DeepEval*

- Engineered a production RAG pipeline over 10,000+ documents with hybrid retrieval and cross-encoder re-ranking, achieving 92% answer relevance (RAGAS), with DeepEval quality scoring automated in CI to monitor p95 latency, token cost, and retrieval-quality drift

AI Voice Agent with Agentic Memory | *Python, Whisper, GPT-4o, ElevenLabs, LangChain, GCP*

- Built a real-time voice agent (Whisper STT, GPT-4o, ElevenLabs TTS) with persistent vector-store memory and multi-tool orchestration across calendar, CRM, and search APIs; achieved sub-1.5s latency via async streaming, deployed on Google Cloud Run with CI/CD and automated quality regression tests

EDUCATION

Malla Reddy Engineering College

Hyderabad, Telangana

B.Tech in Computer Science & Engineering (Data Science)

Nov. 2022 – Jul. 2026

- Relevant Coursework: Data Structures & Algorithms, Database Management Systems (SQL/RDBMS), Machine Learning, Big Data Analytics, Cloud Computing

Narayana Junior College

Hyderabad, Telangana

Intermediate (Class XII)

Jun. 2020 – May 2022