

Ashwin H

AI Engineer / LLM & RAG Systems

 ashwin21rr@gmail.com  +91 9360084775  linkedin.com/in/ashwinaadithan  github.com/ashwinaadithan

AI Engineer with 4+ years of experience building production RAG systems, multi-agent architectures, and LLM infrastructure at scale. Currently building retrieval pipelines over 36K+ document corpora with multi-provider LLM orchestration across OpenAI, Anthropic, and Bedrock. Deep experience in LangChain/LangGraph, vector databases, and reinforcement learning, backed by a strong full-stack and cloud foundation.

Technical Skills

AI & Machine Learning

PyTorch LangChain LangGraph RAG MCP
LLMs Embeddings Tool Calling Deep RL YOLO
VSLAM

Cloud & Infrastructure

AWS Azure CloudFormation Terraform Docker
CI/CD OpenTelemetry Linux Admin

Full-stack Development

Python JavaScript FastAPI React Node.js
Express Next.js MongoDB

Databases & Vector Stores

Qdrant Milvus Chroma Weaviate PostgreSQL
MySQL Neon DynamoDB

Professional Experience

AI Engineer

Feb 2026 – Present

Jagriq

- Built a **production RAG pipeline** over a 36K+ document corpus — semantic chunking, embedding generation, and **Qdrant** vector indexing across ~329K structured units and 550K+ entity relationships
- Designed a **provider-agnostic LLM tool-calling layer** unifying OpenAI, Anthropic, Google, and Bedrock behind a single contract, with cost accounting, response caching, idempotency, and OpenTelemetry tracing
- Engineered **citation-grounded retrieval** that resolves and verifies every source reference against an external corpus API, eliminating unverifiable attributions in generated answers
- Rebuilt **multilingual document ingestion** across 22 languages — migrated OCR to PyTesseract with async parsing, and fixed a race condition corrupting ~22% of ingested documents
- Shipped an internal **LLM cost-observability dashboard** (FastAPI + MongoDB) with per-endpoint pricing, usage tiering, and a metering proxy for token and API spend

TECH STACK:          


Co-Founder & Technical Lead

Aug 2024 – Present

IGIT.in

- Led technical roadmap and architecture for **Memoir**, an AI voice assistant with multi-step task automation using MCP-like architecture for agent coordination and reasoning
- Architected microservices backend and microfrontend architecture enabling **dynamic UI loading** and plugin-based extensibility for rapid feature expansion
- Integrated TTS and speech processing pipelines for multimodal interaction with high-quality voice output and context-aware responses
- Delivered AI-integrated client platforms with chat, audio, and video on containerized AWS infrastructure at **99.9% uptime**

TECH STACK:       

Freelance Full-Stack & Cloud Developer

Jun 2024 – Jan 2025

Self-Employed

- Delivered **5+ client projects** spanning admin dashboards and real-time communication platforms with WebRTC video/audio and WebSocket signaling over TURN/STUN
- Built serverless AWS workloads on Lambda, DynamoDB, Cognito, SNS, and SES, with **CloudFormation IaC** driving repeatable multi-environment CI/CD deployments

TECH STACK: MERN Stack AWS Lambda DynamoDB Cognito CloudFormation WebRTC Socket.io Docker

Autonomous Systems Developer

Jul 2023 – Oct 2024

Ciragu (Government Associated)

- Trained **Soft Actor-Critic** reinforcement learning agents for real-time object tracking on a government-backed autonomous drone program, combined with **YOLO** detection and **VSLAM** localization
- Developed simulation pipelines in **PX4, Gazebo, and AirSim** with custom reward functions for agent training
- Implemented ROS-based communication for real-time sensor processing and scaled RL training runs across AWS and Azure GPU workflows

TECH STACK: Soft Actor-Critic RL YOLO VSLAM PyTorch AirSim PX4 Gazebo ROS

Software Engineer

May 2022 – Jun 2023

Consultadd

- Delivered the CVS automation system integrating **Azure Cognitive Services, Computer Vision OCR, and Form Recognizer** for accurate bill text extraction and structured data generation
- Built MERN backend pipelines for secure document ingestion, validation, and storage with protected AI inference endpoints, provisioned via **Terraform**
- Proposed and implemented a **priority-based request handling architecture** that significantly reduced processing latency at peak load and improved system stability

TECH STACK: Azure Cognitive Services OCR Azure App Services MERN Stack Terraform REST APIs Docker

Database Administrator Intern

Jan 2022 – Jul 2022

Ford Motor Private Limited

- Supported production database operations, handling tickets for **query optimization, indexing, and schema improvements**
- Assisted with **SQL query tuning** and performance monitoring across PostgreSQL and MySQL, including backup strategies and RBAC implementation

TECH STACK: SQL PostgreSQL MySQL Query Optimization Indexing RBAC Linux

KEY PROJECTS

Data Collector – Multi-Agent Data Ingestion for LLM Training

Oct 2025 – Present

github.com/ashwinaadithan/Adis-Data-collector

- Built a multi-agent pipeline that **scrapes and processes data for LLM training** from heterogeneous web and API sources
- Implemented autonomous agents for search, extraction, transformation, and verification with retry logic and error isolation
- Designed semantic preprocessing and chunking for **embedding generation and vector indexing**, backed by Milvus and Neon (PostgreSQL) metadata

TECH STACK: Python Node-based Agents LangChain Milvus Neon (PostgreSQL) Vector Embeddings Docker

NeuroVault – Advanced RAG-Based Memory Engine

Sep 2025 – Present

github.com/ashwinaadithan/NeuroVault

- Developed an intelligent memory system integrating **RAG architecture for long-term knowledge retention** with sub-second contextual recall
- Built multi-document ingestion with semantic chunking and vector indexing using **Milvus and Chroma**
- Implemented dynamic retrieval patterns optimized for context relevance with **LangChain-based answer synthesis** behind a concurrent FastAPI backend

TECH STACK: Python LangChain RAG Architecture Milvus Chroma FastAPI Embeddings

LearnX – Multi-Agent AI Teaching Platform

Jul 2025 – Sep 2025

github.com/ashwinaadithan/LearnX

- Architected a multi-agent learning system with **Planner and Mastery agents** delivering personalized education paths using **LangGraph**
- Implemented persistent agent memory on **Neon PostgreSQL** with dynamic difficulty adjustment driven by live performance signals
- Designed an **MCP framework** for agent communication, event logging, analytics, and external integrations

TECH STACK: LangGraph LangChain FastAPI Neon PostgreSQL Python MCP RAG Components

EDUCATION

B.Tech in Computer Science and Engineering

85%

SRM Institute of Science and Technology, Chennai

2018 – 2022