

Narendra Yechuri

AI Engineer | Generative AI, RAG & Production Backend Systems
satyasurya7@gmail.com | github.com/satyasurya7
linkedin.com/in/narendra-yechuri-a9295891

PROFILE

AI and Software Engineer building production AI applications, retrieval pipelines, backend APIs, and real-time conversational systems. End-to-end ownership from requirements and architecture through development, deployment, evaluation, and production support.

PROFESSIONAL EXPERIENCE

Cerebra Consulting Inc — Software Engineer

Visakhapatnam, India | Dec 2021 – Aug 2026

- Designed and deployed an enterprise knowledge chatbot integrating OpenAI GPT and Claude with structured prompts and context retrieval.
- Architected a RAG backend using Qdrant, Azure AI Search, and OpenAI Embeddings for semantic and vector retrieval. Implemented evaluation and observability with Prompt Flow and custom telemetry.
- Developed PDF/OCR data extraction with Regex, and Azure Function App-compatible backends using SQLAlchemy for database CRUD operations.
- Built a time-series recommendation engine using XGBoost and LightGBM, with Azure Machine Learning Studio and Azure AI Foundry for model selection and deployment.
- Migrated backend APIs from SQL Server to Azure Cosmos DB; developed synchronization scripts and optimized database aggregations. Created Node.js and TypeScript diagnostic scripts for data anomalies and duplicates.
- Developed an AI DevOps Automation Agent POC using LangGraph, MCP, and n8n for user-requested build, test, and deployment workflows. Built computer vision and image analysis POCs using TensorFlow and PyTorch.
- Developed a Headless BI / Dynamic Report Builder using Node.js, TypeScript, React, SQL Server, Azure Cosmos DB, and Playwright.

SELECTED PROJECT — REAL-TIME VOICE AI

Production | Lead Backend Developer

Built a concurrent conversational backend with Pipecat, Gemini 2.5 Flash on Vertex AI, PostgreSQL/pgvector, Soniox STT, Cartesia TTS, Tata Tele Business, and real-time APIs. Implemented horizontal autoscaling; Kubernetes/HPA remains a future architecture plan.

TECHNICAL TOOLKIT

AI: OpenAI GPT, Claude, Gemini, RAG, embeddings, semantic search, LLM evaluation, LangGraph, MCP, n8n.

Backend & data: Python, FastAPI, REST APIs, microservices, async processing, SQLAlchemy, Qdrant, PostgreSQL/pgvector, Azure AI Search, SQL Server, Cosmos DB, MongoDB.

Cloud & ML: Azure Functions, App Service, Blob Storage, Key Vault, Prompt Flow, Azure ML / AI Foundry, GCP / Vertex AI, Terraform, CI/CD, XGBoost, LightGBM, TensorFlow, PyTorch (POCs).

EDUCATION & CERTIFICATION

MTech, Nanotechnology — Jawaharlal Nehru Technological University, Kakinada | 2018–2022

BTech, Mechanical Engineering — Sanketika Institute of Technology & Management, JNTU-Kakinada | 2017

Certification: Artificial Intelligence and Machine Learning