Naga Bala Swamy Vasamsetti

AI/ML Engineer · Generative AI · LLM Systems · RAG Pipelines · Cloud-Native AI Platforms

+91-9381318160 | balaswamyvasamsetti@gmail.com | linkedin.com/in/naga-bala-swamy-vasamsetti | balaswamy.dev | Andhra Pradesh, India

Professional Summary

AI/ML Engineer with hands-on production experience building Generative AI systems, RAG pipelines, multi-agent workflows, and scalable cloud-native backend APIs. Passionate about applying LLMs to automate engineering workflows, boost developer productivity, and build intelligent platform services at scale. Proficient in Python, FastAPI, vector databases (Pinecone, pgvector, FAISS), Azure AI Services, Docker, and CI/CD pipelines — with strong foundations in distributed systems, ML model evaluation, and full-stack AI application development.

Technical Skills

Generative AI & LLMs: GPT-4, Gemini, RAG Pipelines, Prompt Engineering, Multi-Agent Systems (CrewAI), Intelligent Agents, LLM API Integration, Agentic Workflows

Machine Learning & NLP: Supervised/Unsupervised Learning, Neural Networks, NLP, Model Evaluation, Feature Engineering, Transfer Learning

Vector Search & MLOps: Pinecone, FAISS, pgvector (PostgreSQL), Neo4j (Graph RAG), Embedding-based Retrieval, GitHub Actions (CI/CD), Deployment Pipelines, Monitoring & Logging

Cloud & Infrastructure: Microsoft Azure (AI Services, Speech), AWS (Bedrock, ML Foundations), Docker, Vercel; familiarity with GCP and Kubernetes concepts

Backend & APIs: Python, FastAPI, Flask, Node.js, REST APIs, asyncpg, JWT Authentication, Distributed Systems Fundamentals

Data Engineering & Databases: ETL Pipelines, Pandas, NumPy, PostgreSQL, MongoDB, MySQL, Structured & Unstructured Data Processing

Languages & Tools: Python, SQL, Java, JavaScript, Git, GitHub, Postman, React.js

Professional Experience

Machine Learning & AI Engineering Intern Mar 2024 – Jun 2025
BigDataMatica · Virtual *Remote*

- Owned end-to-end design and deployment of production RAG pipelines integrating GPT-4 and Gemini APIs — automating large-scale document understanding and information extraction across unstructured corpora, measurably improving engineering team productivity on recurring data workflows.
- Engineered a semantic search system using PostgreSQL with pgvector and Pinecone vector databases, achieving sub-second embedding-based retrieval over high-dimensional document spaces; exposed AI inference capabilities through FastAPI REST endpoints with CI/CD deployment via GitHub Actions.
- Architected multi-agent automation workflows using CrewAI intelligent agents, eliminating manual data processing bottlenecks and significantly increasing ETL pipeline throughput for recurring jobs across structured and unstructured data formats (CSV, PDF, text).
- Built robust Python ETL pipelines with integrated logging and monitoring for data ingestion, cleaning, and preprocessing — enabling reliable downstream ML model consumption and system observability at scale.

Generative AI Engineering Intern Jun 2025 – Aug 2025
Celebel Technologies · Virtual *Remote*

- Integrated LLM APIs to power intelligent, context-aware content generation features — extending AI platform capabilities and improving user experience with high-quality automated outputs at scale.
- Built and secured RESTful APIs in Node.js and Express with JWT-based authentication, enabling reliable and secure AI feature delivery to frontend consumers across distributed service boundaries.
- Designed MongoDB schemas and CRUD operations optimized for storing AI-generated content, conversation histories, and user interaction logs; monitored application logs and system metrics to identify and resolve backend performance bottlenecks.

Data & Enterprise Systems Apprentice Feb 2025 – Jul 2025
PwC Acceleration Center · Virtual *Remote*

- Developed complex SQL queries and data models for operational reporting and analytics pipelines, translating business requirements into structured, actionable data insights for enterprise decision-making.
- Implemented Apex triggers and validation rules within Salesforce CRM, enforcing data integrity across enterprise pipelines supporting business-critical workflows at scale.

Projects

ResearchAgent | **AI-Powered Document Intelligence Platform**
FastAPI, React.js, Neo4j, PostgreSQL, pgvector, Gemini API, OpenAI API, Docker

- **Overview:** Built a scalable cloud-native AI platform enabling researchers to query and extract intelligence from 100+ documents concurrently with sub-5 second responses — combining speculative RAG, Neo4j graph-based retrieval, and persistent cross-session neuromorphic memory.
- Achieved **0.8+ RAG evaluation scores** across recall, precision, faithfulness, and answer relevancy via a custom MLOps evaluation pipeline; swarm-based multi-agent retrieval with 50 parallel intelligent agents reached **0.92 consensus accuracy**.
- Engineered multi-tenant backend architecture (Docker-deployed) with 80% document compression ratio, quantum-inspired retrieval (coherence threshold 0.7), and persistent cross-session research memory — with real-time system metrics monitoring dashboard.
- Delivered React.js frontend with i18n support for 5+ Indian languages and adaptive workflow visualizations — validated at **90%+ accuracy** via metamorphic testing benchmarks.

IRRIGO | **AI-Driven Precision Irrigation Platform**
Python, FastAPI, React, PostgreSQL, asyncpg, Weather APIs, MUI

- **Overview:** Replaced expensive physical soil sensors with an AI-driven cloud-native backend — computing crop water needs via evapotranspiration models fed by real-time weather data to automate precision irrigation scheduling.
- Delivered **30–50% reduction in water usage** and **15–25% crop yield improvement** through AI-driven ET calculation and crop-specific irrigation scheduling; generated **Rs. 50,000–1,00,000 in annual savings** per field.
- Built async FastAPI backend with automated fallback retry logic powering high-availability multi-field alerts; multilingual React/MUI frontend (Hindi, English, Telugu, Tamil) enabling non-technical farmers to manage AI-driven irrigation without hardware investment.

HumanSQL | **Natural Language to SQL Engine**
Python, Gemini API, PostgreSQL, FAISS, FastAPI

- **Overview:** Built an LLM-powered intelligent agent that translates plain English queries into accurate, executable SQL — enabling non-technical users to explore databases with zero SQL knowledge, directly improving developer productivity and data accessibility.
- Enhanced query accuracy by encoding database schema metadata as FAISS vector embeddings for schema-aware context retrieval during Gemini API LLM prompt construction — a core RAG pipeline design pattern applied to structured data.
- Automated multi-format data ingestion (CSV, Excel, PDF, text) into auto-generated PostgreSQL tables via ETL pipeline, with voice input support and persistent query history for iterative, non-technical database exploration.

Education

Chandigarh University 2024 – 2026
Master of Computer Applications | Specialization: Artificial Intelligence & Machine Learning *CGPA: 8.50 / 10*

Aditya Degree College, Kakinada 2021 – 2024
Bachelor of Science in Computer Science *CGPA: 9.05 / 10*

Certifications

- **AWS** – Build and Evaluate RAG Applications using Knowledge Bases for Amazon Bedrock
- **AWS** – Fundamentals of Machine Learning and Artificial Intelligence | **AWS Academy** – Machine Learning Foundations Training Badge
- **IBM** – Introduction to Artificial Intelligence | **University of Washington** – AI for Decision Makers
- **Infosys Springboard** – ReactJS Developer Certification | **Oracle Academy** – Java Foundations

Achievements

- **First Prize – Samsung Innovation Campus Hackathon** | Best AI Innovation Team Award — competed against teams from 50+ colleges nationally; recognized for delivering a production-ready AI solution under time constraints.
- **PGCET Rank 135** | Ranked in the **Top 1.35%** out of 10,000+ candidates in the state-level MCA entrance examination — demonstrating strong analytical and technical foundations.
- **Technical Leadership** | Spearheaded organization of a college-level robotics competition with 150+ participants — managed end-to-end planning, judging, and execution, demonstrating cross-functional coordination skills.
- **Selected Participant** | GDG On Campus Solution Challenge, Hack2Skill (AI/ML Track) — selected from competitive national applicant pools for hands-on AI solution building events.