+91 8885999458

nareshtinnaluri7365@gmail.com

Hyderabad, IN

[GitHub]

[LinkedIn]

Naresh Tinnaluri

Generative AI Engineer

SUMMARY

Highly skilled Software Engineer with 5+ years of experience in backend development and AI system design, specializing in **production-grade Python APIs**, **LLM integration**, and **cloud-native deployments**. Adept at architecting scalable, asynchronous backend systems and deploying intelligent agent workflows using **LangChain**, **LangGraph**, **FastAPI**, **Docker**, and **AWS**. Proven ability to deliver robust, high-performance systems in mission-critical environments.

KEY SKILLS

Backend & API Development: Python, FastAPI, Django, RESTful APIs, AsyncIO
LLM Orchestration: LangChain, LangGraph, RAG Pipelines, Multi-Agent Systems
Cloud & DevOps: AWS (EC2, Lambda, S3, Bedrock), Docker, Git, CI/CD
ML & NLP: Transformers, GPT, Phi-2, Meditron-7B, Fine-tuning, Prompt Engineering
Data Pipelines & Tools: Playwright, Web Scraping, ONNX, TensorRT
Monitoring & Observability: Logging, Metrics, Health Checks for Production APIs

PROFESSIONAL EXPERIENCE

Generative AI Engineer
Feb '24 – Present

Nirmata Neurotech Pvt. Ltd.
Remote

- Designed and deployed **production-grade Python APIs** for multi-agent health AI system (MOA) integrating LLMs like Phi-2 and Meditron-7B.
- Built scalable, **asynchronous agent orchestration** using **LangChain** and **LangGraph**, enabling seamless contextual sharing.
- Deployed on **AWS** with **Dockerized microservices**, utilizing EC2, S3, Lambda, and Bedrock for inference.
- Integrated advanced **supervisor architecture** for coordination between agents in diagnosis, nutrition, and e-commerce.
- Developed and scaled **RAG pipelines** and Chain of Thought (CoT) enhancements to reduce hallucinations.

AI Researcher (Freelance)
Oct '24 – Feb '25

IAN
Remote

- Designed a **DAG-based orchestration pipeline** for misinformation analysis using Chain of Thought techniques.
- Built decentralized agent communication using IPFS and blockchain-enhanced trust layers.
- Focused on **real-time bias detection** using LangGraph and validator chains.

Generative AI Engineer
May '23 – Jul '24

Workplete (Persist Ventures)
Remote

- Developed backend services using **FastAPI**, LangGraph, and OpenAI APIs to automate VC and fellowship form submissions.
- Integrated **LangChain agents and Playwright** for full workflow automation, achieving 75% time reduction.
- Architected **cloud-first deployments** with AWS and Docker, ensuring low-latency and high availability.

Computer Vision Engineer
May '22 – Aug '23

Swatantra Systems Pvt. Ltd.
Hyderabad

- Deployed object detection and segmentation models (YOLOv5, U-Net) on **Jetson Nano with ONNX and TensorRT**.
- Built **asynchronous command triggers** using RabbitMQ and FastAPI in edge environments.
- Improved tracking accuracy by 40% using custom CNN-based vision modules.

Python Developer
Jan '18 – Jul '20

In Technet Limited
Hyderabad

- Built web backends using Django & FastAPI, deployed with Docker.
- Developed **modular login/auth APIs** and certificate generation systems with a 25% speedup in throughput.

EDUCATION

Postgraduate Diploma in Data Science with a specialization in Deep Learning

Nov '20 - Jan '22

IIIT Bangalore

- Secured 82%

Bachelor of Technology in Computer Science

Jul '13 - May '17

SunRise University

- Secured 72%

Higher Secondary (Intermediate)

Jun '11 - Jul '13

Sri Chaitanya Junior College

- Secured 80%

High School (X standard)

Jun '10 - Jul '11

Bhashyam Public School

- Secured 85%