+91 8885999458
nareshtinnaluri7365@gmail.com
Hyderabad, IN
GitHub
LinkedIn

Naresh Tinnaluri

Generative AI Engineer

SUMMARY

Generative AI Engineer with end-to-end project experience in designing, developing, deploying, and optimizing **LLM-based systems**, **RAG pipelines**, and **multi-agent architectures** for real-world applications. Expertise in **retrieval-augmented generation (RAG)** pipelines, **semantic search**, **vector databases (Pinecone, FAISS)**, **LLM fine-tuning**, **prompt engineering**, and **Chain of Thought (CoT)** techniques to reduce hallucinations and improve reasoning. Proven track record of building scalable solutions in **health tech**, **automation**, and **cognitive AI** using **LangChain**, **OpenAI GPT-4o**, **Gemini AI**, and **AWS**.

KEY SKILLS

Generative AI (LLMs, RAG, Multi-Agent Systems)
End-to-End AI Project Development & Deployment
Retrieval-Augmented Generation (RAG) Pipelines
RAG Components: Document Retrieval, Query Reformulation, Response Synthesis, Re-ranking
RAG Internals: Query Embeddings, Vector Retrieval, Contextual Generation
Vector Databases: Pinecone, Chroma
Semantic Search & Document Chunking
Fine-Tuning LLMs for Domain-Specific Tasks
Prompt Engineering: Chain-of-Thought (CoT), Bias Mitigation, Contextual Prompting
LangChain, LangGraph, OpenAI APIs, Gemini AI
Python, FastAPI, Django, Playwright
Cloud: AWS (Bedrock, EC2, Lambda, S3), Docker, GitHub Actions
AI Techniques: CoT, Prompt Chaining, Bias Detection, Knowledge Graphs
Deployment: Containerization, Edge AI Integration, Scalable Microservices

PROFESSIONAL EXPERIENCE

Generative AI Engineer
Jul '24 - Present

Nirmata Neurotech Pvt. Ltd.
Remote

- Led **end-to-end Generative AI projects**, architecting and deploying **LLM-based multi-agent systems** for health-tech applications (nutrition, diagnosis, wellness).
- Designed and implemented **RAG pipelines** with **Pinecone** for **vector storage**, **semantic search**, and **document retrieval**.
- Engineered **RAG components**: document chunking, embedding generation, vector retrieval, query reformulation, response synthesis, and re-ranking.
- In-depth understanding of **RAG internals**: query embedding generation, semantic similarity search, and LLM response integration.
- Applied **fine-tuning techniques** on domain-specific LLMs (e.g., Meditron-7B, Phi-2) for improved reasoning and reduced hallucinations.
- Developed advanced **prompt engineering** strategies using **Chain-of-Thought prompting**, **bias mitigation**, and **contextual prompting** to enhance LLM outputs.
- Orchestrated **semantic search** pipelines for AI-driven document understanding, improving retrieval accuracy and relevance.
- Scaled deployment using **AWS (Bedrock, EC2, Lambda, S3)**, containerized with **Docker**, ensuring low-latency, high-availability AI services.
- Integrated **LangChain**, **LangGraph**, **OpenAI GPT-4o**, and **Gemini AI** for multi-agent orchestration and task execution.
- Deployed **Python APIs** for inference, integrating **multi-agent orchestration** and **RAG pipelines** in real-world applications.

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

IAN
Remote

- Researched and built **multi-agent architectures** with **LangGraph, MCP (Model Context Protocol)** for decentralized data processing.
- Integrated **IPFS** and **Blockchain** for immutable data tracking and **bias detection** in multi-agent AI systems.
- Improved LLM outputs using **CoT prompting**, **bias mitigation**, and **semantic search** in **RAG pipelines**, reducing hallucination rates by 30%.

Generative AI Engineer
Sep '23 - Jun '24

Workplete (Persist Ventures)
Remote

- Developed and deployed **RAG pipelines** with **vector databases** (Chroma, Pinecone) and **LangChain** for automated document processing and web data extraction.
- Built **agentic workflows** for intelligent form submissions, using **Playwright** and **OpenAI GPT-4o**, reducing task time by 75%.
- Automated over 100+ form submissions, integrating **semantic search** and **retrieval pipelines** for improved accuracy.

- Deployed systems using **AWS**, **Docker**, and **microservices architecture** for scalable, reliable AI solutions.

Computer Vision Engineer
May '22 - Aug '23

Swatantra Systems Pvt. Ltd.
Hyderabad

- Developed an **Autonomous Torpedo** with real-time object detection using **YOLOv5**, **U-Net**, and deployment on **Jetson Nano** via TensorRT and ONNX formats.
- Optimized AI pipelines with Docker, achieving 70% cost reduction and 50% model accuracy boost.

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 – IIIT Bangalore (Deep Learning specialization)
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%**