Ruthran RAGHAVAN

Chief AI Scientist | Generative AI & Autonomous Systems Architect
📞 +91 89 39 56 1000
📧 ruthran@hereandnow.co.in
🔗 LinkedIn | GitHub

PROFILE SUMMARY

Chief AI Scientist with 10+ years of hands-on experience architecting, building, and deploying **production-grade Generative AI systems**. Deep expertise in **LLM development from scratch**, fine-tuning open-source models, **multi-agent architectures**, and **LLM-powered full-stack applications**.

Specialist in **LangGraph-based agent orchestration**, **LangChain application design**, **MCP server & client development**, and **voice-first AI systems**. Proven ability to self-host large models, design end-to-end AI infrastructure, and ship scalable AI products without reliance on managed cloud abstractions.

Known for bridging **deep AI research**, **software engineering**, and **DevOps**, delivering autonomous AI systems that are reliable, explainable, and business-ready.

CORE EXPERTISE

Generative AI & LLM Engineering

- LLM architecture design & training from scratch
- Fine-tuning & alignment (instruction tuning, adapters, LoRA)
- Open-source LLMs: LLaMA, Mistral, DeepSeek, Kimi, GPT-OSS
- Model optimization for latency, cost, and on-prem deployment

AI Agents & Autonomous Systems

- Multi-agent systems using **LangGraph** (parallel, looping, supervisor-agent patterns)
- Tool-augmented agents with memory, planning, and reflection
- Enterprise-grade AI workflows with failure handling & observability
- Agent-based automation for ops, analytics, and decision systems

MCP (Model Context Protocol)

- MCP server design for tool, data, and model exposure
- MCP client implementation for agent interoperability
- Secure context routing between models, tools, and services

Voice AI & Multimodal Systems

- Voicebot architecture (ASR → LLM → TTS)

- Conversational AI with low-latency streaming
- Speech synthesis & voice orchestration for AI assistants
- Multimodal agents (text + voice + tools)

Full-Stack AI Development

- Frontend: React
- Backend: Node.js, Express
- Databases: PostgreSQL
- Real-time AI interfaces & chat-based applications
- API-first design for AI platforms

DevOps & Infrastructure

- Advanced Docker-based deployments
- Self-hosted LLM stacks (GPU & CPU)
- CI/CD pipelines for AI systems
- Observability, logging, and monitoring for AI services
- Cloud fundamentals learning path: **AWS & Azure (in progress)**

PROFESSIONAL EXPERIENCE

Chief AI Scientist

HERE AND NOW AI | 06/2018 – Present

- Architected and deployed **180+ autonomous AI agents** across finance, healthcare, education, and manufacturing.
- Designed **LangGraph-based multi-agent systems** handling complex enterprise workflows with state, memory, and coordination.
- Built **LLM-powered applications end-to-end**, from model selection to UI and production deployment.
- Developed **MCP servers and clients** to enable structured tool and context sharing across AI agents.
- Led development of **voice-enabled AI assistants**, integrating speech recognition, LLM reasoning, and speech synthesis.
- Implemented large-scale **RAG systems**, improving retrieval accuracy by 40%+.
- Designed AI automation frameworks reducing operational costs by up to 50%.
- Built and self-hosted AI infrastructure using Docker, ensuring privacy, cost control, and performance.
- Trained engineers and leaders to build real-world GenAI systems, not demos.

Data Scientist

HERE AND NOW – The Language Institute | 07/2011 – 12/2023

- Built predictive ML systems using classical ML, deep learning, and NLP.
- Designed AI-driven personalization engines for learning platforms.
- Developed analytics dashboards and decision-support systems.
- Applied cognitive science principles to AI system design.

KEY PROJECTS & INNOVATIONS

- Autonomous enterprise operations using LangGraph agent networks
- MCP-based AI tool orchestration platforms
- Voice-first AI assistants for education and operations
- Self-hosted LLM platforms for privacy-sensitive deployments
- RAG systems with hybrid search & agent-driven retrieval
- AI Professor (multimodal, conversational, adaptive learning)

TECHNICAL SKILLS

Languages & Frameworks
Python, JavaScript, TypeScript, PyTorch, TensorFlow

LLM & GenAI
LangChain, LangGraph, Hugging Face, OpenAI-compatible APIs

Agents & Protocols
MCP, Tool Calling, Multi-Agent Orchestration

Infrastructure
Docker, CI/CD, Self-hosted GPUs, Linux

Full Stack
React, Node.js, Express, PostgreSQL

Cloud (Learning)
AWS, Azure

EDUCATION

M.Sc. Psychology – Madras University
B.Sc. Psychology – Madras University

Strong foundation in statistics, cognitive modeling, and decision systems applied to AI.