Vasanthkumar Ramamoorthi

Chicago, Illinois I vas.ram.2620@gmail.com I +1 (425)-457-4619 I Portfolio: www.vasanthkumarr.com

CAREER OBJECTIVE

AI Engineer with expertise in deploying ML/LLM features, including RAG systems built with LangChain, LlamaIndex, and vector databases. Developed and fine-tuned generative AI models using QLoRA, DPO, and RLHF to enhance performance. Also delivered production-grade full-stack applications using React/TypeScript and Java/Spring Boot for high user impact.

PROFESSIONAL EXPERIENCE

WERSEC Inc. **Naperville, Illinois, United States (June 2025 – Current)**
Artificial Intelligence Engineer and Research Intern

- Designed and developing end-to-end AI features framing and data pipelines to production inference using Python, Pandas/PySpark, and scikit-learn/XGBoost, improving baseline accuracy.
- Building LLM-powered assistants with **RAG** (retrieval-augmented generation) using LangChain/LlamaIndex and vector stores (FAISS/Pinecone); cut response latency and boosted answer quality.
- Fine-tuned open-source LLMs (**Llama-3, Mistral**) via **QLoRA** experimented with **DPO/RLHF** to increase the data reflect accuracy.

ByteSimplified, **India (May 2025 – August 2025)**
Microsoft Copilot Developer Intern

- Developed custom AI solutions using **Microsoft Copilot Studio, Power Platform, and Azure OpenAI** Service to build intelligent plugins and orchestration workflows with **RESTful APIs** and semantic kernel, enabling natural language access to enterprise data and automated task execution that improved organizational efficiency by 40%.
- Established data governance frameworks using Azure Active Directory and Microsoft Purview for secure AI operations; collaborated with cross-functional teams to implement Power Automate and Python-based automation initiatives, reducing manual processing time by 60%.

Cognizant Technology Solution, **Bangalore, Karnataka, India (September 2021 – July 2023)**
Program Analyst
Project: Johns Hopkins HealthCare (JHHC)

- Achieved proficiency in full-stack development by developing web application, enabling users to submit insurance claims serving 10,000+ monthly users, using **JavaScript/TypeScript, HTML** and **CSS** for **React** powered **front-end** and **Java** with **Spring Boot** for **back-end**.
- Employed **MSSQL** as the primary database solution, using **JSON** based data exchange between frontend and backend throw **Fetch API** and **RESTful APIs.**
- Managed and monitored Active Batch jobs, resolving errors to ensure system reliability through **PowerShell scripting** and **CI/CD pipelines**.
- Utilized Agile SCRUM methodology for efficient project delivery and maintained version control with Subversion (SVN) for effective code management and collaboration.

Cognizant Technology Solution **Coimbatore, Tamil Nadu, India (June 2021 – September 2021)**
Research Intern

- Developed back-end part of e-commerce web application using **Java, Spring Boot,** and **MySQL**.
- Streamlined database performance by optimizing **SQL** queries.
- Gained expertise in Agile processes, including sprint planning and SCRUM ceremonies.

PROJECTS

AI Coding Agent (Langgraph, Langchain, LLM Model, Python)

• I have built a production-grade multi-agent system using ReAct agent pattern in **LangGraph's** StateGraph framework, implementing three specialized LLM agents (Planner, Architect, Coder) with structured output validation via Pydantic schemas to ensure type-safe agent communication and state transitions.

• Designed and optimized LLM integration pipeline leveraging Groq's high-performance inference API with ChatGroq, implementing temperature controls, token limits (2048), and sequential tool calling to prevent race conditions in parallel file operations

• Engineered prompt engineering strategies with role-specific system prompts and context-aware task descriptions, achieving deterministic structured outputs through with_structured_output() method for Plan and Task Plan objects, enabling reliable agent-to-agent data flow.

Data Speak AI – AI-Powered Analytics Platform (RAG, ChromaDB, LLM Model, Whisper, Plotly/Matplotlib)

• Built conversational analytics platform using **RAG**, **Llama 3.2/Mistral LLMs**, and **Chroma Vector DB** to convert natural language into SQL queries—achieved 92% accuracy and saved 40+ analyst hours weekly.

• Integrated voice-to-visual pipeline with **Whisper STT, PostgreSQL**, and Plotly/Matplotlib—eliminated SQL barriers for 80% of users and reduced analysis time from 3 hours to 10 seconds.

• Led end-to-end product strategy designing modular system from voice input to automated charts, enabling non-technical users to generate insights through simple conversation.

EDUCATION

DePaul University, Chicago, Illinois - Master of Science in Computer Science (STEM) (**September 2023 – November 2025**)
Relevant coursework: Distributed Systems, Algorithms, Objected Oriented Programming Language, Programming Concepts, Data Structures, Artificial Intelligence, Image processing, Image Analysis, Computer Vision.

TECHNICAL SKILLS

Programming Languages: Python, Java, C#, VB.NET, SQL, PL/SQL, PowerShell, HTML, CSS, JavaScript, TypeScript.
AI tech: PyTorch, TensorFlow, scikit-learn, XGBoost, LangChain, LlamaIndex, Hugging Face, FAISS/Pinecone, QLoRA, DPO/RLHF, RAGAS,ONNX, DVC, Airflow/Prefect, AWS, Spark, Pandas, pytest. Generative AI, Machine Learning, Prompt Engineering, OpenAI APIs.
Frameworks: .NET Framework, ASP.NET MVC, Django, Angular.js, React.js, Tailwind CSS, Bootstrap.
Tools: Visual Studio, MSSQL Server, Git, Docker, Jenkins, TFS, Active Batch, Selenium.

CERTIFICATIONS

AWS Certified AI Practitioner certificate
Microsoft Certified Python Developer (MCSD)
Basics of JavaScript (Udemy)
AWS Certified Machine Learning Engineer-Associate
Google - Introduction to AI
Data Science with Python (Udemy)