K.V HITESH KRISHNA

+91 9496019918 | hiteshkrishna.kv2022@vitstudent.ac.in | /hiteshkrishna07 | /hiteshhhh007
Leetcode: /hiteshkrishna43

Education

Vellore Institute of Technology | CGPA: 9.04 — Chennai, India
Bachelor of Technology in Computer Science with Specialization in AI & ML (Transcript) — *Sep 2022 – May 2026*

Bhavans Varuna Vidyalaya (CBSE) | X: 97% XII: 96% — Kochi, India
Class IX - Class XII — *Jun 2018 – Jun 2022*

Experience

Software Intern — Feb 2025 – May 2025
Hewlett Packard Enterprise — *Bangalore, India*

- Managed Kubernetes Cluster by using Open Source CEPH as the storage orchestrator.
- Implemented Time-Series Forecasting Models to identify the duration of Replication and Recovery of CEPH Storage at an early stage.
- Used Prometheus and Victoria Metrics to efficiently create datasets and implement Time-Series Models for a scalable and efficient product.

PRISM Research Intern — Jul 2024 – Dec 2024
Samsung R&D Institute — *Bangalore, India*

- Curated an audio dataset of Human Body Sounds (300Hz-600Hz) for building the iHuman Student Foundation Model.
- Developed a Knowledge Distillation-based Teacher-Student Model, employing adaptive distillation, audio-specific data augmentation, and attention mechanisms.
- Enhanced the student model with model pruning and quantization, processing over 10,000 data points efficiently, leading to improved user satisfaction and interaction rates.

Projects

RAG-Based Cryptography & Network Security Chatbot — **GitHub Link**

∗ **Tech-Stack: Python, Langchain, Ollama, Flask, ChromaDB, AWS S3, Tailwind-CSS, JavaScript**
∗ Built a self-hosted RAG-based Conversational AI agent tailored for cryptography and network security education.

∗ Integrated **Ollama** to run local LLMs like `qwen2.5:7b` and `deepseek-r1:7b` for efficient private inference.
∗ Implemented **S3 version-controlled document sync**, allowing automatic updates, deletions, and additions to the knowledge base.
∗ Used **nomic-embed-text** embeddings and stored them in **ChromaDB** for fast similarity-based context retrieval.

∗ Built a real-time **document upload interface** with Flask and Tailwind-CSS, enabling seamless PDF/PPT ingestion.
∗ Employed **Langchain RAG pipeline** to inject contextual chunks into prompts, improving answer relevance.
∗ Added support for **streaming token-level responses** for enhanced user interactivity and responsiveness.

Enterprise Multi-Cloud AI Agent for Knowledge Management

- **Tech-Stack: Python, LangChain, LangGraph, CrewAI, FastAPI, FAISS, ChromaDB, GPT-4o, Claude, Gemini, Azure OpenAI, AWS Bedrock, GCP Vertex AI, TruLens, PostgreSQL**
- Designed and deployed a multi-agent RAG system integrating **Azure OpenAI, AWS Bedrock, and GCP Vertex AI** for scalable enterprise document Q&A.
- Implemented **modular agents** (Retriever, Summarizer, Verifier) using CrewAI, Autogen, and LangGraph to decompose and solve complex queries.
- Leveraged **FAISS and ChromaDB** for vector-based document retrieval and used **TruLens** for real-time LLM output evaluation.
- Used advanced prompting techniques like **Chain of Thought, Step Back Prompting**, and **Iteration of Thought** to boost reasoning and accuracy.
- Exposed all functionalities via a **FastAPI microservice**, with built-in observability and evaluation pipelines.

Technical Skills

Languages: Java, Python, C/C++, SQL, JavaScript, HTML/CSS
Frameworks: Tensorflow, PyTorch, Flask, FastAPI
Libraries: Pandas, NumPy, Matplotlib, Hugging Face, NLTK, SpaCy, Transformer, ChromaDB, FAISS, VectorDB, TruLens
Developer Tools: Microsoft Office Suite, Git, Docker
Cloud Platforms: AWS, GCP, Azure
Cloud Services: AWS - Sagemaker, Bedrock, DynamoDB, Lambda; GCP - Vertex AI; Azure - OpenAI, Azure Functions
Gen AI/Agentic AI: Langchain, LlamaIndex, Autogen, Crew AI, Bedrock Agents
ML/LLM Concepts: Supervised Finetuning, PEFT, Model Finetuning, RAG, RAG Evaluation, LLMOps, NLP, BERT, Observability, Vector DBs