

EDUCATION

Degree	Institute	CGPA/Percentage	Year
B.E	R.V. College of Engineering	9.01	2021-2025
Senior Secondary	Vidyamandir Ind Pre University College	98.66%	2021
Secondary	Little Lillys English Public School	97.44%	2019

EXPERIENCE

- uCube.ai

AI Backend Engineer

Jun 2025 – Present

– Architected an **event-driven Neo4j metadata graph** for cross-source relationships, enabling hybrid RAG across **100+ sources** via lineage and dimension resolution, improving query latency by **68%** with **87% accuracy**.

– Developed a **deterministic graph-based mapping engine** between PostgreSQL and Weaviate, leveraging intent-aware dimension linking to automate **column-level lineage** and reduce hallucinations in multi-source hybrid queries.

– Fine-tuned **InLegal-BERT** using **LoRA** for sentiment analysis, improving accuracy from **52% to 80%** with low compute overhead; built a **Neo4j knowledge graph** and applied **PageRank-based reranking** to enhance RAG retrieval.

– Built a **self-learning multi-agent system** using LangChain and Selenium; implemented an **Exploration-Execution loop** to reduce manual workflows to **2-minute automated runs** via cached navigational paths.
- Sweet Design Hub

Data Science Intern

Jan 2024 – Jun 2024

– Designed a **Time-LLM** forecasting model achieving **89% directional accuracy** for Indian & US equities.

– Implemented an **LSTM-based volatility model**, outperforming GARCH by **7%** on market risk prediction.

– Fine-tuned a financial news **sentiment engine** across **15+ sources** with **91% stance classification accuracy**.

PROJECTS

- LLM Inference Engine with Continuous Batching

TechStack: Python, PyTorch, FastAPI, Transformers, Trie/Radix Cache

Github

– Built an **LLM inference engine** from scratch with **PagedAttention-inspired KV cache management** and **continuous batching**, supporting **8 concurrent requests** while reducing memory fragmentation.

– Designed an **adaptive chunked prefill scheduler** that dynamically adjusts chunk sizes based on workload pressure, reducing **time-to-first-token** by **61%** for mixed short and long prompts.

– Implemented a **trie-based prefix KV cache** with LRU eviction and shared prompt reuse, improving inference throughput by **44%** and reducing average latency compared to sequential HuggingFace inference.
- Agentic Investment Intelligence Platform

TechStack: Python, FastAPI, PyTorch, PostgreSQL, ChromaDB, Docker, vLLM

Github

– Built an **agentic query router** using **Qwen-3.5-8B** via **vLLM** to classify financial queries and dynamically dispatch them between **Text-to-SQL** and **RAG** workflows.

– Developed a **hybrid retrieval and ranking pipeline** using ChromaDB embeddings, PostgreSQL financial data, and **cross-encoder reranking** to improve relevance of retrieved market/news context.

– Evaluated retrieval quality on a custom benchmark covering **50 NIFTY companies**, achieving **76% top-k retrieval accuracy**; added a **semantic caching layer** to reduce repeated-query latency from seconds to milliseconds.

– Deployed a **Dockerized microservices architecture** with parallel news ingestion and incremental market-data updates to support real-time financial intelligence workflows.

TECHNICAL SKILLS

- Languages: Python, C/C++, SQL, Bash

• AI & ML: PyTorch, Pandas, NumPy, Transformers, Fine-tuning, Retrieval-Augmented Generation, LLM Agents, LangChain, LangGraph, Inference Engines (vLLM, SGLang)

• Backend Development: FastAPI, RESTful APIs, Microservices, Redis, Celery, Vector Databases, PostgreSQL, MongoDB, AsyncIO

• Performance & Systems: Model Serving, Docker, Linux, Logging, Performance Monitoring

ACHIEVEMENTS

- Participated in **Smart India Hackathon Grand Finale** and ranked among the **Top 5** teams. Created a model for detecting asthma attacks, showcasing skills in real-world product building and innovative solutions.

• **Won the Cubestorm competition** by building a pipeline for parsing and vectorizing Janes PDFs for accurate retrieval.

• Achieved **top 1%** in NPTEL's Design and Analysis of Algorithms course and **top 5%** in Data Science course.

• Achieved **200+** days streak badge on LeetCode and solved **500+** Data Structure and Algorithms Problems on LeetCode and GeeksforGeeks (GFG).