Lay Sheth

+91 7000035904 - laysheth1@gmail.com - LinkedIn:laysheth - GitHub:Cloaky233 - Portfolio: cloaky.works

EDUCATION

VIT Bhopal University **09/2022 – 05/2026**

B.Tech in Computer Science - 8.89/10 GPA *Bhopal, India*

TECHNICAL SKILLS

- **Programming Languages & Frameworks:** Python, Rust, TypeScript/JavaScript | FastAPI, Flask, Next.js, Tokio, AsyncIO
- **Databases & Vector Stores:** PostgreSQL, SurrealDB, DynamoDB, Neo4j, Firestore |
- **AI/ML & Agentic Systems:** PyTorch, TensorFlow, Hugging Face Transformers | LangChain, LlamaIndex, MCP SDKs, Agentic Development Kit |
- **Data Engineering & Streaming:** Apache Kafka, Polars, Pandas, NumPy, MLflow, Seaborn |
- **Cloud & DevOps:** AWS, Docker, GitHub Actions, Vercel, Heroku | Production deployment, CI/CD pipelines, containerization

WORK EXPERIENCE

ThePreProdCorp **07/2024 – 12/2024**

Machine Learning Engineering Intern *Bengaluru, India (Remote)*

- Layered AutoML pipeline with scikit-learn RandomizedSearchCV for automated model selection across 15+ algorithms, reducing model development cycle time through integrated hyperparameter optimization workflows.
- Contributed to streaming pipeline processing real-time data ingestion via Kafka consumers with FastAPI model serving endpoints, implementing async request handling for concurrent inference workloads.
- Constructed semantic search system as proof of concept integrating open source Mistral-7B transformer with ChromaDB vector store, achieving document retrieval through 3 metric (cosine, Euclidean, Manhattan) embedding similarity search served via Streamlit dashboard interface.

PROJECTS

Agentic RAG Engine with Iterative Self-Correction **05/2025**

- Engineered production RAG system with 3-cycle iterative self-correction across 40+ LLM models, targeting 85% confidence threshold through multi-LLM orchestration (Meta-Llama-405B, Cohere-command-r) with async streaming and token-aware context management.
- Implemented hybrid vector search combining SurrealDB HNSW indexing (3072D embeddings) with real-time Google Custom Search integration, processing 1000+ document chunks with sub-linear performance and Crawl4AI content extraction.
- Built modular architecture with LRU caching (100-item capacity), YAML prompt templating, dependency injection, and comprehensive error handling supporting concurrent queries at production scale.

Kafka-Driven Distributed Image Classification Pipeline **08/2024**

- Architected Kafka-based asynchronous ML pipeline with producer-consumer topology for image ingestion, CNN inference, and result distribution across decoupled microservices.
- Built CNN classifier processing 256x256 grayscale images using PyTorch & TensorFlow with batch training (size=32) and Adam optimizer.
- Engineered production deployment with ONNX optimization, Flask API serving, MLflow tracking, and cross-framework validation supporting PNG/JPG/WEBP formats.

ACHIEVEMENTS

Smart India Hackathon (SIH) **12/2024**

National Finalist *Team of 6*

- Designed and prototyped a low-cost myoelectric prosthetic hand using EMG sensors and an ESP32 microcontroller, drastically reducing the typical hardware cost [12k> prototype].
- Developed a companion mobile app to process raw EMG signals powered by BLE, enabling real-time gesture recognition and control of the prosthetic.

PreProdCorp Buildathon **12/2024**

Winner *Team of 3*

- Developed a web-based performant AutoML platform using Streamlit and PyCaret to automate model training, evaluation, and comparison for tabular datasets.