Harsh Singh

+(91) 6359120976 | hsingh050803@gmail.com | linkedin.com/in/harsh | github.com/harsh

Education

Birla Institute of Technology and Science Pilani — Pilani, RJ
Bachelor of Engineering in Electrical and Electronics Engineering — *Oct 2021 – June 2025*

Experience

SDE-1 — Jul 2025 – Present
Vegapay — *Bengaluru, KA*

- Designed and developed **Athena**, a multimodal ***AI assistant*** with **real-time voice/text interaction**, leveraging ***LangGraph*** agents, ***RAG*** (***vector*** + ***graph DB***), ***Whisper***, and ***Kokoro-TTS*** to enable intelligent onboarding and contextual guidance.

SDE Intern — Jan 2025 – June 2025
Vegapay — *Bengaluru, KA*

- Engineered a warm-up mechanism post-deployment which **reduced** first-hit **API latency by 76%**, accelerating processing of high-volume financial transactions and improving overall system responsiveness for users
- Built WaveCtrl, **a full-stack automation tool** using ***Streamlit***, ***FastAPI***, and ***Celery***, enabling teams to configure and trigger data ingestion pipelines through a self-serve dashboard; **reduced manual onboarding time by automating backend workflows** and job scheduling via ***RESTful APIs*** and async task queues.
- Migrated financial reporting queries from Amazon Athena to **ClickHouse**, ***reducing query execution time by 70%*** and ***cutting operational costs by 50%***, enhancing fraud detection and transaction analytics.

SDE Intern — Jun 2024 – Aug 2024
Samsung R&D Institute Bangalore — *Bengaluru, KA*

- Collaborated with the Language AI Team to refine Speech Keyword Detection and Verification modules, **boosting recognition accuracy by 35%** through advanced signal processing and model optimization
- Used Tensorflow lite models in C++ to **reduce memory footprint and power usage the model footprint by 40%**

Projects

AI Terminal Autocomplete Engine | *Python, Go, FAISS, LLM* — April 2025

- Engineered a **modular backend in Python** for prediction logic and a **real-time interactive TUI** in *Go* using Bubble Tea, communicating efficiently via *gRPC* for **low-latency suggestions** inside the terminal
- Developed a **context-aware prediction engine** that combines ***Trie*-based prefix matching**, ***FAISS* vector search**, and **LLM-based Retrieval-Augmented Generation (RAG)** to deliver intelligent, ranked shell command suggestions
- Designed a custom ***FAISS*-powered prompt caching system** to store and retrieve semantically similar user prompts, significantly reducing redundant LLM calls and improving system responsiveness

Distributed In-Memory Cache | *Go, Distributed Systems* — Aug 2024

- Created a *Golang*-based distributed in-memory cache focused on **high availability and partitioning**
- Used **leader-write, follower-read architecture** to dynamically elect leaders for replication
- Utilized **Consistent Hashing** technique for sharding and seamlessly handling node failures

BITS Fest Backend and Infrastructure | *Python, Django, Go, Docker, Redis* — Apr 2023

- Created a *Golang* based wallet system for BITS Pilani Fest apps, enabling transactions, ticketing, T-OTP and an ordering system seamlessly handling around **2.49Cr in app transactions**
- Created internal tools for Registrations, Event, and Inventory management using *Django* and *Django REST Framework* accumulating **revenue of 48 Lacs total**
- **Managed 5000+ users** with a **peak of 1000+ concurrent users**, ensuring minimal downtime by implementing a microservice structure with *Docker*, *Redis* with scalable *SSE* servers for real-time app updates

Technical Skills

Languages: JavaScript, Python, Java, Go, C++
Frameworks: Django, Django-REST, Express.js, Springboot, FastAPI, Gin
Developer Tools: Git, Docker, Kubernetes, Redis, PostgreSQL, MongoDB, Redis, Kafka, Linux