Yuvaraj G

+91 6381470364 • gyuvaraj444@gmail.com • github.com/yuva-raj444 • linkedin.com/in/yuvaraj-g-53a1a827b
https://gyuvaraj.dev/

SUMMARY

Computer Science Engineering student with strong interest in software development and artificial intelligence. Experienced in building AI-powered applications including on-device LLM inference, legal assistants and health diagnosis systems. Skilled in Python, Java, Kotlin and Flask with a focus on practical problem solving and scalable system design. Also enjoys photography as a creative hobby.

EDUCATION

Bachelor of Engineering – Computer Science and Engineering *2023 – 2027*
Vel Tech High Tech Rangarajan Dr. Sakunthala Engineering College
Cumulative GPA: 8.0 / 10

TECHNICAL SKILLS

Programming Languages: Python, Java, Kotlin, C
Web Technologies: HTML, JavaScript, Flask
Tools & Technologies: llama.cpp, FAISS, Vertex AI, Gemini API, Hugging Face, Android NDK/JNI, Git
Core Concepts: Data Structures, Algorithms, Machine Learning Basics, SQL Basics

PROJECTS

On-Device LLM Runner – Android App

- Built a Kotlin Android application that runs quantized GGUF language models fully offline on-device by integrating llama.cpp through the Android NDK/JNI, with real-time streaming token output.
- Designed a hardware-aware analyzer that profiles device RAM, CPU cores and ABI to auto-recommend compatible models, with resumable Hugging Face downloads and checksum verification.
- Benchmarked Gemma 3 1B (Q4) on a 4 GB RAM device at 25 tokens/sec decode, 50% faster than PocketPal under identical model, context and thread settings.

Tech Stack: Kotlin, llama.cpp, Android NDK/JNI, GGUF Quantization, Hugging Face Hub, Jetpack Compose

AI Powered Health Diagnosis Web App

- Engineered a full-stack AI health diagnosis platform with Next.js and React.js, integrating Gemini and Vertex AI to generate condition suggestions from user-reported symptoms in real time.
- Architected the REST API layer connecting frontend and Vertex AI backend, reducing average response latency to under 2 seconds across 150+ test queries.
- Validated system output against 100+ symptom scenarios, achieving 85% relevance accuracy in internal testing.

Tech Stack: Next.js, React.js, Gemini API, Vertex AI, REST APIs, JavaScript, HTML, CSS

RAG Based Legal Assistant

- Built a Retrieval-Augmented Generation (RAG) legal assistant covering 500+ sections of Indian law, combining FAISS vector search with Hugging Face embeddings for semantic retrieval.
- Reduced legal-query retrieval time by 40% versus keyword search by indexing 1000+ legal documents with FAISS, achieving 90% top-5 retrieval accuracy across 200 test queries.
- Deployed the pipeline on T4 GPU infrastructure via Flask/Streamlit, cutting inference cost while maintaining sub-2-second response times.

Tech Stack: Python, FAISS, Hugging Face Embeddings, Flask/Streamlit, Indian Legal Datasets, T4 GPU

WORK EXPERIENCE

Python Development Intern *Jul 2025 – Aug 2025*

- Developed a Python web-scraping program to automate data collection, processing 500+ records per run.
- Built a Flask-based web application integrating backend logic with a simple user interface, used by 20+ test users.
- Strengthened programming fundamentals and debugging skills through practical development tasks.

CERTIFICATIONS

- Accenture – Software Engineering Job Simulation
- NPTEL – Privacy and Security in Social Media
- Educative – Java Programming Course
- Star Ideation Award – Selected among the Top 50 projects out of 800 for innovative project development at college level.

RELEVANT ACTIVITIES

- Solving Data Structures and Algorithms problems on LeetCode.
- Studying Java programming and object-oriented concepts.
- Reading technology articles and books to stay updated with emerging software trends.