Vividh Mahajan

548-922-2600 | v7mahaja@uwaterloo.ca | linkedin.com/in/vividhm | github.com/Lasdw6 | Portfolio

Technical Skills

Languages: Python, C++, Typescript, HTML, CSS, SQL, Git
Developer Tools: Docker, AWS, Google Cloud, Pinecone, LinuxCLI, GitHub, REST APIs
Libraries & Frameworks: Pytorch, FastAPI, Django, Langchain, Huggingface, Numpy, React.js, Next.js, MongoDB

Education

University of Waterloo Waterloo, ON
Bachelor of Mathematics in Combinatorics and Optimization, minor in Computer Science. *Sep. 2023 – Present*
Presidents Scholarship Recipient
Relevant Coursework:
Tools for Software Development, Algorithm Design and Data Abstraction,
Object-Oriented Software Development, Linear Programming, Optimization
External Courses: Deep Learning Specialization - Deeplearning.ai (125 Hours - ongoing)

Experience

Software Engineering Intern(Ongoing) May 2025 – Aug. 2025
GOQii *Mumbai, India*

- Developing a full-stack RAG Medical assistant for doctors to understand patients' health

Machine Learning Engineer Intern Dec. 2024 – Feb. 2025
The Innovation Story *Remote, Canada*

- Designed a lightweight, graph-based recommendation algorithm for deployment in a mobile education app
- Trained a YOLOv11 model using PyTorch on a custom PCB component dataset (350 images across 7 classes), **achieving 94.3% mAP@0.5 with ∼206 ms CPU inference**, reducing lab-setup time by 67%

Software Engineering Intern Sep. 2024 – Dec. 2024
Electron Online *Mumbai, India*

- Worked on an Analytical Marketing Platform to process **10,000+ social media posts** monthly to generate reports, helping businesses understand customer sentiment on their products, using **Django** and **React**
- Built an end-to-end data pipeline to collect, process, and store over **100GB of social media data/week** using web scraping and API integrations.
- Achieved **20%** reduction in **Google Cloud** server cost by optimizing workflow and resource allocation
- Optimized **Natural Language Processing (NLP)** model accuracy by **15%**, using LLM prompt engineering and fine-tuning, for sentiment analysis

Projects

Tea Tree Chat [Site] | *FastAPI, Next.js, Postgres, REST APIs* June 2025 - Present

- Built a full-stack AI chat app supporting GPT, Claude, and Gemini via OpenRouter and BYOK architecture
- **Reduced LLM response latency by over 60%** by implementing backend response caching, significantly improving perceived app performance
- Shipped a responsive, minimal UI with dynamic LLM switching using Next.js and deployed the app end-to-end

Agentic Personal Assistant | *Python, FastAPI, Langchain, Pinecone, AWS, Docker, Git* Jan. 2025 – Present

- Build a personal assistant from scratch using **Python and FastAPI**, integrating **retrieval-augmented generation (RAG)** with **Pinecone** for efficient knowledge retrieval
- Designed the architecture using **Model Context Protocol(MCP)** and the **Evaluator-Optimizer workflow**.
- Built and deployed the assistant on **AWS Lightsail** using **Docker**, ensuring scalability and efficient API performance

Job Application Tracker[Site] | *Python, Typescript, React, MongoDB, OAuth2, LLMs, Git* June 2024 – Sep. 2024

- Developed a web application to help users track and manage their job applications efficiently
- Successfully launched a closed beta testing phase with **20 registered users**, using **Google OAuth2**
- Used LLMs to parse and categorize email data
- Utilized **Render** to deploy the Django server and **Vercel** for **Vite + React**