Raghav Yadav

647-892-4001 | raghav.yadav@mail.utoronto.ca | Portfolio | LinkedIn | GitHub

Education

University of Toronto — Toronto, ON
Bachelor of Applied Science and Engineering, Computer Engineering — *Sep 2023 – May 2028*

- Recipient of **UofT International Scholar Scholarship**.

Experience

Software Engineer (Google Summer of Code — Intel OpenVINO) — May 2025 – August 2025
Intel OpenVINO — *Toronto, ON*

- Built and deployed **RESTful + gRPC embedding endpoints** in **OpenVINO Model Server** for **multimodal inputs**, serving **100k+ requests/day**.
- Integrated preprocessing with **OpenVINO runtime**, reducing multimodal pipeline latency by **45%** and scaling throughput by **3×**.
- Optimized inference on **Intel iGPUs**, cutting latency from **3600ms → 55ms (65× speedup)** and improving **model throughput** by **300%**.
- Containerized services with **Docker** and built **CI/CD pipelines**; enabled **cross-modal retrieval** using **HuggingFace + LangChain**, supporting **10+ concurrent model deployments**.

Software Engineer Intern — June 2025 – August 2025
Datacurve (Y-Combinator) — *Remote*

- Engineered **backend microservices** for enterprise AI agents, processing **200k+ fragmented data entries/day** and generating actionable insights, improving decision-making speed by **30%**.
- Contributed **5+ production-ready PRs** to high-profile **open-source repositories** advancing LLM coding capabilities, increasing AI agent code coverage by **20%**.
- Designed and deployed **CI/CD pipelines** with **Docker** and **Kubernetes**, reducing release time by **40%** and enabling **10+ concurrent deployments**.
- Optimized data processing workflows, cutting average inference latency from **450ms → 60ms (7.5× speedup)** and integrating **Google** and **Okta APIs** to reduce setup time by **25%**.

Software Engineering Intern — May 2024 – August 2024
University of Toronto Entrepreneurship Hatchery — *Toronto, ON*

- Engineered a high-throughput RAG system over **100K+ documents** using LangChain, Pinecone, and FastAPI
- Reduced API latency from **1100ms to 180ms** using streaming and async execution
- Deployed scalable services with Docker and CI/CD, enabling **zero-downtime rollouts**
- Built monitoring and logging pipelines to support **20K+ monthly queries** in production

Projects

ResumeAI | *Python, Next.js 14 , TypeScript, spaCy, scikit-learn* — Github

- Built an AI-powered tool with NLP ML to deliver ATS checks, keyword relevance scoring, and FAANG-optimized resume feedback.

ShellPilot | *Python, Javascript, Typer, Linux CLI, AI/LLM APIs* — Github

- **Built** a CLI-based assistant enabling natural language execution of Linux commands with **DeepSeek R1** integration.
- **In Progress:** Expanding support for additional **LLM providers** (OpenAI, Anthropic, Ollama) and web frontend integration.

Technical Skills

Languages: Python, JavaScript, TypeScript, C++, C, SQL, Bash Script, HTML/CSS, Java
AI/ML: PyTorch, HuggingFace, scikit-learn, Keras, OpenCV, NumPy, SciPy, Pandas, Matplotlib, Seaborn, TensorRT
Frameworks: React, NextJS, Node.js, Express.js, Flask, FastAPI, TailwindCSS
Utilities: Git/GitHub, Amazon Web Services (AWS), Google Cloud Platform (GCP), CI/CD, MongoDB, PostgreSQL, Spring Boot