Dev Joshi

437-881-4353 | dev.joshi1@ontariotechu.net | https://linkedin.com/in/devjoshi0 | https://github.com/devjoshi0 |

Technical Skills

Languages: Python, Java, TypeScript, JavaScript, C, C#, GoLang, SQL, Dart, HTML5, CSS3, Bash
AI & Machine Learning: PyTorch, Hugging Face, Ollama, LangChain, NumPy, Pandas, Scikit-learn, CUDA, OpenCV
Frontend: React JS, Next JS, Vue JS, Tailwind CSS, Framer Motion, Flutter, Monaco Editor
Backend & DB: Node JS, Express JS, FastAPI, Flask, Spring Boot, PostgreSQL, Firestore, SQLite, SQLAlchemy
DevOps & Tools: Docker, AWS, Azure, GCP, Vercel, Git, CI/CD, Postman, Jira, Linux/Unix

Education

Ontario Tech University **Oshawa, ON**
Honours Bachelor of Engineering in Software Engineering *2024 – Present*

Work Experience

AI/ML Engineer **Oct 2025 – Present**
Brilliant Catalyst WIL *Oshawa, ON*

- Architected self-hosted LLM infrastructure serving 100+ students by migrating from OpenAI API to Ollama models deployed on AWS EC2, eliminating 100% of recurring inference costs while improving response latency by 30%.
- Increased student engagement by 20% by engineering Socratic teaching agent using LangChain for multi-turn dialogue management, integrating RAG with vector embeddings for contextual code feedback.
- Built production-grade IDE with session persistence using Next.js 15, TypeScript, and Monaco Editor, implementing browser-based storage and WebSocket connections to maintain 500+ active coding sessions with minimal latency.

Machine Learning Researcher **Apr 2025 – Nov 2025**
Ontario Tech University *Oshawa, ON*

- Accelerated medical image segmentation model training by 40% by implementing Mixed Precision (AMP) training on custom U-Net architecture using PyTorch, optimizing GPU memory utilization on CUDA-enabled hardware.
- Achieved 93% segmentation accuracy on imbalanced medical datasets by designing a weighted loss function combining Cross-Entropy and Soft Dice metrics, addressing class imbalance in 10,000+ MRI scan images.
- Developed an interactive 3D visualization pipeline transforming 2D prediction slices into mesh models using Marching Cubes algorithms, processing datasets for clinical review workflows.

Software Engineering Intern **Apr 2025 – Jul 2025**
Intelligent Solutions Lab *Peterborough, ON*

- Eliminated manual data lookups for clinical staff by engineering real-time dashboard in Flutter with WebSocket-enabled data streams, visualizing lab metrics with sub-second latency for concurrent sessions.
- Architected RESTful APIs using Node.js and Firestore optimized for 10,000+ concurrent users, implementing database indexing strategies and connection pooling that maintained under 200ms response time under peak load.
- Designed scalable backend microservices handling 500K+ API requests daily, implementing caching layer with Redis and rate limiting to ensure 99.9% uptime during production.

Projects

VeridianAI - Personalized Newsletter SaaS) | *Python, Firebase, NextJS*, Tailwind CSS | GitHub **2025**

- Engineered scalable news aggregation pipeline processing 10,000+ articles daily using Hugging Face Transformers (BART model) for abstractive summarization, reducing average article length from 800 to 150 words while preserving key details.
- Filtered 95% of redundant content by implementing semantic deduplication using sentence-transformers for vector embeddings and cosine similarity clustering, cutting newsletter size from 50 to 3 unique stories per digest.
- Deployed serverless architecture using Cloud Functions to handle 5K+ daily API requests and SendGrid for transactional email delivery, maintaining 99.5% uptime with automated retry logic and error handling.

CensorIt - Live Audio Moderation System | *Python, Whisper AI, NumPy* | GitHub **2025**

- Achieved less than 500ms end-to-end latency by implementing asynchronous I/O for concurrent audio capture/playback, NumPy vectorization for real-time manipulation, and multi-threaded Whisper transcription on 16kHz audio streams.
- Reduced speech recognition inference by 70% using CTranslate2-quantized Whisper model with 100ms sliding window buffers while maintaining 95%+ transcription accuracy.
- Implemented real-time content moderation system using better_profanity NLP library to detect and censor 200+ profanity patterns with 99%+ precision.