Uday Singh Kanwar

Canadian Citizen | udaykanwar007@gmail.com | linkedin.com/in/uday-kanwar | Portfolio | Github

Education

McMaster University — Expected Graduation, April 2027
Bachelor of Engineering in Software Engineering- GPA: 3.9/4.0, Dean's Honour List — *Hamilton, Ontario*

Experience

Software Developer Intern — Sep 2025 - Present
IBM — *Markham, ON*

- Led end-to-end development of a production-grade content creation platform for **2500+** authors and **10M+** learners, delivering a dual-mode WYSIWYG/Markdown editor built using **Next.js, React**, **TypeScript**, **Tiptap** and **Codemirror.**
- Shipped AI-assisted authoring features with document-level text-to-speech and content suggestions using **GPT-4o-mini-TTS** and **OpenAI API**, achieving **90%** consistency with the source technical content and improved accessibility.
- Built secure **RESTful** backend services with **Express.js** and **OAuth2/JWT**, integrating with a host authoring platform to enable draft, publish, and retrieval workflows with type-safe TypeScript contracts, reducing lab publishing time to less than **10 minutes**.
- Improved developer velocity with dockerized deployments and **Playwright E2E** tests, reducing regression bugs by **40%**.

Open Source Research Developer — Sep 2025 - Present
Google Developers Club: Glassbox LLMs, an open-source Python library for interpreting and evaluating LLMs. — *Toronto, ON*

- Built a transformer attention inspection pipeline on **DistilBERT** (6 layers, 12 heads) using **Hugging Face Transformers** and **PyTorch** to extract per-layer, per-head attention tensors for static and interactive visualizations via **Matplotlib**, **Seaborn** and **BertViz**.
- Investigated long-context attention behaviour, analyzing inter-sentence attention to study how **coherence** and **representations** evolve with increasing sequence length, contributing reproducible analysis to an active area of interpretability research
- Developed a local, API-independent LLM evaluation pipeline by integrating **DeepEval** with a custom **Ollama**-backed model wrapper, enabling explainable LLM-as-a-judge evaluation using **GEval** metrics with thresholded pass/fail signals.

AI Engineer Intern — May 2025 - Aug 2025
Blu Creative - A software company providing AI-powered web solutions across North America. — *Toronto, ON*

- Developed an **Retrieval-Augmented Generation** (RAG) chatbot using **OpenAI API, LlamaIndex, LangChain, and SQLAlchemy**, providing natural language query support over company data and RFPs, deployed across **50+** organizations.
- Engineered AI-powered FastAPI endpoints with structured prompt engineering, chat history summarization, and query caching, enabling context-aware query handling, seamless Next.js integration, and a **30%** improvement in response efficiency.
- Built LLM screening to automate screening of **100+** daily documents into **PostgreSQL**, reducing manual processing time by **85**%.

Machine Learning Engineer — Jan 2025 - Present
McMaster Aerial Robotics and Drone Team: Student Club building drones for SUAS Canada competition. — *Hamilton, ON*

- Compared **blob detection**, **hough circle transform,** and **connected-components** methods for circular target detection, tuning parameters to avoid background false positives and achieving 90–95% accuracy on **100+** validation images.
- Benchmarked detection accuracy and latency across approaches, selecting a **blob-based pipeline** that reduced **inference time** by **30%** while maintaining stable detections under varied lighting conditions in images captured by a drone camera.

Software Engineer Intern — Jan 2025 - March 2025
Learning Mode AI - EdTech startup transforming YouTube videos into interactive learning with LLMs. — *Toronto, ON*

- Built **3+ Go** and **Python** based **RESTful** APIs to enable dynamic quiz generation, concise video summaries, and real-time Q&A by leveraging **Redis** and **GPT-4o LLM** integration, working in a docker based environment.
- Developed **10+ UI features** and resolved **15+ bugs**, improving responsiveness by **30**% using **React** and **JavaScript**.

Technical Projects

AI Travel Agent — LLM-Based Flight Search Assistant | *NVIDIA AIQ Toolkit, Python, NVIDIA NIM, SerpAPI* — May 2025

- Developed an AI-powered travel assistant using **NVIDIA AIQ Toolkit**, implementing a ReAct-based agent with **LLaMA-3.3-70B** deployed via NVIDIA NIM. Built a YAML-based agentic workflow with tool chaining and LLM-guided slot extraction for natural-language flight search.

LearnBridge- AI Agent for Students | *Google Agent Development Kit, Python, Google Classroom & Calendar API* — Dec 2024

- Architected a ReAct-based multi-agent system with sequential/parallel workflows to automate Google Classroom & Calendar integration in a Streamlit UI, extracting key insights and syncing deadlines to cut assignment workflow time by **80**%.

Hobbies and Clubs | *McMaster Muay Thai, Intramural: Hockey, Basketball and Volleyball*

Skills

Languages: Python, JavaScript, Java C++, C, TypeScript
Technical Proficiencies: React, Next.js, Node.js, Express.js, Flask, FastAPI, PostgreSQL, Nvidia AIQ toolkit, TensorFlow, OpenCV, Arduino IDE, MatLab, GitHub, Maven, Docker, Pandas, Numpy, Scikit Learn, AWS, JUnit, Pytorch, OpenAI API, Pinecone, LlamaIndex, LangChain, Retrieval Augmented Generation Pipelines (RAG)