

ARYA GOSAVI

☎ 437-262-3797 ✉ aryagsv@gmail.com 🌐 aryas-portfolio.vercel.app 🔗 linkedin.com/in/arya-gosavi 🐙 github.com/aryagith

Education

York University

Expected April 2028

Bachelor of Engineering in Software Engineering, Big Data Stream

Toronto, ON

- **Relevant Coursework:** Data Structures & Algorithms, Database Systems, Discrete Math, Operating Systems, Advanced Object-Oriented Programming, Embedded Systems, Engineering Statistics, Effective Engineering Communication

Experience

Software Developer Intern

September 2025 – August 2026

FGF Brands

Toronto, ON

- Architected and developed a full-stack document-review platform using Flask, Kendo React, and Azure AI Foundry, integrating OCR extraction, reviewer validation, and persistent cloud storage into an end-to-end workflow. Testing and user teams reported up to 80% less review and extraction time per manual during rollout across 2–3 FGF Brands manufacturing facilities.
- Designed Azure storage for a document-review workflow spanning 150+ manuals and approximately 10–15 reviewers, using **Azure Blob Storage** for source manuals and **Azure Cosmos DB** for structured OCR results.
- Contributed to client-specific **ML demand forecasting**, preprocessing data and engineering features from historical demand, seasonal trends, weather, and retailer data to predict product-order quantity ranges. Approximately 90% of product demands fell within the predicted ranges on held-out evaluation data.
- Optimized **Azure DevOps CI/CD** pipelines for automated testing, model versioning, and deployment, cutting build time from approximately 20–30 minutes to 5–10 minutes. Developed **MCP servers**, vector-search infrastructure, and **RAG** workflows connecting internal applications with LLM tooling.
- Built a **.NET and React** attendance application, integrating Python-based facial recognition and anti-spoofing services.

AI & Web Development Co-op

May 2025 – August 2025

PharmShift (BEST Program Award Recipient)

Toronto, ON

- Implemented a chatbot using LangChain retrieval and the DeepSeek API for an AI-powered OSCE platform supported by a \$4,000 BEST Co-op Program award.
- Developed full-stack features using **React**, **Next.js**, and **Node.js** and deployed Dockerized services on **GCP**.

AI Software Developer Intern

March 2025 – August 2025

GrooveAI (part-time)

Remote – Edmonton, AB

- Designed product- and store-specific assistants using the **OpenAI Agents SDK** and **Responses API**, reducing human-handled support inquiries by **20–35%** across clients.
- Developed a **FastAPI** backend handling hundreds of customer inquiries per day, implementing caching to reduce repeated processing and improve response latency.
- Built a **Next.js** conversational interface and integrated retailer knowledge bases, increasing self-service resolution by **15–30%** across clients.

Software Developer Intern

June 2023 – August 2023

Lela Sacks Agro Ltd.

Remote – Kano, Nigeria

- Maintained and extended a **Java Spring Boot** production-management system supporting **130 weaving machines** and **80+ staff**.
- Improved production-data retrieval by **20%** through operational analysis and optimized **SQL/PostgreSQL** queries.
- Improved software reliability by adding automated **JUnit** tests, **Jenkins CI** workflows, and shell-based build orchestration.

Engineering Design Teams

Lassonde Motorsports – Embedded Systems Team

Ongoing

York University

Toronto, ON

- Developing a **QNX/Raspberry Pi** FSAE real-time dashboard and data-acquisition system for temperature, RPM, and voltage telemetry.
- Collaborating with mechanical and electrical engineers on sensor integration and analyzing telemetry data to inform vehicle improvements.

Arbalest Rocketry – Flight Computer & Imaging Team

Ongoing

York University

Toronto, ON

- Developing flight-computer and **payload software** on **STM32 MCUs** in **C**, including telemetry, control logic, launch fail-safes, and onboard payload operations.
- Leading rocket tracking and developing computer-vision tooling for launch footage, trajectory analysis, and post-launch validation.

Personal Projects

Real-Time Body-Tracking Feedback System | *Unity, C#, Python, MediaPipe, PyTest*

GitHub

- Built an interactive **Unity/C#** application that converts webcam movement into reaching tasks with aligned, cursor-hidden, and +30° cursor-rotation feedback modes.
- Created a Python pipeline that extracts **MediaPipe** landmarks and streams pose JSON over UDP into a deterministic Unity experiment controller.
- Normalized users of different body sizes through landmark-visibility checks and shoulder/hip calibration into a consistent body-relative reference frame.
- Separated physical movement from displayed feedback in session JSON and frame/trial CSV logs; tracking loss pauses, aborts, and requeues incomplete trials.
- Validated calibration, experiment sequencing, and retry behavior with **PyTest** and Unity EditMode test suites.

- Built an OpenAI Codex-optimized AI-assistant plugin with Python/SQLite backend and plugin skills; unified three GitHub internship feeds and public-board leads in a tracker keyed by official URL or requisition, keeping status and documents together without duplicate alerts.
- Built a durable resume-tailoring queue around a hash-checked LaTeX master and PDF layout checks, producing editable job-specific drafts and interview notes from verified facts; plan to optimize for other AI providers, including Claude.

Incident Atlas – Service-Metrics Investigation | *Python, Node.js, scikit-learn, AWS, Azure, GCP*

- Built a service-metrics investigation app for AWS ECS, Azure Container Apps, GCP Cloud Run, and CSV exports; preserved actual timestamps and missing readings while ranking anomalies around a known incident onset.
- Raised top-1 service localization from 73.5% with BARO to 80.7% on 223 held-out RCAEval fault-injection cases by combining BARO ranks with an Extra Trees model.

CUDA Attention Engine | *C++, CUDA, PyTorch, GPT-2*

- Implemented float32 C++/CUDA RMSNorm and tiled causal/cache attention kernels; the prompt tile cut peak live attention allocation from 10 to 2 MiB at 256 tokens versus a score-matrix baseline.
- Integrated custom attention and fused GELU into GPT-2; on an RTX 4060 Laptop GPU, fused GELU cut median eight-token direct-greedy generation by 19.4% (70.58 to 56.88 ms) across 27 varied prompts with matching token IDs.

Real-Time Rocket Tracking System – Arbalest Rocketry | *Python, PyTorch, OpenCV, YOLOv8, TensorRT*

- Trained a domain-specific YOLOv8 detector for launch footage, achieving **mAP50 0.882**, **precision 0.928**, and **recall 0.790** on held-out data.
- Diagnosed smoke-trail false positives and integrated **SORT** tracking, coasting logic, and size filtering to preserve rocket trajectories through temporary detection gaps.
- Optimized GPU inference using **TensorRT**, exceeding **90 FPS** end to end.

Quizzle – Full-Stack AI Quiz Web App | *Next.js, React, TypeScript, OpenAI API, NextAuth, Supabase, Prisma*

- Developed a full-stack **TypeScript/Next.js** quiz application that turns a topic into structured multiple-choice and open-ended questions.
- Integrated the **OpenAI API** for question generation and built an interactive answering and review workflow with user-performance tracking.
- Implemented **NextAuth** authentication and persisted quiz history and progress in **Supabase PostgreSQL** through **Prisma ORM**.

Captioner – Captions Generator Web App | *Next.js, React, AWS S3, AWS Transcribe, FFmpeg*

- Built and deployed a full-stack **Next.js/React** application for generating, editing, and exporting video captions using an **AWS**-based transcription pipeline.
- Designed an asynchronous media-processing pipeline using **AWS S3** and **AWS Transcribe**, with automatic language detection and persisted transcript results.
- Built editing controls for transcript corrections and caption styling before export.
- Implemented browser-based **FFmpeg** processing to render edited captions into standalone video files without additional server-side processing.

Cinemax – IMDb-Style Movie Catalogue | *Java, Swing, SQLite, Maven, OMDb API*

- Led a four-person Scrum team, authored user stories, prioritized the backlog, and aligned iterations with delivery goals.
- Built a **Java Swing** frontend and Java/SQLite backend for users, friends, authentication, browsing, search, ratings, and 800+ movie entries.

E-Commerce Chatbot | *PyTorch, NumPy, NLTK, PyTest*

- Built **PyTorch** classifiers for 17 customer-service intents from 172 example phrases and compared bag-of-words, TF-IDF, bigrams, and learned embeddings under the same three-fold cross-validation protocol.
- Measured validation accuracy of 60.5% for bag-of-words, 57.1% for TF-IDF, 48.0% for bigrams, and 47.3% for embeddings, exposing overfitting despite perfect training accuracy.
- Created a command-line interface for probing the trained intent classifier and added **PyTest** coverage for features, metrics, and responses.

IMDb Sentiment Analysis | *Python, TensorFlow, Keras, NLP*

- Built and fine-tuned Simple RNN, GRU, LSTM, and bidirectional LSTM models for sentiment analysis.
- Applied tokenization, sequence padding, and word embeddings and achieved **88.14% accuracy** with GRU.

Skills

Programming: Java, Python, C, C++, C#, JavaScript, TypeScript, R, SQL (Postgres, SQLite, MySQL), HTML/CSS

Software Engineering & Methodologies: Git, SDLC, Agile/Scrum, OOP, Functional Programming, TDD, JUnit, PyTest, Jenkins, Azure DevOps, CI/CD, REST APIs

Data Science & AI: LLMs, RAG, vector databases, MCP, LangChain, OpenAI Agents SDK, Azure AI Foundry, ChatGPT, TensorFlow, PyTorch, scikit-learn, NLTK, pandas, NumPy, Matplotlib

Web, Cloud & Embedded: React, Next.js, Node.js, Flask, FastAPI, .NET, Java Spring Boot, Azure, AWS, GCP, Docker, CUDA, QNX, STM32, Raspberry Pi

Tools & Soft Skills: FFmpeg, JIRA, RDBMS, technical writing, cross-functional collaboration, problem solving, time management, leadership