

Aaryan V Bhalla

+1 (942) 444 1412 | avbhalla@uwaterloo.ca | bhalla.info | github.com/aaryanved | in/aaryanved

SUMMARY

Second-year Computer Science student at the University of Waterloo with hands-on experience building and shipping production AI systems. Currently working as an AI Engineer Co-op at Decimal Technologies, designing LLM agent pipelines, RAG systems, and data automation tools in Python. National hackathon winner with experience across competitive events at institutions across Canada, consistently delivering working, well-engineered products under time pressure. Strongest in AI and ML engineering, full-stack development, and rapid prototyping with modern AI tools.

EDUCATION

University of Waterloo Sep 2025 – Present
Bachelor of Computer Science (Honours, Co-Op) Waterloo, ON, CA

- **Activities:** Member, Computer Science Club | Chassis Design Team, UW Formula Electric | Undergraduate Advisor, Campus Housing
- **Relevant Coursework:** CS135 (Designing Functional Programs) | CS136 (Elementary Algorithm Design and Data Abstraction) | CS245 (Logic and Computation) | CS246 (Object-Oriented Software Development)

SKILLS

Languages: Python, C++, C, JavaScript, TypeScript, Java, C#
AI & ML: LangGraph, LangChain, OpenAI API, Anthropic API, Google Gemini, HuggingFace Transformers, TensorFlow, PyTorch, scikit-learn, FAISS, RAG pipelines, pgvector, Pydantic, Pandas, NumPy, prompt engineering
Frontend: React, Angular, TypeScript, JavaScript, Next.js, Electron, Tailwind CSS, HTML, CSS, face-api.js, Web Audio API
Backend & APIs: FastAPI, Flask, Node.js, Fastify, REST APIs, SSE, asyncio
Applied Cryptography & Systems: RSA/AES hybrid encryption design, JWT auth, LLVM, CMake, compilers/static analysis, distributed systems design
Databases: PostgreSQL, SQLite, Redis
DevOps & Cloud: Docker, AWS, Vercel, Git, GitHub Actions, Jenkins, CI/CD, Bash, Linux, macOS, Windows

CERTIFICATIONS

Claude Code 101 | Anthropic 2026
Model Context Protocol: Advanced Topics | Anthropic 2026

HACKATHONS & COMPETITIONS

National Winner
IIT Delhi Youth Ideathon Indian Institute of Technology Delhi
– Placed first nationally competing against university teams across India in a technology and innovation challenge

Google "All Things Agentic" Hackathon
Participant, Collaborative Partner track - Built Weaver Google
– Built and deployed a conversational multi-agent workflow system solo in 10 days - 15 Google ADK agents across 16 private Cloud Run services, running in production

PIVOTHACKS
2nd Place - Built Verte Waterloo Tech Week
– Placed 2nd in a 12-hour hackathon on the "Sustainable choices" track, building an AI shopping assistant with a team of three

HackTrack
2-time consecutive winner Scottish High International

ShriTeq Hackathon
3-time participant (1-time winner) Shri Ram School, Moulisari

Hack Canada
Participant - Built Grounded Wilfrid Laurier University
– Designed and shipped a fully functional multimodal AI application in under 24 hours - integrating computer vision (face-api.js), large language model reasoning (Gemini 2.5 Flash), and real-time text-to-speech (ElevenLabs) into a unified live pipeline

Ctrl + Hack + Del 2.0
Participant York University

CXC Data Science Hackathon
Participant University of Waterloo

PROJECTS

Weaver – Conversational Multi-Agent Workflow Builder | *Python, JavaScript, Google ADK, Cloud Run, Firestore, Vertex AI* Aug 2026 – Present

- Built a conversational multi-agent workflow system solo in 10 days for Google’s ”All Things Agentic” hackathon (Collaborative Partner track): 15 Google ADK agents across 16 private Cloud Run services, ~12,250 lines, 87 automated tests, deployed to production and used by real users
- Designed a two-agent architecture separating user interviewing from system design, with a deterministic, model-free validator between them catching dependency cycles, dangling references, and unknown step types before deployment
- Grounded agent capabilities in a live Firestore registry rather than prompt content, so adding video and music generation required a database row and one service each with zero prompt changes
- Diagnosed and fixed a silent-failure class where the orchestrator reported HTTP 200 as success while the underlying action had failed, by parsing raw tool-response objects from the event stream instead of trusting model-generated prose
- Enforced least-privilege security across 16 private services (per-call OIDC tokens audienced to individual service URLs, a 4-role service account, a separate weaker identity for the public proxy) and implemented transactional Firestore scheduling so concurrent instances cannot double-fire an automation, verified by firing 8 concurrent ticks and confirming exactly 1 execution

Verte – AI Sustainability Shopping Assistant | *JavaScript, React, Node.js, Express, Snowflake, OpenAI (GPT-4o-mini, Whisper), Chrome* Aug 2026 – Present

- Placed 2nd at PivotHacks (12-hour, team of three): a Chrome Manifest V3 extension and Node/Express backend that surfaces secondhand and lower-carbon alternatives on Amazon and Best Buy product pages
- Led the architecture refactor (~2,000 lines) that consolidated the team’s parallel workstreams into the final codebase, built around four agreed data contracts so mid-event pivots changed the data producers without touching the UI
- Built an anti-hallucination verification pipeline: every GPT-4o-mini product suggestion is checked against live Amazon.ca search results via a custom HTML parser and token-based matching that rejects sponsored listings and accessories, with budgets enforced against real prices and a retry loop that tells the model what was unavailable - nothing unverified ever reaches the user
- Built hands-free voice search with OpenAI Whisper and a Web Audio silence detector that calibrates to the room’s noise floor and stops recording ~1.2s after speech ends, skipping the transcription call entirely when nothing was said
- Integrated Snowflake (key-pair JWT auth) as the source of a 24-category buy-used knowledge base with row validation, caching and a built-in fallback table, and wrote a rate-limited Amazon.ca scraper (custom HTTPS client, gzip, concurrency queue, 30-minute cache)

VFC – Virtual Fighting Championship | *TypeScript, React, three.js, MediaPipe, vitest* 2026 – Present

- Engineered a real-time computer-vision pipeline (MediaPipe pose landmarker) processing live webcam video entirely client-side to classify boxing techniques (punch type, speed, extension, rotation) into a scored combat system, holding robust detection across camera frame rates from 15-60fps through empirically-tuned adaptive (One Euro) signal filtering
- Diagnosed and fixed a chain of real accuracy bugs in the vision pipeline through direct before/after measurement - including a filtering bug silently discarding up to 65% of fast-punch velocity and a tracking bug letting occluded/lost landmarks score as high-quality punches - each fix backed by a regression test proving it against the prior, wrong behavior
- Built an adaptive performance governor that measures live model inference cost and paces camera/inference frame rate against actual device capability rather than a fixed assumption, with hysteresis to prevent visible stutter
- Designed a fixed-timestep, intent-based combat engine architected for future multiplayer without ever transmitting webcam video - only lightweight serialized game intents cross the network boundary - plus an adaptive AI opponent and an algorithmic ”mirror champion” that profiles and replicates a player’s own measured fighting style
- Built a full 3D rendering layer in three.js (procedural inverse-kinematics rigs, a researched arena environment, instanced crowd rendering within a strict draw-call budget) and maintained 400+ automated tests, including a seeded-RNG harness for reproducible AI behavior, enforced via CI on every change

Order Book Matching Engine | *Price-Time Priority, Distributed Systems Design, Zero-Knowledge Settlement* Aug 2026

- Designed a price-time priority matching engine architecture: a sorted structure over distinct price levels for $O(\log n)$ best-price lookup, a FIFO queue per price level for $O(1)$ time-priority matching, and an order-ID-to-node hashmap for $O(1)$ cancellation
- Modeled the full order lifecycle - validation, insertion, cross detection, partial and full fill handling, and cancellation - with explicit state transitions for each outcome
- Mapped the design onto a real off-chain sequencer / zero-knowledge prover / on-chain settlement architecture, reasoning through the trade-offs between off-chain matching speed and the trust model a zk-proof and L1 settlement layer remove
- Documented the full system as a technical architecture diagram to communicate design decisions and data structure trade-offs clearly

BagCheck – Computer Vision Cabin-Baggage Sizing System | *Python, PyTorch, YOLO11, OpenCV, FastAPI, SQLite, Open3DAu* Aug 2026 – Present

- Built a computer-vision system that measures hand-luggage dimensions from a single camera and flags oversized bags against airline cabin limits - ~8,300 lines of production Python plus 1,500 lines of tests (134 passing) across a segmentation/pose pipeline, three interchangeable measurement backends, and a FastAPI web app
- Designed three swappable dimension-estimation backends behind one shared interface (depth, reference-object homography, and monocular 3D pose recovery) chosen by camera geometry, so adding a fourth backend requires implementing one abstract class rather than touching the pipeline

- Solved full camera-pose recovery from a single image via homography decomposition and fit 3D bounding boxes by maximizing silhouette IoU with a scaled, restarted Powell optimizer - validated to ± 3 cm on a synthetic ground-truth harness
 - Designed a three-verdict compliance model that propagates measurement uncertainty into every decision, explicitly refusing to fail a bag on a number the system isn't confident about
- Melodex – AI-Powered Spotify Intelligence Layer** | *TypeScript, Node.js, Fastify, React, PostgreSQL, OpenAI, Vercel, Supabase* 2026
- Architected a full-stack pnpm monorepo deployed as a single serverless function on Vercel with Supabase Postgres - designing around hard function timeouts via checkpointed resumable jobs, idempotent writes, and fast-fail connection pooling
 - Designed an OpenAI tool-calling agent grounded in real user data through structured tools and a post-generation validator that cross-checks every claim against tool results, with all mutating writes gated behind a preview-confirm-undo workflow
- MiniLang – Compiler and Static Analyzer** | *C++, LLVM, CMake, GoogleTest, ANTLR* Jul 2026
- Built a from-scratch compiler: hand-wrote a lexer and recursive-descent parser producing an AST, a type checker, and a stack-based bytecode interpreter, plus a static analysis pass modeled on tools like Clang-Tidy and an experimental LLVM IR backend
 - Wrote 100+ unit tests with GoogleTest and set up a CMake build with GitHub Actions CI so every commit was automatically tested
- Clueless-ly – Real-Time AI Assistant Overlay** | *TypeScript, Electron, React, OpenAI Realtime API, ScreenCaptureKit* 2026
- Built a cross-platform desktop app with real-time transcription of microphone and system audio, streamed through OpenAI's realtime, transcription, and vision APIs to answer questions in context
 - Designed a platform-contract abstraction isolating all OS-specific behavior behind one adapter file, keeping shared feature code identical across macOS and Windows
- Custom End-to-End Encryption & Auth SDKs** | *WebCrypto, RSA, AES-256-GCM, JWT, Redis, Postgres* Jun 2026
- Designed and built a versioned hybrid encryption protocol with a byte-identical wire format across TypeScript, Python, and Java: an RSA-4096 OAEP/PSS handshake establishes an ephemeral per-session AES-256-GCM key with RSA-PSS server-signature verification and forward secrecy
 - Built a companion stateless JWT auth library with refresh-token rotation and single-use, session-bound tokens to prevent replay; shipped as semver-versioned npm packages with Jest/Vitest test suites
- KnightVision – Autonomous Chess Engine & Lichess Bot** | *Python, python-chess, Stockfish (UCI), threading, REST/NDJSON streams*
- Built an autonomous chess-playing bot integrating the Lichess Board API and a local Stockfish engine over UCI - multi-threaded seek/play/log pipeline with dedup locking to prevent race conditions from redelivered events
 - Designed a human-latency model layering time-control-aware statistical delay distributions on top of engine move selection, and a 410-line opening book across 16 repertoires as an $O(1)$ UCI-move-history-keyed lookup
- Pulse AI** | *Python, LangGraph, LangChain, OpenAI API, FastAPI, PostgreSQL, pgvector, Docker, AWS* May 2026
- Designed and built a production AI agent system using LangGraph - structuring multi-step reasoning as task graphs with defined triggers, routing logic, and failure handling, integrating with PostgreSQL and pgvector for structured and semantic data retrieval
 - Built a full RAG pipeline end-to-end and deployed on AWS with Docker and CI/CD automation, maintaining observability across latency, token usage, and retrieval quality metrics
- Grounded** | *TypeScript, JavaScript, Gemini 2.5 Flash, ElevenLabs TTS, face-api.js, Web Audio API* Mar 2026
- Built at Hack Canada 2026 in under 24 hours - a real-time multimodal AI application that integrates computer vision (face-api.js), LLM reasoning (Gemini 2.5 Flash), and natural text-to-speech (ElevenLabs) into a single live pipeline
- The Originality Engine** | *Python, HuggingFace Transformers, FAISS, Pandas, NumPy* Feb 2025
- Designed and built a large-scale semantic analysis pipeline end-to-end: data ingestion and preprocessing, embedding generation using open-source HuggingFace transformer models, FAISS high-dimensional vector indexing, and similarity-scored analytical output, benchmarked systematically across model architectures
- Disease Detection** | *Python, TensorFlow, Flask, scikit-learn, Pandas, Raspberry Pi, Linux* Aug 2023
- Trained a TensorFlow CNN for medical image classification from scratch - data preprocessing and augmentation, architecture design, hyperparameter tuning, and held-out evaluation - deployed as a Flask REST API on Linux for real-time inference on a Raspberry Pi

EXPERIENCE

-
- Software Engineer** Sep 2026 – Present
Xsolla *Los Angeles, CA, USA*
- Building on the payments infrastructure behind Roblox, Epic Games, Rockstar Games, Valve, and Twitch
- AI Engineer** May 2026 – Aug 2026
Decimal Technologies *Delhi, India*
- Design and build AI-powered automation tools in Python - developing intelligent agent pipelines using LangGraph and LangChain that replace manual multi-step workflows with structured, production-grade automations
 - Build and maintain RAG systems and data ingestion pipelines: structured and unstructured data ingestion, embedding generation, pgvector indexing, and retrieval-augmented generation grounded in live knowledge bases
 - Design and optimize PostgreSQL schemas and SQL queries, transforming operational data into clean queryable formats that support data-informed decision making across the engineering team

- Collaborate cross-functionally to identify automation opportunities, translate requirements into technical solutions, and communicate findings clearly to both technical and non-technical stakeholders

Co-Founder and Event Director

Apr 2021 – Apr 2025

SHISTECH

Delhi, India

- Built and maintained full-stack production systems in Node.js and JavaScript as sole engineer across four years - backend services, REST APIs, relational database design, and web interfaces supporting 6,000+ annual event participants
- Identified operational bottlenecks across logistics, registration, and communications and built automation tools to eliminate them - reducing staff workload and improving service delivery at scale
- Led cross-functional teams across content, logistics, and sponsorship - coordinating diverse stakeholders toward shared outcomes with clear documentation and communication throughout

ICT Infrastructure and Community Technology Contributor

2023

Baqiya NGO

Delhi, India

- Designed and built an E-Library and established ICT Labs across two NGO branches, making educational technology accessible to underserved children
- Conducted stakeholder discovery with the NGO's leadership and community members to understand needs before specifying or building any technology - delivering tools that served real community outcomes