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 a Software Engineer at Xsolla working on payments infrastructure, after an AI engineering term at Decimal Technologies building LLM agent pipelines and RAG systems 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 Expected Graduation: Apr 2030

Bachelor of Computer Science (Honours, Co-Op) – President's Scholarship of Distinction *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, Go, JavaScript, TypeScript, Java, C#
AI & ML: LangGraph, LangChain, OpenAI API, Anthropic API, Google Gemini, HuggingFace Transformers, TensorFlow, PyTorch, scikit-learn, reinforcement learning (PPO), MuJoCo, MediaPipe, ONNX, FAISS, RAG pipelines, pgvector, Pandas, NumPy
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

Cal Hacks 13.0
Winner, Moss Track & Top 20, Hyphenate Track - Built Countersign *San Francisco*
- Built an AI payments gatekeeper in about 35 hours; owned the deterministic Policy Engine, case pipeline and sandbox bank, scoring 0 false payments across a 39-scenario answer key

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

Formula Tech Hacks
1st Place - Built ApexTrace *2026*
- Won 1st place with an F1 driving simulator with a corner-entry BRAKE warning and an AI test bench that found 3.5× more safety failures than random testing, with a team of three in 48 hours

HackTrack
2-time consecutive winner *Scottish High International*

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

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

Countersign – AI Payments Gatekeeper for One-Person Companies | *TypeScript, Next.js, Convex, Claude API, Cartesia, ECDSA*

- Won the Moss track and placed Top 20 in the Hyphenate track at Cal Hacks 13.0, built with a team of four in about 35 hours: AI agents take payment requests by phone, web chat and email, call the real vendor back on the number already on file, and only a rules engine can release money; I owned the Policy Engine, the case pipeline, the sandbox bank and the scenario answer key
- Achieved 0 false payments across a 39-scenario answer key (26/26 scams blocked, 13/13 legitimate bills paid or escalated) and 28/28 scenarios on a live deployed pipeline by designing an 18-rule deterministic TypeScript Policy Engine as the only path to PAY, with Claude limited to parsing requests and writing replies
- Rejected 4/4 forged, tampered or replayed transfers by building a sandbox bank that moves money only with an ECDSA P-256 countersignature over canonical JSON, single-use nonces and a 2-minute expiry, plus import-guard tests that let only one module touch the signing key
- Fixed pipeline runs that died silently in a live rehearsal by building an idempotent Convex step function with a watchdog (90 s, at most 2 restarts, then fail closed) and race-based LLM timeouts, and wrote the Policy Engine's 53 tests and 27 pipeline integration tests in a 1,126-test CI suite

Deja – Real-Time Free-Trial Abuse Decision Service | *Java 21, PostgreSQL, LightGBM, Python, TypeScript, Next.js*

- A service that catches people farming free trials with multiple accounts: it links accounts that share a device, card, phone or email pattern into identity clusters as they form, scores them with a machine-learning model and rules, and returns allow, challenge, review or block in milliseconds; built solo, ∼34,000 lines across Java, Python and TypeScript with 485 passing tests
- Made every live decision reproducible exactly offline - a 100% match on 100,000 events from 32 concurrent clients, identical on x86 and ARM - by running one pure-function engine against Postgres online and in memory offline, with lock-ordered transactions that commit interacting events in log order
- Cut expected abuse cost per 1,000 sign-ups by 68% against a rules-only baseline in backtests on simulated data (PR-AUC 0.78, 84% recall at 1% false positives) by training a calibrated LightGBM model on identity-graph features that more than doubled PR-AUC over per-event features alone
- Hardened it for production before deployment, verified by 21/21 failure modes passing under fault injection, 90,800 generated API contract tests with 0 failures, and 300+ dependency vulnerabilities cut to 0, by building the checks into CI on every push

ApexTrace – F1 Safety-Warning Simulator | *TypeScript, three.js, Web Workers, Python, PyTorch*

- A Formula 1 driving simulator with a safety system on board that tells the driver when to brake, built by a team of three in 48 hours that won 1st place at Formula Tech Hacks 2026: you drive real-layout Monza and Baku in the browser, while the bench deliberately breaks the telemetry link to test whether the warning still arrives in time
- Built the core simulator and safety logic - 2026-regulation car dynamics (bicycle model, Pacejka-style tyres, active aero, hybrid power unit), the look-ahead BRAKE warning that never trusts stale data, and a seeded fault-injection test runner - leading the team with 62 of 111 commits
- Found 3.5× more safety failures than random testing on the same 50-simulation budget (47.0 vs 13.6) by training a PyTorch TCN that forecasts lap failure from the first 12 s of telemetry (ROC-AUC 0.918) and using it to screen scenarios for a Soft Actor-Critic search
- Took the app to zero infrastructure after the event, deleting 26.9k lines of backend, by porting the Python simulator line for line to TypeScript in a Web Worker and proving it tick-for-tick against recorded Python sessions

Weaver – Conversational Multi-Agent Workflow Builder | *Python, JavaScript, Google ADK, Cloud Run, Firestore, Vertex AI* Aug 2

- 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

Marionette – Real-Time Webcam Motion Capture Driving a Simulated Humanoid | *TypeScript, Python, Go, MuJoCo, PyTorch*

- Built a 14,790-line browser app where a MuJoCo-simulated Unitree H1 humanoid copies your upper body from a webcam in real time, entirely on-device: pose estimation, 2D-to-3D lifting, retargeting and IK in a WebWorker, and 500 Hz control and physics on the main thread
- Trained a 1.1M-parameter residual MLP that lifts 2D landmarks to 3D on 1.04M CMU mocap poses, using a randomised virtual camera as augmentation; 74.8 mm MPJPE on 40,000 held-out-subject samples, measured on the exported ONNX model, at 0.3–0.5 ms per inference in the browser
- Wrote closed-form retargeting from human limb axes to H1 joints, verified against MuJoCo forward kinematics to under 10^{-6} on 200 sampled configurations, plus damped least-squares IK that cuts hand position error 28% (83.6 vs 116.7 mm) in 2.3 iterations on average
- Measured that H1 cannot stay upright on planted feet while its arms follow a person (centre of mass 12 mm from the support edge, falls in 3.2 s), then added centre-of-mass feedback inside the 500 Hz physics loop - stable at 500 Hz, unstable at 50 Hz - reaching 83% push recovery on a disclosed support rig that applies 0 N when undisturbed
- Cut knee sag 10× (0.206 to 0.021 rad) with a contact-aware gravity feed-forward, low-pass filtered after finding that the raw term formed a positive feedback loop with ground contact and made the robot fall faster than no feed-forward at all
- Trained a PPO residual balance policy over 37M CPU-only environment steps with domain randomisation; it beat PD alone 3.2× on push recovery but added only 4.4 points over the classical controller against a 5-point bar, so it ships switched off, after fixing clip-fraction, reward-scale and truncation-bootstrap bugs found in the diagnostics
- Held 19 ms p95 camera-to-screen latency and 500 Hz physics even at 8× CPU throttle, cut cold load from 4.34 s to 2.85 s by removing a duplicate 13.6 MB WASM download, and compiled a mesh-free physics model from 56 MB to 51 KB with 0.0 measured dynamics divergence
- Backed it with 110 unit and 9 Playwright end-to-end tests in CI, including an automated check that no request leaves the origin during a live camera session, and cross-language fixtures that caught a gravity-projection bug (matrix column instead of row) before it reached training

Crucible – Compute-in-the-Loop Regression Testing for Autonomous Driving | *C++20, Go, Python, TypeScript, NATS JetStrea*

- Built a regression-testing platform for self-driving planners solo (∼12,000 lines across four languages) that runs each planner as a CPU-quota'd OS process under an enforced millisecond deadline inside a deterministic simulator, then decides with paired statistics whether a code change made driving better or worse - live at crucible.bhalla.info
- Achieved bit-exact reproducibility across four platforms - 0 fingerprint mismatches over 40,000 runs on macOS/arm64, glibc, musl and emulated x86-64 - by writing an IEEE-exact math library, eliminating compiler-dependent argument-evaluation order, and quantising measured planner latency into the virtual clock
- Simulated driving at 501× real time per core (1,904× across 10 cores) with a zero-allocation tick loop enforced by a test that overrides `operator new`, and a POSIX shared-memory SPSC ring at 2μs p99 round trip holding 100 ms deadlines to 1.037 ms p99 overshoot
- Designed the statistics that call a regression: 10,000-sample paired bootstrap over 11 driving metrics with Benjamini-Hochberg false-discovery control and a sequential probability ratio test reaching the same verdict on 84% fewer scenarios, with power and false-alarm rate validated against injected regressions of known size rather than asserted
- Built the distributed execution path in Go - NATS JetStream work queue with leases and redelivery, content-addressed result cache (1,000 scenarios × 2 planner versions in 21.3 s cold, 0.6 s warm), idempotent Postgres writes, S3-compatible trace storage - and ran untrusted planner code under cgroup v2 quotas and a seccomp-locked container with an automated escape-probe test in CI
- Used the platform to find four real planner defects and one map-fidelity defect the metrics had been misattributing, each now covered by a regression test verified to fail when the bug is reintroduced

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

- 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, Open3D* Au

- 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 strea*

- 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*

- Built and shipped an internal Systems Registry dashboard in React 19, TypeScript and Neo4j through about 20 production releases, tracking stewards, builders, operators, risk scores and outstanding actions for 50+ company systems
- Ran a column-by-column data-accuracy audit of the dashboard against its graph sources, finding 11 issues; fixed the 4 app defects and routed the data gaps to their owners
- Designed a rule-based red/yellow/green risk score from Slack, meeting and weekly-update activity; removing a false-positive signal took the dashboard from 48 of 51 programs green to an accurate 47 green / 3 yellow / 1 red split
- Found missing Slack thread replies in the company knowledge graph and drove a pipeline fix that refreshes reply counts, now running in production
- Built OneBox Viewer, a sandbox and recipe-catalog app with a Cloud Run back end that hides secrets from every response, tested with Playwright against recorded real data
- Designed the architecture for Prototype Overlays, a 6-phase system of chained sandboxes, and wrote a Go control-plane feature that saves sandbox runs to an experiments repository for review
- Shipped every change through a gated pipeline (preview, stakeholder approval, pull request, CI/CD, production check and rollback), building alongside AI coding agents while owning the specs and review

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

Bagiya 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