

Yair Korok

🌐 yair.ca | 🐙 github.com/yair-k | 🛠️ yair.design | 📞 647-621-0645 | ✉️ yairkorok@gmail.com

EDUCATION

Harvard University — Researcher in AI & Medical Machine Learning @ Harvard Business School
McMaster University — BEng. Software Engineering — 4.0 GPA

Cambridge, MA
Hamilton, ON

EXPERIENCE

AI/NLP Senior Data Specialist | Cohere

March 2025 — Present | Toronto, ON

- Designed scalable evaluation and data-quality workflows for code-generation & NLP models, building validation pipelines, annotation rubrics, and auto-grading tests that increased model reliability & coverage across diverse tasks.
- Audited high volumes of LLM-generated code (120+ samples/week) across Python, JavaScript, and other languages, identifying syntax/logic bugs and edge cases and feeding structured corrections back into model training.
- Partnered with cross-functional ML teams to fine-tune, benchmark, and debug models, iterating on prompts, datasets, and labeling frameworks to improve code correctness, reduce hallucinations, and strengthen AI-driven evaluation mechanisms.

Founding Engineer | Zaq.ai

June 2025 — November 2025 | Boston, MA

- Architected & deployed AI-powered document processing platform handling PDF/XML extraction with AWS Textract, processing 30,000+ attachments through SQS-driven agent workers and Lambda functions
- Built scalable backend infrastructure using FastAPI, PostgreSQL, SQLAlchemy ORM, and Alembic migrations, implementing 11-state status tracking system with 7 performance indexes
- Developed AI agent pipeline leveraging Strands agents, OpenAI GPT-4, and MLflow experiment tracking, creating 6 reward functions for model evaluation and achieving 85% extraction accuracy.
- Implemented modern frontend with Next.js 15, React, TypeScript, and PDF.js, building real-time status dashboards, type-safe API integrations, and responsive UI components tested with 80+ unit tests.
- Deployed AWS infrastructure using CDK (TypeScript), managing S3, RDS, SQS, Lambda, CloudWatch, and multi-environment CI/CD with Docker containerization and automated database migrations.

Full-Stack Software Engineer | LTV.ai (Early Team Member)

April 2025 — June 2025 | Houston, TX / Remote

- Shipped core product features across React (TypeScript), NestJS microservices, and Terraform-provisioned AWS, forming the backbone of LTV.ai's autonomous AI Email/SMS "Brand Ambassador" platform.
- Engineered a high-throughput segmentation engine on PostgreSQL & ClickHouse powering targeting for 20M+ consumers (10% MoM list growth), with consent tracking and GDPR/CCPA-aligned data retention.
- Architected a streaming data pipeline ingesting on-site, order, and messaging events, processing 5M+ events/day with <2s end-to-end latency to feed real time LLM context-backed agents
- Implemented multi-LLM routing policies that cut inference cost 38% while improving reply relevance by 22% in A/B-tested Email/SMS reply flows.
- Built client-facing analytics dashboards for brands, exposing campaign KPIs, conversation-level sentiment analysis, and LTV/retention insights for non-technical marketers.

Co-Founder & CTO | Flair Social Inc.

March 2024 — August 2025 | Toronto, ON

- Launched an AI shopping platform with real-time scraping, ML recommendations, and "Social Media" UX.
- Architected NestJS, GraphQL, PostgreSQL, deployed on Azure; scaled scrapers to 12K items/hour & 180+ retailers.
- Built the frontend in React Native (Pinterest-style feeds, dynamic filtering, Auth0 accounts, sub-150ms loads).
- Engineered Scrapy + Selenium pipelines with OpenAI and 60K+ pair-trained models hitting 72% top-1 style accuracy.
- Grew Instagram account to 20K+ followers and 150K+ impressions; launched waitlist to 400+ Flair Studio users.

PROJECTS

Lattice | Hack the North Finalist Winner | C++, C#, OpenCV, Kinect, HoloLens SDK

- Built a real-time volumetric capture system with three Xbox Kinect v2 sensors; fused via multi-view extrinsic calibration (marker-based init + ICP) into a single 3D point cloud streamed to HoloLens at ~23FPS.
- Designed an event-driven client-server pipeline (C++ capture clients; C# HoloLens viewer) with OpenCV preprocessing and denoising; mitigated LiDAR interference.

DataBot | LAHacks (UCLA) 2025 Finalist | Go, Google Gemini, Next.js, TypeScript

- Created an AI database analysis platform for PostgreSQL/MySQL/MongoDB; assigns a Green Score and auto-optimizes slow queries, storage bloat, and unused data.
- Used Gemini for live SQL rewrite suggestions, cutting latency 40% and CPU 25% in benchmarks.
- Architected a modular Go backend + streaming pipeline (millions events/day); Next.js + React dashboard visualization

TECHNICAL SKILLS

Languages: Java, C/C++, C#, Python, GoLang, JavaScript, HTML/CSS, Assembly, COBOL

Frameworks: React, Next.js, Node.js, Express, Flask, jQuery, REST API, Tailwind, Bootstrap, Three.js, TypeScript

Developer Tools: Git, Linux, Docker, Figma, Excel, MongoDB, AWS, Google Cloud, Azure, PostgreSQL/MySQL

Libraries: pandas, NumPy, Matplotlib, Keras, OpenCV, Jupyter, TensorFlow, PyTorch

Soft Skills: Communication, Leadership, Team Player, Troubleshooting, Problem Solving, Decision Making, Attention to Detail

Other: Photoshop, Video Editing, Graphic Design, Football