

Devarsh Prajapati

+1 437-264-3155 | devarsh.sp13@gmail.com | linkedin.com/in/devarsh | github.com/devarsh05 | devarsh05.github.io

EXPERIENCE

Automators Lab | AI/ML Intern

May 2025 – Sep 2025

- Built a **LangGraph multi-agent system** that automated the proposal workflow (call transcription, insight extraction, drafting, and task routing), cutting turnaround from **hours to under 10 minutes** by routing each step to GPT, Claude, Gemini, or LLaMA to balance cost and quality.
- Integrated **10+ external services** (Slack, Jira, Google Drive, CRMs) into AI agents via **custom MCP servers** so they could act on live business data, securing each with OAuth 2.0 and AES-256 per-user credential encryption.
- Built a **RAG pipeline** so agents could answer reliably from company documents, raising how often they retrieved the right source from **60% to 84%** by tuning chunking and top-k against a 50-question labeled eval set (OpenAI embeddings, Qdrant).
- Automated lead-generation prep with an **event-driven n8n workflow** linking web scrapers, LLM enrichment, and CRM sync, cutting manual outreach work by **~40%**.

PROJECTS

ContextCode [Live Demo](#) | [GitHub](#) | *FastAPI, Celery, Tree-sitter, ChromaDB, Next.js*

2026

- Built an **AI codebase assistant** that indexes Python/JS/TS repos with **Tree-sitter**, chunking by function, class, and module so retrieval never splits a function mid-body — verified by substring and no-overlap tests across all three languages.
- Grounded every answer in retrieved code with **exact file:line citations** built from the index (never model text), returning honest refusals on weak context and skipping the LLM entirely on empty retrieval (**zero cost and latency**).
- Scaled indexing by offloading CPU-bound parsing and embedding to **Celery workers** behind an async FastAPI, keeping the API responsive while streaming **7-stage progress over SSE**, with swappable local-MiniLM or OpenAI embeddings.
- Deployed a **public demo** (Railway + Vercel) hardened with Cloudflare Turnstile, Redis quotas, and per-IP rate limits, backed by **274 automated tests** (pytest, Vitest, Playwright).

SixRise [Site](#) | [GitHub](#) | *FastAPI, TypeScript, Postgres, Redis, Shopify*

2026

- Built an **AI visibility platform** for Shopify stores that measures how often their products are recommended by AI shopping engines vs. competitors, running a closed audit → fix → verify loop scored on a **24-query buyer-intent panel** via Perplexity Sonar.
- Enforced a **hallucination-safe publishing path** so no AI-generated claim reaches a live store: gpt-5-nano proposes typed candidates (strict Structured Outputs), **deterministic code verifies each value** exists in the merchant's catalog, and published copy is string-templated.
- Architected a **two-service Shopify app** that isolates merchant credentials from the AI agent — a thin TypeScript shell owns OAuth and tokens while the **Python FastAPI + arq** agent fetches short-lived tokens on demand; deployed on Northflank + Neon + Upstash.
- Optimized the async job model to cut idle infrastructure cost, raising the arq worker poll interval **0.5s → 15s** after calculating the default would exhaust Upstash's **500K-command/month** quota in ~3 idle days.

TECHNICAL SKILLS

Languages: Python, TypeScript, JavaScript, SQL, Java, R

AI/ML: LangGraph, LangChain, RAG, Model Context Protocol (MCP), embeddings, vector search, structured outputs, prompt engineering; OpenAI, Anthropic, Gemini, Perplexity, LLaMA/Ollama

Frameworks & Backend: FastAPI, Next.js, React, Node.js, Celery, arq, Flask, Streamlit

Data & Infra: PostgreSQL, Qdrant, ChromaDB, Redis, Docker, Railway, Vercel, Northflank, Google Cloud

Concepts: System Design, Multi-Agent Systems, REST API Design, Distributed Systems, CI/CD

EDUCATION

York University – BSc. Honours in Computer Science

Toronto, ON | Expected Dec 2026

- Relevant Coursework: Machine Learning, Artificial Intelligence, Data Structures & Algorithms, Databases, Operating Systems, Computer Networks, Software Design.