Harshaan Singh Babra

BACKEND & AI PLATFORM ENGINEER · TEAM LEAD

Vancouver, BC · harshaan.babra@gmail.com · harshaan.dev · linkedin.com/in/harshaan-babra

SUMMARY

Backend and AI platform engineer, currently Team Lead at ApyHub, leading 6 engineers and owning release management for a multi-tenant AI platform serving **800–1,000 daily active users**. Builds retrieval (RAG) and document-intelligence services end to end in **Python** and **TypeScript** — including two production microservices designed, built, and owned solo — and shipped the cost-attribution layer that gave the business its first per-tenant view of AI spend.

CORE SKILLS

AI & Data: RAG pipelines, vector search (Milvus), embeddings and chunking, multi-provider LLM integration (Azure OpenAI, OpenAI, Anthropic, Gemini), prompt versioning, LLM cost telemetry, OCR, Whisper speech-to-text, PySpark, Databricks

Backend: Python (Flask, FastAPI, Django), Node.js / TypeScript (Express), REST API design, WebSockets, async workers, job queues, PostgreSQL, Redis, S3, schema migrations

Cloud & Delivery: AWS (EC2, S3, ECR, Glue, Athena), Azure, Docker, GitLab CI/CD, pytest, structured logging and tracing, Agile/Scrum, Jira, code review, release management

Frontend: React, TanStack Router/Query, Material UI, Tailwind CSS, Vite

Security: OAuth2/SSO, TOTP two-factor authentication, JWT issuance and rotation, XSS and dependency vulnerability remediation — CompTIA Security+ (SY0-701)

EXPERIENCE

ApyHub — Vancouver, British Columbia

Software Engineer → Team Lead, Software Engineering — *Feb 2026 – Present*

- **Team leadership.** Lead 6 engineers across the platform's core services — sprint scoping, code review, and every **dev → stage → production** release — after being promoted from an individual-contributor role within the first year.
- **RAG platform.** Lead contributor on the multi-tenant RAG service (**Python/Flask**, **PostgreSQL**, **Milvus**): built the intent router, chat orchestrator, multi-turn threading, and per-tenant isolation behind every AI feature in the product.
- **LLM cost visibility.** Designed and shipped cost attribution across three services — per-request token and dollar emission, per-job costing, and a spend dashboard by team and organization — giving the business its first per-tenant view of AI spend.
- **Model flexibility.** Built multi-provider LLM support (Azure OpenAI, OpenAI, Anthropic, Gemini) with per-bot model selection, versioned prompts, and backoff retry on **429**s, cutting failed calls under load.
- **Security workstream.** Delivered **Google OAuth SSO**, **TOTP 2FA** with recovery, JWT rotation, and password-reuse prevention; remediated reported **XSS** and dependency vulnerabilities across Jira epics ADN-504/512/515/563.
- **Services owned solo.** Designed, built, and own two production microservices: **Phoenix** (transcription — S3 multipart upload, language detection, LLM filtering, per-job costing) and **audio-services** (real-time **WebSocket** transcription, Redis job tracking, spend dashboard).
- **Document intelligence.** Built a SCORM extraction pipeline solo (**FastAPI**, Postgres **LISTEN/NOTIFY** queue, Playwright, ffmpeg, Whisper) and an in-house alternative to Unstructured.io's partitioning API, replacing a paid vendor dependency.
- **Engineering standards.** Introduced pytest and Bun integration suites, structured tracing, and health/version endpoints across owned services; cut container image size by moving inference to a CPU-only build.

Renisu Precious Metals Inc. — Vancouver, British Columbia

Full Stack Software Developer — *Aug 2023 – Nov 2024*

- Built a responsive **Django** e-commerce platform, resolving cart inconsistencies and cutting load times **20%** for a **15%** gain in user retention.
- Designed **RESTful APIs** for third-party integrations, reducing data processing time **20%**; optimized **SQL** indexing to cut query execution **30%**.
- Deployed on **AWS EC2** behind an Application Load Balancer, sustaining **99.9% uptime** through traffic spikes.

Infosys Ltd. **Vancouver, British Columbia**

Associate Software Developer *Aug 2022 – Aug 2023*

- Optimized **ETL/ELT** pipelines on **AWS** (Lake Formation, S3, Glue, Athena), cutting processing time **15%**.
- Automated extraction and loading via a **PySpark** and **Databricks** pipeline, reducing manual processing **30%** and sustaining a **95%** client satisfaction rating.

EDUCATION & CERTIFICATION

University of Alberta **Edmonton, Alberta**

B.Sc. Computer Science, Minor in Mathematics *Sep 2018 – Jun 2022*

CompTIA Security+ (SY0-701) **Sep 2024**