

Suraj Van Verma

+1 514-641-6429 | suraj.verma@mail.mcgill.ca | [linkedin.com/in/bythebug](https://www.linkedin.com/in/bythebug) | github.com/bythebug
Authorized to work in Canada · No sponsorship required

EDUCATION

McGill University

Master of Science, Computer Science

Montreal, QC

Aug. 2023 - May 2025

Veltech University

Bachelor of Technology, Computer Science

Chennai, TN

Aug. 2019 - May 2023

EXPERIENCE

Software Engineer, AI/ML

Plumfind

June 2025 - Mar. 2026

Montreal, QC

- Fine-tuned **Prithvi**, NASA's geospatial foundation model, on multi-spectral satellite imagery to classify **13 crop types** across **46km² per minute**, automating agricultural land analysis that previously required manual expert review.
- Built a **FAISS** RAG pipeline with semantic chunking that reduced knowledge retrieval from **5 min to 15 sec** (95% faster) across **1,000+** documents, adopted by 3 enterprise teams.
- Architected a serverless **AWS Lambda** pipeline ingesting **10TB+** of NASA satellite imagery that slashed ML data prep from **3 days to 2 hours**, enabling automated weekly model retraining.
- Fine-tuned **Stable Diffusion** with a custom LoRA adapter on **5,000+** domain-specific samples to generate on-demand print designs, eliminating manual design iteration at zero marginal cost.
- Integrated **Whisper** ASR with Gemini to ship a bilingual voice AI agent to production, achieving **sub-2s** end-to-end latency across real-time English and French queries.

Software Engineer, Research

McGill University

Aug. 2023 - May 2025

Montreal, QC

- Designed a modular **Spring Boot** backend with 6 loosely coupled service modules so new language plugins extend the system without touching core code, compressing integration time from weeks to days.
- Built real-time collaboration for **20+ concurrent users** using conflict-free sync and role-based access control, passing a formal security audit with zero findings.
- Designed an extensible plugin layer so any LMS platform or grading engine integrates with zero core changes, enforcing **GDPR/FERPA**-compliant audit logging via **PostgreSQL**.

PROJECTS

Keeto: LLM Observability | Python, FastAPI, PostgreSQL, OpenTelemetry, Docker | [PyPI](#)

Jul. 2026

- Engineers have no visibility into LLM costs and latency. Shipped **keeto** to **PyPI**, a zero-config library auto-tracing **12 AI SDKs** with cost tracking and dashboards at **sub-millisecond** overhead.

ContextOS | [pgvector](#), FastAPI, PostgreSQL, Redis, Docker | [GitHub](#) | [PyPI](#)

May 2026

- AI agents lose all context between sessions. Built a model-agnostic persistent memory layer with hybrid semantic search and smart decay scoring, deployed to **GCP** with Python and TypeScript SDKs.

ML Forge | [XGBoost](#), [MLflow](#), [scikit-learn](#), FastAPI, Docker, AWS | [GitHub](#)

Apr. 2026

- Ad-hoc model selection produces irreproducible results. Shipped an ML platform to **AWS ECS** covering **5 model families** with an **MLflow** leaderboard and A/B testing, making every experiment reproducible.

LLM Studio | [PyTorch](#), [HuggingFace](#), FastAPI, Prometheus, Docker | [GitHub](#)

Feb. 2026

- Fine-tuning LLMs requires stitching 5+ separate tools. Built a unified fine-tuning platform across **6 Docker microservices** with **28 API endpoints**, shrinking iteration from days to **4 hours** with full production observability.

TECHNICAL SKILLS

Languages: Python, TypeScript, SQL, Java

AI/ML: RAG, LangChain, LangGraph, PyTorch, HuggingFace, OpenAI, Anthropic, Gemini, LoRA, MLflow

Backend & Infra: AWS, Docker, Kubernetes, FastAPI, GCP, REST APIs, Spring Boot, GitHub Actions

Frontend: React, Next.js, React Native, Tailwind CSS

Databases: PostgreSQL, Redis, MongoDB, pgvector

CERTIFICATIONS

AWS Certified Machine Learning Engineer, Associate

Amazon Web Services

PUBLICATIONS

A Generative AI Empowered Chatbot: System Usability and Student Teachers' Experience

Nikou, S.A., Guliya, A., **Van Verma, S.**, Chang, M. | *Int'l Conference on Intelligent Tutoring Systems (ITS)* | 2024 | **12 citations**

A Modular Backend for a Collaborative Educational Modeling Tool

Van Verma, S. | *M.Sc. Thesis, McGill University* | 2025

LEADERSHIP

Google Developer Student Clubs (GDSC) Lead, Veltech University (2022 - 2023) | Led a **1,000+ member** developer community; organized hackathons and **GDSC WOW**, Google's flagship tech event.