SHREYA SAVANT

shreyasavant10@gmail.com | Phone | Portfolio | github.com/shreyasavant | linkedin.com/in/shreyasavant | Montreal, QC

Professional Summary

Machine Learning Developer with hands-on experience developing production GenAI systems, LLM and multi-agent AI architectures. Proven track record deploying RAG pipelines, fine-tuning LLMs, and building cloud-native ML applications on AWS and Azure. Skilled in translating business challenges into scalable AI solutions through cross-functional collaboration.

Technical Skills

- **GenAI & LLMs:** LangChain, LlamaIndex, Autogen; GPT-4, Claude, Gemini, Llama, Mistral; prompt engineering, RAG, fine-tuning, RLHF; vector databases (Pinecone, Weaviate, Redis)
- **ML Frameworks:** PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers; CUDA optimization
- **Cloud & MLOps:** AWS (SageMaker, Lambda, S3), Azure AI, GCP; Docker, Kubernetes, CI/CD (GitHub Actions); MLflow, Weights & Biases, TensorBoard
- **Data Engineering:** PySpark, Databricks, Apache Kafka, Airflow; PostgreSQL, MongoDB, Solr
- **Programming:** Python, Rust, TypeScript, Go, SQL; FastAPI, Flask, Django, React, Node.js

Professional Experience

Machine Learning Developer — Innovation Lab, Concordia University January 2024 – December 2025

- Led 4-person cross-functional team developing multilingual conversational AI system for cyber-violence prevention, deployed to 200+ youth participants
- Architected production RAG pipeline using Gemini LLMs with Redis-backed conversation state management and real-time context retrieval
- Designed novel prompt-driven persona framework achieving 87% user satisfaction through systematic prompt engineering
- Built a MVP for automated UI/UX testing system using GPT-4o and event log analysis, reducing QA cycles by 50%
- Developed AI-powered assessment tool using Llama-based topic modeling and sentiment analysis, improving document review efficiency by 60%

Machine Learning Engineer — 7Dish (MITACS Accelerate Program) June 2023 – January 2024

- Designed and deployed production-grade RAG architecture using LangChain and GPT-4 for domain-specific nutrition question-answering called 7Chef, Spearheaded the project with design and implementation choices for 10+ user personas
- Evaluated 5+ embedding models (BERT, SentenceTransformers) on Databricks, improving semantic retrieval accuracy by 21%, Preprocessing on large text corpora for enhanced representations of over 100,000+ recipe and nutrition data
- Built a real-time ETL pipeline with structured LLM outputs for automated knowledge base enrichment
- Conducted systematic ablation studies on chunking strategies, retrieval methods, and re-ranking approaches
- Published internal technical report on RAG optimization methodologies adopted across product teams

Software Development Engineer — SpokenWeb, SSHRC/Concordia University June 2024 – January 2025

- Architected cloud-native search infrastructure processing audio recordings using Apache Solr, Ruby on Rails, and Docker
- Developed automated data pipelines for metadata processing and indexing, deployed across 13+ universities
- Led CI/CD workflows and containerization, reducing deployment time by 80%
- Presented technical architecture, demo and scalability solutions at Blacklight Summit 2025

Graduate Research Assistant — Concordia University September 2025 – Present

- Developing multi-agent reinforcement learning systems for distributed decision-making in critical infrastructure
- Investigating coordination mechanisms and optimization algorithms for agent collaboration under uncertainty
- Building prototypes in PyTorch for real-world deployment with industry partners

Data Science Intern — Visive Inc. August 2022 – December 2022

- Built a real-time image search engine using TensorFlow on AWS, reducing false positive rate by 16% for a retail firm
- Developed a fraud detection pipeline using statistical modelling, achieving 92% production accuracy
- Engineered ETL workflows processing 1M+ daily transactions with optimized database schemas

Software Developer Intern — ProfCess May 2021 – August 2021

- Developed ML-driven recommendation system using scikit-learn, improving relevance by 40%
- Enhanced frontend performance and mobile UX across core user flows

Projects

Comparative Analysis of Vision Architectures for Medical Imaging – GitHub

- Evaluated 7 state-of-the-art models (ResNet, EfficientNet, DenseNet, DeiT, Swin, ViT) on 212K tissue images
- Documented performance-efficiency tradeoffs between CNNs and Vision Transformers for pathology classification
- Implemented efficient training pipelines in PyTorch with mixed-precision training and data augmentation

SWin: A Sliding Window Summarization Approach for Coherent LLM-driven Dialogue Systems
Conference on Artificial Intelligence (CAI) 2026

- Proposed novel memory mechanism for maintaining long-context coherence in multi-turn LLM conversations
- Systematic evaluation across GPT-4, Claude, and Llama models demonstrating 23% improvement in coherence metrics (SOTA: F1, BScore, BLEU1/2)

Mitigating Causal Bias in LLMs via Potential Outcomes Framework
European Chapter of the Association for Computational Linguistics (EACL) 2026

- Developed causal inference framework for detecting and mitigating bias in LLM reasoning
- Applied counterfactual analysis to identify intersectional bias patterns across demographic dimensions

SpokenWeb Search Engine: Open-Source Digital Archive Platform
Software Release, Zenodo 2024 — DOI: 10.5281/zenodo.17705608

- Full-stack search platform for academic audio archives using Blacklight, Solr, Rails, Docker
- Presented technical implementation at Blacklight Summit 2025

Education

Concordia University, Montreal, QC
Doctor of Philosophy, Information and Systems Engineering September 2025 – Present

- **Research Focus**: Multi-agent AI systems for infrastructure optimization using reinforcement learning in Energy Systems
- **VoltAge Doctoral Fellowship** – Funded research on distributed AI decision-making for Power Grids

Master of Applied Science (Thesis), Quality Systems Engineering January 2023 – August 2024

- **Thesis**: Memory Mechanisms for Coherent LLM-Based Dialogue Systems (GPT-4, Claude, Llama)
- **MITACS Accelerate Scholar** – Industry-partnered research on production LLM deployment

D. Y. Patil College of Engineering and Technology, India
Bachelor of Technology, Computer Science and Engineering August 2018 – June 2022

- **Advanced Certifications**: Applied ML (BITS Pilani), Data Analytics (IIT Kharagpur)
- **Leadership**: Lead, Google Developer Student Club — Coordinator, Linux Club