

SIMON BOUCHARD

Montréal, QC • 438-392-6707 • simon.bouchard31@gmail.com • simonbouchard.space • linkedin.com/in/simon-bouchard-54580b339

RÉSUMÉ PROFESSIONNEL

Développeur en machine learning avec expérience en production dans la construction de systèmes de recommandation et d'applications d'IA agentique. Développement d'un moteur de recommandation en temps réel avec architecture à deux facteurs (collaborative filtering ALS et attention-pooled embeddings) atteignant une latence inférieure à 50ms, et d'un système multi-agent conversationnel avec routage d'intention et pipeline de recommandation en plusieurs étapes. Certifications AWS Machine Learning - Specialty et Generative AI - Professional. Développeur autodidacte avec solide expertise en développement backend, infrastructure cloud et IA appliquée. Formation en administration des affaires (BAA, HEC Montréal) permettant de faire le pont efficacement entre implémentation technique et exigences d'affaires. À la recherche d'un poste en ML pour appliquer les compétences démontrées en environnement de production.

CERTIFICATIONS

[AWS Certified Generative AI - Professional \(AIP-C01\)](#) | Early Adopter | Jan 2026

[AWS Certified Machine Learning - Specialty \(MLS-C01\)](#) | Score : 890/1000 | Jan 2026

COMPÉTENCES TECHNIQUES

Machine Learning & IA : PyTorch, scikit-learn, FAISS, LangGraph, Sentence-Transformers, collaborative filtering, réseaux de neurones, mécanismes d'attention, systèmes RAG, framework agentiques, déploiement de modèles, inférence en temps réel, MLOps

Cloud & Infrastructure : AWS, Docker, Nginx, CI/CD (GitHub Actions), Linux/Bash, Redis

Développement Backend : Python, FastAPI, Git, APIs RESTful, architecture microservices, tests unitaires

Ingénierie de Données : SQL (MySQL), NoSQL (MongoDB), pipelines ETL/ELT, Kafka, Spark, pandas, NumPy

Observabilité : Prometheus, Grafana, Jaeger, OpenTelemetry

Langues : Français (langue maternelle), Anglais (courant)

PROJETS TECHNIQUES

Système de Recommandation ML en Production | recsys.simonbouchard.space

- Développement d'un API de recommandation en temps réel atteignant une latence inférieure à 50ms pour des recommandations de livres personnalisées, déployé avec une architecture *model server* répartie en conteneurs dédiés, permettant le déploiement et la mise à l'échelle indépendants de chaque composant
- Architecture d'un système de recommandation à deux facteurs : collaborative filtering ALS pour les recommandations warm et la similarité comportementale, subject embeddings avec attention-pooling entraînés avec double perte (régression et contrastive) pour les recommandations cold-start et la similarité de contenu, avec similarité hybride combinant les deux approches
- Construction d'un pipeline MLOps automatisé avec porte qualité recall@30 bloquant la promotion d'artéfacts en cas de régression, versionnage des artéfacts, notifications d'échec, et suivi de la version active par fichier pointeur pour un rechargement à chaud sans interruption sur 5 serveurs de modèles

- Conception et gestion d'une base de données MySQL avec pipelines d'ingestion et optimisation du schéma pour les charges de travail ML
- Intégration de Meilisearch pour moteur de recherche full-text sur 250 000+ livres avec temps de réponse en millisecondes
- Déploiement sur infrastructure cloud avec automatisation CI/CD, sécurité SSL, surveillance, et durcissement pare-feu

Système de Recommandation Conversationnel Multi-Agent |

recsys.simonbouchard.space/chat

- Développement d'un système multi-agent en production utilisant LangGraph avec routage d'intention vers quatre agents spécialisés (recommandation, docs, web, réponse) ; l'agent de recommandation exécute un pipeline en plusieurs étapes (planification → récupération → sélection → réponse) intégrant APIs de recommandation, recherche sémantique, personnalisation utilisateur et sources de connaissances externes avec réponses en streaming
- Implémentation de génération augmentée par récupération (RAG) avec vector store personnalisé en mémoire utilisant l'index FAISS HNSW pour recherche sémantique dans le catalogue de livres, ancrant les réponses du chatbot pour réduire les hallucinations
- Conception d'un pipeline de génération d'embeddings event-driven utilisant Kafka pour la mise en file d'attente de messages et Spark pour le traitement par lots des descriptions générées par LLM
- Mise en œuvre d'une suite de tests et évaluations (unitaire et intégration) assurant la fiabilité et la qualité du système et permettant une itération rapide.

EXPÉRIENCE DE TRAVAIL

Commis d'Épicerie | Métro Renaud Poissant, Montréal | Été 2022

Copropriétaire | Gazonnet Royal (Entreprise d'Entretien Paysager) | 2016-2021

FORMATION

Baccalauréat en Administration des Affaires (BAA), Programme Bilingue | 2021-2025

HEC Montréal – Spécialisation en Technologies de l'Information

DÉVELOPPEMENT PROFESSIONNEL

Machine Learning Avancé : MIT Introduction to Deep Learning, fast.ai Practical Deep Learning for Coders (Parts 1 & 2)

IA & IA Générative : DeepLearning.AI – AI for Everyone, Generative AI for Everyone, Agent AI

Génie Logiciel : MIT The Missing Semester of Your CS Education

Développement Backend : FreeCodeCamp (Backend APIs, Quality Assurance, Information Security, Scientific Computing, Data Analysis, Machine Learning with Python)