

Beedi Goua

 Portfolio |  GitHub |  LinkedIn |  beedi.goua@eleve.ensai.fr |  +33 779 872 812

Profil

Ingénieur Data Science en début de carrière, passionné par le machine learning, le NLP et l'IA générative. Expérience concrète en conception de pipelines RAG, modélisation prédictive et analyse comportementale. Je souhaite appliquer mes compétences à des cas d'usage réels en développant des solutions robustes, explicables et adaptées aux besoins métier.

Expérience Professionnelle

Stage – IA Générative & RAG appliquée aux cas métiers (Finance, RSE, Marketing) Avr 2025 – Aujourd'hui
Square Management – Square Research Center

- Conception d'un pipeline RAG industrialisable avec benchmark de 189 configurations (7 OCR × 3 chunking × 3 embeddings) sur corpus multi-métiers
- Recherche hybride (vectorielle + BM25), reranking LLM et découpage sémantique optimisé
- Évaluation automatique via RAGAS, TruLens, LLM-as-a-Judge (faithfulness, relevance, traçabilité)
- Prototype Streamlit avec interface métier (Finance, RSE, Marketing) et logs de performance intégrés
- **Stack** : Python, LangChain, Docling, OpenAI API, ChromaDB, Streamlit, GitHub Actions

Stage – Analyse comportementale & pollution urbaine Juin 2024 – Sept 2024
Ville de Paris – Agence de la Mobilité

- Détection des flottes les plus polluantes via clustering supervisé (cartographie + scoring sectoriel)
- Appariement automatisé de bases hétérogènes (SIRENE, données régionales) avec 95% de match
- Recommandations opérationnelles pour orienter la politique de mobilité durable
- **Stack** : Python, scikit-learn, pandas, DigDash, API SIRENE

Formation

Diplôme d'ingénieur – Data Science & IA 2022 – 2025
ENSAI – École Nationale de la Statistique et de l'Analyse de l'Information (grande école, concours)

- **Spécialisation IA** : machine learning, deep learning, NLP avancé, LLMs, RAG, vector stores
- **Stats & maths** : GLM, séries temporelles, calcul stochastique, statistiques bayésiennes
- **Ingénierie** : Hadoop, Spark, SQL, Python, Java, APIs

Projets

Système de recommandation musicale hybride – Approche collaborative et par contenu Janv – Fév 2024
→ *pandas, Surprise, Streamlit, fallback dynamique, KNNBasic*

Développement d'un moteur de recommandation combinant filtrage collaboratif (KNN) et similarité de contenu, avec mécanisme de fallback dynamique pour les nouveaux utilisateurs.

ReviewGuardian – Détection de toxicité dans les avis avec explication locale Mars 2024 – Mai 2024
→ *scikit-learn, SHAP, FastAPI, Streamlit*

Modèle MultinomialNB avec explication SHAP, déployé sous forme d'API FastAPI et interface Streamlit interactive.

Calibration bayésienne – Modèle Lorenz-96 Oct 2024 – Mars 2025
→ *Python, NumPy, matplotlib, ABC-SMC, ABC-MCMC*

Inférence bayésienne pour calibrer les paramètres du modèle chaotique Lorenz-96, avec analyse des performances et du coût de calcul.

InsightDetector – Détection d'hallucinations dans des textes générés Déc 2024 – Mars 2025
→ *BART, BERTScore, spaCy, Streamlit, RSS, OpenAI, LLM-as-a-judge*

Pipeline de résumé + fact-checking + annotation utilisateur ; détection d'hallucinations sur 300+ articles (actualité, politique) via interface Streamlit, code prévu en open-source.

FraudTrack360 – Détection de fraude transactionnelle explicable Janv 2025 – Mars 2025
→ *pandas, scikit-learn, LSTM, SHAP, FastAPI, Docker, GitHub Actions, AWS EC2*

Modèle LSTM sur séquences transactionnelles avec scoring d'anomalie ; API conteneurisée, CI/CD GitHub + déploiement EC2 ; tableau de bord explicatif Streamlit.

Compétences

Langages : Python, R, SQL, SAS

Machine Learning / Deep Learning : scikit-learn, XGBoost, TensorFlow, CNN, LSTM, BERT

NLP / IA générative : Hugging Face, LangChain, OpenAI API, Whisper, BART

Déploiement / Ingénierie : FastAPI, Docker, Git, CI/CD, REST API

Cloud / MLOps : AWS, GCP, automatisation de pipelines

Visualisation : Matplotlib, Seaborn, ggplot2, Streamlit

Interprétabilité : SHAP, LIME

Langues

Français (natif), Anglais (professionnel)