

JOY MIKHAEL

AI / Machine Learning Engineer

Bane, North Lebanon | +961 70 213 365 | joymikhael1701@gmail.com

linkedin.com/in/joymikhael17 | github.com/JoyMikhaelLeb

PROFESSIONAL SUMMARY

AI / Machine Learning Engineer with 6+ years of production Python experience across applied ML, LLM integration and large-scale data infrastructure. Hands-on experience building RAG pipelines with ChromaDB and sentence-transformers, integrating LLMs via the Anthropic SDK with native tool use, and developing supervised and unsupervised models with scikit-learn and XGBoost. Strong engineering foundation: FastAPI microservices on GCP, ETL pipelines processing ~1M records per month and multi-database architectures across PostgreSQL, MongoDB, Firestore and Supabase.

TECHNICAL SKILLS

AI / Machine Learning: scikit-learn, XGBoost, TensorFlow, NLP, supervised and unsupervised learning, regression, classification, clustering, feature engineering, model evaluation, hyperparameter tuning

LLMs and RAG: Anthropic SDK, native tool use, retrieval-augmented generation, ChromaDB, sentence-transformers, embeddings, Pydantic-structured outputs, prompt engineering

Languages and Core Libraries: Python, SQL, Pandas, NumPy, Matplotlib

Backend and APIs: FastAPI, REST API design and integration, Streamlit, Docker, Git, GitHub

Databases: PostgreSQL, MongoDB, Google Cloud Firestore, Supabase, SQLite

Cloud and Infrastructure: Google Cloud Platform (GCP), Cloud Run, Firebase

Data Engineering: ETL pipeline design, data validation, data cleaning, data transformation, web scraping (Selenium, nodriver, Playwright, BeautifulSoup, Scrapy)

Analytics and Reporting: Operational dashboards, trend analysis, Microsoft Excel, Google Sheets, Power BI

PROFESSIONAL EXPERIENCE

Python Backend Data Engineer | RanDee (Remote)

June 2020 – Present

- Engineered ETL pipelines processing approximately 1 million records per month, transforming raw web data into analytics-ready datasets across Google Cloud Firestore, Supabase, MongoDB and PostgreSQL.
- Built and maintained scalable data acquisition infrastructure using Selenium, nodriver, Playwright, BeautifulSoup and Scrapy, including bot-detection bypass and structured logging for production reliability.
- Developed FastAPI microservices deployed on GCP Cloud Run to automate data ingestion, transformation and synchronization across internal systems.
- Authored Python scripts with Pandas and NumPy for automated reporting, trend analysis and feature preparation supporting downstream analytics and modeling workflows.
- Implemented error handling and Slack and email alerting to maintain pipeline reliability at scale; collaborated with cross-functional teams to define data requirements and resolve cross-system inconsistencies.

Python Automation and Data Consultant | Freelance

January 2023 – Present

- Designed and evaluated ML models using scikit-learn for classification and regression tasks; delivered prototypes with documented evaluation metrics including accuracy, RMSE and train/test validation.

- Led a research data analysis engagement for a PhD candidate: processed unstructured text with Python NLP techniques, performed word frequency and feature analysis and applied regression modeling to derive final equations.
- Converted a client's Java codebase to Python, refactored logic into clean maintainable scripts and delivered hands-on scikit-learn training including model evaluation workflows.
- Delivered data cleaning, validation and reporting automation solutions for multiple clients using Python, Pandas and Excel.

Junior Software Developer | Menalytix

February 2019 – May 2020

- Built automated data collection pipelines via web scraping, structuring raw data into clean queryable formats for business analytics.
- Integrated REST APIs to automate data ingestion and synchronization across internal systems, reducing manual effort by approximately 70 percent.
- Designed MongoDB schemas to store and manage large volumes of structured datasets.
- Contributed to backend data workflows supporting reporting and business intelligence features.

SELECTED PROJECTS

Tasklet RAG Pipeline | *Open-Source Contribution*

- Contributed retrieval-augmented generation capabilities to Tasklet, an open-source B2B SaaS support agent built on the Anthropic SDK with native tool use, SQLite and FastAPI.
- Implemented a search_knowledge_base tool using sentence-transformers for embedding generation and ChromaDB as the vector store, enabling semantic search across knowledge documents.
- Repository: github.com/JoyMikhaelLeb/tasklet-project (PR submitted to upstream [larawehbe/tasklet-project](https://github.com/larawehbe/tasklet-project)).

Food Price Forecasting with XGBoost | *WFP Lebanon Open Data*

- Built an end-to-end ML pipeline on World Food Programme Lebanon food price data: exploratory data analysis, feature engineering, model development, hyperparameter tuning and model persistence.
- Compared multiple regression approaches and tuned an XGBoost model with documented evaluation metrics.

Resume Screening Pipeline | *LLM Tool Use*

- Developed a deep-dive screening stage using the Anthropic SDK with native tool use and Pydantic-validated structured outputs to evaluate candidate resumes against job criteria.

Street Food Clustering | *Unsupervised Learning*

- Designed an unsupervised learning pipeline using clustering and dimensionality reduction techniques to identify natural groupings in Lebanese street food data.

LinkedIn Activity Monitoring Platform | *Private Project*

- Delivered a TweetDeck-style LinkedIn post monitoring system using Playwright and BeautifulSoup for data acquisition, SQLite for persistence, FastAPI for the backend and Streamlit for the UI.

EDUCATION

MSc Cyber Security and Cyber Defence | USEK / Université Bretagne-Sud 2019 – 2021

BSc Computer Science | USEK – Université Saint-Esprit 2015 – 2018

LANGUAGES

Arabic (Native) | English (Professional) | French (Intermediate)