Tony Boutros

Junior AI Engineer

Lebanon, Kousba | (+961) 70 310 923 | tonyboutros77@gmail.com | GitHub: github.com/Tons-7

ABOUT

Hands-on experience designing and integrating machine learning and LLM-based systems into real-world applications. Strong background in NLP, Retrieval-Augmented Generation (RAG) pipelines, and scalable AI inference architectures using PyTorch and Hugging Face.

EDUCATION

American University of Culture and Education 2022 - 2025

BSc in Computer Science | GPA: 3.75

Certificates:

- *Natural Language Processing (NLP) in Python – Udemy (2025)*

SKILLS

Languages: Java, Python, JavaScript.

Databases: MySQL, PostgreSQL.

AI/ML: PyTorch, Hugging Face Transformers, PEFT, LLMs, spaCy.

Frameworks/Tools: Spring Boot, Hibernate, FastAPI, Flask, React, Git, REST APIs, LangChain, LangGraph.

EXPERIENCE

CollideAI – Junior AI Engineer (Contract) | collideai.app December 2025

- Designed and implemented a Retrieval-Augmented Generation (RAG) memory system using vector databases and embeddings to maintain long-term conversational context in production.
- Built a structured story-state architecture to persist canonical characters, relationships, plot, and scene data across conversations.
- Developed schema-driven LLM pipelines to reliably extract, validate, and apply incremental canonical updates.
- Implemented automated conversation summarization with long- and short-term memory compression to preserve context over extended dialogues.
- Integrated multi-retriever semantic search pipelines to ensure narrative consistency and accurate context retrieval.

PROJECTS

AI Chatbot Backend

- Built complete backend system for AI chatbot with user authentication and persistent conversation storage.
- Integrated Hugging Face text generation model via FastAPI microservice to generate AI-powered responses.
- Implemented secure authentication with JWT tokens, password reset functionality, and user session management.
- Designed relational database schema using Spring Data JPA for users, conversations, messages, and security tokens.
- Implemented a separate FastAPI-based inference service, decoupling AI model execution from the core backend.

- Technologies: Spring Boot, PostgreSQL, FastAPI, Hugging Face, Hibernate.

NLLB Transformer Pipeline for Translation

- Developed an end-to-end machine translation pipeline using the NLLB model with LoRA optimization.
- Implemented spaCy-based text preprocessing and built modular architecture for data preprocessing, training, and inference.
- Technologies: PyTorch, Transformers, PEFT, spaCy, Weights & Biases (WandB).

AI-Powered Image Classification System

- Built full-stack web application with React frontend and Flask backend using CNN+FPN architecture.
- Achieved 95% classification accuracy across 13 categories.
- Integrated batch processing, real-time image manipulation, and REST APIs.

LANGUAGES

Arabic: Native

English: Fluent

French: Basic