Montreal, QC, Canada

rani@baghezza.ai

Education

Computer Science MSc
INSA de Lyon, France

Deep Learning PhD
UQAC, Canada

Languages

French, English

Rani Baghezza, PhD

AI Integration Engineer

I guide organizations through their AI transformations, with a focus on AI Strategy, Internal Assistants, and Automation.
I make sure that your AI initiatives are aligned with your broader strategy. I guide you through the technical choices with a hands-on approach, all the way to the final implementation of a production-ready system.

Technical Skills

Languages: Python, JS, Bash, C++, Java
Gen AI: LLMs, RAG, Vector DB
Databases: Postgres, Neo4j, SQL, Cypher
Deep Learning: Tensorflow, PyTorch, ONNX
API: FastAPI, pydantic
Cloud: Azure, AWS, GCP
CI/CD: Git

Soft skills:
Communication, planning, leadership

Personal project:
Multi-agent systems in video games (last slide)

Professional Experience

CMRE Logiciel - Data Scientist (7 months)
Machine Learning, Big Data and No-SQL

UdS - Deep Learning Researcher (1 year)
Real-time object detection in mixed reality

Genalyte - AI Engineer (6 months)
OCR and embedded object detection on Raspberry Pi

Survue AI - MLOps Engineer (1 year)
Object detection, model optimization and quantization

Velox Operations - Gen AI Engineer (1 year)
AI assistant for medical researcher in Microsoft Azure.
AI business strategy

Velox Operations

Generative AI Engineer

1 year

Project:

Velox Operations had bought a company with over 20 years of historical market research data in the medical field. Their CEO and COO wanted to use AI to build an assistant that could help their market research team to speed up their work, and tap into all of that data.

Outcomes:

Implemented a LLM Assistant using RAG in Microsoft Azure on 100Go+ of medical data.
Sped up market researcher work for consulting in the medical field, and eased access to 20 years of historical data.

Technologies:

Microsoft Azure, OpenAI, Chat-GPT4o, Azure Search, RAG, Python, Uvicorn, Javascript, React, CI/CD

Survue AI

MLOps Engineer

1 year

Project:

Survue AI is a startup that is leveraging embedded vision models and trajectory estimation to make riding bikes safer on the roads. I have worked on improving and optimizing the vision model, including quantization for embedded deployment, as well as the transition to a new tech stack. I have also mentored an intern from Northeastern University during the summer to teach him the ropes of MLOps.

Outcomes:

Doubled mAP for vehicle detection. Translated a TF version of MobileNetV2-SSD to PyTorch, and quantized it manually before exporting it to ONNX. Switched to a new stack.

Technologies:

Tensorflow, PyTorch, ONNX, MLOps

Genalyte

MLOps Engineer

6 months

Project:

Genalyte sells a mobile lab called “Merlin” for sample analysis. Any lab technician can use this lab to analyze samples in a simple way. Inside this Merlin lies a constrained vision system that needed to be reworked. My role was to rethink that system to switch from image classification to object detection, and make sure the end solution could run in real-time, perform OCR to extract error messages, and be deployed on an embedded device.

Outcomes:

Detection of each step of a sample analysis process, and extraction of error messages using OCR in real-time on Raspberry Pi. I built an entire MLOps pipeline from scratch in Tensorflow, and updated the model running in production with minimal, seamless changes (single python script change).

Technologies:

Tensorflow, Python, VastAI, Bash, Raspberry Pi, Linux

Antarctic Foods Aquitaine

AI Integration Engineer

2 months

Project:

Nuxly contacted me for a project with Antarctic Foods Aquitaine. They wanted to use LLMs to automate invoice processing for hundreds of invoices. I built a LLM pipeline using Llamaparse and OCR to extract entities from various invoice formats, and a Pinecone vector database to automatically match entities. A JSON was then built to represent the order in Odoo.

Outcomes:

Fully automated invoice processing. The human in charge of that work now only has to check for AI mistakes.

Technologies:

Pinecone, LangChain, OpenAI, Mistral, Python, FastAPI, pytesseract

Université de Sherbrooke

Deep Learning Researcher

1 year

Project:

A collaboration between the DOMUS lab and VMWare led to a real-time object detection project in mixed reality using both RGB and Depth images on a Microsoft Hololens 2. I used the HL2SS library to collect RGB and Depth images and send them over a backend PC used as a server. Yolov8 nano was used for object detection, and FastSAM for segmentation. The Depth map is aligned with the RGB image, turned into a pointcloud, and clustering is used to extract the boundaries of the object in 3D before reconstructing a final 3D-mesh

Outcomes:

Real-time mesh reconstruction using Yolo, FastSAM, and pointcloud clustering on a python server with a Unity frontend application on the Hololens 2

Technologies:

VisualStudio, Unity, C#, Python, HL2SS, Pointcloud clustering, vectorized functions, multithreading, literature review, scientific publications

Université du Québec à Chicoutimi

Doctoral Student

4 years

Project:

My thesis was carried out in the LIARA lab, which specializes in smart homes and ambient intelligence. The aim of my thesis was to extend assisted living from smart homes to smart cities. For this, I have studied the use of image classification on low-resolution thermal cameras to recognize profiles of vulnerable people that might need assistance in the city.

Outcomes:

2 Generations of prototypes (Arduino, Raspberry Pi), capturing thermal images. A CNN-RNN architecture built from scratch and fine-tuned to detected profiles of pedestrians with a high accuracy. Proved that the thermal signature and gait patterns were sufficient on a small dataset to recognize profiles of people.

Technologies:

Arduino, Raspberry Pi, C, Python, Tensorflow, Keras, scientific writing, publications, literature review

CMRE Logiciel

Data Scientist

7 months

Project:

This project was my last year internship in engineering school. I evaluated Big Data solutions for CMRE Logiciel, to see if transitioning to a MongoDB database made sense. I also implemented Machine Learning techniques to predict the quantity and quality of milk production ahead of time.

Outcomes:

Study of state-of-the-art Big Data techniques, and decision to keep the current SQL approach. Used Random Forest Regression to predict milk quantity and quality indicators with over 85% accuracy

Technologies:

Knime, MongoDB, Python, SQL,

IHMTEK

Game Developer

7 months

Project:

I did two internships at IHMTEK in the game development industry. My first project was to set up the environment to develop a cross-platform game called Ludomuse, using the Cocos2D framework (Java/C++).
The second project was about developing a Unity app used by restaurants for faster and easier orders

Outcomes:

Specified the architecture of the Ludomuse game as a UML and set up the foundations for a cross-platform game. Built a Unity application, linked the frontend with the backend to ingest menu and restaurant data dynamically

Technologies:

Java, C++, Cocos2D, Unity, C#, UML, UI, UX

Personal Project

Multi-Agent systems in Video Games

(ongoing)

Project:

I have been working on using LLMs in video games, as a way to turn NPCs into intelligent agents that can adapt to their environment, and make a game feel realistic. I started with a Pygame prototype, where a LLM and vector similarity were used to generate relevant quests that could be completed in real-time.
I am now working on a Godot game I have built from scratch with several agents each having a role in a small village (blacksmith, farmer, merchant…). They use GOAP (Goal-Oriented Action Planning) to pick their highest priority goal. I am working on adding an emotional module and LLM support to let dialogue and actions influence their decisions

Technologies:

Python, Godot, LLMs, RAG, AI Agents, GOAP