SHARHAD BASHAR

☎ +1 (646) 683-8387

sharhad.bashar@uwaterloo.ca

sharhadbashar.com

Sharhad Bashar

SharhadBashar

Manhattan, NY

EDUCATION

Master of Science (MSc)

University of Waterloo | Waterloo, ON

Computer Science, AI, ML and Computer Vision

TA: AI/ML, Functional, C and Concurrent Programming, Data Structures, and OS

2019 - 2022

Bachelor of Engineering (BEng)

McGill University | Montreal, QC

Honors Electrical Engineering

Minor: Software Engineering

2012 - 2017

Brainstation Bootcamp | New York, NY

Machine Learning and AI Instructor

Instructor for AI, Data Science, Python

2024 - Current

SKILLS

Languages:

- Python
- Java
- NodeJS
- C#

LLMs (Models and Features):

- OpenAI
- Anthropic
- LLama
- PydanticAI
- Tool Calls
- MCP Servers
- Cursor
- Gemini
- Groq
- Ollama
- DeepSeek
- Structured Output
- Agentic LLM
- Colab

Libraries:

- TensorFlow
- OpenCV
- Django
- AWS/GCP
- SciPy
- Keras
- PyTorch
- HuggingFace
- FastAPI
- Firecrawl
- Scikit-learn
- Datadog/Logfire

Database:

- PostgreSQL
- MongoDB
- Vector DB
- Duck DB

Scripting and Tools:

- Terminal
- Bash
- EC2
- Git

SUMMARY

10x Senior AI and ML engineer with expertise in AI, ML, Large Language Models (LLM) and a variety of Data Bases and Cloud Platforms. American Green Card Holder and Canadian Citizen

EXPERIENCES

Senior AI Engineer, Fora Travel | Manhattan, NY | **Python, LLM, AI, AWS, VectorDB,** **2024 - Current**

10x Senior AI Engineer at a Travel Agency. Working with AI, LLM, Python, AWS, NodeJS, VectorDB, SQL

- Designed and built the AI infrastructure from scratch with **Python, LLM, FastAPI, SQL, VectorDB**
- Responsible for performing RND and building POCs using the latest **AI** and **LLM** features
- Built end to end AI prototypes as well as production grade solutions
  - Conversational **AI chatbots** and **RAG systems**
  - Built **Agentic** Systems with **Response API, Tools call, Pydantic Structured Outputs, CoT prompting**
  - Deployed **MCP Servers** to connect to internal agentic services and tools as well as external resources
  - Built and trained **NLP models** on **PyTorch** and **TensorFlow** to work with 1 million+ WhatsApp data
  - Deployed AI solutions using **Docker**, **AWS**, **Kubernates**, **Vercel**, **Retool**
  - **Finetuned OpenAI** models for lead scoring, and to predict and evaluate churn
- Wrote custom eval scripts to test AI products with LLM as a judge, **Hamming**, **LogFire**, **Datadog**, **Pi Labs**
- Built backend features with **Python**, **Django**, **PostgreSQL**, **AWS**, **Docker**
- Designed the AI tech round (coding and system design), and responsible for interviewing candidates

Co-Founder and CTO, InspectlyAI | Manhattan, NY | **Python, LLM, AWS, VectorDB** **2024 - Current**

Cofounder and CTO of a Real Estate based AI company, connecting customers and contractors to work on fixing home issues

- Responsible for designing and built the entire architecture using **Python**, **FastAPI**, **PostgreSQL**, **Alembic, AWS, Stripe**
- Built the AI tools and solutions using **LLM (OpenAI, Anthropic, Ollama)**, **VectorDB (Weaviate, Pinecone)**, **OCR**
  - Built a **RAG system** to allow customers to search and ask questions about Property Reports
  - Built an **OCR system** to extract issues and images from Property Reports
- Managing a team of **six** Engineers and Interns

Senior AI / ML Engineer, Apple | Manhattan, NY | **Python, PyTorch, TensorRT, LLM, AWS** **2024**

Worked on Apple Intelligence team building and training models for Apple TV, Apple Store and Apple Music

- Trained and evaluated Teacher and Student LLMs using **Python**, **PyTorch**, **TensorRT** and proprietary data
- Built HuggingFace like internal library and tools for hosting all of Apples models and data for teams to use

Senior AI / ML Engineer, Azerion | Manhattan, NY | **Python, LLM, AWS, GCP, VectorDB, SQL** **2022 - 2024**

Lead the AI, ML and Data Science initiatives at AdTech firm with **Python**, **LLMs**, **VectorDB**, **LangChain**, **AWS**, **CUDA**, and **SQL**

- Trained an **NLP** pipeline with custom models and **BERT** to categorize podcasts and URLs into over 1000 categories
- Built a similarity search from over 100 million podcasts using **BERT**, **VectorDB**, **Bedrock**, **LLMs** enhancing content discovery
- Podcast and Video Contextualization (over 25,000 a day) in multiple languages:
  - Developed a pipeline using custom **NLP** models, **WhisperAI**, **AWS**, and **SQL**
  - Implemented a feature for limitless custom topics targeting using **GPT-4**, **LLaMA2**, **WhisperAI**, **AWS**, and **SQL**
  - Auto generate comprehensive Brand Safety reports using **LLaMA2**, **GPT-4**, and **SQL**, ensuring content adheres to IAB guidelines
  - Deployed on **EC2** instance with GPU and is constantly monitored, and periodically trained and upgraded
- Created a local chat bot, summarizer, and translator for sensitive 100+ page docs using **LLaMA2**, **LangChain** and **Pinecone**
- Utilized **LLaMA2**, **GPT-4**, **LangChain**, and **VectorDB** to autonomously generate detailed company information from websites
- Implemented an email chat bot using **GPT-4**, **LLaMA2**, **VectorDB**, and **GCP**, optimizing communication and response times
- Developed financial and forecasting AI model for bidding optimization generating an additional $1 million in revenue
- Devised solutions to curate database of 7 million+ user preferences and trends for targeted advertising from 100+ sources:
  - Built scripts to scrape user preference and trends data from social media sites using **Python**, **DataBricks**, **PySpark** and **AWS**
  - Used **GPT-4** and custom **NLP** models to find connections between user preferences and trends for targeted advertising
- Automated tests, system diagnoses, and generated detailed statistical reports for internal teams and shareholders
- Delivered several presentations on products, solutions and data-driven insights to executives and senior management

AI/ML Research Intern, Microsoft | Montreal, QC | **Python, TensorFlow, PyTorch, Keras, SQL** **2020 - 2022**

Researcher, working with **NLP**, financial and traffic data to improve machine learning model training and accuracy

- Developed patent pending AI for next generation navigation based on traffic and cellular data
  - Scraped and cleaned over 32 million data points using **DataBricks** and **PySpark** for training AI models
  - Trained AI models in **Python** using **Keras**, **TensorFlow**, **PyTorch**, and **SkLearn**
  - Built and deployed the ML pipeline on **EC2** instance. Pipeline included periodic automatic updates and re training of models
- Built custom Federated Machine Learning models using TensorFlow and trained on Fintech, Financial, and Statistical data
- Researched new strategies and models to improve Automatic Speech Recognition (**ASR**)
- Implemented **BERT** models from **HuggingFace** to analyze millions of online product reviews

ML and Software Engineer, Montrium | Montreal, QC | **C#, Python, Java, NodeJS, SQL** **2017 - 2019**

Worked as a ML and Backend Engineer for a new health tech platform, used by over a million customers globally

- Created APIs and primary functions using **.NET Core**, **C#,** and **MongoDB**
- Developed a Recurrent Neural Network (**RNN**) to translate company's products to several languages

THESIS

Graduate | ***Semantic Segmentation*** | University of Waterloo | **Python, PyTorch, OpenCV, TF** **2017 - 2019**

- Thesis title: Volumetric Weak Supervision for Semantic Segmentation
- Used image level data and approximate class sizes to improve accuracy of Weakly Supervised Semantic Segmentation
- Improved the accuracy by over 14% mean Intersection over Union (**mIoU**)

Undergraduate | ***Image Captioning*** | McGill University | **Python, TensorFlow, OpenCV** **2016 - 2017**

Used over 1 million images to train **CNN** and **RNN** to generate image captions with text descriptions for the Autour app

PROJECTS

Applied Machine Learning and Artificial Intelligence | **Python, LLM, JavaScript, SQL, AWS** **2016 - Curr**

- Built Agentic RAG system with MCP servers for SEC fillings using Python, OpenAI, Weaviate
- Built various **PyTorch** and **TensorFlow** models for Vision, Audio, Classification and Regression tasks
- Built a ML model using **OpenCV** and **RNN** to identify exercise activities from videos
- Build a ML classifier with **PyTorch** and **OpenAI** to detect wake words for Spotify. Wake Word: **Hello Spotify**. Demo
- Created a **CNN** for unsupervised single image depth prediction and a **RNN** for speech recognition
- Programmed several different optimization methods, including **Gradient** and **Coordinate Descent**, **ALM** and ADMM