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 and ML

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

2019 - 2022

Bachelor of Engineering (BEng)

McGill University | Montreal, QC

Honors Electrical Engineering

Minor: Software Engineering

TA: Power Engineering, Statistics

2012 - 2017

SKILLS

Languages

- Python
- TypeScript
- React
- NodeJS

Large Language Models

- OpenAI
- Anthropic/Claude
- Gemini
- Grok
- DeepSeek
- Ollama (Open Source)
- Perplexity
- AssemblyAI
- VoyageAI
- Cohere

Agentic AI and Frameworks

- Pydantic AI
- Agents SDK
- LangChain
- LangGraph
- RAG
- MCPs
- AI Evals (Braintrust)
- LLM Fine tuning
- DSPy
- Multi Agent System
- Loop Engineering
- Graph Engineering

Agentic AI coding tools

- Cursor
- Claude Code
- Codex
- GitHib Copilot

Libraries, Packages and Frameworks

- TensorFlow (TF)
- PyTorch
- Pydantic
- HuggingFace
- Django
- FastAPI
- AWS/GCP
- OpenCV
- Datadog
- Docker

Databases

- PostgreSQL
- NoSQL
- Vector DB
- Graph DB

Scripting and Tools

- GitHub
- Terminal/Bash

SUMMARY

10x AI and ML Engineer in NYC with expertise in Agentic and Applied AI, ML, LLMs, Evals and a variety of Databases and Cloud Services. Experienced with Agentic Coding Platforms like Cursor, Claude Code, Codex. American Green Card and Canadian Citizen

EXPERIENCES

Senior Applied AI Engineer, Fora Travel | Manhattan, NY | **Python, TypeScript, AWS, SQL** **2024 - Curr.**

10x Senior AI Engineer, building and researching Agentic AI solutions and tools with LLMs for travel entrepreneurs

- Designed and built the Agentic AI service from scratch using **Python**, **Pydantic AI**, **LLM**, **FastAPI**, **SQL**, **VectorDB**
- Built end to end AI features, from POC and prototypes to production grade solutions for over 45,000 advisors:
  - Lead engineer for Via, a multi agentic chat bot with access to over 40 internal MCP tools, file upload, memory, websearch, deep research, knowledge base, skills, personalization, Gmail integration, guardrails, Speech to Text. Built with **Python**, **Pydantic AI**, **Perplexity**, **AssemblyAI**, **OpenAI** and **Claude**, **Braintrust**, **Turbopuffer**
  - Built a multi agentic system to generate itineraries, saving advisors over 4 hours on average
  - Built an agentic system to generate hotel summaries, transcribe and summarize notes for advisors
  - Built and maintained conversational AI chatbots and RAG systems using **TypeScript**, **Pinecone** and **Braintrust**
  - Finetuned OpenAI models for lead scoring, and to predict and evaluate churn
- Wrote custom evals and simulations to test AI products with LLM as a judge on **Braintrust**, **LogFire**, **Datadog**
- Built backend features and **MCP** tools with **Python**, **Django**, **PostgreSQL**, **AWS**, **Docker**
- Lead R&D initiatives; building Agentic AI and LLM POCs that shape product direction and adoption both externally and internally
- Designed the AI tech round (coding and system design), and responsible for interviewing candidates

Co-Founder and CTO, InspectlyAI | Manhattan, NY | **Python, Pydantic AI, AWS, VectorDB** **2024 - Curr.**

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

- Designing and building the entire BE architecture using **Python**, **FastAPI**, **SQL**, **AWS**, **Stripe**
- Building the Agentic AI layer using **Python**, **Pydantic AI**, **LLMs**, **VectorDB**, **OCR**
  - Built an **AI Agent** with **MCP** tools to allow users to perform all the platform actions and intake process
  - Built a **RAG** to allow customers to search and ask questions about their Reports and local law using **Weaviate**
  - Built an Agentic **OCR** system to extract issues and images from Reports using **Pydantic AI**, **LLMs, SQL**
- Managing a team of six Engineers and Interns

AI / ML Engineer, Apple | Manhattan, NY | **Python, PyTorch, TensorRT, JAX, 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**, **JAX** and proprietary data
- Built **HuggingFace** like internal library and tools for hosting all of Apple's models and data for teams to use

Data Scientist, 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**, **AWS**, **CUDA**, and **SQL**

- Built a Podcast and Video Contextualization (over 25,000 podcasts a day) in multiple languages:
  - Developed a pipeline using custom trained **NLP** models, **WhisperAI**, **AWS**, and **SQL**
  - Implemented custom topics targeting using **GPT-4**, **WhisperAI**, **AWS**, and **SQL**
  - Auto generate comprehensive Brand Safety reports using **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
- Trained an **NLP** pipeline with custom models and **BERT** to categorize podcasts and URLs into over 1000 categories
- Built similarity search for 100 million podcasts using **BERT**, **VectorDB**, **Bedrock**, **LLMs** enhancing content discovery
- Created chat bot, summarizer, and translator for sensitive 100+ page docs using **LLaMA2**, **LangChain** and **Pinecone**
- Utilized LLaMA2, GPT-4 and LangChain to autonomously generate detailed company information from websites
- Implemented an email chat bot using GPT-4, VectorDB, and GCP, optimizing communication and response times
- Delivered several presentations on products, solutions and data-driven insights to executives and management

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

Researcher, working with **NLP** 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
  - Trained AI models in Python using Keras, TensorFlow, PyTorch, and SkLearn with 32 million data points
  - Built and deployed the ML pipeline on EC2 instance
- Researched new strategies and models for Automatic Speech Recognition using **BERT** models from **HuggingFace**

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

Developed ML and backend for health tech platform serving 1M+ users, improving performance and latency

- Created APIs and BE functions using **.NET Core**, **C#**, and **MongoDB**. Created automations using Selenium and Java
- Developed a Recurrent Neural Network (**RNN**) using **Python** to translate company's products to 33 languages

THESIS

Graduate | **Semantic Segmentation** | University of Waterloo | **Python, PyTorch, OpenCV, TensorFlow** **2020 - 2022**

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

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

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

PROJECTS

Applied AI, Machine Learning and Deep Learning | **Python, TypeScript, Agents, SQL, AWS** **2016 - Curr.**

- Building a Stock Trader AI bot: **Python**, **Pydantic AI**, **LangGraph, LangChain, MCP**, **Turbopuffer, Brantrust, SQL**
  - Multi agent orchestration with websearch, deep research, MCP tools on Robinhood, Liquid, NasDaq
  - Uses custom skills and instructions, user memory, message compaction and various capabilities
  - Evals and multi turn simulations on **Braintrust** and logging on **Logfire**
- Built **Agentic RAG** system with **MCP** servers for SEC fillings using **Python**, **Pydantic AI**, **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
- Built and trained a ML classifier with **PyTorch** and **LLM** 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 optimization methods, including **Gradient** and **Coordinate Descent**, **ALM** and **ADMM**