Shanjidul Islam Sadhin

Linkedin: linkedin.com/in/sadhiin/
Github: github.com/sadhiin
Portfolio: sadhiin.github.io

Email: sadhin.aiub.cse@gmail.com
Mobile: 01746-362426

Education

- **American International University-Bangladesh** — Dhaka, Bangladesh
  Bachelor of Science - CSE — *January 2020 - December 2023*

Skills

- **Languages**: Python, C++, C#, JavaScript, PHP, SQL
- **Library**: Pytorch, TensorFlow, Keras, NumPy, Pandas, Scikit-Learn
- **Frameworks**: FastAPI, LangChain, Flask, Django
- **Tools**: GIT, DVC, MLflow, DagsHub, MySQL, PostgreSQL, ChromaDB, FAISS
- **Platforms**: Linux, Windows, Docker
- **Cloud-Platforms**: AWS, vLLM
- **Soft Skills**: Leadership, Writing, Public Speaking, Time Management

Experience

- **Applied AI Engineer — Askturing.ai** — Remote
  (Full-time) — *June 2025 – Present*
  - **Reduced LLM token costs by 40%** by designing an intelligent context management and chunking strategy for large PDF documents within the financial RAG pipeline.
  - **Developed a vision retrieval system** for optimized, cost-efficient information extraction from complex financial documents, reducing operational costs while increasing retrieval accuracy.
  - **Transitioned from text-based LLMs to Vision-Language Models (VLMs)**, enhancing correctness and robustness of financial document understanding within the RAG pipeline.
  - **Architected and refined a RAG evaluation framework** to systematically measure and improve retrieval accuracy, directly supporting critical financial decision-making.
  - Implemented an **ephemeral file management feature in the Askturing Chat** application, enabling users to seamlessly incorporate new uploads and contextual information into ongoing conversations enhancing coherence and knowledge-grounded responses from the project's existing knowledge base.
  - Developed multi-step hybrid reasoning workflows combining similarity search, web search, query reformulation, and context prioritization—improving answer correctness and reducing hallucination rates.
  - Implemented **AI Faithfulness & Fact-Validation pipelines** using multi-model cross-verification (GPT, Gemini, Claude) to detect misinformation, validate claims, and provide source-grounded responses across diverse financial and user-generated content.
  - Enhanced end-to-end retrieval consistency by orchestrating multi-agent workflows that route complex user queries through RAG, ephemeral files, web search, and project vectors to guarantee high-confidence answers.
  - Implemented an **event-driven action trigger system** (file uploads, scheduled tasks, email events) that automatically executes user-selected or custom prompts, resulting in **30% higher new-user adoption** and improved task automation and problem resolution.
- **AI Engineer — TechnyX.ai** — On-site
  (Full-time) — *December 2024 - May 2025*
  - Designed and architected an AI system for image generation and virtual try-on for models and clothing, enabling real-time inference with scalable deployment using NVIDIA Triton Inference Server.
  - Configured and optimized NVIDIA Triton Inference Server for scalable model inference, ensuring high-performance deployment of computer vision and multimodal AI models in production environments.
  - Led development of advanced computer vision and multimodal AI solutions, address and solved real-world industry challenges with 98% accuracy.
  - Architected and deployed vision-language models for in-house API development, optimizing inference cost by 40% and reducing computational costs through efficient model deployment.
  - Engineered state-of-the-art computer vision algorithms and end-to-end AI applications, integrating LLMs with computer vision systems for building robust solution.
  - Collaborated with cross-functional teams to design and deploy scalable AI software solutions, ensuring seamless integration of multiple AI technologies across the platform
- **Freelancer — UpWork** — Remote
  (Part-time) — *October 2024*
  - Developed a state-of-the-art super-resolution model for image up-scaling, leveraging advanced techniques to enhance image quality and detail.
  - Modified the architecture of the "Dual Attention Transformer" to effectively upscale images from two distinct sources, ensuring optimal performance and quality.

- Designed a robust architecture that utilizes channel-wise concatenation of images, focusing on the most efficient features to improve the overall up-scaling process.
- Conducted extensive testing and validation of the model to ensure high accuracy and reliability in various real-world applications.
- Collaborated with clients to understand their specific needs and provided tailored solutions, resulting in a 30% increase in client satisfaction ratings.

- **Junior AI Developer — CRTVAI** Remote
  (Full-time) *April 2024 - September 2024*
  - Designed and implemented the system architecture and API endpoints of an automated quality control AI application to enhance customer care services.
  - Fine-tuned speech-to-text models, deploying server-less models to minimize third-party transcription costs.
  - Improved the Voice Activity Detection pipeline and conducted a detailed analysis of speech metrics to enhance accuracy and efficiency.
  - Achieved a **25% reduction** in GPT API calling costs through strategic optimization.
  - Implemented Background task processing techniques to significantly enhance system performance and reliability.

- **Research Assistant — Internship** On-site
  Dept: Computer Vision and Bio-Robotics (Full-time) *Sep 2023 - Jan 2024*
  - **Research Collaboration**: Actively participated in a Computer Vision research project, leveraging my skills and knowledge to contribute to the team's success.
  - **Research Dataset**: Contributed to the creation of a research dataset containing 12,000+ labeled images, supporting future computer vision research and applications.
  - **Research and Data Analysis**: Reduced research time by 20% by implementing efficient data analysis and documentation practices.
  - **Documentation & Presentation**: Successfully presented research progress and insights to the research group and advisor, ensuring clear communication and project transparency.

Projects

- **EasyQuery-Conversational Database Query System**: Built an AI-powered conversational interface that enables users to query databases in natural language. The system **generates and executes SQL queries dynamically** based on user requests, retrieves results from the database, and presents them in a human-readable format—eliminating the need for manual query writing. This streamlined workflow significantly improves accessibility and efficiency for data-driven decision-making.

  Tech-stack: FastAPI, LangChain, SQLAlchemy, PostgreSQL, LLMs (Openai, Gemini, Anthropic, Groq), Docker (August 2025 - Present) Github-Link Live-Backend
- **DeepCrawl-Chat Intelligent Web Crawler and RAG System**: Developed "DeepCrawl-Chat," an intelligent system for advanced web crawling, information extraction, and Retrieval Augmented Generation (RAG). This allows users to crawl websites and interactively query the content using AI language models, facilitating tasks such as ***competitor website analysis and effective LLM-integrated data analysis***.

  Tech-stack: FastAPI, LangChain, FAISS, Docker, Uvicorn, NVIDIA AI (embeddings), Hugging Face, Groq, SQLAlchemy, Redis, (March 2025 - Present) Github-Link
- **YouTube Video Summarizer**: Developed a Python application that downloads YouTube videos, transcribes them using Groq's Whisper model, and generates concise summaries using LLM models. Features include semantic search with FAISS vector database, interactive chat with video content using LangChain memory, and multiple interfaces (web UI, API, CLI). **Tech-stack:** FastAPI, Streamlit, LangChain, FAISS, SQLAlchemy, Redis, Groq API, Nvidia-inference API (April 2025) GitHub-Link
- **[Chat-bot] MediChat-Assistant**: Developed a medical chat-bot application using open source LLM Llama-2 7B-chat with RAG for accurate source based response, with FastAPI for user-friendly interaction and Restful API's access.

  Tech-stack: LLM, RAG, VectorDB, Python, FastAPI, LangChain (Oct - Nov 2024) GitHub-Link
- **[Python Package] SpectraClassify**: Developed a web-based application to train a custom image classification model without writing code and published it as PyPI Package. **Tech-stack:** Python, TensorFlow, Flask, GitHub-Action (Dec 2023 - Present) Github-Link PyPi page
- **Kidney Tumors and Stones Classification**: An end-to-end deep learning model for classifying kidney tumors and stones from medical images, with CI/CD pipeline to ensure efficient data management and streamlined workflow. **Tech-stack:** Python, TensorFlow, DVC, Git, MLOps, DagsHub (November - December 2023) GitHub-Link
- **Real-Time Emotion Classification from Face Images**: Developed a real-time emotion classification system utilizing a custom deep learning model and Flask API. Analyzes facial expressions via webcam for high-accuracy emotion classification with a user-friendly web interface.

  Tech-stack: Python, TensorFlow, OpenCV, Flask, JavaScript, HTML, CSS, Git. (June - August 2023) GitHub-Link
- **Community Platform for AIUB Students - AIG**: Developed an ASP.Net application for the AIUB student community platform implementing SOLID principals and 3-tier architecture, which lets users create resumes, post job openings and apply for jobs with one click. The role included use-case analysis, Database designing, authentication and authorization, and other features including API endpoints. **Tech-stack:** ASP.NET, C#, Entity-Framework, (Jul - Aug 2023) GitHub-Link

- **Podcast Web App (PHP)**: Developed a user-friendly podcast web application for managing and enjoying podcasts. Features include subscription management, playlist creation, and intuitive playback. **Tech-stack:** PHP, JavaScript, Bootstrap, CSS, HTML, Git (January - April 2023) Github-Link
- **Fashion Recommendation System Using ResNet-50**: An end-to-end fashion recommendation system using ResNet-50 for feature extraction and KNN for similarity matching and image-based product suggestions, which takes an apparel image and suggests related products from inventory. **Tech-stack:** Python, TensorFlow, Numpy, Scikit-learn, Streamlit, Git, Google-Colab. (July-August 2022) Github Link
- **Chrome Dingo Clone Build with OpenGL**: A 2D Chrome Dino-like game using OpenGL(freeglut) and C++ which challenges players in a infinite horizontally scrolling background with randomly appearing obstacles to avoid them by jumping. Responsibilities including Low-level design with OpenGL, implementing game level difficulty controls, text rendering **Tech-stack:** C++, OpneGL, Git (April 2022) GitHub-Link

Research Experience

- [**Draft**] **Research: A Large-Scale Action Dataset**: Tech: Python, NumPy, Pandas, Sci-kit, TensorFlow, Keras (January 2024 - Present)
- **Research: High-Accuracy Image Segmentation for Self-Driving Cars**: Tech: Python, NumPy, Pandas, TensorFlow, Keras (August 2023 - Present)

Publications

- [**Pre-Print**] **Thesis: Speech Emotion Recognition using Transfer Learning Approach and Real-Time Evaluation in English and Bengali Language**: Tech: Python, NumPy, Pandas, Sci-kit, TensorFlow, Keras (January 2023) Researchgate

Honors and Awards

- Dean's List Award - for outstanding academic performance. August 2023

Extra-curricular Activities

- **Problem Solving**
  Solved 800+ problems in different Online Judges. stopstalk Profile *Feb 2020 - Present*
- **12'th National Undergraduate Mathematics Olympiad 2021 Dhaka North Region** Dhaka
  Participate the Mathematical Olympiad *November- 2021*
- **Hacktoberfest-2023** Remote
  Programming Problem solving *October 2023*
  - **Problem Discussion**: Refactor the code for improve readability, time and space complexity.
  - **Programming-Solution**: Added solution of DSA problems from LeetCode with detailed implementation approach.