

MUHAMMAD ARHAM

✉ arhamm40182@gmail.com [in LinkedIn](#) [on GitHub](#) ☎ +18253334209

Skills Summary

Languages: Python, C++, Java, JavaScript

ML Frameworks: PyTorch, HuggingFace, Unsloth, LangChain, LangGraph, OpenCV, TensorRT, Numpy, Pandas

Backend Technologies: FastAPI, Git, SQL, NoSQL, Vector Databases, Redis, Celery, Docker, AWS, Terraform, CI/CD

Domain Knowledge: Classification, Segmentation, Detection, CNNs, RNNs, GANs, Transformers, LLMs, Stable Diffusion

Industrial Experience

ServeDen

April 2025 - Present

Machine Learning Engineer

Remote, USA

- Building no-code platform for general-purpose LLM-agents supporting RAG, code execution, and MCP tools.
- Developing a custom agentic framework for secure, auto-provisioned cloud deployment of LLM agents.

Traversaal.ai

September 2024 - February 2025

Senior NLP Researcher

Remote, USA

- Engineered a production-grade platform combining *RAG pipelines* with *agentic multi-modal systems* for context-aware information retrieval, report synthesis, and automated generation of insights from proprietary data.
- Provided B2B consultancy to clients on RAG systems (>\$200K contracts), including Ikona Analytics, JIB, & Kraft.
- Core member of development of *Urdu-Llama* under Traversaal.ai's LARGE initiative, designing scalable training and evaluation pipelines for low-resource language modeling; won the META AI APAC Accelerator Grant (USD 100K)

CERN

June 2024 - September 2024

CERN Summer Student

Meyrin, Switzerland

- Contributed to the CMSSW project under the Patatrack team at CERN, for the next phase of HL-LHC.
- Profiling and optimizing heterogeneous CUDA code for the event data reconstruction of the TIDL project for CMS.
- Porting CPU-only algorithms to heterogeneous CUDA and ROCm devices using the alpaka framework.
- Official report titled *Optimization of Event Reconstruction within the TIDL Framework in CMSSW* available on CDS

Vyro

September 2022 - April 2024

Machine Learning Engineer

Islamabad, Pakistan

- Developed and maintained PhotoShot, ImagineAI, Chatly, ReShot, ReTouch and RemoveIt for millions of users.
- Optimized Stable Diffusion Generative AI pipelines, resulting in a *fivefold acceleration in inference speed and a 50% reduction in GPU consumption, using torch compilation, ONNX, TensorRT, and Oneflow computational graph caching*.
- Implemented and deployed custom image generation pipelines using Diffusers and ComfyUI for inpainting, outpainting, face-swapping, and avatar creation within the ImagineAI application. Integrated support for SD1.5, SD2.1, and SDXL.
- Developed custom nodes for AI background swap, business headshot generation, image segmentation (*foreground & background*), and facial detection to enhance image understanding and image processing workflows.
- Trained and deployed image segmentation models, including DeepLabV3 and U2Net, for image editing application.
- Enhanced GAN-based image upscaling models, including SwinIR, Codeformer & ESRGAN, to elevate user satisfaction.
- Deployed QwenVL for image question answering application, integrating Amazon S3 and Redis for efficient computation.
- Implemented and improved a RAG pipeline for document-based question answering with HuggingFace and PineconeDB.

Publications

LLM-Informed Discrete Prompt Optimization

LLM Cognition Workshop @ ICML

Muhammad Arham, Zeeshan Memon, Adnan-ul-Hasan, Faisal Shafait

Published

NLU of Devanagari Script Languages

CHIPSAL @ COLING

Jebish Purbey, Siddhartha Pullakhandam, Muhammad Arham

Published

Query Attribute Modeling: Improving search relevance with Semantic Search

IEEE

Karthik Menon, Muhammad Arham

Accepted

Alif: Advancing Urdu LLMs via Multilingual Synthetic Data Distillation

MRL @ EMNLP

Ali Shafique, Kanwal Mehreen, Muhammad Arham

Accepted

Education

University of Alberta

Sep 2025 - May 2027

MS Computing Science, Supervisor: Dr. Martha White, RLAIM, Amii

National University of Sciences & Technology (NUST)

Sep 2020 - May 2024

BE Software Engineering, CGPA 3.86 / 4.00, MITACS & CERN Intern, *Rector's Gold Medal*