

Marwan AbdElhameed

linkedin.com/in/marwan-abdelhameed | marwan@nyu.edu | +971565377128 | github.com/Marwanabdelhameed

EDUCATION

Polytechnique Montréal

Montreal, Canada

PhD in Software Engineering

Starting August 2026

New York University Abu Dhabi

Abu Dhabi, UAE

Bachelor of Science in Computer Science, Minor in Interactive Media

GPA: 3.887 (Cum Laude)

• **Honors:** University Honors Scholar, Founders Day Award

• **Global Study:** NYU New York, NYU Florence, NYU Athens

• **Graduate-level coursework:** Deep Learning Systems, Operating Systems, and Open Source Development

Languages: Arabic (Native), English (Fluent), German (Beginner)

RESEARCH PUBLICATIONS

• Kassab, M., **AbdElhameed, M.** (2025). Detecting and Repairing Incomplete Software Requirements with Multi-LLM Ensembles *Accepted: ASE 2025 NIER*

• Kassab, M., **AbdElhameed, M.** (2025). Mapping the Landscape of Requirements Completeness: Definitions, Techniques, Tools, and the Emerging Role of AI. *Accepted: Euromicro SEAA 2025*

• **AbdElhameed, M.**, Halim, P. (2024). Inference Scaling vs Reasoning: An Empirical Analysis of Compute-Optimal LLM Problem-Solving. *New York University*

PROFESSIONAL EXPERIENCE

Whiteshield

Hybrid - Dubai, UAE

Senior Full Stack Engineer

November 2025 – present

• Managing **30 client entities**, conducting requirements meetings, troubleshooting platform access, and proposing new agentic AI system implementations

• Rearchitected data ingestion system by replacing SSIS file uploads with **SqlBulkCopy** in **C#/Blazor**, achieving **120x** speed improvement and significantly enhancing user experience

• Designed and implemented Medallion architecture (Bronze/Silver/Gold) using **SSIS** and **Power BI**, establishing a scalable ETL pipeline deployed on **Azure** and **IIS**

D&V Herrmann Services

Dubai, UAE (Hybrid)

Full Stack Engineer

July 2025 – November 2025

• Developing full-stack applications using **C#**, **.NET Core**, **Blazor** (Web/MAUI), and **SQL Server** for enterprise solutions

• Optimizing database performance through query optimization, indexing strategies, and deadlock resolution, improving system response time

• Co-managing **Bitbucket** CI/CD pipelines and deployments, ensuring seamless integration and delivery processes

• Implementing comprehensive unit and integration testing frameworks, reducing production bugs by identifying issues early in development cycle

• Collaborating with frontend and backend teams on **API integrations** for WinMan, Kiosk applications, and SSRS reporting solutions

RAI Institute

Remote

Open Source Software Engineer

March 2025 – May 2025

• Improved UI responsiveness and reduced CPU load by **90%** by identifying and resolving a critical animation loop bottleneck, optimizing update rate from 1000Hz to 60Hz in a PR merged by the BD AI Institute team

• Engineered high-performance backend service using **FastAPI** and **ROS2**, enabling real-time teleoperation with low-latency translation of user inputs into direct hardware commands for the Spot robot

• Architected full-stack robotics control and animation platform using **React**, featuring comprehensive joint controls and an advanced keyframe editor for creating complex choreographed sequences

NYUAD, Center for Interacting Urban Networks (CITIES)

Abu Dhabi, UAE

Research Assistant

February 2024 – August 2024

• Engineered data pipeline processing **1.4M** audio spectrograms using **NumPy**, and custom **CUDA** kernels, boosting efficiency by **35%** and reducing processing time by **50%**

• Deployed scalable ML models with **Singularity** in HPC environments, reducing deployment time by **30%**

• Led data analysis with **Numpy** and **Matplotlib**, improving research insights by **25%**

NYU Tandon, High-Performance Computing Group

New York, NY

Research Assistant

August 2023 – Dec 2023

• Optimized 3DMHD benchmark simulation performance by **40%** through integration of **Parallel processing**, using **C++**,

and MPI

- Improved **GCP** on-prem cluster utilization and job scheduling efficiency by **25%** by implementing comprehensive performance tuning and intelligent workload distribution.
- Led code optimization, GPU utilization, and parallel computing practices, reducing error rates by **20%**

MACHINE LEARNING & AI PROJECTS

Multi-Tool Financial & Tech Analysis Agent

Python, LangChain, OpenAI, SerpAPI

- Architected a ReAct (Reasoning and Acting) agent using **LangChain** to autonomously decompose and execute multi-step queries, such as comparing financial assets or technology products.
- Developed and integrated a suite of three custom, LLM-driven tools: a **SerpAPI** web search tool for data retrieval, a custom **Comparison tool** for feature analysis, and an **Analysis tool** to synthesize results.
- Implemented an end-to-end agentic workflow that intelligently chains tools to provide comprehensive answers, successfully handling complex queries like comparing top smartphones on camera and battery life.

AI-Powered Requirements Analysis Platform (ASE 2025 paper)

Python, Flask, JavaScript, Multi-LLM APIs

- Architected intelligent SRS analysis platform integrating **3 LLM APIs** with advanced aggregation strategies, achieving **95%** consensus accuracy through meta-analysis
- Engineered document processing pipeline supporting PDF/DOCX/TXT with Mistral AI OCR integration, achieving **99%** text accuracy and automatic requirements extraction from **100%** document types
- Developed UML domain model generator creating class diagrams from natural language requirements, reducing manual modeling time by **80%** with **100%** PlantUML compatibility
- Implemented requirements completeness analyzer detecting missing requirements and inconsistencies, improving SRS quality by **60%** through automated issue detection and intelligent suggestions
- Built interactive web interface with session management, enabling users to accept/reject AI suggestions with real-time updates and persistent state tracking

ML Infrastructure Optimization

Python, Ray, TorchX, W&B

- Architected batch prediction system using Ray actors and TorchX, processing **250** requests per batch with 0.5 GPU allocation
- Implemented multi-worker data loading system and distributed hyperparameter optimization, identifying optimal configurations
- Developed threaded execution pipeline using Python's threading and ZMQ, reducing model serving latency by **28%**

Seq2Seq Conversational AI

PyTorch, NLTK, Neural Networks

- Engineered sequence-to-sequence model processing **220,000+** movie dialogue exchanges with nn.GRU and nn.Embedding
- Implemented Luong attention mechanism, improving response relevance by **30%**
- Developed custom text preprocessing pipeline handling **18,000+** unique vocabulary tokens efficiently

Distributed ML Training System

PyTorch, CUDA, Parallel Computing

- Developed distributed training system using PyTorch's DataParallel, achieving **1.92x** speedup with 4 GPUs
- Optimized batch size scaling to achieve **31,063** samples/second throughput, a **7.3x** improvement
- Reduced I/O bottlenecks by **40%** through custom data loading pipeline optimization

Math Problem-Solving LLM System

Python, Large Language Models

- Fine-Tuned Mistral-7b achieving **35%** accuracy on GSM8K dataset
- Enhanced solution accuracy by **25%** from base model through advanced inference scaling techniques
- Created systematic evaluation framework for optimal prompting and inference strategy selection

Multilingual RAG System

Python, mBART, Vector DB

- Architected multilingual RAG system with mBART-large-50, enabling cross-lingual content generation
- Improved content relevance by **40%** through context-aware document fetching
- Scaled system to handle **100k+** document pairs with optimized memory usage

SOFTWARE ENGINEERING & DATA ANALYSIS PROJECTS

Distributed Model Serving Platform

Flask, Ray,serve, TorchX

- Built scalable REST API handling concurrent model inference requests, achieving **99%** uptime
- Integrated Ray with TorchX for distributed computing, processing **10,000+** concurrent requests across multiple nodes
- Implemented efficient task distribution and monitoring system, reducing inference latency by **20%**

Photo Sharing Web Application

Python, Django, SQL, PostgreSQL

- Built full-stack app with user authentication, photo upload, and album management, increasing retention by **25%**
- Optimized data relationships and queries, improving retrieval efficiency by **40%**