**CHHALMA CHHAYA**

[chhaya.mist@gmail.com](mailto:chhaya.mist@gmail.com) | 07578956460 | LinkedIn: [ChhalmaLinkedin](https://www.linkedin.com/in/chhalma-chhaya-0640741b/) | GitHub: [ChhalmaGitHub](https://github.com/Chhalma)

East Dulwich, London, SE22 9JW | Full Driving License

**profile**

MSc Software Engineering graduate (awaiting final results) with distinction-level performance across all completed modules, specializing in AI and cloud systems. During my internship at AI-Aware Ltd, I developed and optimized multi-batch AI pipelines using **Docker** and **FastAPI** for the DeBERTa V3-Large model with PyTorch, quantized (INT8/INT4), ONNX, and TensorRT, achieving up to **4× FP16** inference speedup. Proficient in Python, Java, AI/ML, Docker, and cloud platforms (AWS/GCP).

**TECHNICAL skills**

* **AI/NLP & LLMs:** BERT/DeBERTa fine-tuning, zero-shot detection (Binoculars), adversarial robustness, SPARQL QA
* **Optimization:** ONNX Runtime, quantization (FP16/INT8/INT4), dynamic batching, CPU/GPU inference, TensorRT, Docker
* **ML Tools:** Hugging Face, PyTorch, Optimum, TextAttack,ONNX, TensorRT
* **Web & Cloud:** Docker, FastAPI, React.js, Next.js, NestJS, AWS (SageMaker), Google Cloud (App Engine, Pub/Sub)
* **Big Data:** Spark, partition optimization
* **Languages:** Python, Java (Oracle Certified), C/C++/C#, SQL, TypeScript
* **Databases:** MySQL, PostgreSQL, Oracle DB, MongoDB (NoSQL concepts)
* **Dev Tools:** GitHub, Git Bash, Visual Studio Code, IntelliJ, Jupyter, Eclipse, Django

**education**

|  |  |  |  |
| --- | --- | --- | --- |
| **City, University of London** | **MSc Software Engineering with Cloud Computing** | Sep 2023 - Present | |
| * Class Representative | Expected Distinction * **Key Modules:** Big Data (**86%**), Advanced Databases (**93%**), Semantic Web Technologies and Knowledge Graphs (**96.1%**), Algorithm (**88%**) | |  | |
| **University of Dhaka** | **MSc in information technology** | Jul 2009 - Dec 2010 |
| * CGPA: **3.76/4.00** UK Distinction Equivalent. | |  |

|  |  |  |
| --- | --- | --- |
| **Bangladesh University of Professionals** | **BSc in Computer Science and Engineering** | Feb 2005 - Dec 2008 |
| * CGPA: **3.35/4.00** UK 2:1 Equivalent. |  |  |

**CERTIFICATIONS**

* AWS Machine Learning Engineer Nanodegree – Udacity, Sep 2025
* Programming with Python Nanodegree – Udacity, Feb 2025
* [AWS Academy Graduate - AWS Academy Cloud Developing](https://www.credly.com/badges/d085cefc-b89f-4f89-a97d-e638f91721b7/linked_in_profile) – AWS, Aug, 2025
* Oracle Certified Associate Java Programmer – Oracle, Dec 2022

**projects**

**Optimized Spark Pipeline for Image Processing** Mar 2025 - May 2025

* Boosted Spark cluster performance by 34% using dynamic partitioning and RDD caching.
* Reduced processing time by 19% on cost-effective 4-node clusters.
* Demonstrated TFRecords’ linear scalability (R²=0.98) over JPEG bottlenecks.

**Convert Tabular Data to Knowledge Graph** Mar 2024 - May 2024

* Built a software component generating knowledge graphs from tabular data using OWL2Vec and SPARQL.
* Performed reasoning, ontology alignment, precision/recall evaluation, and embedded external vocabularies (Google KG, Wiki KG).

**Share Market Adviser** Oct 2023 - Jan 2024

* Developed GCP-based web application (Spring Boot + MySQL + App Engine).
* Integrated Pub/Sub messaging and Cloud SQL for secure database management.

**work experience**

|  |  |  |
| --- | --- | --- |
| **AI-Aware Ltd** | **AI Research Intern** | Jul 2025 - Present |

An AI detection Research Company

* Developed a multi-stage benchmarking pipeline for the Desklib model across PyTorch, quantized (INT8/INT4), ONNX, and TensorRT.
* Optimized multi-batch inference on CPU and GPU, measuring latency, throughput, and memory usage.
* Quantitative analysis of full-precision vs. quantized models, highlighting trade-offs in speed and accuracy.
* Containerized inference pipelines using Docker for reproducible CPU/GPU deployments.
* Performed experimental validation of multi-batch inference strategies, achieving **up to 4× speedup** in TensorRT FP16 deployments.

|  |  |  |
| --- | --- | --- |
| **Geta Apps** | **Volunteer Software Developer** | Nov 2024 -Feb 2025 |

An Investor-founder matching platform

* Built NestJS APIs and SQlite for an AI-powered investor-founder matching platform.
* Collaborated remotely via Slack in an Agile team to integrate AI decision-making features.

|  |  |  |
| --- | --- | --- |
| **Orbund Systems** | **Java Developer** | Jan 2012 - Jun 2012 |

A Student Information System (SIS) provider.

* Developed student info system modules with Java Spring & MySQL.
* Delivered features according to client requirements.

|  |  |  |
| --- | --- | --- |
| **Reve Systems** | **Graduate Software Developer** | Jan 2011 - Dec 2011 |

A Telecommunication and Software provider.

* Built e-commerce platform (LeReveCraze) from scratch in Java
* Collaborated across teams to align UI and backend features

|  |  |  |
| --- | --- | --- |
| **Career Break and Upskilling** | **Full-Time Parent** | 2016 - 2022 |

* Self-studied Python, Java, and cloud technologies.
* Returned to tech via postgraduate studies and volunteer roles.