

Kyrillos Isaak

+49015204228019 | kyrillosishak16@gmail.com | [linkedin.com/in/kyrillos-ishak](https://www.linkedin.com/in/kyrillos-ishak) | github.com/kyrillosishak

EDUCATION

Technische Universität Darmstadt

Masters of Science in Computer Science

Darmstadt, Germany

Oct. 2024

Faculty of Engineering, Alexandria University

Beng of Computer & Communication

Alexandria, Egypt

Oct. 2019 – Oct 2024

EXPERIENCE

AIML SWE Intern

Apple

February 2026 – present

Munich, Germany

- Owned the end-to-end design of a **probabilistic time-series forecasting** pipeline for modem profiling data, capturing temporal and feature-space correlations across multivariate sensor streams.
- Developed **anomaly detection** models on top of probabilistic forecasts, using predictive uncertainty to distinguish genuine anomalies from normal variation across multivariate sensor signals.
- Achieved real-time on-device inference under $\leq 10 \mu\text{s}$ latency and $\leq 64 \text{ KB}$ memory (L1 cache-resident) by evaluating model compression strategies (**quantization, knowledge distillation**) implemented in **Embedded C** with **NEON engine optimizations** and **Fast Math**.
- Built an **evaluation framework** to benchmark model performance without ground-truth labels, combining **synthetic anomaly injection** with domain-expert validation and bootstrapping labels via an **LLM-based RAG** pipeline that generates initial ground-truth candidates for expert review.
- Explored scaling a **Mixture-of-Experts (MoE)** model in tandem with the data — quantifying multivariate **subdomain shift** (**Wasserstein**, per-feature KS) and growing expert capacity as subdomains were introduced — to study specialization under on-device latency constraints.

Software Engineer Intern

Lexemo GmbH (Legal AI startup for automating legal workflows)

September 2025 – February 2026

Frankfurt, Germany

- Reduced disaster recovery RTO from 40+ hours to under 4 hours by auditing **Terraform & K8s**, enabling MySQL zone-redundant HA with automatic failover, extending backup retention from 7 to 30 days, and parameterizing **Helm** charts to bootstrap a standby cluster in a secondary AZ via a single **helm install**.
- Cut engineer onboarding time by 40% by building a unified **docker-compose** environment for backend, AI services, and MinIO S3 storage, enabling fully reproducible local development.
- Built automated integration tests for backend and AI server endpoints across multiple regions, catching configuration issues early and reducing deployment-related errors by 25%.

Software Development Engineer Intern

Amazon Web Services (AWS)

March 2025 – June 2025

Berlin, Germany

- Migrated Redshift Spectrum Canary Testing infrastructure from **CloudFormation** to AWS CDK (TypeScript, Java), reducing template duplication and enabling type-safe, reusable constructs across regions.
- Architected the test pipeline using Redshift for the target clusters under test, S3 for query/result artifact storage, **ECS-on-Fargate** with Docker for periodic isolated test execution, SQS to decouple scheduling from result processing, and **IAM** cross-account roles to securely federate access between the canary and target test accounts.
- Instrumented end-to-end observability with **CloudWatch** metrics, logs, and alarms across canary executions, enabling on-call engineers to detect regional Spectrum regressions within minutes instead of hours.
- Reduced deployment setup time by 50% by automating multi-region rollout, enabling reliable scaling of Redshift Spectrum Canaries across 40+ global AWS regions.
- Built network monitoring tooling (latency, packet loss, DNS resolution) for canary clusters, surfacing transient cross-AZ issues that were previously misattributed to Spectrum query failures.

Machine Learning Research Fellow

University of California, Santa Cruz

June 2024 – September 2024

Remote

- Conducted an in-depth research into irreproducibility in machine learning, analyzing 10+ research papers to identify and address data leakage issues.
- Resolved critical data integrity issues in HAM10000 skin cancer classification dataset by implementing proper data grouping strategies for 10,015 images, preventing overfitting and improving model generalization for clinical deployment.

- Enhanced fake news detection pipeline performance by identifying and fixing TF-IDF preprocessing errors, eliminating 20% performance inflation.
- Deployed and managed 30+ ML experiments on **ChameleonCloud** HPC infrastructure, orchestrating distributed training workflows across multiple GPU nodes to validate reproducibility findings.

PROJECTS

Financial Chatbot | *Python, Llama 3, QLoRA, FastAPI*

- Fine-tuned Llama 3 8B model using QLoRA for domain-specific Q&A in financial services, achieving 87% accuracy across English and Arabic queries on regulatory compliance topics.
- Developed Named Entity Recognition pipeline to extract financial entities (institutions, instruments, regulations) from mixed-language documents, improving information retrieval by 3x.

Writing Tools for Diverse Scripts | *SwiftUI, Core ML, Python*

- Built native ios application replicating intelligent writing assistance features (rewriting, summarization, key point extraction) both in Arabic and English.
- Implemented on-device inference using quantized Llama 3.2 1B model via Core ML, achieving 120ms average response time while maintaining privacy through local processing.
- Addressed code-switching challenges in mixed-script environments, enabling seamless transitions between different writing systems.

Distributed Consensus System (Raft Algorithm) | *Golang, TCP/IP, Docker*

- Implemented Raft consensus protocol to ensure fault-tolerant leader election and distributed coordination across nodes.
- Engineered a reliable TCP-based communication layer handling message ordering, retries, and network partition recovery.
- Automated failure detection + recovery mechanisms to maintain cluster availability under network disruptions.
- Deployed and tested a multi-node Docker cluster, demonstrating practical distributed infra skills.

PROBE | *Python, NixOS, Docker*

- Contributed to PROBE (Provenance for Replay Observation Engine) a system for capturing and analyzing execution data from computational experiments to enable **reproducible** research.
- Implemented the Replay phase, transforming provenance graph into reproducible workflows using CWL, Nextflow, Docker and Markdown, addressing complex challenges like missing metadata and nested dependencies.

Needle | *Cython, Cuda, C++*

- Created a machine learning framework simulating PyTorch features, supporting 10+ tensor operations for deep learning experimentation. Proposed arrayNd, a high-performance backend leveraging modern C++ advancements, improving matrix multiplication speed by 40%.
- Implemented an automatic differentiation engine, capable of training CNNs that reached 90% accuracy on the Cifar-10 dataset; extended support to RNNs, LSTMs, and attention-based models for sequential learning.

PUBLICATIONS

Learning from Irreproducibility: Data Leakage Case Studies for Machine Learning Education

ACM REP Conference

2025

- Developed educational framework to help researchers identify and prevent data leakage, contributing to reproducible machine learning research practices.
- Curated 4 case studies from reputable sources and created hands-on learning materials, enhancing comprehension of data integrity challenges in ML pipelines.

TECHNICAL SKILLS

Languages: Java, Python, C/C++, SQL (Postgres), JavaScript, HTML/CSS, R, Go, Cuda

Front/Back-End Frameworks: React, Next.js, Gradio, Node.js, Flask, FastAPI, JUnit

Data/Infrastructure Tools: Docker, Terraform, Serverless, AWS CDK, Selenium, Prometheus, Grafana

Distributed Systems: Raft Consensus, gRPC, HTTP/HTTPS, WebSocket, Network Fault Tolerance, Multi-node Coordination