Tamir Yirga

Seattle, WA | tamirkyirga@gmail.com | (206) 785-5928 | linkedin.com/in/tamir-yirga | github.com/tamirkifle | tamir.info

Education

Northeastern University Seattle, WA
Master of Science in Computer Science (GPA: 4.0 / 4.0) **Expected Dec 2026**

- **Graduate Teaching Assistant (5 terms)**: Scalable Distributed Systems, Cloud Computing, Computer Vision
- **Coursework**: Scalable Distributed Systems, Machine Learning, Computer Vision, Algorithms, Advanced Software Dev
- **Honors & Leadership**: Research Apprenticeship Award; Student Director, NU App Lab
- **Publication**: ***"Dreaming Up Authors: Factual Prevalence & LLM Hallucinations"*** - *EMNLP 2026 Main Conference* | Paper
  - Built a 5-stage ETL pipeline distilling 10M academic records into a 2M-paper corpus and a 9K balanced experiment set
  - Evaluated 5 LLMs via output metrics and hidden-state embeddings, linking training-data prevalence to hallucinations

Addis Ababa University Addis Ababa, Ethiopia
Bachelor of Science in Civil Engineering

Skills & Certifications

Languages: Java, Python, Rust, C++, Go, TypeScript, SQL
Distributed Systems: Apache Spark, Kafka, RabbitMQ, Redis, Apache Arrow, gRPC
Backend Frameworks: FastAPI, Spring Boot, Node.js
Databases: PostgreSQL, DynamoDB, MySQL, MongoDB
Cloud & Infrastructure: AWS, GCP, Linux, Docker, Kubernetes, Terraform, Prometheus, Grafana

Certifications: Google Cloud Certified Professional Cloud Architect

Work Experience

LakeSail **San Francisco, CA**
Software Engineer Intern - Open-Source Rust Query Engine Jan 2026 - May 2026

- Shipped Spark 4.1 SQL TIME support across the query parser, analyzer, Arrow conversion, and Spark Connect layers, clearing 46 gold-data test failures
- Owned the Arrow-native vectorized runtime for Python UDFs/UDTFs, adding 6 eval types and keyword-arg dispatch, eliminating Pandas serialization overhead
- Benchmarked CPU/GPU execution paths against JVM Spark across 5 AI workloads, measuring 80x lower serialization overhead and 3.5x faster embedding pipelines

Canaria **New York, NY**
Software Engineer Intern - Labor Market Intelligence Platform Jul 2025 - Sep 2025

- Designed a PostgreSQL analytics layer with 28-column materialized views and composite GIN/GiST indexes, removing multi-table JOINs for sub-second PostGIS and time-series queries
- Shipped a Kafka-to-PostgreSQL ingestion pipeline with automated partition pruning and rolling-window retention, powering real-time analytical dashboards
- Added schema-scoped Redis caching for text-to-SQL queries with TTL eviction, cutting repeat latency from 2.8s to 22ms

Scandiweb **Riga, Latvia**
Full Stack Engineer - Enterprise eCommerce May 2022 - Aug 2024

- Delivered features across 10 enterprise platforms and 5+ stacks, driving requirements directly with client stakeholders
- Rebuilt multi-step checkout and payment integrations for high-volume storefronts, cutting cart abandonment 18%

Projects

InferRS – *LLM Inference Engine* | *Rust, LLM Inference, Quantization* | GitHub *Spring 2026*

- Engineered a zero-copy Rust LLM engine with memory-mapped GGUF parsing, KV-cache, and grouped-query attention
- Implemented INT8 quantization and parallel ARM NEON/AVX2 SIMD kernels with size-aware runtime dispatch

LedgerKV – *Distributed Key-Value DB* | *Java, gRPC, Raft, Kubernetes, Prometheus* | GitHub *Fall 2025*

- Built a distributed KV database with custom LSM-tree storage engine, tunable quorum consensus, and vector clocks
- Added Prometheus/Grafana observability and a Jepsen linearizability checker; mitigated tail latency via hedged requests

Liftline – *Distributed Ski Tracker* | *Java, Spring Boot, RabbitMQ, DynamoDB, Redis* | GitHub *Spring 2025*

- Engineered a distributed lift-ride event ingestion and aggregation service using Spring Boot, RabbitMQ, and single-table DynamoDB, layered with cache-aside Redis reads and queue-depth admission control for graceful load shedding
- Cut DynamoDB write requests 96% via client micro-batching; verified zero event loss across 4 consumer-kill faults