

FATIMA FAROOQ

Arlington, TX 76013 | +1-817-970-3202 | fatimafarooq2310@gmail.com
[linkedin.com/in/fatimaa-farooq](https://www.linkedin.com/in/fatimaa-farooq) | github.com/fatimaafarooq | fatimafarooq.com

SUMMARY

Third-year PhD researcher in the [Large-Scale Intelligent Data Systems \(LIDS\) Lab](#) at UT Arlington, working on machine learning models that operate directly over relational databases and web-scale knowledge graphs. Co-first author on two manuscripts under review: a relational foundation model that transfers across unseen databases without retraining, and **gHAWK**, a structural encoding method that topped the [Open Graph Benchmark](#) leaderboard for 3 knowledge graph benchmarks. Spent 2.5 years as a senior AI developer building and deploying real-time machine learning systems in production at 800 to 1,000 events per second. Strongest in PyTorch and graph learning, C++ and database internals, and Linux GPU infrastructure. Seeking a **Summer 2027 Research or ML Engineering internship**.

EDUCATION

Ph.D., Computer Science

Aug 2024 to Present

[University of Texas at Arlington](#)

CGPA 4.00/4.00

Advisor: Prof. Ashraf Aboulmaga, Large-Scale Intelligent Data Systems (LIDS) Lab. Merit-based STEM scholarship.

M.S., Data Science

Sept 2020 to Aug 2022

[FAST National University of Computing and Emerging Sciences, Pakistan](#)

CGPA 3.74/4.00

Silver Medalist, graduated 2nd in cohort.

B.S., Software Engineering

Oct 2015 to Aug 2019

[University of Engineering and Technology, Taxila, Pakistan](#)

CGPA 3.48/4.00

PUBLICATIONS

[1] **Fatima Farooq***, Humera Sabir*, Ashraf Aboulmaga. "FORTE: A Relational Foundation Model with Foreign-Key-Role Awareness." *Manuscript under review*, 2026.

One model that learns directly from multi-table databases and transfers to new databases and tasks without retraining.

[2] **Fatima Farooq***, Humera Sabir*, Ashraf Aboulmaga. "gHAWK: Structural Encoding for Scalable Training of Graph Neural Networks on Knowledge Graphs." *Manuscript under review*, 2026. Preprint: [arXiv:2512.08274](https://arxiv.org/abs/2512.08274).

Topped the OGB leaderboard for 3 knowledge graph benchmarks: MAG240M (244M nodes), OGB-MAG, OGB-WikiKG2.

*Equal contribution.

RESEARCH EXPERIENCE

Graduate Research Assistant

Aug 2024 to Present

[Large-Scale Intelligent Data Systems \(LIDS\) Lab](#), UT Arlington

Arlington, TX

- Designed **FORTE**, a zero-shot generalized model that learns directly from multi-table databases and predicts without hand-built features for unseen downstream task. Supports classification, regression, recommendation, and link prediction in one architecture.
- Engineered the model to distinguish the role of each link between tables. Where a buyer and a seller both reference the same users table, it preserves their separate meanings instead of collapsing them, improving accuracy on nearly every task in a public benchmark of real-world business databases.
- Developed two pretraining methods, **Contrastive Link Prediction** and **Partial Schema Perturbation**, so the model learns general database structure and stays accurate when a new database is missing tables or laid out differently.
- Co-developed **gHAWK**, a preprocessing step that summarizes graph structure before training, reducing GPU memory and accelerating training on graphs with hundreds of millions of nodes. Integrates with PyTorch Geometric and DGL. Topped the OGB leaderboard for 3 benchmarks. Owned all node classification method design and experiments.
- Researching how to select the most informative slice of a large database per query, keeping training and inference within practical memory and latency budgets.
- Benchmarked large graph models at web scale: MAG240M (244M nodes), OGB-MAG, OGB-WikiKG2 (2.5M entities, 17M triples).
- Administer the lab's shared GPU cluster**: user accounts and permissions, CUDA and driver upgrades, Python environments, storage and dataset management, job scheduling, and uptime monitoring.

Graduate Student Researcher

AIM Lab, [FAST NUCES](#)

Sept 2021 to Aug 2022

Islamabad, PK

- Built a transformer-based speech recognition system for Urdu by fine-tuning wav2vec 2.0, reaching **16% word error rate** and **8% character error rate**, competitive with well-resourced languages at similar corpus size.
- Curated the audio and transcription corpus from scratch. Implemented dialect detection using word2vec and transformer encoders.

Theses

Transformer-based Automatic Speech Recognition for Low-Resource Languages

2022

M.S. Thesis, FAST NUCES. Supervisor: [Supervisor Name]

Islamabad, PK

- Addressed the absence of usable speech recognition for Urdu, a language with over 230 million speakers and almost no labelled training data.
- Collected and transcribed an Urdu speech corpus from scratch, defining the annotation protocol and handling speaker and dialect variation across recordings.
- Adapted the self-supervised **wav2vec 2.0** architecture through fine-tuning on the limited labelled corpus, tuning the CTC decoding head, learning rate schedule, and freezing strategy for a low-resource regime.
- Achieved **16% word error rate** and **8% character error rate**, competitive with results reported for well-resourced languages at comparable corpus size.
- Built a dialect identification model using word2vec and transformer encoders over the same corpus.

Autonomous Driving Simulator using OpenCV and Deep Learning

2019

B.S. Thesis, UET Taxila. Supervisor: [Supervisor Name]

Taxila, PK

- Designed and built an end-to-end driving simulator to generate perception training data without the cost and risk of real-world vehicle data collection.
- Implemented object detection by fine-tuning **MobileNet-SSD**, and lane detection using camera calibration with perspective transform and curve fitting.
- Developed the simulation environment in Delphi, a sparsely documented language, working from limited reference material.
- Awarded **First Prize** at UET Taxila Open House 2019; the underlying methodology received the **GRAND PRIX AWARD** at **CPWC 2018**, Tokyo.

INDUSTRY EXPERIENCE

Senior AI Developer

Jan 2022 to Jul 2024

Center of AI and Computing (CENTAIC), [NASTP](#)

Islamabad, PK

- Led design and production deployment of real-time ML pipelines classifying aircraft trajectories from streaming radar data, sustaining **800 to 1,000 tracks per second** under live operational load.
- Built anomaly detection models over continuous sensor streams and integrated inference output with PyQGIS for real-time geospatial visualization used by operational analysts.
- Engineered a modular, sensor-agnostic decision support system combining trajectory classification, anomaly alerts, and geospatial context into one operator-facing pipeline.
- Deployed a **Memgraph** graph database over noisy geospatial data with Cypher queries, enabling multi-hop relational analysis the prior tabular schema could not support.
- Designed a Multiple Hypothesis Tracking framework for multi-target tracking under uncertainty. Built the distributed IPC architecture and stress-testing harness.
- Led a multidisciplinary team of engineers and scientists. Shipped cross-platform Qt applications.

Lab Instructor

Jan 2019 to Jan 2022

[FAST NUCES](#)

Islamabad, PK

- Taught Data Analysis, Data Visualization, Databases, Object-Oriented Programming, and Digital Logic Design over 3 years to hundreds of students per semester. Mentored student ML and data pipeline projects.

Software Engineer, Deep Learning

Apr 2018 to Jul 2019

VR and Motion Simulation Lab, [UET Taxila](#)

Taxila, PK

- Built an autonomous driving simulator using OpenCV and deep learning (MobileNet-SSD object detection, camera-calibration lane detection) to generate low-cost perception training data.
- Won **1st Prize** at UET Open House 2019 and the **GRAND PRIX AWARD** at CPWC Tokyo 2018. Organized VR workshops with 100+ attendees.

TECHNICAL SKILLS

Languages	Python, C/C++, SQL, Java, R, MATLAB, Cypher
ML and DL	PyTorch, PyTorch Geometric, DGL, Hugging Face Transformers, TensorFlow, Keras, scikit-learn
Research Areas	ML on databases and graphs, Graph Neural Networks, Transformers, Self-Supervised Pretraining, Few-Shot and Zero-Shot Transfer, Anomaly Detection, Time-Series Classification, Speech Recognition
Data and Systems	Distributed training, Apache Spark, Hadoop, Hive, Pig, Memgraph, pthreads, IPC, CMake, gdb/valgrind/perf, Qt
Infrastructure	Linux and GPU server administration, CUDA and driver management, Conda and Docker environments, job scheduling, shared storage, user access management, cluster monitoring
Database Internals	B+-tree and LSM indexing, buffer pool management, WAL and redo logging, query execution (joins, projection, aggregation)
Benchmarks	OGB (MAG240M, WikiKG90M, WikiKG2), RelBench , FB15k-237
Tools	Git, OpenCV, D3.js, PyQGIS, QGIS, $\LaTeX$

SELECTED PROJECTS

- **B+-Tree Index and Query Executor (C++)**. Implemented B+-tree indexing, buffer pool management, WAL and redo logging, and a query executor with nested-loop join, projection, and aggregation from scratch.
- **Recommendation System at Scale**. Collaborative filtering and content-based hybrid engine over 25M records, with a D3.js dashboard for revenue and popularity analysis.
- **Few-Shot Event Detection**. Black-box meta-learning for event detection where labeled training data is scarce.
- **Darknet2020 Network Traffic Analysis**. Large-scale traffic analysis for threat detection using distributed analytics.
- **DBLP Bibliometric Prediction**. Predicted publication venue trends with neural models and t-SNE visualization, achieving **93% accuracy**.
- **Big Data Text Mining (Pokec)**. Association rules, frequent-pattern mining, and LSH over unstructured social graph data.
- **COVID-19 Spread Dashboard**. Interactive D3.js dashboards integrating multi-country pandemic data.
- **Freelance AI and ML Consulting**. Completed **200+ projects** in deep learning, computer vision, and applied data science for global clients over 5+ years. Top-rated seller.

AWARDS AND HONORS

- **2nd Place, [Innovation Day 2026](#)**, UTA College of Engineering. Graduate category, awarded for gHAWK, selected university-wide.
- **GRAND PRIX AWARD (Winner)**, [6th Cloud Programming World Cup](#), Forum8, Tokyo, Japan, 2018. International competition with teams from multiple countries.
- **Silver Medal**, MS Data Science, [FAST NUCES](#), 2022. Graduated 2nd in the entire cohort.
- **First Prize**, Final Year Project, Open House and Job Fair, UET Taxila, 2019. Selected from all graduating engineering projects university-wide.
- **Merit-based STEM Scholarship**, [University of Texas at Arlington](#), 2024.
- **HEC Government Merit Scholarship**, [Higher Education Commission of Pakistan](#), 2017 to 2018.
- **3rd in Province**, International Kangaroo Mathematics Contest, 2014. Administered in 90+ countries.
- **Microsoft Student Partner**, UET Taxila, 2017 to 2018. Competitively selected campus ambassador.