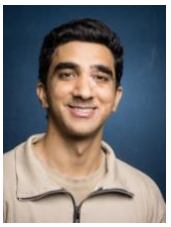
Asfand Yar Khan

a.yar86.ay@gmail.com | +491634871130 | https://linkedin.com/in/asfand-yar-ahmad-khan | github.com/AsfandYar98 | asfandyar.vercel.app | Nürnberg, Deutschland



Summary

Data Scientist and **Software Engineer** with 3+ years of experience across **machine learning**, **backend development**, and **MLOps**. Skilled in designing **scalable APIs**, deploying **ML models** in cloud-native environments such as **Kubernetes, Docker, and Google Cloud Platform**, and building robust data pipelines. Strong command of **Python**, **Java**, **Spring Boot**, and ML frameworks including **PyTorch** and **TensorFlow**. Experienced in time series forecasting, model tracking with **MLflow**, and full SDLC delivery.

Professional Experience

AI *Engineer*

Körber Supply Chain Logistics GmbH, Nürnberg *02/2025 - Present*

- Built a **synthetic data pipeline** with ControlNet inpainting (diffusion models), boosting dataset diversity **+40%** and detection accuracy **+10%** on rare damages.
- Tackled severe **class imbalance** (~5% damages) with focal/class-weighted loss and balanced sampling, significantly reducing **false negatives**.
- Deployed a **GPU autoscaling inference service** on Kubernetes (p99 ≤ 300 ms) with Airflow-driven training, MLflow registry, canary rollouts, and KPI-based auto-rollback.
- Implemented metrics & drift monitoring (recall, confusion matrix, CLIP-based embedding drift) to ensure **real-time** model performance and trigger proactive retraining.

Master's Thesis – Data Science Researcher

Fraunhofer IIS, Nürnberg *01/2025 – 07/2025*

- Engineered an **end-to-end forecasting pipeline** (Airflow, Kafka, FastAPI, Redis, MLflow, GluonTS) for household heat demand with scalable ingestion, feature computation, and training.
- Developed **stream processing** (Kafka Streams) for validation, feature generation, and DWD joins; persisted curated data in a **Data Lake**.
- Benchmarked Moirai, Amazon Chronos, ARIMA, TFT → achieved **22% lower RMSE** and **30% faster inference**; extended Chronos with weather+thermal features for **15–18% MAE/RMSE gain**.
- Delivered a **Forecast API** (p95 ≤ 200 ms) with Triton CI/CD deployment, observability (Prometheus/Grafana, ELK), and GDPR-compliant security stack.

Software *Engineer*

Körber Supply Chain Logistics GmbH, Nürnberg *05/2023 – 02/2025*

- Built a **logistics simulation platform** (Angular SPA + Java Spring Boot Orchestrator on Kubernetes Jobs) with **real-time feedback** via PostgreSQL.
- Secured APIs with Gateway, WAF, and KMS; scaled DB with batching, partitioning, and replicas powering Grafana/BI dashboards.
- Automated **CI/CD (GitHub Actions + Helm)** with JUnit/PyTest, static analysis, canary rollouts, and observability (Prometheus, Grafana), cutting shipped bugs **~40%**.
- Enabled **10× scalability** with stateless Orchestrator, HPA, FIFO scheduling, and concurrency caps.
- Actively contributed to **code reviews**, **sprint planning**, and end-to-end release processes, ensuring **code quality**, **sprint velocity**, and smooth **CI/CD-driven deployments** using **GitLab and Docker**.

Software *Engineer*

Exper Labs *07/2020 – 10/2022*

- Built a **polyglot microservices platform** (Rails BFF + Spring Boot) on EC2 with PostgreSQL and API Gateway (WAF, OIDC, JWT, RBAC, TLS, KMS) powering 3+ high-traffic client platforms with **99.9% uptime**.
- Designed a **low-latency booking pipeline** (Redis, PostgreSQL, Elasticsearch), achieving P95 ≤ 350 ms for writes and ≤150 ms for search.
- Delivered **real-time observability** (Prometheus, Grafana, OpenTelemetry) with CI/CD rollouts, health checks, and auto-rollback, sustaining **99.9% uptime**.
- Tuned Elasticsearch (shards, caching, routing) to optimize throughput and relevance under scale reducing latency by 30%,

Education

M.Sc. Data Science

Friedrich-Alexander-Universität Erlangen-Nürnberg 2023 – 2025

B.Sc. Computer Science

FAST-NUCES, Lahore 2016 – 2020

Skills

- **Languages & Frameworks :**
  Python, Java, Ruby on Rails, Spring Boot, SQL, REST API
- **Data Science & ML :**
  TensorFlow, PyTorch, Scikit-learn, GluonTS, Numpy, Feature Engineering, LangChain, Transformers, Retrieval-Augmented Generation (RAG), GenAI, Diffusion Models, Foundation Models, LLMs
- **Data Engineering & MLOps :**
  MLflow, Docker, Kubernetes, Terraform, CI/CD, Apache Spark, Redis, ETL Pipelines
- **Cloud & Analytics Tools**:
  Google Cloud Platform, Grafana, Kibana, Power BI, Google Analytics
- **Databases & Tools:**
  PostgreSQL, MongoDB, Elasticsearch , Git, Jira, Agile/Scrum

Personal Projects

Medical Chatbot Assistant

- Built a **medical Q&A system** using **BioBERT**, **Pinecone**, **OpenAI GPT-4**, and **Streamlit UI** delivering citation-backed answers.
- Designed **retrieval pipeline**: PII/PHI redaction → embeddings → semantic search in Pinecone → GPT-4 generation with inline citations.
- Implemented **document ingestion** (PDF/HTML) with GPU batch embeddings and idempotent upserts into Pinecone + PostgreSQL + S3.
- Achieved **scalable low-latency RAG** via Redis caching, autoscaling, and streaming LLM responses; secured system with **OIDC/OAuth2, RBAC, WAF, TLS**.

Personalized Tutoring Assistant

- Designed an **AI tutoring system** with **Llama 3.2 (vLLM)**, **Chroma DB**, **PostgreSQL**, and **Redis**, supporting chat Q&A, content ingestion, and quiz generation.
- Implemented **hybrid retrieval** (semantic + keyword search with MMR diversification) and Redis caching to cut token costs and latency.
- Built ingestion workflow (GPU embeddings + idempotent upserts) with **sharding, replication, and snapshots** for reliability.
- Delivered **quiz generation** via RAG Orchestrator + Llama 3.2 producing structured quizzes with mastery tracking.
- Ensured **scalability & compliance**: autoscaled GPU pools, GDPR-ready data handling, audit logging, and moderation filters.

Languages

- English – Fluent (C1)
- German – Intermediate (B1, improving)