Talha Rashid

AI Evaluation Analyst — Model Benchmarking & Applied ML

Talha.rashid344@gmail.com — +92 333 9512443 — Remote (Islamabad, PK – UTC+5)
rtalhaa.github.io/portfolio — github.com/rTalhaa

Profile

Data Science graduate whose work centres on measuring model quality, not only building models: controlled fine-tuning benchmarks, multi-model comparison under a single held-out metric, sliced-metric validation in TFX, and groundedness checks on LLM outputs. I define what "better" means for a model, build the harness that measures it reproducibly, and write the result up so a non-technical reader can act on it.

Evaluation Skills

Evaluation	Benchmark design, held-out validation, model selection, ROC-AUC, precision, recall, F1, accuracy, confusion-matrix analysis, evaluation loss, error analysis, class-imbalance handling
LLM evaluation	Groundedness and source-attribution checks, retrieval-quality inspection, hallucination reduction, fine-tuning ablations, quality/cost trade-off reporting
Python	pandas, NumPy, scikit-learn, PyTorch, TensorFlow Extended (TFX), matplotlib, Jupyter, pytest
ML / NLP	Supervised classification, ensemble methods, embeddings, RAG, QLoRA/LoRA adapters, DeepSpeed ZeRO-2, PubMedBERT, BioMistral, Qwen2.5
Pipelines	TFX components (ExampleGen, StatisticsGen, ExampleValidator, Trainer, Evaluator, Pusher), TensorFlow Model Analysis, Apache Airflow, MLflow
Reproducibility	Git, GitHub Actions CI, Docker, MLflow tracking, fixed seeds and matched-budget comparisons, Prometheus/Grafana monitoring, technical write-ups for non-technical stakeholders

Evaluation & Analysis Projects

Fine-Tuning Strategy Benchmark: QLoRA vs. DeepSpeed ZeRO-2 *PyTorch, Qwen2.5-0.5B-Instruct, CodeAlpaca*

- Designed a controlled benchmark comparing two fine-tuning strategies on the same base model and dataset, holding the training budget matched so the comparison isolated the variable of interest.
- Evaluated single-GPU QLoRA (Tesla T4) against dual-GPU DeepSpeed ZeRO-2 (RTX A4000) across evaluation loss, wall-clock runtime, memory footprint, and trainable-parameter count.
- Quantified the parameter-efficiency trade-off directly – 8.80M trainable adapter parameters versus 494.03M full parameters – and wrote up where each strategy is the right choice, with harness and configs committed alongside the results.

Multi-Model Benchmarking on Tox21 Molecular Toxicity *scikit-learn, RDKit, MLflow, FastAPI, GitHub Actions*

- Benchmarked Logistic Regression, Random Forest, and Extra Trees under identical RDKit Morgan-fingerprint features on the Tox21 SR-ARE endpoint, selecting Extra Trees on a held-out ROC-AUC of 0.7648.
- Chose ROC-AUC deliberately over accuracy, which is misleading on a task this heavily imbalanced.
- Made the evaluation repeatable rather than one-off: tracked every run in MLflow, containerised the pipeline, added API tests and CI validation via GitHub Actions, and exposed live prediction metrics through Prometheus and Grafana.

Automated MLOps Pipeline with TensorFlow Extended (TFX) *TFX, TensorFlow Model Analysis, Apache Airflow, MLflow*

- Built an Airflow-orchestrated TFX pipeline so data validation, training, and evaluation run as versioned, repeatable stages, using StatisticsGen and ExampleValidator to catch schema drift and training/serving skew before those faults reached the model.
- Applied the TFX Evaluator with TensorFlow Model Analysis to compute sliced metrics across data segments, surfacing subgroups where aggregate accuracy hid materially worse performance.
- Gated promotion on TFMA thresholds against the production baseline, so a candidate reached the Pusher stage only if it measurably beat the incumbent.

SkinSync: RAG Output Quality & Groundedness Evaluation *LangChain, Qdrant, PubMedBERT, BioMistral*

- Built a retrieval-augmented skincare assistant and evaluated it on what matters in a health-adjacent domain: whether each answer is actually supported by retrieved evidence.
- Traced weak answers back to retrieval failures versus generation failures across chunking strategies and 768-dimensional biomedical embeddings, then cut unsupported claims by constraining prompts and returning source-attributed responses auditable against the cited passage.

Amazon ASIN Search Performance Assessment *pandas, KPI diagnostics, stakeholder reporting*

- Decomposed 60-day search data across three products into visibility, click-through, conversion, and profitability so each failure mode was addressable separately.
- Identified the strongest performer on 1.19% CTR, 11.07% CVR, and 23.36% ACOS, then wrote non-technical recommendations with defined success KPIs.

Bay G Pharma ERP – LLM Analyst in Production *ERPNext v16, LLM workflows, Python*

- Delivered a production ERP running a live pharmaceutical distribution business, with an LLM analyst layer answering operational questions over real company data.
- Validated LLM responses against ground-truth ERP records before release – evaluation under real commercial stakes rather than on a benchmark set.

Education & Certifications

BS Data Science, FAST NUCES 2021–2025
Coursework: Machine Learning, Natural Language Processing, Statistics, Data Structures, Databases, Linear Algebra.
Perform Cloud Data Science with Azure Machine Learning, Microsoft — **Google Data Analytics Certificate**, Google