TAIMOOR KHAN

AI Engineer | LLM & RAG Systems | Applied AI
+92 348 484 2212 | taimoorkhaniajaznabi2@gmail.com | github.com/TaimoorKhan10

PROFESSIONAL EXPERIENCE

Ninth Dev Solutions — *AI Engineer* Nov 2022 – Dec 2024

- Engineered an LSTM-based predictive analytics platform processing EHR data from 15 hospitals, forecasting patient health trajectories and adverse events at scale.
- Built a stacked ensemble model (XGBoost + LSTM) for multivariate solar-energy forecasting, deployed to production with automated retraining via MLflow pipelines.
- Designed GRU-based sequence models over 5TB of retail transaction data to extract purchase-propensity and churn signals, feeding downstream business decision systems.
- Applied K-Means + PCA on hospital operational data, surfacing efficiency bottlenecks that cut operational costs by 12% and improved resource allocation by 18%.
- Led end-to-end delivery of YOLOv4 edge deployment (mask detection) and BERT-based NLP tools — owned model quantization, pruning, and CI/CD integration.

FLAGSHIP PROJECT

CliniSynapse AI | *Clinical Decision Support System* | *RAG · LLM · BioBERT · FAISS · FastAPI* 2025 – 2026

- Architected and deployed a production-grade RAG system over 167K+ real patient case records (PMC-Patients), delivering sub-200ms end-to-end clinical query responses.
- Designed a 3-layer hybrid retrieval pipeline: FAISS IVFFlat semantic search (BioBERT embeddings, ~10ms, ~97% recall, 20–50× speedup) → clinical feature overlap scoring with domain-aware normalization → cross-encoder reranking for deep contextual relevance.
- Integrated PubMed live evidence retrieval and a strict RAG constraint layer ensuring all LLM outputs (Gemini / GPT-4 class) are grounded in retrieved evidence — hallucinations explicitly blocked via engineered prompts.
- LLM output pipeline generates structured clinical reports: diagnosis, differential, clinical reasoning, evidence validation, confidence scoring, and hallucination checks.
- Built a 5-layer multi-level caching system and GPU+CPU parallel execution pipeline, eliminating redundant API calls and cutting latency by >40% under production load.
- Shipped full-stack system: real-time doctor-facing query interface, multi-turn Syna AI assistant, admin panel with eval metrics, and auth with secure session management.
- Built fine-tuning pipelines for BioBERT bi-encoder and cross-encoder reranker with automatic weight override on improved checkpoints.

SELECTED PROJECTS

Enterprise-RAG-Framework | *Modular RAG Architecture* 2024

- Engineered a production-grade, modular RAG system with pluggable retrieval backends, configurable reranking stages, and multi-tenant context isolation for enterprise deployments.

MLOps-Forge | *ML Lifecycle Automation* 2023

- Built end-to-end ML pipelines covering experiment tracking (MLflow), model versioning, automated evaluation gates, and deployment workflows on AWS.

AI-Fairness-Explainability-Toolkit | *Model Interpretability* 2023

- Developed a bias detection and explainability framework (SHAP, attention maps, fairness metrics) for auditing production ML models across sensitive domains.

RESEARCH & PUBLICATIONS

Distractor-Aware Memory (DAM) for Robust Visual Object Tracking
Lead Author | PAF-IAST, Haripur, Pakistan | 2026

- Designed a dual-branch memory architecture on SAM 2 with sigmoid-gated cross-attention fusion; achieved 12.4% relative EAO improvement on the VOT2022 benchmark at ~45 ms/frame with only ~3% computational overhead.

TECHNICAL SKILLS

Languages: Python (Advanced), C++, JavaScript, TypeScript
AI / ML: RAG Systems, LLM Integration, Prompt Engineering, Transformers, BioBERT, Sentence Transformers, FAISS, Cross-Encoders, Federated Learning, GANs, RL (PPO, DDPG)
Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, HuggingFace
Backend & Systems: FastAPI, Uvicorn, SQLite, Docker, AWS
MLOps & Tools: MLflow, Git, Jupyter, Linux, MongoDB

EDUCATION

Bachelor of Science in Artificial Intelligence (BSAI) Completed