Nadeem Akhtar

AI Researcher & Engineer · Bengaluru, India
nadeemnns2000@gmail.com · LinkedIn
GitHub · Hugging Face · Kaggle · Portfolio

Summary

AI Researcher & Engineer specializing in **LLM fine-tuning, model compression, RAG, and enterprise document intelligence**. 1.5+ years of production experience at **Synaptics**, with AI automation live for the Korea division and rollout underway across 8 global regions. **First author of 2 accepted Scopus-indexed papers and 2 more under review** (ICLR 2027, ARR/NAACL), with public datasets and models on **Hugging Face**.

Experience

AI Researcher & Engineer – Synaptics *June 2025 – Present*
Bengaluru, India

- Led end-to-end design and rollout of an **Accounts Payable Invoice Automation** platform – **live for the Korea division**; rollout underway across 8 global regions.
- Built the original production POC (Python, FastAPI, Next.js/React, Oracle DB, Redis, Graph API) with live KPI dashboard; presented at Synaptics AI Day and recognized as the best AI Work.
- Architected the enterprise stack after the POC: Oracle Integration Cloud ingestion, Azure Content Understanding extraction, OCI Functions transformation into Oracle staging and AP interface tables.
- Designed a scheduled CAD data extraction pipeline (Alteryx Server + embedded Python + Azure Document Intelligence custom model), replacing a week-long manual process with a few-day automated flow; JSON/CSV outputs feed Jira costing and price negotiations.
- Built a Procurement chatbot embedded in Oracle EBS (Python, FastAPI, Redis, RBAC, Oracle federated Microsoft auth, read-only SQL agent) enabling Natural Language queries for PO status, open POs by department, and pending approvals.

Python Developer Intern – Qedge Tech *July 2022 – Nov 2022*

- Built backend applications with Python, Flask, and Django; implemented APIs, business logic, and database integrations.

Publications & Research

- **FedQuant: Mixed-Precision Federated Fine-Tuning of LLMs under Android Hardware Heterogeneity.** *Accepted (in press), PCCDA 2026, Springer LNNS (Scopus).*
- **Fine-Tuning LLMs for Indian Legal NLP: A Multi-Task Comparative Study.** *Accepted (in press), CVR 2026, Springer LNNS (Scopus).*
- **Beyond Zero-Padding: Nested-SVD Aggregation for Heterogeneous-Rank Federated LoRA.** *Under review, ICLR 2027.*
- **Beyond Answer Accuracy: Step-Level Evaluation of Compressed Small Language Models.** *Under review, ARR Oct 2026.*

Public models and datasets: huggingface.co/nadeem172.

Projects

ModelBench – Open-Weights & Frontier AI Benchmark Platform Live Demo

- Built a dynamic benchmark and evaluation platform tracking 890+ frontier and open-weights AI models across 270+ benchmarks with traceable source citations, context window comparisons, and live pricing.

Multi-Agent Conversational Data Analyst Live Demo

- End-to-end analytics engine: NL→SQL via LLM, context-aware memory, Chart.js visualizations, SQL-injection prevention, Docker.

AI Resume Analyzer Live Demo

- FastAPI + Sentence Transformers resume–JD matching via cosine similarity, with LLM-generated feedback; deployed on Hugging Face Spaces.

Technical Skills

- **LLM/SLM:** PyTorch, Hugging Face (Transformers, PEFT, bitsandbytes), LoRA, QLoRA, Diffusion LMs, Federated Learning, RAG, Autogen.
- **Compression & Optimization:** 4-bit NF4, GGUF/llama.cpp, AWQ, GPTQ, SmoothQuant, SparseGPT, Wanda, QAT, Pruning.
- **Engineering & Cloud:** Python, FastAPI, Docker, Redis, Kafka, Oracle SQL, MongoDB, Vector DBs (FAISS, Chroma); Azure (ML, Document Intelligence), OCI, AWS, GCP, NVIDIA H100 HPC.
- **Evaluation & Stats:** McNemar, Wilcoxon, Cohen's d, Human Annotation, LLM-as-Judge.

Education

M.Tech in Data Science *2024 – 2026*
National Institute of Technology Jalandhar *CGPA: 7.80 / 10.0*

B.Tech in Computer Science and Engineering *2019 – 2023*
Shri Mata Vaishno Devi University *CGPA: 7.27 / 10.0*