Md Nadeem Akhtar

AI Researcher and Engineer
Bengaluru, India · nadeemnns2000@gmail.com
Hugging Face · GitHub · LinkedIn · ORCID ID · Portfolio

Summary

Researcher in efficient large language models, covering parameter-efficient fine-tuning, federated learning, model compression, and reasoning evaluation. Two papers accepted (in press) at Springer LNNS venues, two under review, and one in preparation. Models and datasets are public on Hugging Face.

Research Interests

Legal NLP; LoRA; QLoRA; DoRA; prefix tuning; adapter tuning; instruction tuning; supervised fine-tuning; RLHF; DPO; reward modeling; continual learning; catastrophic forgetting; federated learning; FedAvg; heterogeneous clients; differential privacy; quantization; pruning; knowledge distillation; mixed precision; model compression; edge AI; retrieval-augmented generation; vector search; embeddings; named entity recognition; text classification; semantic similarity; document AI; optical character recognition; chain-of-thought reasoning; multilingual NLP; cross-lingual transfer; low-resource languages; evaluation of LLM reasoning; step-level evaluation; LLM-as-judge; diffusion language models; mechanistic interpretability.

Education

M.Tech in Data Science *2024–2026*
Dr. B.R. Ambedkar National Institute of Technology Jalandhar *CGPA: 7.80/10*
Dissertation: *Fine-Tuning Large Language Models for Indian Statutory Legal NLP using Low-Rank Adaptation.* Supervisor: Dr. Amritpal Singh.

B.Tech in Computer Science and Engineering *2019–2023*
Shri Mata Vaishno Devi University (SMVDU) *CGPA: 7.27/10*

Publications

Accepted (in press)

- **Akhtar, M. N. and Singh, A.** "FedQuant: Mixed-Precision Federated Fine-tuning of Large Language Models under Android Hardware Heterogeneity." *PCCDA 2026, Sejong University, Seoul. Springer LNNS (Scopus-indexed).*
- **Akhtar, M. N. and Singh, A.** "Fine-Tuning Large Language Models for Indian Legal NLP: A Multi-Task Comparative Study." *CVR 2026, NIT Goa. Springer LNNS (Scopus-indexed).*

Under review

- **Akhtar, M. N.** et al. "Beyond Zero-Padding: Nested-SVD Aggregation for Heterogeneous-Rank Federated LoRA." *Under review, ICLR 2027.*
- **Akhtar, M. N.** et al. "Beyond Answer Accuracy: Step-Level Evaluation of Compressed Small Language Models." *Under review, ARR October 2026 cycle.*

In preparation

- **Akhtar, M. N.** et al. "Cross-Lingual Reasoning Elicitation Gaps in Diffusion Language Models." *Targeting ACL 2027 (ARR submission deadline 4 January 2027).*

Research Projects

NSVD-Agg: Heterogeneous-Rank Federated LoRA *2026*

- Proposed a nested-SVD aggregation method that removes zero-padding communication waste when clients train LoRA adapters at different ranks. Achieved a **53.1% communication reduction** with lower final loss across 7 seeds on Llama-3.2-3B and Qwen2.5-3B (Wilcoxon $p = 0.0156$; Cohen's $d = 5.76$–7.89).
- Derived a stationarity bound under non-convex assumptions, with a four-term error decomposition that isolates rank-induced projection error.

Step-Level Evaluation of Compressed Small Language Models *2026*

- Designed a study of about 49,000 generations testing whether 4-bit quantization and 50% pruning degrade reasoning validity beyond final-answer accuracy.
- Introduced a step-level annotation schema covering error propagation, recovery, and compensating-error masking. Validated it with two human annotators and an LLM judge. Evaluated 7 model families across 6 compression methods.

Cross-Lingual Reasoning in Diffusion Language Models *2025–2026*

- Built a 722-item Hindi/English dataset derived from GSM8K, with syntactically scrambled variants. Manually validated all perturbations as a fluent Hindi speaker.
- Compared Dream-v0-Instruct-7B with Qwen2.5-7B-Instruct on NVIDIA A100. Found a **90.3-point reasoning-elicitation collapse**, from English (91.6%) to Hindi (1.3%). Applied exact McNemar tests and an error taxonomy.

Professional 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.

Public Research Artifacts

Hugging Face: huggingface.co/nadeem172

- **FedQuant models (LoRA):** `fedquant-mistral-7b-lora`, `fedquant-phi3-mini-lora`, `Llama-3.1-8B-FedQuant-lora`
- **Indian legal models:** `llama3-indian-legal`, `mistral-7b-indian-legal`, `Microsoft-phi3-mini-indian-legal`
- **Datasets:** `fedquant-indian-legal-nlp`, `indian-legal-raw-pdfs`, `indian-legal-phi3-dataset`, `indian-legal-mistral-dataset`, `indian-legal-llama3-dataset`

Technical Skills

- **LLM and SLM:** PyTorch, Hugging Face (Transformers, PEFT, bitsandbytes, Datasets), LoRA, QLoRA, federated learning.
- **Compression:** Quantization (NF4, AWQ, GPTQ, SmoothQuant, SparseGPT, Wanda, QAT), pruning.
- **Theory and methods:** SVD and subspace tracking, non-convex optimization, convergence analysis, statistical hypothesis testing (McNemar, Wilcoxon, Cohen's d).
- **Engineering and cloud:** Python, FastAPI, Docker, Oracle SQL, TensorFlow, scikit-learn, Azure, OCI, NVIDIA H100 HPC.
- **Languages:** English, Hindi.