# NEERAJ KUMAR SINGH

+91-8787040049 ◇ neeraj1909@gmail.com ◇ https://www.linkedin.com/in/neeraj1909

## ABOUT

LLM/GenAI Engineer with 4+ years in NLP and deep learning, delivering transformer- and MoE-based model analysis, RAG applications, and evaluation-driven AI systems. Focused on AI alignment, model safety, quantization, and production-oriented LLM workflows.

## WORK EXPERIENCE

**Lexsi AI, Mumbai, India**
Research Scientist *April 2024 - Present*

- Shipped DLBacktrace v2 as a production-grade explainability stack, enabling reliable relevance tracing across tabular, vision, NLP, and LLM workloads.
- Expanded model support via PyTorch `torch.export`/FX graphs and improved relevance propagation across transformer attention paths and common operators.
- Added robust LLaMA-family support and validation utilities for consistent large-model backtrace behavior.
- Implemented end-to-end MoE compatibility (JetMoE, OLMoE, Qwen3-MoE, GPT-OSS), including routing-aware relevance propagation for expert-level attribution.
- Improved inference efficiency with FP4/FP8/FP16 relevance quantization, achieving up to 17–93× compression with MAE $< 0.001$.
- Built evaluation and pipeline tooling (`dlbacktrace-viz`, ModelRegistry, HF-compatible sampling/beam search), and improved long-sequence efficiency through token-wise delayed relevance, caching, and compressed trace storage (gzip/lzma/7z).

**CVPR Unit, Indian Statistical Institute, Kolkata, West Bengal**
Research Assistant, NLP Lab *September 2021 - February 2024*

- Top contributor to "TrustED: Evaluating Trustworthiness of Deep Learning Systems"; developed explainable hate speech detection system for Indic languages with intensity reduction.
- Built web application integrating tCNNs, Precily, and Paccmann models for IC50 prediction using drug properties.

## EDUCATION

**Indian Statistical Institute, Kolkata, West Bengal** *2019 - 2021*
Master of Technology Percentage: 77.45

**Bundelkhand Institute of Engineering & Technology, Jhansi, Uttar Pradesh** *2015 - 2019*
Bachelor of Technology Percentage: 71.82

## PUBLICATIONS

- Interpretability-Aware Pruning for Efficient Medical Image Analysis [Accepted]
  Interpretability-guided pruning for VGG19/ResNet50/ViT on MURA, KVASIR, CPN, and Fetal Planes; achieved ∼**65**–**80%** pruning (up to ∼**85%** on ViTs) with minimal accuracy drop.
- DLBacktrace: A Model Agnostic Explainability for any Deep Learning Models [Accepted]
  Introduced a model-agnostic explainability method for MLPs, CNNs, and LLMs in PyTorch/TensorFlow; benchmarked against SHAP, LIME, Grad-CAM, and Integrated Gradients across NLP, vision, and tabular tasks.

- XAI-Evals: A Framework for Evaluating Post-Hoc Local Explanation Methods [arXiv]
  Developed `xai_evals`, a Python framework for unified XAI generation and benchmarking with faithfulness, sensitivity, and robustness metrics.
- Convolutional Neural Networks Exploiting Attributes of Biological Neurons [Under Review / arXiv]
  Proposed a PP-CORF + ResNet-18 two-tower CNN, improving accuracy by **5–13%** on CIFAR-10, CIFAR-100, and ImageNet-100.
- SafeSpeech: A Three-Module Pipeline for Hate Intensity Mitigation of Social Media Texts in Indic Languages [Accepted]
  Built a three-stage hate speech mitigation pipeline; validated on Hindi, Marathi, Tamil, Telugu, and Bengali with BERTScore **0.96–0.99** (automated + human evaluation).
- Analysis of Transformer-based Models for Code-mixed Conversational Hate-speech Identification [FIRE 2022]

## PROJECTS

- **RAG Chatbot** *Oct 2024 - Nov 2024*
  Built a RAG chatbot with LangChain, FastAPI, Streamlit, and Chroma-based retrieval, enabling source-grounded conversational QA and document workflows. GitHub: RAG_Chatbot

## PROGRAMMING LANGUAGES AND TOOLS

- LLM / GenAI: LLMs, RAG, Agentic Engineering, LangChain, LlamaIndex, Hugging Face Transformers, prompt engineering, AI alignment/model safety
- LLM Architectures: BERT, RoBERTa, T5, LLaMA, Qwen, Mixture-of-Experts (JetMoE, OLMoE, Qwen3-MoE, GPT-OSS)
- ML / NLP / CV: Python, PyTorch, TensorFlow, Scikit-Learn, FastAPI, OpenCV, spaCy, NLTK, NumPy, Pandas, Matplotlib
- Retrieval / Data: embeddings, vector retrieval (Chroma), similarity search, document chunking, retrieval-grounded response generation
- Evaluation / Research: XAI benchmarking, reproducible experiments, SHAP, LIME, Grad-CAM, Integrated Gradients, Layer-wise Relevance Propagation
- Developer Tools: Git, GitHub, CI/CD, Docker, Jupyter, VS Code, PyCharm, Plotly