NEERAJ KUMAR SINGH

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

ABOUT

Applied AI/ LLM engineer with 5 years of experience building NLP, explainable AI, RAG, and agentic AI systems. Experienced in transformer interpretability, retrieval evaluation, LangGraph workflows, and production AI tooling using Python, PyTorch, Hugging Face, FastAPI, Docker, and cloud-ready ML pipelines.

WORK EXPERIENCE

Lexsi AI, Mumbai, India
Research Scientist *April 2024 - March 2026*

- Shipped DLBacktrace v2, a PyTorch/TensorFlow-compatible explainability framework for NLP, LLM, transformer, and MoE models, enabling relevance tracing, visualization, and model-debugging pipelines.
- Built ML/LLM explainability pipelines for graph tracing, activation capture, relevance propagation, evaluation, compressed trace storage, and visualization across large transformer workloads.
- Expanded support for LLaMA-family and MoE models using PyTorch torch.export/ FX graphs; implemented routing-aware relevance propagation for JetMoE, OLMoE, Qwen3-MoE, and GPT-OSS.
- Used NumPy and Pandas for preprocessing, numerical validation, benchmarking, relevance-score analysis, compression-error analysis, and pipeline output comparison.
- Built cloud-ready pipeline components for model artifacts, batch execution, caching, logging, and trace storage.

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

- Worked on SafeSpeech, a three-module NLP pipeline for Indic hate-speech mitigation covering hate-speech detection, word-level hate-intensity estimation, and text rewriting to reduce harmful content while preserving semantic meaning.
- Built and evaluated transformer-based NLP models for multilingual and code-mixed hate-speech detection, using BERT-style architectures, token-level analysis, and evaluation workflows for Indic-language social media text.
- Proposed a two-tower CNN architecture combining a shallow PP-CORF-inspired tower with ResNet-18 to exploit biologically inspired visual feature extraction.
- Improved classification performance over ResNet-18 by combining biologically inspired shallow feature extraction with deep residual learning, achieving around 5%–10% average improvement on public benchmark datasets.
- Used Pandas and NumPy for dataset cleaning, annotation analysis, preprocessing, label distribution checks, experiment tracking, metric calculation, and error analysis across Indic-language NLP datasets.

PROJECTS

- **Explainable Agentic RAG** – LangGraph workflow with typed tools, structured outputs, retrieval attribution, verifier/retry loops, human review, tracing, and faithfulness/context/factuality metrics. [Link]
- **Indic Research Agent** – Chainlit assistant using LangGraph, LiteLLM, `query-kit` provider search, PostgreSQL persistence, Redis caching, streamed progress, and typed research tools. [Link]
- **query-kit** – Provider-agnostic Python CLI/library for ACL Anthology, arXiv, PubMed, Semantic Scholar, and OpenReview with normalized text/JSON results and sync/async APIs. [Link]
- **XAI-RAG** – Production RAG with ChromaDB, Elasticsearch, RRF, BGE re-ranking, retrieval attribution, DeBERTa NLI faithfulness checks, RAGAS, OpenTelemetry, Redis, and async FastAPI/SSE endpoints. [Link]

PUBLICATIONS

- **IAP-Med** – Interpretability-Aware Pruning for Efficient Medical Image Analysis [Accepted]
- **DLBacktrace** – model-agnostic interpretability for MLPs, LLMs, and custom DNNs; benchmarked against SHAP, LIME, and IG. [Accepted]
- **XAI-Evals** – Python framework for explanation evaluation with faithfulness, sensitivity, and robustness metrics. [arXiv]
- **SafeSpeech** – Indic NLP hate-speech mitigation pipeline validated on five languages with BERTScores **0.96–0.99**. [Accepted]

TECHNICAL SKILLS

- **Languages & Core ML:** Python, NumPy, Pandas, PyTorch, Hugging Face Transformers, OpenAI API, FastAPI

- **LLMs & NLP:** Large Language Models (LLMs), Transformer Models, Prompt Engineering, Fine-Tuning, PEFT, LoRA, Model Evaluation, Explainable AI, Model Interpretability
- **RAG & Agentic AI:** Retrieval-Augmented Generation (RAG), LangChain, LangGraph, LlamaIndex, Agent Workflows, Tool Calling, Structured Outputs, Retrieval Attribution
- **Search, Retrieval & Evaluation:** Vector Databases, ChromaDB, Elasticsearch, Hybrid Search, Re-ranking, RAGAS, BERTScore, NLI-based Faithfulness Evaluation
- **Backend, Data & Deployment:** FastAPI, REST APIs, PostgreSQL, Redis, Docker, Git, GitHub, AWS

EDUCATION

Indian Statistical Institute, Kolkata – M.Tech *2019 - 2021*
Bundelkhand Institute of Engineering & Technology, Jhansi – B.Tech *2015 - 2019*