

Shreyansh Tripathi

+4917685911333 | shreyanshtripathi03@gmail.com | Google Scholar
LinkedIn | GitHub | Website

SUMMARY

Machine Learning Engineer and Researcher specializing in **large-scale generative models, distributed ML systems, and GPU-optimized inference**. Experienced in designing end-to-end machine learning pipelines, **optimizing Mixture-of-Experts architectures on multi-GPU clusters**, and deploying production ML systems. **Published researcher in IEEE Transactions on Networking**.

EDUCATION

Universität des Saarlandes

MSc., Data Science and Artificial Intelligence

Oct 2023 - Present

Cluster Innovation Centre, University of Delhi

Bachelor of Technology, IT and Mathematics

July 2017 - July 2021

EXPERIENCE

AI Security Research Fellow

AI Safety Saarland

Nov 2025 - Feb 2026

Saarbrücken, Germany

- Red Teamed against unlearned models to reveal uptick in PII data leakage from 8% to 20%.
- Completed AI Safety Fellowship at AI Safety Saarland

ML Student Research Assistant (HiWi)

Max Planck Institute for Informatics

May 2024 - Dec 2025

Saarbrücken, Germany

- Designed parallel routing and communication strategies for **Mixture-of-Experts generative models running on multi-GPU clusters**.
- Improved inference latency by $1.5\text{--}2.5\times$** while achieving **up to 98% GPU utilization** across heterogeneous GPU systems.
- Extended a packet-level simulator to model **distributed ML workloads** and refactored components to support reusable experimentation workflows.
- Co-authored a peer-reviewed research publication in IEEE Transactions on Networking on optimizing distributed inference for Mixture-of-Experts models.

Data Scientist - Freelance

Jindal Healthcare

May 2023 - Aug 2023

Delhi, India

- Developed a **GBDT-based ranking system** to prioritize healthcare **insurance A/R follow-up** cases using historical payment data.
- Automated prioritization workflows that **reduced manual operational workload by 2.5 hours** per day.

Computer Vision Engineer - Freelance

Novaeavenue, PapliLabs

July 2022 - March 2023

Delhi, India

- Built a pothole detection system for Indian road infrastructure using **Mask R-CNN with PyTorch**.
- Developed **road asset segmentation and water-logging detection models** using U-Net architectures.
- Implemented **MLOps pipelines** including **model registry, experiment tracking, and monitoring for model deployments**.

Data Scientist

Aidetic Software Pvt. Ltd

July 2021 - July 2022

Bengaluru, India

- Improved NLP classification models for finance-related problems. **Achieved close to 20% improvement in f1-scores**.
- Built pipelines to extract and digitize graphs from corporate financial PDF documents **using hybrid transformer and rule-based methods**.
- Worked on data engineering systems involving **Elasticsearch and PostgreSQL**.
- Developed and deployed machine learning systems using **Docker-based deployment workflows**.

- Developed internal scripts and CLI-based utilities to streamline experiment execution and model deployment workflows.

Research Assistant - Intern

November 2020 - July 2021

Basque Centre for Applied Mathematics

Bilbao, Spain

- Worked towards creating systems that would help in creating a novel drug for Alzheimer's disease
- Merged techniques like **Molecular Dynamics and Topological Data Analysis** to produce new insights.
- Used complex data structures and **algorithms to calculate graph cliques in a time and space-efficient manner.**
- Published a **peer-reviewed research paper in a respected journal**

PROJECTS

Research Agent Framework | *LangChain, LlamaIndex, ChromaDB, Python* | [GitHub Link](#) | [App Link](#)

- Developed a multi-agent academic paper discovery system using **LangChain ReAct, LlamaIndex, and ChromaDB** to automate literature reviews and map evolving research interests.
- Designed an **LLM-powered orchestrator utilizing Chatgroq** to query **arXiv and Semantic Scholar APIs**, applying strict topic re-ranking and broad contextual expansion.
- **Implemented a local vector-based memory module** with exponential temporal decay ($\lambda = 0.05$), allowing the system to continuously personalize recommendations based on historical session data.
- Developed a modular, environment-driven configuration **supporting parallel agent execution and drop-in replacements for LLMs and vector stores.**
- Deployed the app using **Streamlit**

Robustness Analysis of Vision State Space Models (VMamba) | *PyTorch, State Space Models* | [GitHub Link](#)

- Evaluated robustness of **VMamba** against **21 image transformations** using a modified ImageNet-1K validation dataset containing $\sim 15K$ images per transformation.
- Benchmarked **VMamba-Base ($\sim 89M$ parameters)** against **Swin-Transformer** and **ConvNeXt** to analyze model performance under distribution shifts.
- Demonstrated that VMamba consistently outperformed baseline architectures across transformations with **top-1 accuracy up to 83.6%.**
- Fine-tuned **U-Mamba** for road segmentation achieving **>80% Dice score** and proposed a **gradient-freezing strategy** that reduced training time by $\sim 20\%$.

Interpreting Domain-Specific Knowledge in Vision-Language Models | *PyTorch, VLLMs* | [GitHub Link](#)

- Analyzed how **domain-specific knowledge is encoded in multimodal models** including **LLaVA** and **InstructBLIP.**
- Implemented **Domain Activation Probability Entropy (DAPE)** to identify domain-specific neurons across five domains including medical, remote sensing, and autonomous driving.
- Applied **LogitLens probing techniques** to track how hidden representations evolve across layers toward final predictions.
- Found that **document understanding tasks activate significantly more domain-specific neurons**, suggesting specialized internal processing pathways in VLMs.

PUBLICATIONS

Optimizing Mixture-of-Experts Inference Time via Model Deployment and Communication Scheduling
IEEE Transactions on Networking, vol. 34, pp. 2478-2497, 2026, doi: 10.1109/TON.2025.3645806.

Targeting Impaired Antimicrobial Immunity in the Brain for the Treatment of Alzheimer's Disease.
Neuropsychiatr Dis Treat. 2021;17:1311-1339

A review of Indian seismic zone based on historical earthquake data. AIP Conference Proceedings
2061,020038 (2019)

TECHNICAL SKILLS

Languages: Python3, C/C++, SQL

Infrastructure & MLOps: Docker, FastAPI, Flask, Git, Azure

Machine Learning/AI: PyTorch, Scikit-learn, Gradient Boosted Trees, Transformers