

Vishal Pandey

+91-6289730108 | pandeyvishal.mlprof@gmail.com | [LinkedIn](#) | [GitHub](#) | [Google Scholar](#) | [Portfolio Site](#)

PROFESSIONAL SUMMARY

Early-career Research Engineer specializing in transformer optimization, reinforcement learning, and interpretable AI systems. Experienced in building research prototypes (AgenticCyberOps, AgroSense) and drafting papers on LLM optimization and spiking decision transformers. Seeking opportunities in AI labs or RA positions to advance efficient, responsible, and explainable AI.

KEY PROJECTS

SmartTraffic-RL: Gym-style RL Environment for Urban Traffic Control (GitHub)

- Developed an RL benchmarking environment for adaptive traffic signal control using macroscopic flow models and SUMO integration.
- Implemented a PPO agent that outperformed fixed-time baselines in simulation experiments.
- Reduced vehicle queue lengths with tuned hyperparameters (reward scaling, learning rate, n_steps).
- Built visualization tools to track metrics such as average delay, KL divergence, and explained variance.

AgenticCyberOps: Autonomous Cybersecurity Pipeline (GitHub)

- Engineered a LangGraph + LangChain multi-agent framework for autonomous penetration testing and security audits.
- Designed custom safety metrics to evaluate agent performance and robustness.
- Demonstrated feasibility of scalable, autonomous red-teaming for AI and safety-critical systems.

AgroSense: Soil Classification & Crop Recommendation (GitHub)

- Built a Vision Transformer + data pipeline tool, deployed via Streamlit.
- Achieved 94.3% soil classification accuracy and 97–98% crop recommendation accuracy.
- Added explainability features for reproducibility and interpretability.

RESEARCH PROJECTS

FlashMoE: Lightweight Sparse MoE for Edge Devices (arXiv Draft in Progress)

- Proposed FlashMoE, a lightweight sparse Mixture-of-Experts architecture, enabling $4\times$ faster inference and $2.5\times$ lower memory footprint on edge GPUs compared to dense baselines.
- Designed expert routing with token sparsity + quantized experts, achieving near-parity accuracy on language modeling while cutting energy usage by 40%.

Spiking Decision Transformer (Research Draft)

- Designed a spiking neural variant of Decision Transformers for long-horizon environments.
- Achieved $2\times$ reduction in inference energy consumption with performance parity on Gym benchmarks.
- Draft paper in preparation, focusing on energy efficiency and neural dynamics.

RESEARCH EXPERIENCE

Independent ML Researcher

Oct 2024 – Present

Independent Research Fellowship

Remote

- Dedicated this period to deepening my understanding of machine learning, with a focus on transformer optimization, neuro-symbolic reinforcement learning, and autonomous agent systems.
- Engaged in self-directed study and discussion on the development and enforcement of ethical policy guidelines for large language models, contributing to a nuanced understanding of responsible AI deployment, while also exploring the mathematical underpinnings of AI safety and control.
- Developed and refined AI research prototypes, including LangChain-powered multi-agent pipelines, with an emphasis on transforming cutting-edge ML research into practical tools for potential internal productivity gains in dynamic, ambiguous environments
- Organized and maintained extensive research documentation and internal write-ups, facilitating knowledge sharing and future research initiatives.
- Explored methodologies for constructing ecologically valid benchmarks to assess the agentic propensities and capabilities of AI systems, with a focus on real-world applicability
- Pioneered the development of practical AI tools like AgenticCyberOps and AgroSense, demonstrating the ability to translate advanced ML concepts into deployable solutions and improve reproducibility for real-world application.

National Institute of Technology, Mizoram

Nov 2022 – Apr 2023

ML Research Internship

Aizawl, IN

- Developed a CNN-based signature verification model, reducing false positives by 20% and improving overall reliability.
- Mentored C-DAC PG-DAI students and validated 15 ML models, enhancing their learning and improving model performance consistency by 25%.
- Assisted with scheduling research meetings and managing project documentation for the ML research team, ensuring efficient collaboration and information access.
- Facilitated communication between interns and senior researchers, acting as a point of contact for project-related queries and resource sharing

WORK EXPERIENCE

Cube Construction Engineering Ltd.

July 2023 – Aug 2024

Graduate Engineer Trainee

Vadodara, IN

- Streamlined report generation and distribution processes, reducing turnaround time by 30% and enhancing client satisfaction through improved information flow and document management.
- Resolved critical GST-related issues in Farvision ERP, implemented address mapping for accurate vendor information, and designed document management for Purchase and Work Orders.
- Managed vendor inquiries and maintained accurate vendor information within Farvision ERP, ensuring efficient communication and operational continuity.
- Supported cross-functional team coordination for critical GST-related issue resolution, ensuring timely communication and progress tracking.

TECHNICAL SKILLS

Programming & Frameworks: Python, C++, PyTorch, TensorFlow, JAX, Scikit-Learn, Keras, Typescript (for full-stack integration)

LLMs & Retrieval: Hugging Face Transformers, LangChain, FAISS, Elasticsearch, RAG pipelines

Optimization & Fine-Tuning: TRL, LoRA/QLoRA, Axolotl, Distributed Training, CUDA

Evaluation Metrics: nDCG, BLEU, ROUGE, BERTScore, QuestEval

Infrastructure & DevOps: Docker, Linux/Unix Shell, FastAPI, Prometheus, Grafana

Research Focus: Reinforcement Learning, Spiking Neural Networks, Multimodal Models, Explainability & Interpretability

CERTIFICATIONS & SELF-LEARNING

Machine Learning (Stanford Online - Coursera): Completed Dec 2022 Scored 92%; covered supervised learning, neural networks, and regularization techniques. Taught by **Andrew Ng**.

McKinsey Forward Program - Level 1: Completed July 2025 Trained in structured problem-solving, communication, digital tools, and self-leadership. Selected among top global early-career talents.

Cohere ML Summer School: Completed July 2025 Gained hands-on experience in large language models (LLMs), representation learning, and practical ML workflows through expert-led sessions and applied projects.

EDUCATION

Maulana Abul Kalam Azad University of Technology

Kolkata, IN

Bachelor of Technology, Information Technology (CGPA: 9.28/10.0)

July 2023

Relevant Coursework: Data Structures & Algorithms, Machine Learning, Deep Learning, NLP

Thesis: “Diabetes and Its Type Prediction” - Developed and evaluated machine learning models for diabetes classification and subtype prediction on the Pima dataset.