APRATIM MISHRA

Irving, TX | apratim941208@gmail.com

LinkedIn: https://www.linkedin.com/in/apratim94/ | **Personal Page**: https://apratim-mishra.github.io/

SUMMARY

Applied Machine Learning Ph.D. professional with expertise in developing and deploying petabyte-scale machine learning, NLP, and LLM-based models. Optimized ML pipelines for real-world applications with a strong foundation in analytics, experimental design, statistical testing (A/B), ML at scale, agentic AI orchestration, and model productization.

PROFESSIONAL EXPERIENCE

AI Researcher – Toyota

Mountain View, CA | Feb 2026 – Present

- Research, develop, and prototype multi-agent AI frameworks for mobility using Python, PyTorch, and foundation models, focusing on agent orchestration, tool integration, fine-tuning, evaluation, and alignment.

Machine Learning Engineer – Verizon

Irving, TX | July 2025 – Feb 2026

- Built and owned production AI systems, leading a team of four, a network and server QA platform for automated testing, diagnostics, and validation of distributed infrastructure, and a configuration validation system for testing and enforcing configuration logic in production codebases. Together, these systems improved reliability and reduced manual intervention across network operations and release workflows.
- Implemented both systems using agentic AI architectures and custom LLM modeling. Designed multi-agent workflows to orchestrate tools, APIs, and knowledge sources, and trained instruction-tuned models via SFT and RFT on logs, test outputs, configuration artifacts (YAML, JSON, TOML), and incident data using PyTorch, DeepSpeed, Accelerate, and LoRA/PEFT. Deployed scalable training and inference pipelines with CI/CD and monitoring, achieving 24 percent higher training throughput and 15 percent lower p95 inference latency.

Graduate Research Assistant – University of Illinois Urbana-Champaign

Champaign, IL | June 2024 – May 2025

- Processed and modeled 15 million PubMed abstracts using Databricks and PySpark to build large-scale NLP pipelines for feature extraction and entity disambiguation. Trained and evaluated LLM-based classifiers in Python and PyTorch, improving F1 by 12 percent for hype detection and releasing a dataset with 300-plus external downloads.
- Developed a hybrid citation recommendation system combining GNN with LLM-based semantic embeddings, achieving a 25 percent lift over BM25. Deployed pipelines on AWS SageMaker using Airflow for orchestration and experimentation.

Data Science Intern (Applied ML) – AstraZeneca

Gaithersburg, MD | May 2023 – Aug 2023

- Improved protein–property prediction performance by 10 percent F1 by training transformer-based models on A100 multi-GPU clusters using PyTorch and Slurm, applying fine-tuning with LoRA (PeFT), DeepSpeed, and Accelerate.
- Reduced model retraining time by 30 percent by optimizing ETL pipelines in PySpark and Spark SQL, introducing intermediate caching and parquet-based storage layers to improve data reuse and pipeline stability.
- Deployed and maintained production inference services using TorchServe on AWS SageMaker, supporting approximately 1K daily predictions. Tracked experiments with MLflow and automated, zero-downtime deployments via GitHub Actions.

NLP Engineer Intern – The Cline Centre for Advanced Social Research

Champaign, IL | May 2021 – August 2021

- Improved entity and quotation extraction precision by 18 percent by fine-tuning BERT and XLNet models with hard-negative mining, conducting systematic error analysis to improve robustness across diverse news sources.
- Built distributed NLP pipelines processing millions of news articles using Apache Spark and MLlib for clustering and segmentation. Reduced infrastructure cost by 25 percent through spot instances and Dockerized Airflow workflows.
- Trained models with Distributed Data Parallel (DDP) and deployed real-time inference services via FastAPI on Kubernetes, maintaining P99 latency under 150 milliseconds in production environments.

Data Analyst – Reliable Power Alternatives Corporation

Garden City, NY | June 2018 – Aug 2019

- Saved 10 % in energy procurement costs (≈ $MM annually) by deploying PySpark-based load-forecasting models.
- Improved forecast MAE 15 % via temporal cross-validation and SHAP-driven feature pruning, processing structured energy data with SQL and Snowflake data cloud analytics, and back-tested forecasting strategies for model evaluation.
- Built Tableau dashboards consumed weekly by ~ 4 - 8 executives; reduced manual reporting hours by 30 %.

Research Intern – Delaware Army National Guard

Wilmington, DE | Sep 2017 – May 2018

- Raised time-series accuracy by 15 % in load forecasting using time-series models (ARIMA, XGBoost, LightGBM, and LSTM).
- Verified gains with rolling t-tests (p less than 0.01) and presented statistical findings to a 12-member energy task force.

EDUCATION

- **Ph.D. in Information Sciences** | University of Illinois at Urbana-Champaign | Champaign, IL | (Aug 2019 – June 2025)
- **Master's in Energy and Environmental Policy** | University of Delaware | Newark, DE | (Aug 2016 – May 2018)
- **Bachelor's in Chemical Engineering** | Birla Institute of Technology and Science | Pilani, India | (Aug 2012 – May 2016)

PROJECTS AND RESEARCH

- **Phone Calling Agent:** Built a real-time voice AI agent for property inquiries that handles inbound/outbound calls, performs natural-language understanding, and executes live property search. Implemented a Twilio-to-FastAPI streaming pipeline with local MLX Whisper STT, Kokoro TTS, and Groq-accelerated LLM reasoning for sub-second responses: designed LangGraph-based conversational flows, semantic search over Pinecone, and a call-logging system backed by SQLite.

- **Prediction Market Agent:** Developed an AI agent for prediction markets using Coinbase AgentKit, integrating natural-language commands with smart-contract operations. Implemented market creation, betting, resolution, and claims through LangChain tools, Solidity contracts, and a FastAPI backend. Added multi-source price oracles, a React frontend, and a test suite of 26 cases to support reliable end-to-end interaction.

- **Notion Agentic AI Assistant**: Built a semantic RAG system for Notion that performs intelligent document retrieval, ranking, and response generation using LangChain, OpenAI embeddings, and Qdrant vector search. Added intelligent chunking, duplicate removal, multi-tool augmentation, and an evaluation suite using OpenAI Evals. Automated ingestion and indexing pipelines with real-time updates and exposed a REST API for search and RAG queries.

- **Sound AI:** Created a multi-agent music generation platform using CrewAI and Suno AI that performs topic research, lyric writing, and song synthesis end-to-end. Implemented a three-agent workflow for research, lyric generation, and music creation, with Docker-based orchestration and a Streamlit interface.

- **PartSelect Chat Agent**: Developed a dual-mode RAG assistant for appliance part discovery with FastAPI, WebSockets, and React. Added part search, compatibility checks, troubleshooting guides, and function-calling tools powered by DeepSeek and OpenAI fallbacks. Enhanced mode includes a routed multi-agent system with guardrails, triage logic, and data-grounded validation. Built a Docker CI/CD pipeline and optimized the UI for mobile and real-time chat.

TECHNICAL SKILLS

- **Programming and Analytics**: Python (Pandas, NumPy, Scikit-learn, NLTK), PyTorch, SQL, NoSQL, R, TypeScript, Node.js, Go
- **Machine Learning**: Transformers, LiteLLM, GPT models (Claude), PySpark, MCP, LangChain, LangGraph, Unsloth
- **MLOps and Tools**: W&B, FAISS, Pinecone, Weaviate, DeepSpeed, MongoDB, Onnx, TensorRT, Kubeflow, OpenAI SDK, vLLM
- **Big Data and Cloud**: AWS (EC2, Lambda, SageMaker, Beam, Bedrock, IAM, Step Functions), GCP (BigQuery, Vertex AI)