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

Machine Learning Engineer – Verizon

Irving, TX | July 2025 – Present

- Developed a unified enterprise agent platform for domain teams to compose, version, and deploy multi-agent workflows across network, server, and business applications. Components include an Agent Builder (drag-and-drop node editor), a Connector Registry for data and tool integration (APIs, vector databases, RAG endpoints), and a ChatKit UI toolkit for portals. (AWS Bedrock & SageMaker, ECS/Fargate, EKS, AgentCore, FastApi, Postgres + Redis memory stores, GitLab for version control, JSON/YAML/TOML/JS schema validation via configuration pipelines.)
- Designed and implemented end-to-end LLM fine-tuning and inference pipelines tailored for network/server diagnostics and application-level automation. Fine-tuned instruction models via SFT and RFT on domain logs, KB transcripts, and incident-ticket datasets (using SageMaker with DeepSpeed & Accelerate, LoRA/PEFT for tuning). Deployed quantized models with ONNX Runtime and bitsandbytes into Kubernetes (EKS) via Triton/TGI, with autoscaling and full CI/CD. Presented the model-training DAGs, production-inference microservice architecture, and live monitoring dashboards showing +24% training throughput and –15% p95 latency.

Graduate Research Assistant – University of Illinois Urbana-Champaign

Champaign, IL | June 2024 – May 2025

- Parsed 15M PubMed abstracts on Databricks with PySpark to extract features (NER and Named Disambiguation). Trained LLM models with Python/PyTorch, raising F1 by 12 % for 'hype' detection, producing a dataset downloaded 300 times.
- Developed a hybrid citation recommendations system combining GNN-based link prediction with LLM-derived semantic embeddings (25 % lift over BM25). Deployed on AWS SageMaker (real-time or serverless) via Airflow orchestration.

Data Science Intern (Applied ML) – AstraZeneca

Gaithersburg, MD | May 2023 – Aug 2023

- Improved protein–property prediction F1 by 10 % by training transformer-based LLM models on A100 multi-GPUs (Slurm) and fine-tuning with parameter-efficient fine-tuning techniques (LoRA) and optimization tools (DeepSpeed, Accelerate).
- Cut model retraining time by 30 % by refactoring ETL in PySpark and Spark SQL and caching intermediate parquet layers.
- Deployed a TorchServe REST microservice via Sagemaker that handles ~1 K predictions per day in production; tracked experiment with MLflow and delivered zero-downtime updates through GitHub Actions.

NLP Engineer Intern – The Cline Centre for Advanced Social Research

Champaign, IL | May 2021 – Dec 2021

- Boosted entity-quotation extraction precision by 18 % fine-tuning BERT and XLNet with hard-negative mining.
- Processed 250 M petabytes of data (news articles) on a 20-node Apache Spark cluster (clustering to segment articles); using MLlib; reduced pipeline cost by 25 % by migrating to spot instances and Dockerized Airflow tasks.
- Fine-tuned scaled data using Distributed Data Parallel (DDP) and served real-time inference with FastAPI orchestrated on Kubernetes, keeping P99 latency under 150 ms.

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

- **Notion Agentic AI Assistant**: Designed a RAG-based semantic search system using LangChain and FAISS with OpenAI/HuggingFace SDKs for LLM-based retrieval, ranking, and generation. Automated serverless deployment and LLMOps pipelines for end-to-end ingestion, vectorization, and response generation.

- **Sound AI:** Built an end-to-end song-generating platform based on topics via web search, lyric writing, and music synthesis. Utilized CrewAI to orchestrate multi-AI agents (Suno AI) and React, using MLOps practices for versioning and deployment.

- **PartSelect Chat Agent**: Architected a dual-mode RAG-based AI assistant for appliance parts with FastAPI, WebSockets, and React. The multi-agent mode adds triage, specialist agents, hallucination guardrails, and function-calling tools for part search, compatibility, and troubleshooting. Implemented DeepSeek-to-OpenAI model fallback and Docker-based CI/CD.

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)