

Shashank Pandey

+1 (716) 970-9684 | snkp.globalcollege@gmail.com | [LinkedIn](#) | [GitHub](#)

SUMMARY

Lead Engineer with 3+ years building and shipping production ML systems — LLM fine-tuning infrastructure, real-time RAG pipelines, and low-latency inference APIs. Experienced across the full stack: model training, versioned MLOps workflows, containerized serving, and CI/CD-gated deployments across AWS, Azure, and GCP. Strong foundations in PyTorch, FastAPI, and software engineering practices that make ML systems reliable in production. IEEE-published.

EDUCATION

Masters in Data Science | **University at Buffalo, USA**

Dec 2024

EXPERIENCE

T-Mobile (Via Techsico), Seattle, WA, USA | Lead AI Engineer

May 2026– Present

- Optimized LLM prompts and agent workflows, reducing token consumption by **25%** and inference costs by **20%**, while improving GPU/PTU utilization.
- Managed Azure Cosmos DB and Azure ML infrastructure, reducing query latency by **40%** and using Azure Application Insights to monitor application performance and failures.
- Developed **5 MCP servers** and integrated **10 APIs** to connect AI agents with enterprise tools and telecommunications data sources.

Synergetics AI, Remote, USA | AI/ML Engineer – Software Developer

Dec 2025 – May 2026

- Designed and built LangTrain, a multi-cloud LLM fine-tuning platform with a cost-aware GPU scheduler that dispatches training jobs across AWS, Azure, and GCP compute instances; reduced per-run GPU spend by ~40% (est. ~\$8K/month savings at current load).
- Applied LoRA fine-tuning and INT8/INT4 quantization (bitsandbytes) on 7B–13B parameter models (LLaMA-2, Mistral); validated quality via ROUGE-L and task-specific benchmarks on held-out splits, cutting VRAM footprint by 35% within 5% of full-precision baseline accuracy.
- Shipped RESTful FastAPI microservices with pytest unit and integration test suites; automated Docker builds and cloud deployments via GitHub Actions CI/CD; delivered a real-time training metrics dashboard with sub-200 ms SSE update latency.
- Engineered backend services for the TWIN digital twin platform using FastAPI and MongoDB; built a multi-tenant agent registry with capability-based routing and async job queues, enforcing schema validation and contract testing with 90%+ backend test coverage.

Jacob's Medicine and Biomedical Sciences, Buffalo, USA | ML Engineer – GenAI

Jan 2025 – Dec 2025

- Built a LangGraph + LangChain retrieval pipeline over 15+ biomedical research corpora using FAISS vector search and GPT-4/Claude-3; evaluated with RAGAS metrics (faithfulness, answer relevancy, context precision), achieving 91% citation precision on held-out benchmarks and cutting researcher literature-review time by 35%.
- Ran PyTorch DDP fine-tuning across 4x A100 GPUs with MLflow experiment tracking and Optuna hyperparameter sweeps, cutting training wall-clock from 18 hrs to 11 hrs and reducing iteration cycles by 40%.
- Designed a human-in-the-loop review workflow where domain researchers annotated and corrected model outputs on ambiguous biomedical queries; used feedback to iteratively refine retrieval ranking and generation quality, improving inter-annotator agreement from 0.71 to 0.86 (Cohen's kappa) across two review cycles.

Baldwin Richardson Foods, New York, USA | ML Engineer – Data Science

Aug 2024 – Dec 2024

- Trained an XGBoost + LSTM + Prophet ensemble on 14 years of sales data across 50+ SKUs; engineered lag, rolling-window, and seasonal features, validated with walk-forward backtesting across 6 quarterly windows, reducing MAPE from 18% to 12.6% and enabling \$1.5M/year in procurement savings confirmed by finance.
- Implemented end-to-end ML pipelines on AWS compute with Databricks, MLflow model registry, automated retraining triggers, and Kubernetes serving; enforced pytest and model evaluation gates in GitHub Actions CI/CD; authored optimized PostgreSQL queries across 10M+ row operational datasets for feature extraction.

KPIT Technologies, India (Remote) | Machine Learning Engineer

Dec 2021 – Jul 2023

- Productionized BERT-based NLP pipelines for text classification and NER over 50K+ automotive incident tickets per month using PyTorch and TorchServe; precision 0.88 / recall 0.85, eliminating ~2 FTE-hours of manual triage daily.
- Developed speech processing pipelines for in-cabin voice command classification using Wav2Vec2 and MFCC feature extraction on automotive audio datasets; achieved 93% command recognition accuracy across 8 intent categories under real-world cabin noise conditions.
- Developed real-time anomaly detection pipelines on 1M+ automotive IoT sensor events per day, serving predictions via a REST API gateway backed by TensorFlow Serving; introduced model versioning, A/B testing harnesses, and drift monitoring adopted across 4 product teams.

Indian Space Research Organization (ISRO), India | Data Scientist – Software Intern

Sep 2021 – Aug 2022

- Built PyTorch deep learning classifiers on satellite telemetry and laser-induced plasma spectra, improving anomaly detection accuracy by 32% on mission-critical signals — results published at [IEEE CONECCT 2022](#); scaled PySpark batch pipelines to 10x throughput to support high-volume telemetry processing.
- Delivered end-to-end solutions by building a Flask REST backend and React dashboard for mission operations teams, enabling real-time signal monitoring and alert visualization and consolidating workflows previously split across 3 separate tools.

SKILLS

Programming & Frameworks: Python, C++, SQL, R, PySpark | PyTorch, TensorFlow, HuggingFace Transformers, LangChain, LangGraph, FastAPI, Flask, React, Pandas, Scikit-learn

ML & AI: LLM Fine-tuning (LoRA, QLoRA, INT8/INT4), Distributed Training (DDP, FSDP), RAG Pipelines (FAISS, RAGAS), NLP (BERT, T5), Speech (Wav2Vec2), LLMs (GPT-4, LLaMA, Mistral, Gemini), CNNs, LSTMs, XGBoost, Forecasting, Anomaly Detection

MLOps & Infra: Docker, Kubernetes, AWS, Azure, GCP, Databricks, MongoDB, PostgreSQL, MLflow, TorchServe, TF Serving, GitHub Actions CI/CD, GPU Optimization

Visualization: Tableau, Streamlit, Superset, Power BI

CERTIFICATIONS & PUBLICATION

Certifications: AWS Machine Learning Specialty, Microsoft Power BI, AWS Certified Cloud Practitioner

Publication: "An Automatic Peak Finding and Fitting Aspects of Laser Induced Plasma Spectra acquired in High Vacuum: Tradeoff Simulations and Statistics" — IEEE CONECCT 2022 (DOI: 10.1109/CONECCT55679.2022.9865112)