

Prem Babu Kanaparthi

New York | prem.b.kanaparthi@gmail.com | 585-290-3036 | [LinkedIn](#) | [GitHub](#)

EDUCATION

Rochester Institute of Technology <i>Masters in Artificial Intelligence</i> Specialization: Machine Learning Applications, Distributed Systems, Engineering Scalable Systems	August 2024 - May 2026 Rochester, NY
National Institute of Technology (NIT), Silchar <i>Bachelors in Computer Science and Engineering</i>	August 2020 - May 2024 India

SKILLS AND CERTIFICATIONS

Machine Learning & AI: PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Computer Vision, NLP/NLU, Generative AI, Prompt Engineering, Deep Learning

Cloud & MLOps: AWS (Bedrock, SageMaker, EKS, Lambda, CloudWatch, DynamoDB), Docker, Kubernetes, MLflow, Terraform, CI/CD Pipelines, Model Context Protocol (MCP), Palantir Foundry, Snowflake

Programming & Data: Python, PySpark, SQL, LangChain, FastAPI, Pandas, NumPy, PostgreSQL, MongoDB, Git, GitHub Actions

EXPERIENCE

Concentrix + Webhelp <i>Generative AI Engineer</i>	Newark, CA <i>February 2024 – July 2024</i>
<ul style="list-style-type: none">Built a multi-model LLM inference platform delivering 1.5s latency and 18% lower cost across 25+ enterprise deploymentsDesigned LiteLLM-based routing with AWS Bedrock and SageMaker inference pipelines to maintain 95%+ response accuracyImplemented automated guardrails, policy enforcement, and CloudWatch monitoring cutting incidents by 42%Built production evaluation pipelines to continuously validate model quality, data drift, and system reliability in live traffic	
AlphaBits Technologies <i>Data Science Intern</i>	India <i>August 2023 – January 2024</i>
<ul style="list-style-type: none">Reduced search model iteration time by 90% while improving ranking relevance by 10%Built point-in-time correct PySpark ETL pipelines and a centralized feature store to support reproducible experimentationDeveloped TF-IDF and Naive Bayes ranking models through systematic feature engineering and evaluationAuthored an ML Playbook adopted by 3 teams, reducing integration errors by 35%	
iNeuron.ai <i>ML Engineer Intern</i>	India <i>June 2023 – August 2023</i>
<ul style="list-style-type: none">Built a phishing detection system achieving 92% accuracy and 20% fewer false positives on enterprise datasetsImplemented a TensorFlow/Keras MLP with dropout and batch normalization for improved generalizationPerformed F1-driven hyperparameter tuning to optimize precision-recall tradeoffs for real-world security dataDeployed the model with monitoring and validation pipelines to ensure stable production performance	

PUBLICATIONS

Lightweight Channel Attention for Efficient CNNs - ([Paper Link](#))

- Designed and evaluated a lightweight channel attention module (LCA) achieving competitive accuracy with negligible parameter and latency overhead on ResNet-18 and MobileNetV2

PROJECTS

PaperMind - Autonomous arXiv Research Assistant - ([GitHub](#))

- Built a multi-agent RAG system over arXiv to generate literature reviews, method comparisons, and research gaps from papers
- Implemented specialized LLM agents coordinating over a shared vector database for citation-grounded research output
- Designed PDF ingestion, chunking, and embedding pipelines to improve retrieval precision and reduce hallucinated citations

Real-Time Feature Store & ML Serving Platform - ([GitHub](#))

- Architected streaming feature pipeline with Feast and Kafka processing 5,000 events/hour at 45ms latency
- Deployed 3 models (fraud: 88%, recommendation: 0.76 NDCG@5, forecasting: 12% MAPE) sharing 8 features via unified store
- Built monitoring with MLflow, Kubernetes metrics, triggering auto-retraining on performance degradation

Multi-Agent Document Intelligence System - ([GitHub](#))

- Built multi-agent system automating text extraction; achieved 82% accuracy on 150-document test set with processing time < 5s
- Implemented RAG with Qdrant processing 500+ chunks; enabled semantic search with 85% top-3 relevance across 50 queries
- Developed production FastAPI service handling 3 document types; reduced manual entry time 84% reduction
- Deployed the model with monitoring and validation pipelines to ensure stable production performance

CERTIFICATIONS

- Stanford Machine Learning Specialization – ([Link](#))