Venkatalakshmi Kottapalli

2019544144 |venkatalakshmik23@gmail.com| LinkedIn | GitHub | Portfolio | Research

SUMMARY

AI/ML Engineer with **4+ years** of experience designing ML models, **RAG pipelines, LLM fine-tuning**, and **AI agents** across Azure, AWS, and GCP. Delivered a **40% ROI uplift and 95%** model accuracy by optimizing pipelines and contributed to **5 research papers**. Seeking roles focused on **machine learning**, generative AI, computer vision, and **AI automation** to build scalable AI systems.

EDUCATION

Master's Degree in **Artificial Intelligence** | Yeshiva University | New York, United States | GPA: 4.0/4.0 **Jan 2024 - Dec 2025**

Bachelor's Degree in **Mathematics** | Adikavi Nannaya University | India | GPA: 3.8/4.0 | (**University 1st Rank**) **Jun 2015 - May 2018**

EXPERIENCE

AI/ML Engineer | Peblink | New York, USA **Sep 2025 - Present**

- Implemented AI-driven optimization models for **Goldman Sachs' $500M Dallas campus initiative**, identifying redundant management layers and improving operational efficiency across multiple business units.
- Built ensemble ranking models (**XGBoost, Random Forest**) to prioritize relocation candidates, reducing analysis time by **40%**.
- Developed **GPT-4 LangChain agents** for relocation and retention scoring, automating **30+ workflows** and cutting project turnaround time by **25%**.
- Created executive dashboards, heatmaps, and cost-benefit visualizations using **Power BI**, enabling real-time insights.

Machine Learning Engineer | ZSAnalytics | Boston, USA **May 2024 - Aug 2025**

- Built **MLOps pipelines** on Azure ML and AKS with MLflow integration, streamlining deployment cycles and reducing model release time by **30%**.
- Performed extensive **EDA** and **feature engineering**, generating 10+ custom features that increased precision from **76% - 88%.**
- Developed a retrieval-augmented generation (**RAG) pipeline** using LangChain, GPT-4, and vector database ChromaDB to index 100K+ documents, achieving **Recall@K = 0.82** and reducing query latency from 6.2 s to 2.1 s.
- Collaborated with cross-functional stakeholders to present results and ROI improvements through interactive dashboards.

Data Scientist | Cutso LLP | Hyderabad, India **Mar 2019 - Dec 2021**

- Executed large-scale **EDA and statistical analysis (Chi-Square, ANOVA)** on 1M+ transactions, uncovering 8 behavioral clusters that enhanced marketing segmentation strategies.
- Applied **K-Means & DBSCAN clustering** to classify customers into 5 groups (Silhouette = 0.68), increasing retention by **18%**.
- Automated KPI dashboards in **Tableau**, saving 50+ hours monthly through real-time performance monitoring of 30+ metrics.
- Delivered predictive inventory analytics, reducing stockouts by **25%** and boosting overall profitability by **12%**.

RESEARCH & PROJECTS

Transformer-based LLM Fine-tuning with DEKF | Large Language Models (LLMs) Optimization Research **Dec 2025**

- Fine-tuned transformer-based large language models using **Decoupled Extended Kalman Filters (DEKF)** for adaptive uncertainty estimation. Implemented in **JAX, NNX, and Dynamax** on Google Cloud TPUs.
- Achieved **42% lower computing**, **38% reduced memory**, and **1.8× faster convergence** compared to LoRA.

AI Hackathon: Multi-modal Agent Conversational AI | Generative AI | GitHub **July 2025**

- Developed a **GPT-4-**powered multi-modal conversational agent using LangChain, LangGraph, and hybrid RAG (ChromaDB + SQLite) to manage 10K+ documents, improving response accuracy by **35%** and ranking **12th among 150+ teams**.

Cardiomegaly Detection for Health Care | Computer Vision | Research Paper **May 2025**

- Developed a **DeepCNN model in PyTorch** and **TensorFlow** for automated detection of cardiomegaly using chest X-ray images. Optimized with data augmentation, achieving **72% accuracy** and **30% lower cost** compared to VGG16 over 100 epochs.

SKILLS

- **Programming Languages**: Python, SQL, R
- **ML/Gen AI**: Classification, Regression, Clustering, Transformers, BERT, GPT, RAG, CNNs, RNNs, GANs, Prompt Engineering
- **Frameworks:** PyTorch, TensorFlow, Keras, Scikit-learn, LangChain, LangGraph, NLTK, JAX, LoRA, Optuna, Dynamax
- **MLOps & Cloud:** MLflow, Azure ML, AWS (S3, RDS), Google Cloud Platform (GCP), VertexAI, CI/CD, Git, Docker, Kubernetes
- **Data Visualization & Databases:** Power BI, Tableau, Matplotlib, Seaborn, PostgreSQL, Neo4j, FAISS, ChromaDB, SQLite
- **Other Competencies**: OOP, REST APIs, Agile (Scrum), Distributed Training, LLM finetuning, Model Deployment Automation