Venkata Shashank Kowtharapu

AI & Machine Learning Engineer | Generative AI Specialist

kvshashank10081998@gmail.com
https://www.linkedin.com/in/shashank-kowtharapu/
https://medium.com/@kvshashank10081998
https://github.com/Shashank-KVS
+13439894078
Mississauga, Ontario, Canada

Innovative AI and Machine Learning professional with 4+ years of experience specializing in Generative AI, MLOps, and large-scale model deployment. Proven expertise in designing and automating AI workflows, optimizing model performance, and implementing scalable solutions using Azure ML Studio, Hugging Face, and Llama. Certified Azure Cognitive Services Specialist and AZAI Generative AI Specialist, with deep proficiency in LLM development, cloud-based ML pipelines, and production-ready AI systems.

Work Experience

Associate Data Scientist | *LTIMindtree | Canada* Dec 2023 - Present

- Architected and optimized Azure Machine Learning (AML) workflows, enhancing LLM accuracy by 30% and reducing training latency by 25% through advanced AutoML and prompt flow optimization.
- Developed and deployed scalable MLOps pipelines using Azure ML SDK, Designer, and CI/CD automation, achieving a 15% improvement in model training efficiency and ensuring high availability in production environments.
- Used PySpark on Azure Spark compute clusters for distributed processing and feature engineering of large-scale datasets, enabling scalable analysis in ML pipelines.
- Designed and troubleshot complex AI pipelines, integrating Azure OpenAI Services, AutoML, and fine-tuned LLMs, boosting NLP processing speed by 30% while improving content safety filtering.
- Optimized Azure ML environments, managing DNS lookups, VNet integrations, and Compute Instance provisioning, leading to a 20% increase in system performance and enhanced scalability.
- Designed and deployed RESTful APIs for integrating ML models into production pipelines, enabling seamless interaction with business applications.
- Leveraged Azure Document Intelligence and AI Vision to automate document processing and image recognition, improving data processing efficiency by 20% and recognition accuracy by 35%.
- Led and mentored a team of 10 engineers in Generative AI and MLOps, fostering expertise in LLM fine-tuning, deployment strategies, and AI model evaluation, increasing team productivity by 30%.
- Collaborated with business stakeholders and cross-functional teams to align AI solutions with business objectives, ensuring transparency and measurable impact.

AI Developer Internship | *En Solution | Canada* Sep 2023 - Nov 2023

- Developed an AI-driven chatbot using Llama and Hugging Face, integrating custom fine-tuned LLMs to improve user engagement by 40% and enhance multi-turn conversation handling.
- Optimized chatbot adaptability and response accuracy by implementing advanced NLP techniques, data preprocessing pipelines, and dynamic prompt engineering, achieving a 25% increase in response accuracy.
- Built a scalable Retrieval-Augmented Generation (RAG) pipeline using LangChain and vector databases for context-aware query retrieval, reducing latency by 30%.

Associate Software Engineer | *Oracle | US-Remote* Oct 2021 - Apr 2022

- Designed and automated CI/CD pipelines for microservice applications using Jenkins, Bash, and Python, reducing deployment time by 40% and enhancing operational efficiency.
- Implemented end-to-end build automation by integrating SVN to Git migration, ensuring version control integrity, artifact management, and seamless CI/CD workflows.
- Developed and optimized Python-based backend services, resolving infrastructure bottlenecks, improving system scalability, and ensuring reliable client-facing deployments through automated troubleshooting.

Research Assistant | *Amrita Vishwa Vidyapeetham | India* Oct 2020 - Sep 2021

- Developed AI-powered IoT frameworks with Edge Computing, optimizing real-time data processing pipelines for smart agriculture, improving operational efficiency by 30% through on-device inference and federated learning.

- Engineered a hybrid Edge-Cloud architecture, integrating low-latency communication protocols and optimized data transmission, reducing latency by 25% and enhancing AI-driven decision-making for automated farming.
- Designed FPGA-based hardware accelerators for Edge AI, optimizing CNN-based crop monitoring models, boosting inference speeds by 15%, and achieving energy-efficient AI deployment in IoT-driven environments.

IoT & AI Engineer | *Indian Government Project* | *Department of Science & Technology* | *India* Jun 2018 - May 2019

- Developed an AI-powered traffic signal optimization system, reducing congestion by 15% at major intersections.
- Designed machine learning models to predict peak-hour congestion patterns, improving traffic light synchronization by 20%.
- Built an end-to-end data pipeline for real-time traffic feeds, enhancing prediction accuracy by 30%.

Projects

AI Financial Risk Monitor Mar 2025 - Mar 2025

- Designed a full-stack AI system for real-time stock risk monitoring with anomaly detection and risk classification using hybrid LOF-Random Forest models.
- Built an asynchronous ingestion engine with asyncio, tenacity, and yfinance, and developed a modular ML pipeline with auto-evaluation and risk labeling.
- Integrated DistilBERT embeddings and auto-refreshing risk dashboards with technical indicators, news feeds, and financial metrics.

DeepSeek-Azure-ML- Serverless-Custom-Filtering Feb 2025 - Feb 2025

- Developed and deployed an AI filtering system using Azure Machine Learning and Deep Seek models.
- Implemented scalable serverless workflows to efficiently process large AI datasets.
- Optimized computational efficiency while ensuring real-time inferencing and seamless deployment.

Phi-4-Fine-Tuning-NLP-Custom-Dataset-Training Jan 2022 - Jan 2025

- Fine-tuned the Phi-4 model on a custom NLP dataset for domain-specific text generation.
- Enhanced model accuracy using optimized hyper parameter tuning and LoRA for low computational fine tune.
- Implemented robust validation strategies to ensure reliable deployment in production environments.

Core Skills

Core professional skills: Python, NumPy, Pandas, Scikit-learn, PyTorch, TensorFlow, Bash, Jenkins, SQL, Docker, Git, Modular Python Code Practices.
AI & Machine Learning: LLM Development, Retrieval-Augmented Generation (RAG), Agents, LangChain, Hugging Face, NLP & NLU, Time Series Forecasting, Predictive Analytics, Optimization Models & Techniques, Computer Vision, AutoML, ML Pipelines, Content Filtering, Basic familiarity with SAS and R for statistical modeling and data exploration.
Cloud & MLOps: Azure Machine Learning Studio & SDK, PySpark, Azure Cognitive Services, Azure Data Factory, CI/CD Pipelines, Machine Learning Operations, Experiment Tracking (MLflow), Data Management, Compute Resource Management, Custom Environment Build, KQL, PySpark
Other Expertise: Data Integration, ETL Processes, Prompt Engineering, Model Deployment & Optimization, API Development (RESTful), Tableau Dashboards, Data Visualization, Stakeholder Communication, Business Intelligence (BI), Team Leadership, Communication & Cross-functional Collaboration, Problem-solving skills, Algorithms, Cloud Computing, Flexibility

Education

St. Lawrence College | Post Graduate Certificate | Business Analytics Dec 2023
Amrita Vishwa Vidyapeetham | Masters | Artificial Intelligence Aug 2022

Publications

Smart Farming based on AI, Edge Computing and IOT Dec 2022
IEEE XPLORE
Credit Risk Assessment for Home Credit Group Aug 2021
IRJET