Shilpa Kuppili

AI/ML Engineer

shilpakp333@gmail.com | +1 (984) 318-4336 | Harrison, NJ | www.linkedin.com/in/shilpa-kuppili-a80ba2126/

SUMMARY

AI/ML Engineer with 5+ years of hands-on experience designing, building, and deploying machine learning models and data-driven systems. Skilled in solving real-world problems using supervised and unsupervised learning, natural language processing, and deep learning techniques. Strong background in Python, TensorFlow, and cloud platforms like AWS and GCP. Proven track record of delivering production-ready ML solutions that improve decision-making, automate workflows, and support business goals.

SKILLS

Languages & Programming: Python, Java, C, C++, R, SQL, TypeScript, Bash/Shell Scripting

Machine Learning & Deep Learning: Supervised Learning, Unsupervised Learning, Neural Networks, Deep Learning, Time Series Forecasting, Anomaly Detection, Transfer Learning, Active Learning, Semi-Supervised Learning, Model Evaluation, Cross-Validation, Feature Engineering, Hyperparameter Tuning, Ensemble Methods (Bagging, Boosting)

Generative AI & LLMs: Generative AI, Large Language Models (LLMs), Transformers, Fine-Tuning, Retrieval-Augmented Generation (RAG), LangChain, Autogen, Tokenization, Embedding Models, Context Optimization

Prompt Engineering: Few-shot Prompting, Chain-of-Thought, Instruction Tuning, Tool Use Integration, Advanced Prompt Design, Chaining Strategies, Context-Aware Prompting

GenAI Tools: OpenAI API, Hugging Face Transformers, Claude, ChatGPT, Gemini, Google NotebookLM, Cursor, Copilot Studio

Libraries & Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, Statsmodels, XGBoost, LightGBM, NLTK, SpaCy, OpenCV, Dlib, Matplotlib, Seaborn, Plotly

Data Engineering & Processing: NumPy, Pandas, Dask, PySpark, Apache Spark, Polars, Airflow, Kafka, DBT

Cloud & MLOps: AWS (SageMaker, EC2, S3, Lambda), GCP (Vertex AI, BigQuery), Azure (Azure ML), Docker, Kubernetes, MLflow, DVC, Weights & Biases, TensorBoard, Git, GitHub Actions, Jenkins, Argo Workflows, CircleCI

Model Deployment & Serving: Flask, FastAPI, gRPC, ONNX, TorchServe, TensorFlow Serving, Streamlit, Gradio, REST APIs, GraphQL, API Rate Limiting, Caching Strategies

Databases & Storage: PostgreSQL, MySQL, MongoDB, Redis, DynamoDB, Cassandra, Pinecone, FAISS, ChromaDB, Weaviate, HDFS, Snowflake, Google BigQuery, Amazon Redshift

Data Visualization & BI: Power BI, Tableau, Looker, Matplotlib, Seaborn, Plotly, Dash

Collaboration & Productivity: Microsoft 365 Suite, Notion, JIRA, Confluence, Slack, Trello, Figma

EXPERIENCE

AI/ML Engineer, Humana **Aug 2025 – Present**

- Engineered machine learning models using scikit-learn, XGBoost, and TensorFlow to identify high-risk individuals for chronic conditions, resulting in a 22% improvement in early intervention accuracy across Medicare Advantage plans.
- Built NLP pipelines using Hugging Face Transformers and spaCy to extract diagnoses, medications, and symptoms from unstructured physician notes, enabling downstream analytics that supported more personalized patient care strategies.
- Fine-tuned domain-specific large language models (LLMs) for automating clinical summary generation and insurance claims review workflows, reducing manual processing time by 40% and improving internal audit consistency.
- Implemented end-to-end MLOps pipelines integrating MLflow for experiment tracking, DVC for dataset versioning, and Jenkins for automated retraining, ensuring reproducibility and compliance with HIPAA data governance standards.
- Monitored production-grade models using FastAPI and TensorFlow Serving on AWS SageMaker, implementing scalable CI/CD pipelines via GitHub Actions and enabling real-time support for clinical decision-making tools.

AI Full Stack Developer, Thoughti **Feb 2025 – Aug 2025**

- Developed AI-powered chatbot on Microsoft Azure using Copilot Studio and Bot Framework, delivering personalized behavioral health support and reducing user dropout rates by 25%.
- Trained CLU and CQA models in Azure AI Language Studio to classify user intent across five behavioral change stages, improving interaction relevance by 40%.
- Streamlined the AI model lifecycle and secured data workflows using Azure Foundry, reducing deployment time by 30%.
- Designed reusable prompt templates and chaining logic to align chatbot dialogue with user behavioral stage, improving coherence and emotional relevance.
- Orchestrated dynamic multi-step conversation flows in Copilot Studio and integrated 3+ external APIs for personalized response generation.
- Constructed language understanding pipelines to detect substance use indicators and symptoms, achieving 87% precision in early detection.

- Devised a multi-agent research pipeline using advanced prompt engineering and dynamic context injection, cutting response time by 45%.
- Deployed ML models via Azure Foundry and integrated secure API endpoints to ensure scalable 24/7 access for 10K+ monthly users.
- Led the development of an automated proposal drafting tool using Azure Foundry, Microsoft Fabric, and Autogen, reducing proposal creation time.
- Architected a five-phase agent framework using Autogen and RAG for automated document analysis, improving accuracy and speed by 50%.
- Mitigated LLM hallucinations by 35% and inference costs by 40% using tiered deployment of GPT-4.1 and o4-mini with prompt optimization.

ML Intern, TechComb | Texas **Aug 2024 – Dec 2024**

- Engineered and implemented an LLM integrated bird detection and classification system using YOLOv8 for real time bird localization and Efficient Net B0 for precise species identification.
- Pioneered LLM such as GPT-2 to provide detailed species-specific deterrent strategies. Integrated Gradio and optimized the end-to-end pipeline, reducing inference time by 25%.

AI/ML Intern, ZSAnalytics | New York **May 2024 – Aug 2024**

- Built a Custom GPT Model and fine-tuned transformer-based language models on enterprise call transcript data to extract relevant knowledge and summarize the transcript.
- Implemented a Mistral LLM Chatbot for varied NLP tasks to generate coherent and contextually relevant information.

Senior Software Engineer, Motherson Technology Services Limited | Bangalore, India **Jun 2022 – Aug 2023**

- Applied recursive feature selection techniques for document understanding using LLM-based prompts.
- Developed an AI-powered document classification model, automating the categorization of records with an accuracy of 92% using a combination of CNN and LSTM architectures.

Software Engineer, Tata Consultancy Services | Bangalore, India **Dec 2018 – Jun 2022**

- Created a conversational AI chatbot for financial advice, reducing customer query response time by 15%.
- Designed an image classification pipeline using pre-trained ResNet50 model to categorize fashion products into various styles, increasing product recommendation accuracy by 30%.

EDUCATION

Yeshiva University | New York, NY **Aug 2023 – Dec 2024**

Master of Science in Artificial Intelligence