

SRINATH KOYI

San Jose, CA | +1 (657) 877-9052 | srinath.koyi@applywizard.ai | [LinkedIn](#)

PROFESSIONAL SUMMARY

- AI/ML Engineer with 4+ years of experience building machine learning, NLP, and Generative AI solutions that improve document retrieval, issue classification, knowledge discovery, and decision support for enterprise teams.
- Hands-on experience developing models and AI workflows using Python, SQL, Scikit-learn, Pandas, NumPy, Hugging Face Transformers, LangChain, LlamaIndex, OpenAI API, embeddings, FAISS, ChromaDB, and FastAPI.
- Skilled in solving real-world problems through RAG pipelines, text classification, predictive modeling, semantic search, recommendation workflows, model evaluation, prompt engineering, and AI application development.
- Experienced in working with product, QA, data, and engineering teams to define AI use cases, validate model outputs, reduce manual review effort by up to 35%, and deliver production-ready AI features.

TECHNICAL SKILLS

Programming & Data Handling: Python, SQL, Pandas, NumPy, Data Cleaning, Data Preprocessing, Exploratory Data Analysis, Feature Engineering

Machine Learning: Scikit-learn, XGBoost, LightGBM, Classification, Regression, Clustering, Recommendation Systems, Predictive Modeling, Hyperparameter Tuning

Deep Learning: PyTorch, TensorFlow, Keras, Neural Networks, CNN, RNN, LSTM, Transfer Learning, Model Optimization

Natural Language Processing: Hugging Face Transformers, BERT, Tokenization, Text Classification, Named Entity Recognition, Sentiment Analysis, Semantic Similarity

Generative AI & LLMs: OpenAI API, Azure OpenAI, LangChain, LlamaIndex, RAG, Prompt Engineering, Embeddings, Vector Search, LLM Evaluation

Model Evaluation & Experimentation: Cross-Validation, Precision, Recall, F1-Score, ROC-AUC, Confusion Matrix, A/B Testing, MLflow

AI Application Development: FastAPI, REST APIs, Model Inference, Batch Prediction, Real-Time Prediction, Streamlit, Gradio

Databases & AI Storage: PostgreSQL, FAISS, Pinecone, ChromaDB, Vector Databases, Similarity Search

PROFESSIONAL EXPERIENCE

AI/ML Engineer

Apr 2025 - Present

Infinite Computer Solutions Pvt Ltd | USA

- Built a RAG-based internal knowledge assistant using Python, LangChain, OpenAI API, embeddings, and FAISS to solve slow document retrieval, improving search accuracy by 32% for support teams.
- Developed a ticket classification model using Scikit-learn, Pandas, NumPy, and SQL to reduce manual issue categorization, improving triage speed by 28% across recurring operational requests.
- Created FastAPI inference endpoints for NLP models so internal applications could consume real-time predictions, reducing manual review effort by 30% during workflow validation.
- Improved LLM response quality by refining prompts, testing failure cases, and validating answers with business users, reducing inaccurate responses by 25% in pilot use cases.
- Built Hugging Face Transformer-based text classification workflows to categorize support tickets and knowledge articles, improving routing accuracy by 27% for internal service teams.
- Partnered with product, QA, and data teams to define model inputs, acceptance criteria, and validation rules, reducing AI feature rework by 22% before release.
- Tuned classification models using cross-validation, precision, recall, F1-score, and confusion matrix analysis, improving prediction reliability by 24% across validated datasets.
- Designed Streamlit AI prototypes for search and prediction use cases, helping stakeholders review model behavior early and reducing requirement clarification cycles by 26%.

AI/ML Engineer

Jun 2024 - Apr 2025

Lumen Technologies | USA

- Built predictive ML models using Python, Scikit-learn, XGBoost, Pandas, and SQL to identify service-risk patterns, improving early issue detection by 30% across operations data.
- Developed Generative AI summarization workflows using Azure OpenAI, LangChain, embeddings, and vector search to reduce technical note review time by 35%.
- Engineered structured features from operational datasets using Pandas, NumPy, and SQL to improve model input quality, reducing data preparation time by 34% before training.
- Created FastAPI-based inference services for classification and recommendation outputs, improving AI integration speed by 29% for internal application teams.
- Improved NLP classification accuracy by refining tokenization, preprocessing, and labeled dataset quality, increasing model precision by 23% across text-based issue records.
- Worked with business analysts, data teams, and engineering groups to convert operational problems into AI use cases, improving delivery alignment by 25%.

- Validated model performance using ROC-AUC, precision, recall, F1-score, and A/B testing feedback, improving release confidence by 27% before stakeholder approval.
- Developed Gradio model demo interfaces for stakeholder review and AI output validation, reducing feedback turnaround time by 31% during early solution testing.

Machine Learning Engineer

May 2021 - Sep 2023

SureshBy Technologies Pvt. Ltd. | India

- Developed supervised ML models using Python, Scikit-learn, Pandas, NumPy, and SQL to classify recurring operational patterns, improving analysis speed by 27%.
- Built reusable ML pipelines for data cleaning, preprocessing, feature engineering, training, and evaluation, reducing repeated model development effort by 30%.
- Created NLP workflows using tokenization, TF-IDF, and classification models to analyze support notes, improving issue grouping accuracy by 24% across records.
- Improved model performance through hyperparameter tuning, cross-validation, and error analysis, increasing prediction consistency by 22% across test datasets.
- Built recommendation logic using clustering and similarity search to suggest relevant knowledge articles, reducing manual lookup time by 26% for support users.
- Collaborated with engineering and support teams to define target variables, review model outputs, and reduce misaligned model requirements by 21%.
- Prepared model evaluation reports using confusion matrix, precision, recall, F1-score, and ROC-AUC, improving technical review efficiency by 25%.
- Developed Python-based Streamlit proof-of-concept AI applications with REST API integration, reducing manual analysis effort by 28% for internal users.

PROJECTS

Enterprise Document Intelligence Assistant

- Built an AI document search assistant using Python, LlamaIndex, OpenAI API, ChromaDB, and embeddings to help users retrieve policy, ticket, and knowledge-base answers faster, improving retrieval accuracy by 31%.
- Created chunking, metadata filtering, and semantic search logic to reduce irrelevant results, cutting manual document review time by 34% for internal support users.
- Validated generated answers against source documents and reviewed failure cases with QA teams, reducing unsupported LLM responses by 26% before rollout.

Predictive Issue Classification System

- Developed a machine learning classification system using Python, Scikit-learn, LightGBM, Pandas, and SQL to predict recurring issue categories, improving ticket prioritization accuracy by 29%.
- Engineered features from structured issue history, resolution notes, and service patterns to improve model readiness, reducing analyst data preparation effort by 32%.
- Measured model performance with precision, recall, F1-score, ROC-AUC, and confusion matrix analysis, increasing prediction consistency by 24% across test datasets.

AI Recommendation Workflow for Knowledge Articles

- Built a recommendation workflow using NLP preprocessing, sentence embeddings, FAISS similarity search, and Streamlit to suggest relevant knowledge articles, reducing lookup time by 30%.
- Improved article matching logic by comparing user queries, issue descriptions, and historical resolutions, increasing recommendation relevance by 27% for support teams.
- Partnered with support users to test incorrect matches and refine search thresholds, reducing repeated manual escalations by 22% during validation.

EDUCATION

Bachelor of Technology in Mechanical Engineering

Jawaharlal Nehru Technological University | Anantapur, India

CERTIFICATIONS

- Machine Learning Specialization - **Coursera**
- Deep Learning Specialization - **Coursera**
- Natural Language Processing Specialization - **Coursera**
- Generative AI with Large Language Models - **Coursera**
- IBM AI Engineering Professional Certificate - **Coursera**
- IBM Generative AI Engineering Professional Certificate - **Coursera**