Yuvaraja Reddy Avuthu

+1 716-910-7802 | yuvaraja.avuthu@gmail.com | linkedin | huggingface | Portfolio | Github

Summary

With 4+ years of experience as an AI/ML Engineer specializing in Generative AI, Large Language Models (LLMs), NLP, and machine learning, with expertise in AWS and GCP cloud platforms. Skilled in deploying scalable AI solutions using FastAPI, Docker, and frameworks like LangChain, OpenAI, and Hugging Face, delivering impactful applications in conversational AI, text generation, and real-time analytics.

Technical Skills

Generative AI (Gen AI): LangChain, GPT-3.5 Turbo, OpenAI, Hugging Face Transformers, Vector Databases (ChromaDB, Pinecone, FAISS), Llama3, Reinforcement Learning (RLHF), Prompt Engineering, Few-Shot Learning, RAG, LoRA, Fine-Tuning

NLP: SpaCy, NLTK, Gensim, Hugging Face, Stanford NLP, TextBlob, Word2Vec, TF-IDF, BERT, Transformers, RNN, TensorFlow, PyTorch

Machine Learning & Deep Learning: PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, CNN, RNN, LSTMs, Transformers, ARIMA, Transfer Learning, Regression, Decision Trees, Random Forest, SVM, KNN, K-Means, Hierarchical Clustering, PCA

Cloud Platforms & DevOps: AWS (S3, SageMaker, Lambda, Glue, EMR, RedShift, IAM, Kinesis, SNS, Athena, CloudWatch, Bedrock, API Gateway), GCP (Vertex AI, BigQuery), Docker, Kubernetes, Jenkins, Bitbucket, Fast API, RESTful API

Big Data & Databases: Apache Spark (SQL, Streaming), Kafka, Snowflake, Hadoop, Apache Airflow, ETL/ELT, PostgreSQL, MySQL

Programming & Tools: Python, SQL, Git, Pandas, NumPy, Jupyter, Anaconda

Experience

AI/ML Engineer Aug 2024 – Present
SMA Tech llc *Nashua, NH*

- Developed a secure RAG ChatBot for a health insurance client using API Gateway and Lambda for real-time, serverless processing.
- Integrated AWS Bedrock with the LLAMA-3 (70B/8B) model and employed a RAG approach using LangChain and ChromaDB for dynamic context retrieval.
- Deployed the RAG system on AWS SageMaker for efficient inference and monitoring, bypassing traditional fine-tuning.
- Leveraged AWS EC2 and S3 to host microservices and securely store data and model artifacts, with AWS IAM roles ensuring strict access control.
- Built an IDD pipeline with AWS Lambda to automate document ingestion, pre-processing, and extraction for enhanced real-time insights.

Data Scientist Analyst Jan 2021 – Jan 2023
Accenture AI *Bengaluru, India*

- Led end-to-end NLP pipeline development for aviation insurance underwriting, fine-tuning spaCy's (en_core_web_lg) pre-trained NER model on 10K+ documents (PDF/DOC/images) to extract 8+ entity types with 92% accuracy.
- Collaborated with computer vision team to preprocess 5K+ scanned documents using OCR for downstream NLP tasks.
- Built decision tree classifier (scikit-learn) using NER and web-scraped risk metrics, automating 80% of insurance decisions.
- Engineered web scraping module with bs4 & Selenium, extracting real-time metrics from 4+ sources to enhance risk models.
- Deployed FastAPI microservices with Docker, reducing API response time to ¡500ms for NER and recommendation predictions.
- Designed cross-sell recommender system using KMeans clustering (customer segments) and KNN (policy similarity), building Excel dashboards with pivot tables to visualize recommendations, increasing upsell conversion by 35% for aviation clients.
- Spearheaded Agile workflows via Jira/Confluence, documenting model versions and maintaining 99% uptime for production systems through Git-based CI/CD pipelines.

Machine Learning Engineer Nov 2020 – Dec 2021
Cluzters.ai *Remote, India*

- Developed document classification model using BERT and TF-IDF for text categorization, deployed on GCP Vertex AI.
- Optimized feature engineering using PCA and feature selection techniques, improving model performance by 25%.
- Automated model training and deployment workflows using AWS SageMaker, S3, IAM, Lambda, and Step Functions.

- Deployed ML models using MLflow, Docker, Kubernetes, integrating them into production environments with FastAPI and Flask.
- Built and optimized machine learning pipelines with scikit-learn, XGBoost, and LightGBM, ensuring efficient feature engineering and model evaluation.
- Integrated Apache Spark for large-scale data processing, handling terabytes of structured and unstructured data.

Research Collaborator Oct 2019 – Nov 2020
Amrita School of Engineering *Kerala, India*
- Conducted extensive data preprocessing using Pandas and NumPy, handling missing values, outliers, and text normalization.
- Implemented text tokenization and vectorization techniques using NLTK and Scikit-learn (TF-IDF, Word2Vec).
- Designed and optimized deep learning architectures in TensorFlow and PyTorch for multi-class text classification.
- Applied LDA and LSTM-based topic modeling to analyze large corpora of academic and industry-related documents.
- Automated dataset labeling using active learning techniques, reducing manual annotation efforts by 50%.

Projects

RAG Conversational AI Chat Bot Sep 2024 – Present
Try it *LangChain, Llama, ChromaDB, HuggingFace, Streamlit, FastAPI, Docker, SQLite*
- Built RAG system using LangChain, Llama-3 (70B/8B), and ChromaDB via Groq API for context-aware responses.
- Designed Streamlit/FastAPI document QA system with dynamic model switching across 8 Llama-3 variants.
- Engineered NLP pipeline for PDF/DOCX/HTML using LangChain splitters and HuggingFace embeddings.
- Optimized LLM outputs via prompt engineering (NO PREAMBLE) and 5-doc context-aware retrieval chains.
- Integrated Llama Guard 3-8B for moderation and metadata-based document security in ChromaDB.
- Reduced search latency by 40% through ChromaDB chunk indexing and batch deletion workflows.

LinkedIn Content Generator Using LLM Jul 2024 – Aug 2024
Try it *Llama, Groq API, LangChain, Streamlit, Few-Shot Learning, Semantic Tagging*
- Built LLM-driven content generator using Llama-3-70B and Groq API, automating LinkedIn post creation with customization.
- Engineered metadata extraction pipeline with LangChain and JSON parsing to classify posts by tags, language and line count.
- Implemented few-shot learning by dynamically selecting high-engagement examples to guide LLM output style and tone.
- Deployed Streamlit UI with tag unification logic, reducing duplicate categories by 60% through semantic merging.

Medical NER System Feb 2024 – Jun 2024
Try it *spaCy, Transition-Based Parsing, Hugging Face Hub, Gradio, Tesseract OCR, PyMuPDF*
- Developed medical NER model using spaCy Transition-Based Parser, achieving 85% F1-score on 10+ entity types.
- Built multi-format text extractor supporting PDF/DOCX/images via PyMuPDF and Tesseract OCR, handling 500+ medical reports.
- Optimized spaCy training with tok2vec and Adam optimizer, reducing training time by 25% through efficiency-focused configs.
- Deployed Gradio UI with entity visualization, achieving 90% accuracy on test cases like ”Aspirin” and ”pneumonia”.

Publications

Performance Comparison of Machine Learning Algorithms in Symbol Detection Using OFDM Jan 2022
Inventive Communication and Computational Technologies

MSE and BER Analysis of Text, Audio and Image Transmission Using ML Based OFDM Jan 2021
IEEE International Conference for Innovation in Technology (INOCON)

Education

University at Buffalo, SUNY Buffalo, NY
M.S. in Data Science *Jan 2023 – May 2024*

Amrita Vishwa Vidyapeetham Kerala, India
B. Tech. in Electronics and Communication Engineering *Aug 2017 – May 2021*

Certifications

AWS Cloud Certification *(verify)* | Gen AI Certification *(verify)* | Python Certification *(verify)*

NLP Certification *(verify)* | Deep Learning Certification *(verify)* | Machine Learning Certification *(verify)*