Varun Billuri
Github: https://github.com/varunbiluri/
Linkedin: https://www.linkedin.com/in/billurivarun/
Email: varunreddy.billuri@gmail.com
Mobile: +91-770-2661-217

Summary

Machine Learning Engineer experienced in building and deploying multilingual AI systems, model fine-tuning pipelines, and retrieval-augmented generation (RAG) applications. Skilled in Python, PyTorch, Hugging Face, FastAPI, and Airflow, with hands-on experience in dataset preparation, model evaluation, and scalable inference. Developed production conversational AI and real-time speech systems for low-resource languages.

Experience

Worldish
Software Engineer Dec 2025 - Present

- Built supervised fine-tuning (SFT) pipelines for small language models (SLMs), covering multilingual dataset preparation, training, evaluation, and deployment.
- Evaluated fine-tuned models against baselines and benchmarked LLM, STT, and TTS providers on multilingual accuracy, response quality, and inference latency to guide model selection.
- Developed an Arabic dialect evaluation platform integrating LLM prompting, context engineering, and speech workflows.
- Built real-time healthcare conversational AI integrating translation, speech recognition, and neural speech synthesis for Tigrinya, Wolof, Bambara, and Arabic dialects.
- Redesigned production AI pipelines to improve inference scalability, multilingual orchestration, and cross-service reliability.
- Applied reinforcement learning from human feedback (RLHF) to align multilingual model responses with human preferences.

ThoughtSpot
Member of Technical Staff 2 Jan 2025 - Dec 2025

- Developed an ML-driven incident response pipeline that routed Slack alerts and created Jira tickets, reducing mean time to resolution (MTTR) by approximately 60%.
- Built an anomaly detection service for Airflow execution metadata, surfacing pipeline failures and bottlenecks through PostgreSQL-backed dashboards.
- Engineered an agentic AI framework using LangChain and LangGraph to triage Airflow failures through rule-based logic, LLM classification, and guarded automated remediation.
- Deployed LLM-based log summarization with LangChain and Ollama, converting raw Airflow logs into actionable debugging insights.
- Extended petabyte-scale ETL orchestration across AWS and GCP, modularizing data pipelines that supplied ML training datasets.
- Contributed to supervised fine-tuning of Azure OpenAI models for internal workflows, including dataset preprocessing and model evaluation.

ThoughtSpot
AI/ML Intern Jun 2024 - Dec 2024

- Developed a retrieval-augmented generation (RAG) prototype using LangChain and Haystack, indexing pricing and metrics data in vector stores for contextual question answering.
- Experimented with embedding models to improve retrieval relevance and support contextual responses.
- Built backend services and data pipelines for pricing and metrics datasets, enabling downstream machine learning workflows.
- Prototyped AI workflows combining retrieval, summarization, and alerting to automate recurring analysis tasks.

Key Projects

Agentic AI System for Data Pipeline Reliability

- Built LangGraph workflows for Airflow failure triage, log summarization, root cause analysis, and remediation.
- Integrated vector search over historical incidents to provide retrieval-augmented context, with guardrails for automated actions.

MediScan: Precision Medical Record Digitization with OCR and LSTM

- Integrated Google OCR, image preprocessing, and LSTM-based terminology correction to digitize medical records.
- Built a Flask interface for uploading medical document images and extracting text.

VisioVoice: Multimodal Conversational AI for Accessibility

- Built a voice-based assistant combining image analysis, conversational memory, and real-time speech-to-text and text-to-speech.
- Converted images into structured JSON representations to support follow-up questions without repeated image processing.
- Developed conversational pipelines using FastAPI, React, Azure OpenAI Vision, and Azure Speech Services.

Technical Skills

- **Programming Languages** — Python, SQL, Java, JavaScript.
- **Machine Learning** — PyTorch, TensorFlow, Hugging Face Transformers, scikit-learn, NumPy, feature engineering, model training and evaluation.
- **LLMs & Fine-tuning** — Large and Small Language Models (LLMs/SLMs), Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF).
- **Generative AI** — RAG, vector embeddings, LangChain, LangGraph, prompt engineering, agentic workflows.
- **Evaluation** — Model benchmarking, task-specific evaluation, multilingual accuracy, speech recognition evaluation, inference latency, and RAG retrieval relevance.
- **Speech & NLP** — Multilingual NLP, conversational AI, speech-to-text (STT), text-to-speech (TTS), OCR.
- **Data & MLOps** — Apache Airflow, MLflow, Snowflake, PostgreSQL, ETL/ELT pipelines, Grafana, Prometheus.
- **Backend & Cloud** — FastAPI, Flask, Django, REST APIs, AWS, Azure, GCP
- **Developer Tools & Coding Assistants** — Git, GitHub, VS Code, Cursor, Claude Code, OpenAI Codex, Gemini Code Assist

Education

Jawaharlal Nehru Technological University, Anantapur *Feb 2020 - May 2024* India

Bachelor of Technology - Computer Science and Engineering & Business Systems; GPA 7.6/10

- Secured Rank **1973** in CodeKaze Coding Competition among **200k+** participants

Publications

Digitization of Medical Records Using OCR

B. Varun Kumar Reddy et al. | International Journal of Advanced Research in Computer and Communication Engineering, 2024
DOI: 10.17148/IJARCCE.2024.134141

Certifications

- AWS Certified Data Engineer – Associate (DEA-C01) — Amazon Web Services
- Machine Learning Specialization — DeepLearning.AI & Stanford University
- Generative AI with Large Language Models
- Java Programming and Software Engineering Fundamentals Specialization — Duke University