Mohnish Sai Prasad

+1-831-869-4373 | mohnish2401@gmail.com | linkedin.com/in/mohnish7 | github.com/mohnish-7

Summary

Graduate student seeking FTE roles in SWE/ML Engineering, available June 2025. Skilled in LLMs, NLP, CV, cloud technologies - AWS, TensorFlow, and PyTorch, with a passion for AI-driven automation and scalable systems.

Skills

- **Programming Languages & Technologies:** : Python, JavaScript, Java, C++, HTML, CSS, SQL
- **Machine Learning & AI** : LLM, RAG, LlamaIndex, TensorFlow, Keras, PyTorch, Computer Vision, NLP
- **Frameworks & Tools:** FastAPI, Flask, Vue.js, MySQL, PostgreSQL, ChromaDB, Docker, jQuery
- **Cloud & MLOps:** AWS Lambda, AWS EC2, DynamoDB, Amazon Connect, AWS S3, AWS SQS, KubeFlow

Education

University of California, Santa Cruz Sep 2023 - June 2025
Master of Science, Computer Science and Engineering GPA: 4.0 / 4.0
Coursework: Machine Learning, Generative AI, Data Structures and Algorithms, Computer Networks, Database Systems

Madras Institute of Technology, Anna University Aug 2019 - July 2023
Bachelor of Engineering, Electronics and Communication Engineering GPA: 9.15 / 10.0
Coursework: Data Structures and Algorithms, Software Engineering, AI and ML, Internet of Things

Work Experience

Sandeza Inc Aug 2024 - Dec 2024
Machine Learning Engineer Internship Alpharetta, GA

- Built an LLM-based agent on Amazon Connect leveraging live stream data from AWS Kinesis Data Streams with the help of AWS Lambda and OpenAI Whisper for automated issue resolution, cutting human workload by 50%.
- Built a custom conversation summarizer using FastAPI and GPT-4o-mini to generate post-call summaries of agent-customer interactions, with structured storage in DynamoDB for analysis.
- Designed and implemented an intelligent IVR system using GPT-4o-mini for efficient call routing, enhancing customer engagement and experience.
- Migrated older JavaScript API calls and event handlers to Python for better compatibility and ease of integration with new intelligent AI features
- **Tech Stack:** Python, JavaScript, FastAPI, AWS S3, DynamoDB, AWS Lambda, AWS Kinesis, GPT-4o

Projects

- **Enhanced Educational Video Lecture Summarization** Jan 2024 - Mar 2024
  - Explored two summary enhancement methods: a RAG system with LlamaIndex + Llama 3, and a few-shot prompt-based GPT-3.5 model using additional reading material.
  - The RAG system produced limited improvements, while the few-shot approach generated more informative summaries.
  - A user survey showed 64% of participants preferred the enriched summaries (with reading material) over those without it.
  - **Tech Stack:** RAG, ChromaDB, Llama 3, GPT-3.5, LlamaIndex, and Python.
- **Forest Fire and Human Detection from Images for Alert Systems** Sept 2022 - Jan 2023
  - Vision Transformers were fine-tuned using PyTorch to detect wildfires from videos with 98.9% accuracy.
  - Used YOLO algorithm to detect the presence of humans inside a wildfire and sent their coordinates to the fire department.
  - **Tech Stack:** Vision Transformers, PyTorch, YOLO, and Python.
- **Flash Cards Web App** Jan 2022 - Apr 2022
  - Developed an online flashcard web app with a responsive UI using HTML, Bootstrap, and Vue.js and the backend with Flask and SQLite, integrating RESTful APIs for seamless data interaction.
  - Implemented background task handling with Celery and optimized performance using Redis caching.
  - **Tech Stack:** HTML, Bootstrap, Vue.js, Flask, SQLite, and Python

Publications

- **A Survey on Human-AI Teaming with Large Pre-Trained Models**, *arXiv*, 2024. (Link) *Explores the impact of large pre-trained models in enhancing Human-AI collaboration.*
- **Deep Learning based Forest Fire Detection and Alert System**, *IEEE Xplore*, 2022. (Link) *Proposes a CNN-based model for fire detection with real-time alerts via Raspberry Pi; contributed a custom dataset of 2500 images per class.*
- **DKMI: Diversification of Web Image Search Using Knowledge Centric Machine Intelligence**, *Springer*, 2022. (Link) *Introduces the DKMI framework for improved web image recommendation using knowledge-based techniques.*