NIKHIL DEEKONDA

5512295780 ◇ deekondanikhil100@gmail.com ◇ Linkedin ◇ Github ◇ Portfolio

EDUCATION

Yeshiva University, Katz School of Health and Science, NY, USA Expected May 2025
Masters of Science in Artificial Intelligence **CGPA** 3.9

WORK EXPERIENCE

Lumamind - REVIV AI, North Hollywood, CA May 2025 – Present
Agentic AI Engineer

- Led the design and development of an iOS application (SwiftUI) for a virtual rehabilitation clinic, enabling secure and conversational patient engagement via both text and voice modalities.
- Architected a multi-agent orchestration framework on AWS Bedrock, enabling dynamic collaboration between persona-based agents for personalized care journeys.
- Designed, implemented, and maintained a FastAPI backend on AWS EC2, delivering secure REST APIs for patient voice uploads, question-answering, and agent orchestration; enabled low-latency integration between the iOS client and AWS Bedrock Agents.
- Developed a clinician-facing dashboard that visualizes patient insights, monitors medication adherence, and displays real-time behavioral analytics, allowing clinicians to intervene early.
- Built and deployed a real-time alerting system to notify clinicians immediately when a patient misses medication or exhibits relapse indicators, enhancing proactive patient care.
- Implemented secure authentication and user profile management with AWS Cognito, enforcing privacy and HIPAA compliance for all patient data and session flows.
- Developed custom AWS Lambda functions and OpenAPI schemas for knowledge retrieval, patient profile access, and tool invocation, powering agent-driven workflows and patient-specific recommendations.
- Integrated Retrieval-Augmented Generation (RAG) flows, leveraging AWS Bedrock knowledge bases for accurate, context-aware agent responses and clinical recommendations.
- Engineered audio processing flows using AWS Transcribe and Polly for real-time voice chat, transcription, and text-to-speech, providing a multimodal user experience.

LTIMindtree, Pune, Maharashtra, India Jul 2022 - Jul 2023
Senior Software Engineer

- Developed guardrails, data pipelines, and validation flows for ML/DL models including LLMs; fine-tuned models like GPT, Falcon, LLaMA-2, and Mistral for real-world use cases (PII removal, tone classification, toxic word detection) using LoRA and QLoRA techniques.
- Built a scalable platform to fine-tune LLMs with multiple techniques and implemented output validation for LLM agents utilizing frameworks such as AutoGen and LangGraph.
- Designed and executed complex Retrieval-Augmented Generation (RAG) workflows integrating LLMs with vector databases; developed hallucination detection systems leveraging Small Language Models (SLMs) and BERT-based models.
- Trained a proprietary 350M parameter language model from scratch and deployed fine-tuning, RAG, and validation workflows on AWS using services like EC2, S3, SageMaker, and Bedrock

National Highway Authority of India, Tirupati, Andhra Pradesh, India Jun 2021 - Aug 2021
Machine Learning Intern

- Applied various machine learning algorithms to NHAI road dataset of half a million instances, evaluating models based on R-squared values, successfully predicting road rutting factors, based on the best performing model.

PROJECTS

TaskFin: Agent-Based Financial Task Automation

- Designed and implemented a multi-agent, AI-powered system for end-to-end bill payment automation, leveraging Large Language Models (Claude 3.7 Sonnet, Mistral 7B) and the LangChain Agents + ReAct framework for secure, reliable orchestration of complex financial workflows from natural language input.
- Developed a robust Orchestrator Agent to coordinate specialized sub-agents for Authentication, Financial Transactions, and State Management, each using LangChain memory (buffer, entity, summary) to maintain contextual state across multi-turn user interactions.
- Built and integrated synthetic banking APIs and simulated multi-factor authentication (MFA) flows, delivering safe, production-like testing environments for LLM-powered financial automation.
- Deployed the complete solution as a fully interactive Streamlit conversational app, enabling real-world task automation, secure data handling, and rapid prototyping of AI-driven fintech assistants.

LungAware: AI Lung Cancer Detection App

- Engineered a deep convolutional neural network (CNN) for early lung cancer detection, achieving high classification accuracy through advanced image preprocessing, transfer learning, and hyperparameter optimization.
- Developed cross-platform (Android/iOS) mobile applications with Swift (iOS) and Kotlin (Android), integrating TFLite-quantized models and Grad-CAM for explainable AI and transparent diagnostic support.
- Implemented a seamless user interface and cloud-based backend for automated inference, patient management, and result notification, ensuring compliance with healthcare privacy and security best practices.

Enhancing LLM Reliability: Automated Fact-Checking with Cross-Encoder Model

- Built and fine-tuned a Microsoft DeBERTa v3 large language model as a fact-checking cross-encoder, achieving an F1 score of 85% on the Facebook Fact Checking Dataset and significantly reducing hallucination in downstream NLP systems.
- Automated real-time fact-checking in conversational and document QA scenarios, increasing the credibility and accuracy of LLM-generated content for end-users.

Optimizing Visual Bird Sound Denoising with Deep Learning: A Segmentation Approach

- Designed a custom encoder-decoder deep neural network architecture for visual bird sound denoising, outperforming state-of-the-art models with an IoU of 66.24 on challenging real-world datasets.
- Integrated advanced image segmentation and noise reduction algorithms, demonstrating measurable improvements in ecological acoustic monitoring and bioacoustics research.

AI-Driven Assistive Communication with FSR402 Sensors

- Prototyped an AI-powered system using FSR402 pressure sensors and machine learning algorithms to interpret finger movements for communication support, aimed at empowering semi-paralyzed or stroke patients with limited speech ability.
- Developed signal processing and classification pipelines in Python to translate physical gestures into actionable digital communication outputs, enhancing accessibility for patients.

CERTIFICATIONS

- AWS Certified Machine Learning Engineer – Associate
- AWS Certified AI Practitioner
- Python
- SQL