ABHINAY KOTLA

+1 469 674 1021 ⋄ abhinaykotla@gmail.com
LinkedIn GitHub Portfolio

SUMMARY

Software Engineer with over 3 years of experience building and deploying AI systems. Skilled in **Machine Learning, Data Structures, and Algorithms**. Developed AI-powered chatbots, optimized computer vision pipelines, and fine-tuned NLP models for production use. Experienced in full-stack software development, integrating ML with scalable systems, and delivering real-world business impact through data-driven automation.

SKILLS

Programming	Python, C++, Java, SQL, Bash, C#, .NET
Core CS	Data Structures, Algorithms, Systems Design, Complexity Analysis
LLMs & NLP	Transformers, T5, Prompt Engineering, Embeddings, RAG, Tokenization, Evaluation
ML & AI	Deep Learning, Generative AI, Transfer Learning, Representation Learning
Model Training	Fine-tuning, Distillation, Quantization, Inference Optimization
Frameworks	PyTorch, TensorFlow, Hugging Face, scikit-learn
Data & Pipelines	Data Preprocessing, Feature Engineering, ETL, Model Pipelines
Systems & Cloud	Docker, Kubernetes, Azure, CI/CD, Model Serving
Web & APIs	REST APIs, React, Node.js, Backend Integration

EXPERIENCE

IT Operations – AI Support Automation, UT Arlington OIT Sep 2024 – Present

- Built an AI-powered IT support assistant using Python, NLP, and LLM-based responses for query resolution.
- Reduced ticket resolution time by 18% using intent detection, retrieval-grounded answers, and ServiceNow integration.
- Developed automation scripts and internal tools to streamline IT workflows and operational reporting.

Full Stack Developer, Saintechinc Aug 2022 – Nov 2023

- Built and deployed a production-ready full-stack application using React, Node.js, and REST APIs.
- Implemented secure authentication, role-based access control, and scalable backend services.
- Delivered 99.9% uptime through modular architecture and optimized deployment pipelines.

Machine Learning Engineer, 1StopAI Nov 2021 – Jan 2022

- Developed and deployed ML models for emotion and gender detection in customer support audio streams.
- Optimized inference pipelines to improve routing accuracy and real-time response efficiency.
- Engineered audio feature extraction and model evaluation workflows for production deployment.

PROJECTS

Efficient CV Models with Knowledge Distillation *Model Optimization*

- Compressed CNN model by 99% (669MB to 6.5MB) while retaining 97% accuracy via pruning and knowledge distillation.
- Optimized for real-time inference using FP16 quantization, achieving 10× reduction in GPU memory usage.

News Summarization using T5 Transformer *T5, LLMs, NLP*

- Fine-tuned T5 for abstractive summarization; achieved ROUGE-1 of 0.53 with high semantic coherence.
- Built an inference pipeline with preprocessing, decoding control, and evaluation metrics.

TuneParams.ai Community Platform *C#, .NET Core, React, AI Systems*

- Architected a scalable community platform using C# for backend services and React for the frontend.
- Designed REST APIs and admin workflows supporting role-based access and real-world user interaction at scale.

FinAI – AI-Powered Personal Finance Assistant *LLMs & AI Systems*

- Built a Python-based finance assistant using Plaid data for automated tracking, budgeting, and personalized insights.
- Integrated LLM-based news summarization and sentiment analysis to contextualize user portfolio with market signals.

Edge- and Color-Aware Adversarial Image Inpainting *Deep Learning*

- Designed a dual-stage GAN with edge and color guidance, reducing perceptual loss by 7%.
- Trained G1 on masked Canny edges and G2 for structure-texture fusion, improving fine-grained details.

EDUCATION

Masters in Computer Science
University of Texas at Arlington
GPA: 4.0/4.0

BE in Computer Science
Gandhi Institute of Technology and Management, Hyderabad
CGPA: 8.47

Certifications
Neural Networks and Deep Learning (DeepLearning.AI)
Robotic Process Automation (RPA)
IBM Big Data with Spark and Hadoop
Wordcloud Using NLP and TF-IDF
Google Technical Support Fundamentals