

TANVI DESHPANDE

Boston, Massachusetts, 02120

857-506-5188 tanvideshpandedc@gmail.com LinkedIn github.com/Tanvi-15 U.S. Citizen

Education

Northeastern University <i>Master of Science in Computer Science (GPA: 3.5/4.0), Boston, Massachusetts</i>	09/2024 – 05/2026
Somaiya Vidyavihar University <i>Bachelor of Technology in Computer Engineering (GPA: 8.74/10), Mumbai, India</i>	09/2020 – 05/2024

Technical Skills

Generative AI: Agent Architectures, RAG, Prompt Engineering, Langchain, LlamaIndex, Vector Databases, Huggingface, Large Language Models (LLM), CTranslate2
ML & Deep Learning: , Computer Vision, NLP, TensorFlow, PyTorch, NLTK, NumPy, Pandas, Librosa
Programming: Python, Java, JavaScript, Go, SQL
Databases: MongoDB, MySQL, PostgreSQL
Cloud & DevOps: GCP, Git, Linux, Shell Scripting, Amazon Web Services(AWS), Terraform

Experience

Northeastern University <i>Research Assistant - Software Engineer - AI/ML, Boston, Massachusetts</i>	09/2024 – Present
<ul style="list-style-type: none">– Led a team of 4 researchers to design and deploy scalable multi-agent systems with task delegation and real-time monitoring, improving workflow efficiency by 40%.– Engineered prompt optimization frameworks incorporating course guidelines, rubrics, and reference materials, increasing AI response accuracy by 35%.– Automated rubric generation and batch grading with LlamaIndex, cutting grading time from 8+ hours to under 30 minutes for 200+ submissions.– Developed enterprise-ready agent architectures integrating multiple generative AI models (GPT, Claude, Llama) and benchmarked AI-only, human-AI hybrid, and traditional grading across 100+ submissions.– Contributed to full-stack development by building React-based front-end components and streamlined AI development pipelines, reducing project delivery timelines by 25%.	
Capgemini <i>Data Analyst Intern, Mumbai, Maharashtra</i>	01/2024 – 06/2024
<ul style="list-style-type: none">– Developed a Scheduler Analyzer leveraging LangChain and LLaMA-3B to enable natural language queries over complex database logs, intelligently mapping keys to scheduling workflows and improving efficiency by 10%.– Designed a Generative AI Assistant for aircraft engineers using LangChain and Mistral-8x7B, applying RAG over 1TB+ of technical data to surface relevant resolutions, reducing issue troubleshooting time by 25%.– Built multiple proof-of-concept solutions across data analysis, generative AI, and robotics, demonstrating measurable value and supporting data-driven decision making for 10+ enterprise clients.– Conducted market research on AI adoption in transportation and retail industries, delivering strategic pitch decks and ideating future use cases to inform client innovation roadmaps.	

Projects

Automated Speech Feedback System Whisper ASR, Praat/Parselmouth, Mistral-7B	09/2025
<ul style="list-style-type: none">– Built an end-to-end speech coaching pipeline analyzing clarity, pacing, cadence, and prosody from audio recordings using Whisper ASR, Praat, and Librosa.– Engineered a multi-modal scoring rubric mapping 10+ acoustic features (pitch, jitter, shimmer, pause ratio, WPM, spectral centroid) to qualitative dimensions like clarity, tone, and confidence.– Integrated a local LLM (Mistral-7B via Ollama) for personalized feedback and deployed an interactive Streamlit dashboard for audio upload, timeline visualization, rubric scoring, and JSON export.	
URL Shortener AWS, Docker, Go, Terraform, EC2, Distributed Systems, Gin Framework	09/2025
<ul style="list-style-type: none">– Developed and deployed a scalable backend service using Golang for URL shortening, capable of serving 500M+ requests/month deployed on AWS EC2 as Infrastructure as Code (IaC) using Terraform Docker, ensuring reproducible cloud deployment.	
Wildfire Image Classification PyTorch, TensorFlow/Keras, NumPy, OpenCV	01/2025
<ul style="list-style-type: none">– Built a binary wildfire detection model on RGB satellite imagery, implementing a from-scratch PyTorch CNN (hand-coded Conv/BN/Pooling) and a Keras baseline; achieved 98% vs 93% accuracy respectively.– Developed a reproducible pipeline (128×128 inputs, data augmentation, precision/recall/F1, confusion matrices) demonstrating feasibility without multispectral/thermal bands.	