# Tanvi Deshpande

US CITIZEN | linkedin.com/in/tanvideshpande1505/ | github.com/Tanvi-15 | Portfolio | Boston, MA | tanvideshpandedc@gmail.com

## SUMMARY

I believe GenAI without ML is just a chatbot - real impact lives at that intersection. I build agentic, RAG-based, and multimodal systems with cross-functional teams, turning complex AI into solutions researchers, faculty, and clients actually use.

## EDUCATION

**Master of Science - Computer Science,** *Northeastern University | (GPA: 3.6/4.0)* **Aug 2024 – Present**

**Bachelor of Engineering - Computer Engineering,** *Somaiya VIdyavihar University | (GPA: 8.74/10)* **Aug 2020 – Jun 2024**

## TECHNICAL SKILLS

**Generative AI** : Agentic AI, vLLM, MCP, LangChain, LangGraph, RAG, Ollama, Hugging Face, Prompt Engineering
**Machine Learning** : NLP, TensorFlow, PyTorch, NLTK, Pandas, NumPy, Librosa, Streamlit, Computer Vision, scikit-learn
**LLMOps & Cloud** : AWS, Lambda, EC2, S3, ECR, CloudWatch, SQS SNS, Docker, CI/CD, Terraform, Git, Github, RabbitMQ
**Programming & Web** : Python, JavaScript, TypeScript, React, Next.js, Express, Flask, FastAPI, Shell Scripting, Go, SQL
**Databases** : Vector DB (FAISS, ChromaDB), MongoDB, PostgreSQL, MySQL, Redis
**Core Competencies** : Cross-functional collaboration, agile iteration, technical communication, production system ownership

## PROFESSIONAL EXPERIENCE

**DMSB AI Strategic Hub,** Boston, MA **Jan 2025 – Present**

***AI/ML Software Engineer***

- Architected an **end-to-end GenAI grading solution**, using LangGraph agents to automate rubric creation and bulk essay evaluation with anti-gaming detection, **saving faculty grading time by 95% .**
- Engineered fault-tolerant LlamaIndex **RAG pipeline** (Hugging Face embeddings + Qdrant) with **custom guardrails,** processing and **indexing 30+ PDFs per assignment** for *accurate, context-aware essay grading*.
- Optimized GPU inference using **vLLM docker container** on NVIDIA Blackwell infrastructure after benchmarking against Ollama, achieving **5x faster throughput** for concurrent requests.
- Launched PresBot, an **AI Voice coach** using Flask backend (F5-TTS, Whisper) and Next.js frontend with specialized **ML libraries** (Praat, Librosa) for real-time audio analysis with feedback for **80+ student presentations.**
- Deployed applications using **Nginx, Github Actions CICD** with self-hosted runners reducing deployment time by 70%.
- Developed a comprehensive speech assessment system with composite scoring models combining acoustic and linguistic features (Clarity Index, Tone Variability) and rhythm assessment using PauseAnalyzer.

**Capgemini,** Mumbai, India **Jun 2023 – Jul 2024**

***Data Analyst Intern***

- Collaborated with **cross-functional teams** and **5+ client stakeholders** to gather requirements and **deliver GenAI PoC** solutions in data analysis, robotics, and process automation
- Reduced aircraft issue **troubleshooting time by 25%** by designing a **Generative AI Assistant** for engineers using **LangChain** and Mistral-8x7B, applying **RAG** over 1TB+ of technical data to surface relevant resolutions.

## PROJECT WORK

AI Roommate Finder — FastAPI, WebSocket, MongoDB, Ollama, Streamlit, Python

- Built a roommate matching platform where **AI clones of users autonomously negotiate compatibility** via WebSocket, featuring a **4-phase conversation architecture** with live streaming and structured compatibility analysis
- Implemented prompt engineering pipeline for persona generation from user questionnaires with hard/soft constraint encoding, async Python for concurrent agent orchestration, and dual authentication via OAuth 2.0 + bcrypt/JWT.

**AI Virtual Garment Try-On System** — PyTorch, Diffusers, U2NET, Hugging Face, Streamlit, Python

- Developed end-to-end clothing analysis pipeline using **U2NET** for **4-class garment segmentation** and **CatVTON-Flux** diffusion models for garment extraction.
- Built modular inference workflows integrating segmentation, mask generation, and diffusion-based processing with a Streamlit UI, optimized for Apple Silicon (MPS) hardware acceleration.

**Flash Sale Order Processing System —** Docker, SNS, SQS, Lambda, Go, Terraform, Distributed Systems

- Designed an **event-driven backend** in **Golang** for high-volume e-commerce using **AWS SNS and SQS** for asynchronous microservice communication. Deployed **Docker containers** on ECS behind an ALB with **IaC via Terraform**, and transitioned to a serverless architecture with **AWS Lambda** for automatic scaling and reduced operational overhead.