

# KHUSHAL MANDAVIA

+1(717) 623-3902 — New York Metropolitan Area — [khushal.mandavia72@gmail.com](mailto:khushal.mandavia72@gmail.com) — [LinkedIn](#) — [GitHub](#)

## EDUCATION

---

### Bachelor's of Science in Computer Science

August 2020 - May 2024

Pennsylvania State University

DSA, Systems Programming, Operating Systems, Deep Learning, NLP, DBMS, Computation Thoery, Tech Entrepreneurship

## EXPERIENCE

---

### Full Stack Developer

May 2024 - June 2025

FedEx Services

*Remote*

- Modernized IBM Maximo asset-management UI, wrote **Python automation scripts**, and built **Db2-backed REST APIs**—raising **network efficiency by 12%** under the company-wide DRIVE program.
- Refactored the Service Provider Vehicle Portal with **Angular** with **Jenkins** tests, cutting late-delivery incidents by 10%.
- Showed agency and lead a team of 10 to design and deploy critical API endpoints in collaboration with IBM, enabling FedEx to gain deeper insights into application usage efficiency and increasing cost savings by 20%.

### Perplexity AI Business Fellow

March 2025 - August 2025

Perplexity

*Remote*

- Selected for Perplexity AI Business Fellowship focused on AI business strategy; developed and deployed two new AI applications during hands-on workshops leveraging **Replit** and **Perplexity Sonar API**, driving 150+ API calls.
- Gained practical insight into aligning AI capabilities with product-market fit, emphasizing how **data quality and domain specificity** influence the design of effective **AI integrations** in real-world customer and exnterprise experiences.

### Software Development Intern

June 2023 - May 2024

FedEx Services

*Remote*

- Revamped legacy APIs within the Service Provider Vehicle Portal, resulting in a 30% faster data retrieval rate and streamlined workflows for over 500 daily active users.
- Revitalized the backend development with **REST APIs** accelerating operations to improve response time by 20%.

## PROJECTS

---

### Multimodal Voice Research Agent with STT, RL Fine-Tuned LLM, and TTS ([Voice Agent](#)), ([Code](#)) -

- Deployed a Multimodal Voice AI agent combining voice and screen context using **Whisper (STT)**, **RL Finetuned Qwen2.5 1.5B Model**, and **gpt-4o-mini-tts (TTS)** through EC2 on AWS.
- Enabled **WebSocket screen/audio streaming** with a **FastAPI** processing inputs at 2s intervals and ~94% **VAD accuracy** while achieving sub-1000ms latency with real-time voice response across screen-interactive queries.
- Powered the agent using a **custom fine-tuned Qwen2.5 (1.5B)** model aligned for first-principles reasoning and factual accuracy, optimized via SFT + RL on conceptual questions.
- Integrated the **Perplexity Model Context Protocol Server (MCP)** to avail **search and deep research** tools enabling real time on screen research via voice.

### Fine-Tuned Qwen2.5-1.5B with SFT Warmup & GRPO RL ([Model](#)), ([Code](#))

- Fine-tuned a **Qwen 2.5 (1.5B)** model using **Hugging Face TRL** with **LoRA**-based SFT warmup & **Reinforcement Learning via GRPO**, allinging model outputs to **first-principles reasoning** and tone using custom reward trainer across ~750 annotated interactions using RLAIF. Also applied **Curriculum Scheduling** with varying system prompt.

### Podcast Knowledge Engine with RAG-based Summarization and QA ([Podnotes](#)) -

- Designed a RAG-based Podcast summarization using **OpenAI Whisper** to transcribe audio to text (**achieving over 95% accuracy on clean speech**) and **Gemma3** for summarization & RAG powered chat for podcast questions.
- Embedded **Chroma** vector store and **Hybrid RAG** with **Semantic** and **BM25** retrieval through **LangChain** alongside **pvfalcon** for speech diarisation achieving accurate speaker labeling on multi-speaker podcasts, and structured storage in **DynamoDB** for scalable retrieval.

## TECHNICAL SKILLS

---

### Programming Languages

### ML/DL & LLM Frameworks

### LM Techniques

### MLOps & Infra

### Web & APIs

Python, JavaScript/TypeScript, Java, C/C++, SQL, HTML/CSS

PyTorch, Transformers, LangChain, LangGraph, DSPy, TRL, Unsloth, Ollama, LlamaIndex

LoRA, SFT, RLHF, RLAIF, PPO, GRPO, Quantization, RAG, Flash Attention, MCP

Docker, Kubernetes, GitHub Actions, Redis, AWS (EC2, SageMaker, S3, DynamoDB)

React, FastAPI, REST APIs, WebSockets, WebRTC