MANISH BISTA

manishbista.ai@gmail.com | (929) 373-9892 | Brooklyn, NY | F-1 / OPT
linkedin.com/in/manish-bista | github.com/codereyinish | manishbista.com

AI/ML-focused CS graduate with shipped production experience in backend development and cloud deployment. Product-minded, startup-oriented, and passionate about building scalable software grounded in machine learning.

EDUCATION

St. Joseph's University, New York — Brooklyn Expected May 2026
B.S. Mathematics / Computer Science | Minor in Data Science | GPA: 3.6 / 4.0

PROJECTS

ClassRec — Lecture Transcription App classrec.com
Python · FastAPI · Modal · WebSocket · GitHub Actions · DigitalOcean · Clerk · Sentry

- Built and deployed a full-stack web app enabling students to upload or live-record lectures and receive AI-generated transcripts
- Eliminated ~30% of hallucinated transcript segments by integrating Silero VAD via ONNX Runtime for silence detection before passing audio to Whisper
- Deployed faster-whisper large-v3 on Modal serverless GPU (T4), cutting per-lecture inference cost ~65% vs OpenAI Whisper API while improving accuracy on accented speech
- Automated deployment to DigitalOcean via GitHub Actions CI/CD, reducing releases from manual SSH to zero-touch on push to main

CloudBot — AI Food Ordering Chatbot github.com/codereyinish/CloudBot
Python · FastAPI · Dialogflow ES · Docker · GCP Cloud Run · Cloud SQL · Cloud Build

- Built a food ordering chatbot using Dialogflow ES for NLP intent handling, connected to a FastAPI webhook backend with Cloud SQL (MySQL) for real-time order management
- Set up CI/CD pipeline via Google Cloud Build reducing deployment build time by 60%; containerized with Docker and deployed to Cloud Run replacing local ngrok tunnels

MBGPT — Fine-tuned LLM for YouTube Comment Replies github.com/codereyinish
Python · Mistral-7B · QLoRA · PEFT · HuggingFace · Gradio

- Quantized Mistral-7B using QLoRA and PEFT to run inference on a MacBook CPU with 16GB RAM — reducing model storage by ~75% with minimal accuracy loss
- Fine-tuned on a custom YouTube comment dataset to generate context-aware reply suggestions, achieving human-like response quality
- Deployed interactive Gradio demo on HuggingFace Spaces for public access and portfolio demonstration

TECHNICAL SKILLS

AI / ML: PyTorch, HuggingFace, Transformers, LangChain, faster-whisper, Streamlit
Backend & Cloud: FastAPI, Node.js, Docker, Modal, GCP (Cloud Run, Cloud SQL, Cloud Build), DigitalOcean
Languages: Python, JavaScript, Java, C, R, HTML5/CSS, SQL
Data Science: Pandas, NumPy, Jupyter, Conda
Tools: Git, GitHub, GitHub Actions, Linux, VS Code, IntelliJ IDEA, Maven, Sentry, Clerk, Dialogflow ES
Frontend: React, Astro, Vanilla JS, Jinja2

RELEVANT COURSEWORK

Human Computer Interaction · Principles of Data Science · Computer Organization & Assembly · Advanced Computer Programming · Software Engineering · Python Programming