Noel Gurivindapalli

404-644-8492 | nagurivindapalli@gmail.com | Boston, MA | LinkedIn

Education

Wentworth Institute of Technology, Boston, MA
B.S. Computer Science | GPA: 3.45 / 4.0 | Dean's List (4 semesters)

Skills and Experience

Languages: Python, Java, JavaScript, TypeScript, C, C++, C#, SQL, R
Frameworks/Tools: React, Flask, FastAPI, PyTorch, Pandas, NumPy, Streamlit, Git, Docker, Kubernetes
Cloud/MLOps: AWS (SageMaker, Bedrock, Redshift, Lambda, S3), Azure, MLflow, vLLM
Databases: MySQL, PostgreSQL, MongoDB, Pinecone, ChromaDB
Generative AI/LLMs: GPT, Gemini, Anthropic, Hugging Face, LLaMA, LoRA Fine-Tuning, RAG/Agentic RAG, LangChain, LangGraph, AI Agents
ML: NLP, Transformers, Deep Learning, Classifier Models

Aigis | San Francisco
Generative AI engineer and AI Marketing Head | September 2025

- Developed a full-stack AI application using FastAPI and React that analyzes social media videos and generates marketing content, reducing content creation time by 80%

- Integrated OpenAI APIs (GPT-4, Whisper, Sora, Vision) with Structured Outputs to ensure type-safe, consistent AI responses and eliminate parsing errors

- Implemented Google Cloud Vertex AI integration for high-quality image generation using Imagen 4 Ultra, enabling automatic visual content creation for marketing posts and Veo 3.1 to create high quality videos

- Built real-time video generation pipeline using OpenAI Sora API with progress tracking, polling mechanisms, and automatic MP4 download functionality

- Designed multi-user fusion mode that combines 2-5 creators' video styles into unified AI-generated content using GPT-4 Vision API for visual analysis

- Created RESTful API architecture with FastAPI serving 15+ endpoints for video analysis, transcription, image generation, and social media posting

- Developed responsive React frontend with Tailwind CSS, implementing real-time status updates, video players, and interactive UI components

- Integrated X (Twitter) and Instagram APIs for automated video scraping, analysis, and social media posting with OAuth authentication

- Implemented Hyperspell memory system for persistent context across user sessions, enabling personalized AI-generated content suggestions

- Optimized image generation service with fallback model hierarchy (Imagen 4 Ultra → Imagen 4 → Imagen 3 → Imagen 2) for maximum reliability and quality

- Built comprehensive error handling and retry logic for external API calls, ensuring 99%+ uptime for critical video and image generation workflows

- Designed scalable architecture supporting batch processing of multiple videos with async/await patterns for improved performance

Boehringer Ingelheim, Ridgefield, CT
Associate Generative AI Engineer | Feb – Jul 2025

- Combined an LLM with a retrieval system (RAG) that fetches relevant schema information or examples of text-to-SQL pairs from a knowledge base to augment the LLM's generation process. The LLM has access to the database schema (table and column names, data types, relationships) to understand the structure of the data it needs to query..

Designed data pipelines for ingestion & processing; applied LoRA fine-tuning and PySpark for large-scale insights.

Created POCs on Agentic RAG, Graph RAG, AI Agents, and MCP servers; deployed LLMs via vLLM APIs to production.

Integrated Tableau dashboards for AI-powered insights, improving user data exploration and reporting efficiency.

Automated ingestion of pharmacy claims across multiple PBMs (CVS, Optum, Humana, ESI) into Amazon Redshift.

Standardized and validated data with SQL scripts, ensuring alignment with Redshift

schemas. **GPT-Based Presentation Maker**, Wentworth | Prompting Engineer | Sep – Dec 2023

Built UI with Gradio to summarize PDFs into presentations using GPT-3.5, LLaMA-70B, and Azure

LLMs. Developed REST APIs in Flask and FastAPI for scalable document processing.

Co-authored research paper with faculty on LLM prompting strategies and performance.

Harvard University – Lichtman Lab, Cambridge, MA
Data Analyst Assistant Researcher | Jan – May 2023

Processed and analyzed brain synapse data using Python (Pandas, NumPy, PyTorch,

Matplotlib). Identified neuron pattern correlations and produced visualizations for peer-reviewed

publications. Delivered presentations summarizing insights across multiple datasets.

Datasets were 10's of thousands with duplicates and dead cells, created script to filter and group the relevant neurons

Harvard University – IT, Cambridge, MA
Technical Support Engineer | May – Aug 2022

Created barcode inventory system for printer items, reducing tracking errors.

Resolved hardware/software issues at Science Center help desk.

Built ServiceNow pipeline to automate ticketing workflows.

Certifications

Advanced Java, Python, MySQL, Pandas, Prompt Engineering, NodeJS, HTML
Projects

GPT4 Stock analyzer - Created a full stack app that allows users to input stock tickers to receive the analysis of GPT4 using LangChain to see why the price moved from a certain date and what the model recommends based off the research done in previous nodes. The GUI lays out each individual node, showing the thought process of the Large Language Model.

Social Media Hook Generator - Created a full stack app that scraped popular social media sites like X and TikTok to find the most popular content on pages the user inputs, generate the hook of the scraped content and recreate it using Sora AI and upload it to the users page of choice. Used OpenAI Responses API, Structured Outputs and Realtime API.

MailGuard Pro – Capstone mobile app with spam detection (TensorFlow, PyTorch) + OAuth 2FA.
WIT Roommate Finder – Django/PostgreSQL web app with MongoDB backend for student matching.

Leadership

Founder & President, Wentworth Coding Club – Organized hackathons, ideathons, and LeetCode competitions. **Member:** CS Society, Robotics Club, Asian Student Association