Varshitha Gogineni

AI Engineer — Applied AI · LLM · Voice

+1 (940) 300-9984 | varshithagogineni007@gmail.com | linkedin.com/in/varshitha-gogineni | github.com/Varshithagogineni | varshitha-portfolio-rho.vercel.app

SUMMARY

AI Engineer specializing in production LLM applications, agentic AI workflows, and real-time voice systems. Hands-on experience shipping customer-facing AI — a self-healing browser-automation agent and LLM-powered voice assistants — using OpenAI and Gemini APIs, LangChain, RAG, tool/function calling, and the Model Context Protocol (MCP). Strong in Python and TypeScript, with a focus on optimizing latency, token cost, and reliability for deployed, forward-facing AI products.

TECHNICAL SKILLS

AI & LLM: LLM Application Development, RAG, Agentic AI, Tool / Function Calling, Model Context Protocol (MCP), Prompt Engineering, LangChain, Conversational AI, Voice AI, Gemini & OpenAI APIs, AWS Bedrock Agents & Knowledge Bases, Whisper STT, NLP, Model Evaluation

Languages: Python, TypeScript, JavaScript, SQL

Automation & Tools: Playwright MCP (browser automation), n8n, Twilio telephony, REST APIs, Docker, Git / GitHub

Cloud: AWS (Bedrock, Lambda, S3), GCP, Azure AI

Databases: PostgreSQL, MongoDB, MySQL, Pinecone Vector DB, Firebase

Frameworks: React, Node.js

EXPERIENCE

Cambo Box Aug 2025 – Dec 2025
AI Engineer Intern Remote

- Developed a voice AI system for restaurant order booking and management, automating 70–80% of inbound call workflows by integrating LLM-driven conversational pipelines with Twilio telephony and external reservation systems.
- Reduced invalid AI tool execution by 90% through structured schema validation, workflow optimization, and rigorous prompt engineering across multi-turn conversational flows.
- Engineered end-to-end voice pipelines using Whisper STT and LLM-based intent routing, enabling context-aware responses and dynamic tool calling for order placement and real-time status tracking.
- Delivered seamless booking management through real-time voice interactions, coordinating telephony, LLM reasoning, and backend reservation APIs into a single production flow.

University of North Texas — CMHT IT Services Dec 2024 – May 2026
Web Developer Denton, TX

- Developed and maintained API-driven web applications and internal data tooling used across multiple CMHT departments, improving load times and reliability on the team's most-used internal tools.
- Built ETL utilities and n8n automations that auto-generate reports and route form submissions, replacing manual multi-step staff workflows with one-click processes that cut report turnaround from hours to minutes.
- Improved backend performance and REST API reliability through system optimization, while maintaining SOPs and providing technical support across CMHT.

PROJECTS

Self-Healing Browser Automation Agent | *Playwright MCP, OpenAI Function Calling, TypeScript, Agentic AI, Node.js*

- Built a carrier-agnostic agentic loop that completes an 8-page workers' compensation insurance quote end-to-end with zero carrier-specific code, using an OpenAI reasoning model that drives a Playwright MCP server controlling a real Chromium browser via function calling.

- Engineered a self-healing core that auto-captures the live page snapshot on any tool failure and reinjects it into the conversation, letting the model diagnose and recover from complex Angular/PrimeNG widgets (date pickers, autocompletes, reactive-form validation) in real time without hard-coded selectors.
- Delivered a validated live submission (priced $539 instant quote, 0 errors, ~9-minute run) and added recipe record-and-replay so routine runs cost cents instead of dollars, cutting projected cost at scale by ~10x.
- Designed a modular TypeScript architecture (agent loop, MCP client, per-carrier prompt/data mapping, CLI) with token/cost tracking, history trimming, and rate-limit backoff; adding a new carrier requires only a single new file.

CMHT Voice Agent | *Gemini APIs, LangChain, Python, RAG, Tool Calling*

- Built a real-time, LLM-powered speech-to-speech voice assistant for UNT's College of Merchandising, Hospitality and Tourism using Gemini APIs and LangChain, helping students and faculty locate rooms and access front-desk assistance.
- Designed modular tool-calling pipelines with session memory for dynamic function routing and context-aware responses across extended multi-turn interactions.
- Optimized API latency and token usage via prompt caching, batched requests, and efficient context-window management.

VoiceFit — AI Voice Fitness Assistant | *Whisper STT, React, n8n, OpenAI, Supabase*

- Built a hands-free, voice-enabled AI fitness companion with Whisper STT and a responsive React front-end, integrating speech-to-text with backend logic for natural-language task execution; selected Top 14 of 32 teams at Dallas AI 2025 and deployed live at voicefit.vercel.app.
- Designed n8n automation workflows that let users complete routine tasks hands-free by voice instead of multi-step manual input.

EDUCATION

University of North Texas May 2026

Master of Science in Information Science — GPA: 4.00 / 4.00 Denton, TX

Relevant Coursework: Natural Language Processing, Information Retrieval, Machine Learning, Data Mining, Database Systems, Software Development for AI

BV Raju Institute of Technology May 2024

Bachelor of Technology in Information Technology Hyderabad, India

CERTIFICATIONS

Anthropic — Introduction to Agent Skills | Introduction to Model Context Protocol (MCP) | Building with the Claude API | Claude Code in Action

ACHIEVEMENTS

- Dallas AI 2025 — Top 14 of 32 teams for VoiceFit, an AI-powered hands-free fitness companion (OpenAI Whisper, n8n, Supabase); live at voicefit.vercel.app.
- Google Gemini × Pipecat Hackathon at Y Combinator (SF Tech Week 2025) — co-built an AI Mock Interview Platform with Gemini Live and Pipecat enabling real-time voice, screen-share, and AI code review.