Anish Gillella

anishgillella@gmail.com | *469-867-4545* | *LinkedIn* | *GitHub*

EXPERIENCE

Founding Software Engineer - Theus **January 2025 - November 2025**

- Implemented **multi-agent orchestration** with tool calling to coordinate ingestion, **OCR**, and enrichment tasks, reducing per-client processing cost by 45% and **accelerating customer onboarding** timelines.
- Developed a data enrichment engine using **React** with **Pydantic** models to enrich master data from **10K+ real estate investors**, directly supporting a **seed raise**.
- Automated 1,000+ client calls using **voice** agents using **VAPI** with **LiveKit** for real-time audio streaming, **Deepgram** (STT), and **ElevenLabs** (TTS), reducing manual post-call review time by 30+ hours weekly.
- Engineered **agentic browser automation** with **TypeScript** and integrated **Stagehand** for real-time monitoring, compressing package processing timelines by **70% (8 hrs to 1 hr)**.
- Trained models on **H100 GPUs** using **CUDA** to distill GPT-5 into smaller models via **contrastive learning** using **PyTorch**, improving retrieval semantics and improving **inference** by 40%.

Founding AI Engineer - AI/ML, AIRRIVED Inc. **January 2024 - December 2024**

- Collaborated with the founding team to design and deploy **AI agents** for cybersecurity triage and threat intelligence, reducing incident resolution time from **15 minutes to under 5 minutes**.
- Built **RAG pipelines** using **Pinecone** and improving **inference** pipelines**,** cutting retrieval latency from **45s to 12s** and accelerating real-time threat detection.

Software Engineer Intern, Cohezion.ai (Founding Team) **June 2023 – October 2023**

- Architected **LLM workflows** with **LangChain** and **OpenAI APIs**, enabling dynamic queries over **1M+ community analytics records** and reducing response latency by **40% (4.8s → 2.9s)**.
- Fine-tuned **LLaMA-2** models using **LoRA adapters (PEFT)** on domain-specific chat data, raising **F1 score from 0.72 to 0.83** and improving semantic relevance in analytics responses.

EDUCATION

The University of Texas at Dallas, Master’s in Business Analytics (GPA 3.7/4.0) **May 2024**

Manipal Institute of Technology, Bachelors in Mechatronics Engineering (GPA 3.6/4.0) **April 2022**

PROJECTS

AI-Powered Insurance Sales Acceleration Platform

Simulated an insurance automation platform integrating **voice agents** with **inference engines** and **TensorRT** GPU acceleration, leveraging **RAG** and **OCR** for document understanding and agent-driven automated form filling, reducing insurance sales cycles by 90% and eliminating manual prospecting..

Reinforcement Learning for Math Reasoning using GRPO & verl

Implemented **GRPO-based reinforcement learning** on Modal using the verl framework to train LLMs on **GSM8K**, achieving 90% accuracy in math reasoning. Automated dataset prep and checkpointing, and deployed via **vLLM** for 4× faster inference using distributed Ray training on H100 GPUs.

SKILLS

- **Languages & Frameworks:** Python, TypeScript, Next.js, Node.js, SQL, Streamlit, Langgraph, Google ADK
- **AI/ML & LLMs:** RAG, AI Agents, Voice Agents, MCP, PyTorch, HuggingFace, vLLM, SFT, CUDA
- **Data Engineering:** ETL, Spark, PostgreSQL, NoSQL, Supabase, Neo4j, Web Scraping, APIs, Vector Database
- **Cloud & DevOps:** AWS, Azure, Docker, CI/CD, Databricks, Kubernetes, Temporal
- **Tools & Platforms:** Git, OpenRouter, Modal, Runpod, DigitalOcean, Stagehand, Beam, vllm, TensorRT
- **AI Tools:** Claude Code, Cursor, Conductor, Lovable, Vercel

ADDITIONAL INFORMATION

- **Research Member**, Harvard Business Review Advisory Council
- **Brand Ambassador**, Raw Nutrition