Narasimha Karthik J

(440)723-0268 narasimhajwalapuram2026@u.northwestern.edu Website

Education

Northwestern University, Master of Science in Artificial Intelligence *Expected Graduation: Dec 2025*

PES University, BTech in Electronics and Communication Eng *Graduated: Sept 2022*

Skills

Languages: Python, SQL, C, Pandas, NumPy
Frameworks: PyTorch, TensorFlow, Transformers, LangChain, FastAPI, ChromaDB, Linux, Git, BASH
Research Interests: LLM Training & Inference, Multi-modal AI, NLP, Causal Inference, Mechanistic Interpretability, Agentic AI, CI/CD pipelines, Model deployment, Data preprocessing pipelines

Experience

Relativity Chicago, Illinois
Applied Science Intern *June 2025 - Aug 2025*

- **Led cross-functional initiative with Product, UX, and Engineering** through 10+ customer interviews to define 6 AI-powered insights (title, summary, structured summary, doc type, red flags, quality score), transforming manual document review into automated triage with 60% time reduction.
- **Built insights extraction pipeline via systematic prompt engineering** (3 iterations, MLFlow tracking) with LLM-as-Judge validation achieving 87.5% accuracy, eliminating 100% hallucinations through structured outputs, and processing 46,864 legal documents.
- **Created 600+ document benchmark dataset** combining real legal corpus (200 samples) with synthetic generation (400+ documents), improving model coverage from 60% to 95% for underrepresented categories.
- **Implemented scalable Databricks labeling system** with 95% inter-rater agreement, delivering reusable YAML configuration framework and evaluation rubrics adopted by 3 subsequent Applied Science projects.

CCL Lab - Northwestern University Evanston, Illinois
Research Assistant *Sept 2024 – Present*

- Developing research on LLM-driven evolution of agent systems with **publication accepted at GECCO workshop 2025**
- Developed a framework integrating **genetic programming** with **LLMs** via **LangChain** and **LangGraph**, improving agent-based code generation performance by **30%**.
- Engineered verification and performance tracking systems that reduced error rates in LLM-generated models by **25%**.

The Boeing Company Bengaluru, India
Data Scientist *July 2022 – Aug 2024*

- Fine-tuned foundation models (Llama, Mistral) using **LoRA** and **RLHF** techniques on domain-specific data with A100 GPUs
- Implemented a **RAG** system with **ChromaDB**, enabling real-time document creation and reducing manual drafting by **80%**.
- Designed and implemented CI/CD pipelines for trained LLM deployment, reducing deployment time by 60%
- Secured **$200k** in funding by demonstrating the business value of AI-driven document automation, processing over **20,000+ PDFs**.
- Developed operational automation tools including: a **BERT**-based Keyword Extraction Model (**95% accuracy**), a fine-tuned **T5** summarization model (Bleu-score of **25**), and an **NLTK**-based intent detector (**95% accuracy**).
- Led **3 technical sessions** for the NLP-LLM community and managed hiring to onboard **6 ML Engineers** from a pool of **50**.

Projects

Medhastra AI Mar 2025 – Present

- Developing Diagnostic and Treatment planning reasoning system to assist doctors take informed decisions.
- Designed an agent orchestration framework with LangGraph to perform complex reasoning chains with explainability
- Built production-grade ML infrastructure for model deployment with FastAPI and React

Agentic Blogging Assistant Mar 2025 – Mar 2025

- Designed and built a multi-agent system that synthesized my notes and code files, generating detailed outlines, section-by-section content for entire blogs, and social media shareable posts.
- Includes generating multiple iterations based on the quality threshold set. Created built-in caching mechanism support with **ChromaDB** vectorstore.
- Developed agent orchestration using **Langchain and LangGraph**, serving backend using **FastAPI**.

AdVocate - Tartanhacks @ Carnegie Mellon University Feb 2025

- Engineered an end-to-end solution that reduced campaign creation time by **90%**, generating **25+ unique campaigns** in under 24 hours
- Designed an **Agentic system** with a microservices architecture using **LangChain** that integrates **GPT-4o on Azure** as a chat endpoint, utilizes **ChromaDB** for caching, and leverages **Stable Diffusion** to generate dynamic images, processing **100+ market queries**
- Optimized API costs with a two-tier caching system, reducing API calls by **60%** and achieving **45% faster** response times