Narasimha Karthik J

(440)723-0268 narasimhajwalapuram2026@u.northwestern.edu jnk234.github.io

Education

Northwestern University | M.S. in Artificial Intelligence — GPA: 3.89/4.0 *Sept 2024 – Dec 2025*

PES University | B.Tech in Electronics & Communication — GPA: 3.6/4.0 *Aug 2018 – Sept 2022*

Experience

Medhastra AI | Co-Founder & CTO Chicago, IL | *May 2025 – Present*

- Architected multi-agent simulation environment (Patient, Tutor, Evaluator) using LangGraph, sustaining 30+ turn clinical dialogues with consistent medical context retention.
- Engineered stateful agent personas with dynamic symptom progression, delivering adaptive Socratic feedback that improved student diagnostic engagement in user trials.
- Optimized real-time streaming pipeline (FastAPI/WebSockets) to serve 50+ distinct medical case scenarios, ensuring ¡500ms latency for interactive training sessions.

CCL Lab - Northwestern University | Research Assistant Chicago, IL | *Sept 2024 – Present*

- Built LangGraph framework using GPT-4/Claude/Ollama to evolve NetLogo agent rules via genetic algorithms.
- Ran 1.5M simulations across 3 environments, discovering effective policies without explicit state annotations.
- Evaluated LLM-guided evolution versus hand-crafted baselines using food collection and task-completion metrics.
- Developed Scala 3 NetLogo extension; 10× faster multi-agent LLM orchestration now underpinning LEAR/QD-LEAR work.

Overtone | AI Engineer (Capstone) London, UK | *Sept 2025-Dec 2025*

- Engineered low-latency (¡200ms) Text-to-SQL agent enabling natural language queries on TB-scale datasets via Google BigQuery.
- Architected async event-driven system (Pub/Sub) to decouple inference from chat UX, ensuring sub-second response times.
- Deployed paragraph-level contextual analysis models, replacing keyword blocking with emotional/safety scoring for clients

Relativity | Applied Science Intern Chicago, IL | *June 2025 - Aug 2025*

- Engineered LLM-as-a-Judge pipeline processing 1M+ documents to detect prompt injections and red flags, reducing time by 60%.
- Synthesized 600+ adversarial red-teaming examples, establishing a company-wide safety benchmark with 95% expert agreement.
- Architected Azure Databricks active learning workflows, integrating feedback from 10+ experts to refine synthetic labels.
- Enforced structured decoding on GPT-4o Mini, achieving 100% schema adherence and 0% hallucination on complex tasks.

Boeing Research & Technology | SDE-2 & Data Scientist Bengaluru, India | *July 2022 – Aug 2024*

- Built end-to-end pre-training/RLHF stack on 100B+ tokens, enabling domain-adaptive generation for aircraft maintenance.
- Scaled alignment via PPO/DPO on expert preferences, improving model adherence to safety-critical repair protocols.
- Architected RAG system over 20k+ manuals, automating technical drafting and reducing manual authorship effort by 80%.
- Secured $200k internal funding by demonstrating LLM viability, leading to a production team of 6 engineers.

Selected Publications & Awards

- LEAR: LLM-Driven Evolution of Agent-Based Rules — ACM GECCO '25 Companion Proceedings
- QD-LEAR: Quality-Diversity Tradeoffs in LLM-Evolved Rules — Poster Acceptance, ALife 2025 (Kyoto)
- 3rd Place, Y Combinator Agents Hackathon — Anthropic Sponsor Track Winner for "VouchAI" (Agentic Insurance)

Projects

Quibo AI: Agentic Blogging Platform | *LangGraph, FastMCP, Supabase, Multi-Provider LLMs*

- Architected hierarchical multi-agent orchestration system (5 agents) using HyDE retrieval and semantic caching to autonomously convert code to blogs, reducing generation costs by 40%. Developed MCP server integration for agentic usage via Claude Desktop.

Lung Tumor Detection in 3D CT Scans | *PyTorch, MONAI, H100, TensorRT*

- Built production 3D segmentation ensemble (UNETR+YOLO3D) on H100 clusters, achieving 75.6% sensitivity and 2× faster convergence via novel patch sampling.

Financial Interpretability | *Transformers, Qwen 7B, Attention Rollout*

- Quantified causal influence of financial terms via attention rollout on Qwen 7B, using attention steering to improve balance sheet analysis accuracy from 65% to 85%.

Skills

- **Model Training**: PyTorch, RLHF/DPO, LoRA/QLoRA, FSDP, Distributed Training, Custom Tokenizers, H100 Optimization
- **Agents & RAG**: LangGraph, MCP, HyDE, Semantic Caching, Vector DBs (Chroma), LLM-as-a-Judge
- **Engineering & Languages**: Python (Expert), Scala 3, SQL, FastAPI, Databricks, BigQuery, Weights & Biases, Docker, CI/CD