Narasimha Karthik J

(440) 723-0268 narasimhajwalapuram2026@u.northwestern.edu Portfolio GitHub LinkedIn

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*

Skills

- **Agents & LLMs**: Agentic Systems, Tool Calling, LangGraph, Claude Agent SDK, MCP, Structured Output, Model Routing
- **Agentic Development**: Claude Code (custom skills, subagents, hooks, MCP servers, multi-agent workflows), Codex
- **Evals**: LLM-as-a-Judge, Eval Gating, Red-Team Benchmarks, Regression Sets, Cost/Latency Tracking, MLflow, W&B
- **Backend & Full-Stack**: Python, TypeScript, SQL, Scala 3, FastAPI, React, Postgres, MySQL, OpenAPI, CI/CD, Docker
- **Retrieval, Data & Cloud**: RAG, HyDE, Semantic Caching, BigQuery, Pub/Sub, Databricks, GCP, Azure, PyTorch

Experience

WhoBlinked | **Co-Founder & CTO** Remote | *July 2026 – Present*

- Architected a multi-tenant competitor-intelligence SaaS end to end, from data model and API contracts to the React product, turning rivals' public catalogues into a live feed of launches, price, and stock changes.
- Designed a resilient extraction pipeline that falls back from store APIs to structured markup to headless browsing, with block detection so failed scans never surface as false changes.
- Built LLM features (competitor discovery, gap analysis, ownership research) that verify model output against live search results before users see it, with per-call cost tracking and an automated test-and-release pipeline.

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

- Built and shipped NetLogo's official multi-provider LLM extension (Scala; OpenAI, Anthropic, Gemini, Ollama, Groq), installable from the NetLogo Extension Manager, with full documentation.
- Built for many-agent workloads: structured JSON output, probabilistic decisions, batched per-tick calls, retries, and cost tracking.
- Designed its flagship demo models, set up the NetLogo Conference workshop, and built the code generator behind LEAR.

Elevate Packaging | **Software Developer (Contract)** Remote | *Apr 2026 – June 2026*

- Built a rollstock traceability system from scratch (FastAPI, MySQL on Cloud SQL) tracking every roll from supplier to shipment, with idempotent loaders merging invoices, QuickBooks, and warehouse data.

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

- Built a LangGraph multi-agent clinical training product with an Evaluator agent grading diagnosis over 30+ turn dialogues.
- Shipped 50+ case scenarios with evolving symptoms, streamed over FastAPI/WebSockets at under 500ms latency.

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

- Deployed an LLM-as-a-Judge pipeline over 1M+ legal documents; GPT-4o/4o-mini routing cut review time 60% and cost 40%.
- Built a 600-example adversarial red-team benchmark that became Applied Science's default release gate across 3 projects.
- Enforced schema-constrained decoding on GPT-4o Mini; ran active-learning label refinement with Databricks and MLflow.

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

- Built a RAG system over 20k+ aircraft manuals with a custom parser for messy PDFs, cutting manual authorship effort 80%.
- Took it from a solo prototype to $200k internal funding and a production team of 6 engineers.
- Built the end-to-end RLHF pipeline behind it (100B+ tokens, reward model, PPO/DPO), gating releases on citation grounding.

Projects

Overtone Industry Capstone | *GCP, BigQuery, BQML, Pub/Sub, Cloud Functions*

- Built a sub-200ms Text-to-SQL agent over TB-scale BigQuery, with Pub/Sub and Cloud Functions decoupling inference from chat.
- Trained BQML ranking models on LLM-generated labels; replaced keyword blocking with paragraph-level contextual safety scoring.

Quibo AI | *LangGraph, HyDE, Semantic Caching, MCP*

- Built a 5-agent LangGraph system turning code repos and notebooks into blog posts; HyDE and semantic caching cut cost 40%.
- Shipped an MCP server so Claude Desktop can drive the full pipeline.

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

- Built a 3D UNETR + YOLO3D tumor segmentation ensemble on H100s (75.6% sensitivity) served with TensorRT.

Publications & Awards

- LEAR: LLM-Driven Evolution of Agent-Based Rules | ACM GECCO '25 Companion Proceedings (second author)
- QD-LEAR: Quality-Diversity Tradeoffs in LLM-Evolved Agent Rules | ALife 2025 Late-Breaking Poster, Kyoto
- 3rd Place, Y Combinator Agents Hackathon | Anthropic Sponsor Track Winner for "VouchAI" (agent-to-agent insurance)