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

- **Evaluation**: Benchmark Design, LLM-as-a-Judge, Red-Teaming, Multi-Turn Agent Evals, Failure Analysis
- **Post-Training & Interp**: SFT, RLHF, Reward Modeling, DPO/GRPO, Constitutional AI, Linear Probing, Attention Rollout
- **Training & Infra**: Python, PyTorch, Hugging Face, TRL, LoRA, vLLM, Multi-GPU (H100), W&B, MLflow, Databricks, GCP
- **Agents & Engineering**: Claude Code (skills, subagents, MCP servers, workflows), Claude Agent SDK, LangGraph, FastAPI

Independent Research

Sycophancy Recovery Study | *PyTorch, TRL, vLLM, DPO/GRPO/CAI, Linear Probing* *Jan 2026 – May 2026*

- Asked whether alignment fixes remove a learned failure or only hide it: induced sycophancy in Qwen3-8B with LoRA SFT, then compared DPO, SimPO, IPO, GRPO, and two Constitutional AI variants, scored by an LLM judge over 20K+ samples on 4×H100.
- Found behavior and internals diverge: per-layer linear probes showed the best behavioral fix (DPO-CAI) sharpened the sycophancy direction rather than removing it; released 12 models, 2 datasets, and a 6-part write-up.

Experience

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

- Built the core of LEAR (GECCO '25): an LLM-as-mutation-operator pipeline evolving NetLogo agent rules at the pseudocode level, with a LangGraph code generator and syntax verifier.
- Rebuilt BALROG, MARBLE, and LLMArena-style experiments as LLM-agent NetLogo models (knowing-doing gap, consensus vs. network topology, spy detection); showed calibrated probabilistic decisions prevent herding.
- Shipped NetLogo's official multi-provider LLM extension behind these experiments; set up the NetLogo Conference workshop.

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

- Co-founded a competitor-intelligence SaaS; built LLM features that verify output against live search before users see them.

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

- Built a LangGraph multi-agent medical simulation where an Evaluator agent grades diagnostic reasoning over 30+ turn dialogues.
- Designed 50+ evolving-symptom scenarios testing how agents handle ambiguous, misleading, and safety-critical presentations.

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

- Built a 600-example adversarial red-team benchmark that became Applied Science's default gate across 3 projects (95% agreement).
- Deployed an LLM-as-a-Judge pipeline over 1M+ legal documents; GPT-4o/4o-mini routing cut review time 60% and cost 40%.
- Ran active-learning label refinement with 10+ domain experts (Azure Databricks, MLflow).

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

- Built an end-to-end domain-adaptation and RLHF pipeline over a 100B+ token aerospace corpus: trained a reward model from SME preference data and aligned models with PPO/DPO.
- Caught a fabricated repair procedure that passed standard metrics; redesigned evaluation to gate releases on citation grounding.
- Took it from solo prototype to $200k funding and a team of 6; built its RAG layer over 20k+ manuals, cutting authoring 80%.

Projects

Financial Statements Interpretability | *Transformers, Qwen 7B, Attention Rollout, Integrated Gradients*

- Traced which balance-sheet items drive Qwen 7B's earnings predictions with attention rollout and integrated gradients; attention steering raised accuracy from 65% to 85%.

Memory for Robotic Manipulation (SAM2Act) | *PyTorch, Imitation Learning, RLBench*

- Extended SAM2Act's memory architecture with a learnable step embedding and phase-keyed memory retrieval; the step embedding took the stack-and-swap task from 0% to 100% success in ablation.

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

- Built a 3D UNETR + YOLO3D segmentation ensemble (75.6% sensitivity) with patch sampling that halved convergence time.

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)