

Production *AI* systems, shipped.

Gen AI Developer · Backend Engineer

A showcase of **RAG / GraphRAG pipelines**, **LLM agents**, fine-tuned models, and **FastAPI** backends — designed, deployed, and operated in production.

ramharikrishnan.dev

github.com/shankerram3

ramsankarharikrishnan@gmail.com · (602) 218-2159



ABOUT

Production-grade AI, *end to end*.

I'm a Gen AI developer and backend engineer with 5+ years of experience, focused on getting LLM systems into production — not demos, but systems that serve real users and keep working under load. I architected a GraphRAG chatbot serving 50K+ monthly queries, won 1st place at the Meta AI / AMD / PyTorch Hackathon, co-authored peer-reviewed research (LERN 2026), and contributed to Meta's open-source PyTorch tooling. I hold an M.S. in Computer Engineering from Arizona State University.

5+

Years shipping

50K+

Monthly queries

1st

Meta AI Hackathon

+28%

Acc. vs flat RAG

LERN

Published 2026

Services

RAG & GraphRAG systems

Chatbots & search that cite your data — chunking, embeddings, vector + graph retrieval, hallucination guardrails.

LLM agents & automation

Agentic workflows that use tools and finish multi-step tasks — LangGraph, MCP, structured output.

LLM fine-tuning

Custom domain models — LoRA/QLoRA & GRPO on Llama 3, Qwen 3, Phi-4 — with synthetic data & eval.

Python / FastAPI backends

Async & streaming endpoints, auth, caching, Prometheus metrics — Dockerized & deploy-ready.

Cloud deployment & MLOps

AWS / Azure / GCP, CI/CD, containerization, model versioning, A/B testing, monitoring.

Computer vision

Detection & segmentation on drone / ground / street-view imagery — benchmarking & reproducible training.

Arizona WaterBot

A statewide multilingual chatbot answering water-resource questions with relationship-aware, grounded retrieval.

50K+ monthly queries

+28% accuracy

AWS • Azure

ASU engagement



GraphRAG retrieval pipeline – relationship-aware context beats flat vector RAG.

THE CHALLENGE

Public users needed accurate, multilingual answers where flat RAG hallucinated and missed entity relationships.

WHAT I BUILT

A GraphRAG pipeline (LangChain + Neo4j) with hybrid vector + graph traversal, an async FastAPI backend with streaming, and output guardrails.

IMPACT

Served 50K+ monthly queries on AWS & Azure with +28% accuracy over flat-RAG; basis of a peer-reviewed LERN 2026 paper.

Stack: LangChain • Neo4j • OpenAI • FastAPI • PostgreSQL • AWS • Azure

Startup Intelligence Engine

Turns TechCrunch articles into a queryable Neo4j knowledge graph of companies, investors & funding.

200× faster cached

70%+ coverage

40+ endpoints

live demo

THE CHALLENGE

Turn unstructured news into queryable intelligence — entities, relationships, funding trends — fast enough for interactive use.

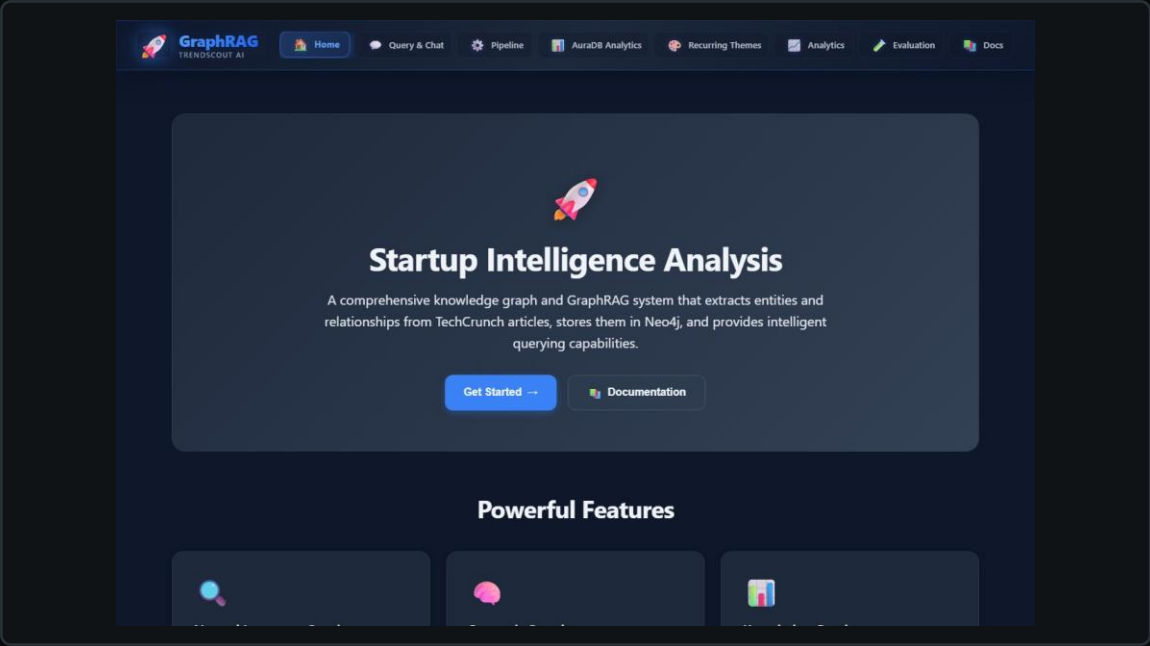
WHAT I BUILT

A 5-phase pipeline (scrape → GPT-4o extraction → enrich → Neo4j → embeddings) with a FastAPI backend, JWT auth, Redis cache, and 15+ Prometheus metrics.

IMPACT

200× faster cached responses (2000ms→10ms), 70%+ test coverage across 150+ tests, GitHub Actions CI/CD.

Stack: FastAPI
 • Neo4j
 • GPT-4o
 • Redis
 • React
 • Docker
 • [GitHub](#)



Live app – startup-intelligence-analysis-app.vercel.app

🏆 1ST PLACE • SYNTHETIC DATA • LORA FINE-TUNING

Meta AI Hackathon — Reasoning Models

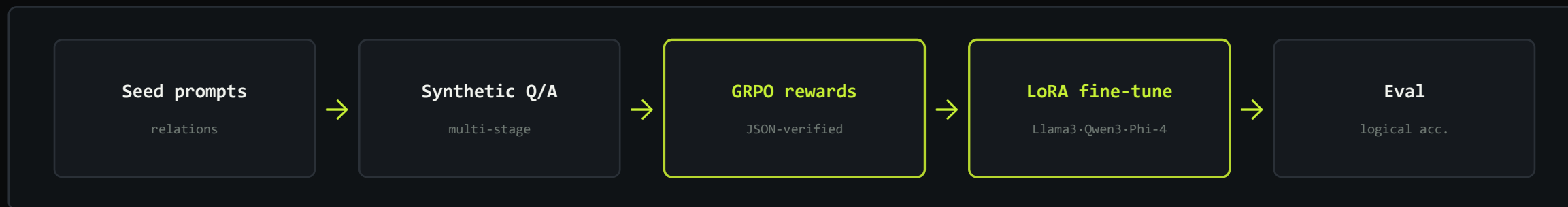
1st place at the Meta AI / AMD / PyTorch Synthetic Data Challenge: a synthetic-data pipeline + fine-tuned reasoning models.

1st place

16K+ LOC

95%+ coverage

4+ models



Synthetic-data → reward-shaping → fine-tune loop.

THE CHALLENGE

Generate high-quality training data for complex relational-reasoning tasks and lift model accuracy.

WHAT I BUILT

A multi-stage synthetic-data pipeline with structured-JSON enforcement, GRPO reward functions, and LoRA/SFT fine-tuning across Llama 3 70B, Qwen3-Next 80B, Phi-4.

IMPACT

1st place overall; identified & mitigated a major LLM reasoning flaw, improving logical consistency. 16K+ LOC, 95%+ coverage.

Stack: PyTorch • LoRA • GRPO • Llama 3 • Qwen 3 • Phi-4 • Unsloth • [deep dive ↗](#)

Wildfire Control RL

A physics-informed wildfire-containment environment + RL agent, contributed to Meta's open-source PyTorch OpenEnv.

Final Four

16K+ demos

OpenEnv PR

WHAT I BUILT

An RL environment with fire-spread physics; trained Llama 3.2 with LoRA on 16,000+ demonstrations using GRPO + behavior cloning on AMD MI100.

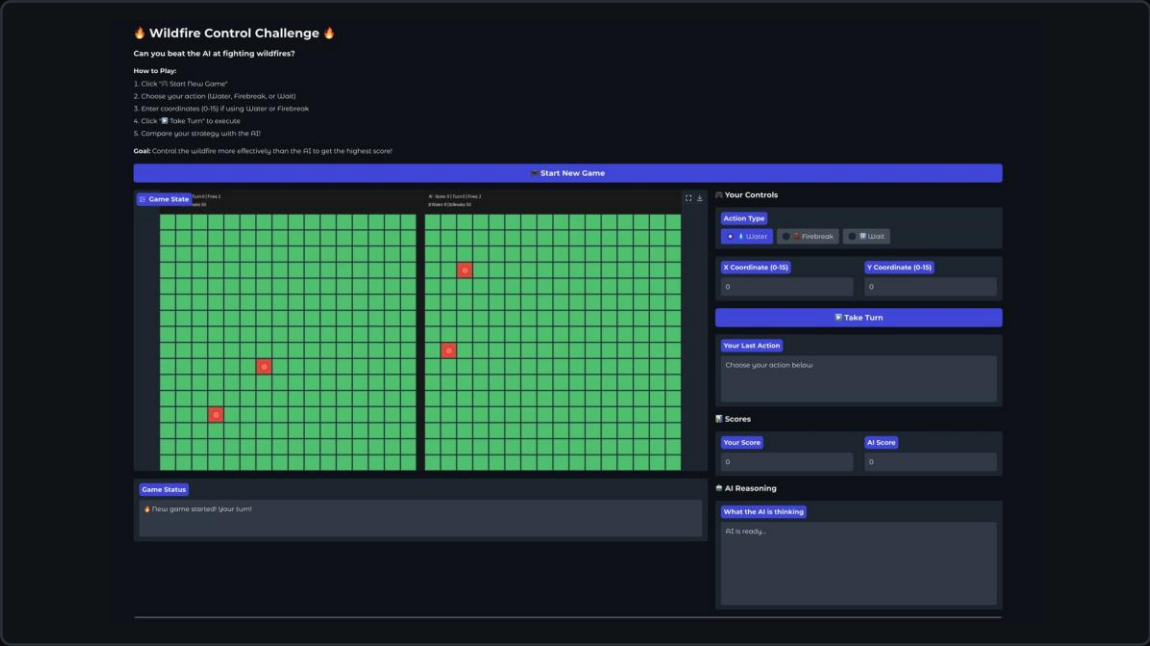
OPEN SOURCE

Contributed the environment to Meta PyTorch OpenEnv with docs & unit tests; deployed inference to Hugging Face Spaces.

IMPACT

Reached the Final Four; shipped an interactive human-vs-AI game demonstrating the trained policy.

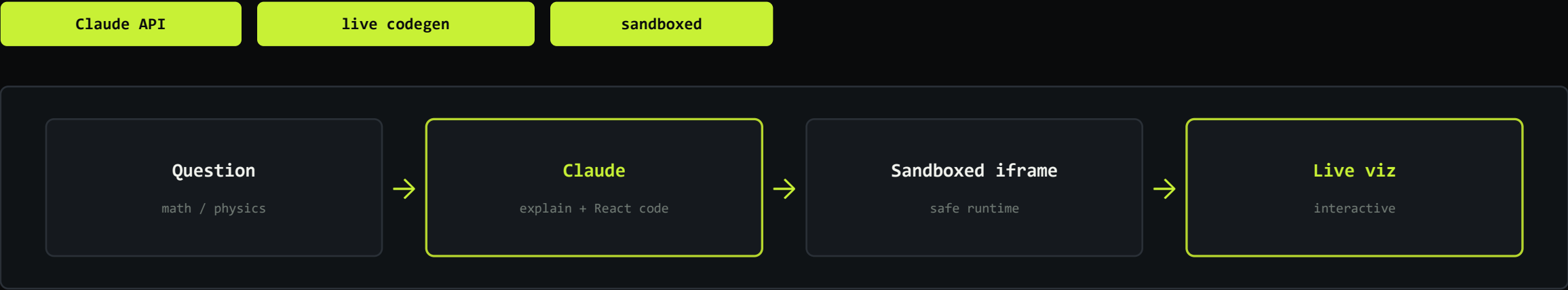
Stack: PyTorch • Llama 3.2 • LoRA • GRPO • Hugging Face



Interactive wildfire-containment game – human vs. trained RL agent.

MakerBOT — AI Tutor

Claude returns explanations plus executable React components, safely rendered in a sandboxed iframe — live STEM visualizations.



Question → Claude (explain + code) → sandbox → interactive visualization.

THE CHALLENGE

Make abstract STEM concepts interactive without pre-building every visualization by hand.

WHAT I BUILT

A FastAPI + React app where Claude returns explanation plus executable component code, rendered safely in a sandboxed iframe at runtime.

DOMAINS

Geometry, algebra, physics & calculus — real-time interactive visualizations generated on demand.

Stack: Anthropic API · FastAPI · React 18 · Three.js · Tailwind · [GitHub ↗](#)

ALSO BUILT

More work

NomadSync

Collaborative AI trip planner — a LangGraph agent with trip-memory extraction, plan versioning, live flight data, and real-time multiplayer over SSE.

LangGraph · OpenAI · Express · MongoDB · React

RagPiper

A flexible, production-ready GraphRAG engine with a pluggable architecture for swapping retrievers, stores, and models.

Python · GraphRAG · Neo4j

Graph Processing at Scale

Large-scale graph pipeline (NYC taxi data) with PageRank & BFS on Neo4j GDS, orchestrated on Kubernetes + Kafka.

Neo4j · Kubernetes · Kafka · AWS · Terraform

AI-Image Classification

ResNet50v2 / EfficientNetV2 model distinguishing AI-generated vs. real images, evaluated with AUC/F1 across CFAKE & ArtiFact.

PyTorch · CNNs · Transfer Learning

5+ YEARS BUILDING PRODUCTION SYSTEMS

Professional experience

ML / Computer Vision Engineer

Mesa Associates, Inc.

Knoxville, TN • Jan 2026 – Present

- Built a production computer-vision pipeline for utility-pole inspection — detecting conductor / wire polylines from drone & ground-level imagery (EPRI dataset).
- Benchmarked YOLOv8-seg, SAM 2 & HAWP; built a Mapillary scraper and Dockerized PyTorch training / inference workflows.

AI Engineer / Full-Stack Developer

Arizona Water Innovation Initiative • ASU

Tempe, AZ • May 2024 – Dec 2025

- Built & deployed a statewide FastAPI / React system on AWS (EC2, S3) with 50,000+ user interactions.
- Architected the GraphRAG WaterBot (+28% accuracy); CI/CD via GitHub Actions; multilingual NLP support.

Software Engineer — AI Integration

PIT Solutions Pvt. Ltd.

Trivandrum, India • May 2022 – Oct 2023

- Engineered Django microservices for AI / data workflows; led a .NET monolith → microservices migration.
- Built NLP intent-classification & AI search (~25% manual review); secured REST APIs with OAuth2 / JWT.

Software Engineer & Co-Founder

MakerLabs Edutech Pvt. Ltd.

Kochi, India • Dec 2020 – Apr 2022

- Co-founded an IoT-enabled STEM platform; built the IRIS AI pipeline across 50+ school deployments.
- Designed ARM Cortex-M smart educational toys with real-time MQTT / WebSocket device-to-cloud comms.

HANDS-ON, IN PRODUCTION

Technical skills

LLMS & AGENTS

Claude API, OpenAI, Gemini, LangGraph, MCP, prompt engineering, LoRA / QLoRA, GRPO, synthetic data

RAG & RETRIEVAL

GraphRAG, LangChain, LlamaIndex, Neo4j, Pinecone, ChromaDB, pgvector, semantic search, Cypher

BACKEND & CLOUD

Python, FastAPI, Django, Node.js, Docker, Kubernetes, AWS, Azure, GCP, CI/CD

DATA & MLOPS

PostgreSQL, MongoDB, Redis, Weights & Biases, MLflow, Prometheus, model versioning, A/B testing

ML & COMPUTER VISION

PyTorch, TensorFlow, Hugging Face, YOLOv8-seg, SAM 2, HAWP, CNNs, RLHF

FRONTEND & SYSTEMS

React, TypeScript, Next.js, Tailwind; C / C++, ARM Cortex-M, embedded firmware

LET'S BUILD SOMETHING

Have an AI idea? Let's *ship* it.

Available for freelance & contract work — RAG, agents, fine-tuning, and FastAPI backends. Also open to full-time ML & software engineering roles.

ramsankarharikrishnan@gmail.com · (602) 218-2159 · ramharikrishnan.dev

github.com/shankerram3 · linkedin.com/in/ram-harikrishnan