

PRAGNYAN RAMTHA

AI/ML Engineer · Founder, Agent7 · Prev. #1 ARC-AGI-2 Public Leaderboard

pragnyanramtha@gmail.com | +91 83318 62473 | pragnyanramtha.dev

github.com/pragnyanramtha | linkedin.com/in/pragnyanramtha

PROFILE

AI/ML engineer specializing in **LLM fine-tuning**, **RAG systems**, and autonomous agent infrastructure. Built a no-code AI agent platform solo from zero to production; previously ranked **#1 on the ARC-AGI-2** public leaderboard, with **70+ merged PRs** across the open-source AI ecosystem.

EXPERIENCE

Founder · [Agent7](#) — Remote

2025 – Present

- Shipped a no-code AI agent platform solo from zero to production, connecting **1,000+ app connectors** and orchestrating **20,000+ tools** across the full stack.
- Enabled teams to offload **20+ hours** of weekly repeat work with a long-running, approval-first agent orchestration system.

Software Engineer · [Learnable India](#) — Remote

Apr 2026 – Present

- Built assistive, **screen-reader-compatible** learning software for blind and visually impaired students; delivered accessible live sessions with educators and TEDx speakers.

Agentic AI Developer · [Reputation Dao](#) — Hyderabad / Remote

Aug 2025 – Jan 2026

- Architected a GCP serverless backend (Cloud Functions, Cloud Run) achieving **99.9% uptime**; cut inference latency **50%** via optimized Gemini API prompt caching.
- Built a RAG pipeline with semantic search for real-time documentation retrieval and source attribution.

ML Engineer (Freelance) · Six Axis Studios — Remote

Feb 2025 – May 2025

- Researched world-model approaches to generate CAD-style design outputs from spatial context and architectural intent.

PROJECTS

[AIMO-3: Efficient Reasoning via LLM Fine-Tuning](#)

- Fine-tuned Phi-4 (14B) on CoT/TiR datasets, achieving **90% accuracy** on reasoning benchmarks — rivaling 125B-parameter models at a fraction of the compute.

[Personality Clone](#)

- Fine-tuned an LLM with PEFT (QLoRA) and contrastive learning via a siamese network, achieving **92% accuracy** replicating personal response style (**28% improvement** over baseline).

[Autopilot](#)

- Built an AI-driven OS automation agent using function-calling and ReAct-style tool-use to execute natural-language commands across **15+ apps/sites** (Gmail, GitHub, X, LinkedIn) via the Gemini API.
- Added command sandboxing to the agent loop, cutting execution errors and achieving **45% faster** task completion than manual workflows.

[Deep Research — Dify Plugin](#)

- Built and published a Dify plugin integrating the **Tavily API** for real-time web search, letting LLM agents pull live information mid-workflow beyond their training data.

TECHNICAL SKILLS

Languages: Python, TypeScript, Rust, C, SQL, Bash, Go

AI/ML: PyTorch, Transformers, Unsloth, PEFT/QLoRA, RAG Pipelines, AI Agents, CUDA

Full-Stack: Next.js, React, Node.js, Tailwind CSS, PostgreSQL, REST APIs

Infrastructure: GCP, Docker, Linux (Arch), Nix, Git, CI/CD

CERTIFICATIONS

- Machine Learning Specialization** — Stanford University (DeepLearning.AI)
- CS50: Introduction to Computer Science** — Harvard University

ACHIEVEMENTS

- Prev. #1**, ARC-AGI-2 Public Leaderboard (Kaggle) · **70+ merged PRs** across major open-source AI projects
- Winner** — IEEE Summer of Code, Empathy Encryption Hackathon, Daydream Hyderabad @ Hackclub, NIAT RAG Challenge (2025)
- Finalist, Smallest AI × IIT-G · **Top 0.5% Finalist**, Shell AI Hackathon 2025
- Authored **2 research papers** on LLM context-window scaling and ML-driven fraud detection (Academia.edu)