

Ishaan Reddy

AI Engineer | LLM Systems | RAG | Agentic AI

Ishaanred7@gmail.com | +91 9618373310 | ishaanreddy.in | github.com/Ishaanred | linkedin.com/in/ishaan-reddy

PROFESSIONAL SUMMARY

AI Engineer building production RAG, agentic AI, and document intelligence systems for enterprise workflows. Cybersecurity background in vulnerability assessment, audit, and secure system design adds an edge for deploying LLM systems in risk-sensitive environments.

PROFESSIONAL EXPERIENCE

AILACDS — AI Intern → Junior AI Developer

Oct 2025 – Present

- Reduced client drafting time by ~90% by building SalesTaxGST with hierarchical RAG, cross-chat conversational context, and drafting assistance across a ~90GB legal/tax document corpus.
- Enabled curriculum-grounded lesson generation by engineering AIPathshala's Docling ingestion pipeline and retrieval system for 100k+ document-scale retrieval, with customer rollout pending.
- Owned cloud migration and bi-directional sync across client document workflows, moving ingestion and retrieval infrastructure to AWS for enterprise deployment.

PROJECTS

TextStruct - Local-First Textbook-to-RAG Engine

Python, DocTR, PyTorch, LanceDB, CLIP

- Created an open-source local-first pipeline that converts scanned textbooks into structured, LLM-ready data through PDF-to-image conversion, layout-aware OCR, cleanup, section detection, semantic chunking, and embeddings.
- Implemented document understanding with pedagogical role classification for definitions, explanations, procedures, examples, questions, results, references, evidence, visual proxy, and context, with optional Ollama fallback.
- Engineered LanceDB retrieval with ANN search, SQL metadata filtering, CLIP-based vision embeddings, hybrid text and vision retrieval, and image-aware queries for textbook RAG.

FrameWeaver - Multi-Agent Video Generation Pipeline

Python, FastAPI, Celery, Redis, React

- Designed a multi-agent orchestration pipeline for research, planning/script generation, image generation, voice synthesis, animation/rendering, and video compilation.
- Integrated llama.cpp local inference, FLUX image generation, IndexTTS-2/Qwen3-TTS voice synthesis, Godot, FFmpeg, checkpoint/resume, FastAPI status APIs, and Celery/Redis asynchronous jobs.
- Orchestrated GPU-heavy stages around single-consumer-GPU VRAM constraints, with websocket progress updates and a React monitoring UI for long-running jobs.

SLM From Scratch - Single-GPU Language Model Training

Python, PyTorch, Hugging Face

- Implemented a single-GPU GPT-style transformer trainer, handling tokenization, data pipelines, checkpointing, and generation.
- Optimized training experiments with OpenWebText, Hugging Face tokenizers, mixed precision, torch.compile, SDPA/Flash Attention, cosine learning-rate scheduling, gradient accumulation, gradient clipping, AdamW, and Weights and Biases logging.
- Structured 50M–500M parameter experiments to study practical LLM training limits without a GPU cluster.

EARLIER EXPERIENCE

MarketsandMarkets Research Pvt. Ltd. - Technology Intern

January 2025 – July 2025

- Delivered an end-to-end Lead Classification System that prioritized inbound business leads and achieved ~89% accuracy on real operational data.
- Designed feature engineering, model experimentation, and evaluation workflows that converted messy business records into classification-ready datasets.
- Shipped internal automation tools that accelerated research workflows, including an FAQ chatbot, GeoSearcher vendor discovery, Yahoo Finance ingestion, and LLM/regex data cleaning pipelines.

Grant Thornton Bharat - Cyber Intern

October 2023 – December 2023

- Conducted vulnerability assessments and security analysis as part of enterprise audit engagements, contributing findings to risk review workflows.
- Documented assessment findings and structured evidence for threat-aware review, improving traceability across security workflows.

ALSO BUILT

LLM Cookbook – 26-module Markdown/MkDocs reference covering transformers, tokenization, SFT, LoRA, QLoRA, RLHF, DPO, GRPO, evaluation, distillation, quantization, serving, and optimization.

DataVault – OCR and LLM pipeline for structured document extraction, cleanup, reminder generation, and cloud processing.

Tracium – Rust/Tauri/SQLite cross-platform network monitor with latency, jitter, DNS diagnostics, and bufferbloat grading.

Yahoo Finance Scraper – Standalone ~81,000-ticker financial dataset spanning 93 countries and 11 GICS sectors.

TECHNICAL SKILLS

AI / LLM: LLMs, Generative AI, RAG, Hierarchical RAG, Agentic AI, Voice Agents, Prompt Engineering, Prompt Orchestration, Tool Calling, Function Calling, LLM Integration, Fine-Tuning, QLoRA, LLM Evaluation, Multimodal AI

ML / Retrieval: PyTorch, Hugging Face, Transformers, Embeddings, Sentence Transformers, Semantic Search, Hybrid Search, Vector Search, Vector Databases, Reranking, Knowledge Graphs, Contextual Retrieval, Document Grounding, Citation Tracking, Vision-RAG, CLIP, OCR, DocTR, Docling, LanceDB, Mem0

Engineering: Python, Rust, TypeScript, JavaScript, React, FastAPI, Docker, Linux, Bash, SQL, REST APIs, Celery, Redis, PostgreSQL, pgvector, Neo4j, SQLite, Tauri, Godot, FFmpeg

Infrastructure / Security: AWS S3, AWS RDS, GPU Inference, Local LLM Inference, Model Deployment, MLOps, CI/CD, Scalable Architecture, llama.cpp, Ollama, Self-Hosted AI Infrastructure, Vulnerability Assessment, Audit, Threat Modeling, Secure System Design, Linux Security, Network Diagnostics, CEH, eJPT

EDUCATION

Vellore Institute of Technology

B.Tech in Computer Science, CGPA: 7.55

Graduated July 2025

Vellore, Tamil Nadu

CERTIFICATIONS AND TRAINING

- Certified Ethical Hacker (CEH) - EC-Council
- eJPT - INE
- DefAI - LLM Tuning + AI Safety - CLAWS and Hacktify