

Jayadithya R

Chennai, Tamil Nadu | +91 63696 05616 | rjayadithya11@gmail.com | linkedin.com/in/jayadithya-r | github.com/barryallen16 | jayadithya.is-a.dev

GENAI ENGINEER, RAG AGENTS AND LLM EVALUATION

Ships working AI systems across the full stack: 1.18M-chunk data pipelines, LoRA fine-tunes, a local RAG agent with typed tool calling, and eval harnesses that publish their own failures.

SKILLS

AI/ML: Retrieval-Augmented Generation (RAG) | LLM tool calling (JSON schema) | LoRA fine-tuning (Unsloth) | Eval harnesses for agentic systems | Hugging Face Hub | ChromaDB | CLIP/transformers.js

Languages: Python | Rust | SQL | TypeScript

Backend/Data: FastAPI | Docker | SQLite | PostgreSQL | Meilisearch | scrapy | yt-dlp

Spoken Languages: Tamil (native) | English (proficient) | Hindi (basic) | Telugu (basic)

EDUCATION

B.Tech Artificial Intelligence and Data Science, Dr. M.G.R. Educational and Research Institute, Chennai, TN |

2026

PROJECTS

hr-rag-agent | Python, FastAPI, ChromaDB, SQLite, Docker (github.com/barryallen16/hr-rag-agent)

- Built a local-first RAG agent over ChromaDB whose typed JSON-schema tool calls book leave in SQLite, and scored 9/9 on a 9-case eval harness that separates tool selection from answer match
- Hardened it with a retrieval-before-answer system prompt and a booking guard that rejects leave requests missing an explicit intent or a YYYY-MM-DD start date, so employee records never leave the machine

fitcheck-family | Qwen3-VL, LoRA, FashionCLIP, Next.js, Postgres (github.com/barryallen16/fitcheck-website)

- Ran the teacher-student distillation behind a first-author IEEE paper: a Qwen3-VL-8B teacher annotated 106K garment images, yielding 538K caption-and-QA training pairs
- Distilled the teacher into an on-device student: 3.1x less VRAM at 92% of teacher quality, validated against a blind human study

VulnChecker-Java | Python, Unsloth, LoRA, Qwen2.5 (github.com/barryallen16/VulnChecker-Java)

- LoRA fine-tuned Qwen2.5-7B over a 7-stage corpus pipeline, reaching 97% on a held-out Java vulnerability set; datasets and eval harness published on Hugging Face
- Self-audited the evaluation set and found 31 of 200 rows leaked fix-explanation prose, then published a neutralized set to replace the confounded benchmark

tamil-tokenizer | Rust, rayon, Python, Hugging Face (github.com/barryallen16/tamil-tokenizer)

- Designed a 1,135-token Tamil-aware base vocabulary in Rust around conjunct stems and ligatures, after measuring that conjuncts, not characters, drive Tamil compression
- Published Tamil BPE models that beat mBERT and Llama-3 on held-out compression; re-indexed the merge table to keep the larger vocabulary fast

kaggle-automation | Python, FastAPI, Docker, GitHub Actions, AWS EC2 (github.com/barryallen16/kaggle-automation)

- Dockerized a FastAPI control plane on Amazon Web Services (AWS) that pools 19 Kaggle credentials to shard GPU jobs across accounts, with quota meters and Telegram alerts

sivabharani-comments | Python, Gemma, Hugging Face Datasets (github.com/barryallen16/sivabharani-comments)

- Built a resumable pipeline over 2,755 Tamil tech videos, then ran local Gemma 4-bit inference across 1.18M transcript chunks

WORK EXPERIENCE

Blubridge Technologies, Chennai, TN | AI Research Intern |

07/2025 - 10/2025

- Implemented operator-fusion passes in MLIR for the team's custom tensor dialect, cutting kernel launch overhead by roughly 20% on benchmark workloads
- Wrote 40+ lit tests covering pass correctness and IR round-trip stability, wired into the team's CI