Kabyik Kayal

Kolkata, West Bengal, India ✉ mail@kabyik.dev Kabyik.dev Kabyik-kayal Kabyik-Kayal

Projects

Cogito — Self-Correcting Graph RAG *Github*

- Architected a cyclic state machine using LangGraph to replace linear RAG pipelines with an autonomous ”Trust but Verify” loop, implementing an AuditNode that programmatically detects and rewrites hallucinations.
- Engineered a Graph-Augmented Retrieval layer using NetworkX to traverse document relationships, capturing multi-hop context that traditional vector-only searches often omit.
- Developed a hardware-aware local inference engine via llama.cpp, optimizing memory usage for quantized GGUF models (Ministral-3-8B) to achieve up to 60 tokens/sec on consumer-grade hardware.
- Streamlined production deployment by designing an optimized 700MB Docker image with persistent multi-volume storage for ChromaDB and implementing a real-time FastAPI orchestration layer with Server-Sent Events.

UnArxiv — Science For All *Github*

- Engineered a memory-efficient fine-tuning pipeline for Qwen 2.5-3B-Instruct on Intel Arc A750 (8GB VRAM), utilizing Intel Extension for PyTorch and LoRA adapters (Rank 2) to overcome XPU hardware constraints while maintaining bfloat16 precision.
- Architected a Knowledge Distillation pipeline using the Groq API (Kimi K2) to synthesize a custom instruction-tuning dataset, enabling a 3B parameter ”Student” model to achieve +12.5% ROUGE-1 score improvement over the base model.
- Built a fault-tolerant training orchestration system with subprocess isolation to solve VRAM fragmentation issues inherent to the Intel XPU stack during long training sessions.
- Deployed the fine-tuned model via a FastAPI backend with Server-Sent Events (SSE) for real-time streaming, reducing output complexity by 2 grade levels (Flesch-Kincaid) to make scientific literature accessible to a younger audience.

Technical Skills

Languages: Python, Go, SQL

AI/ML Engineering: PyTorch, LangChain, LangGraph, Transformers, DeepEval, Scikit-learn, Pandas

MLOps & Systems: Docker, Jenkins, MLFlow, DVC, FastAPI, Circle CI, ZenML, Prometheus, Grafana, Git

Data Infrastructure: ChromaDB, NetworkX, PostgreSQL, Power BI, Tableau

Education

Indian Institute of Technology Madras (IITM) *Jun 2024 to Present*

BS in Data Science and Applications

- Currently Pursuing Diploma in Programming with a focus on core software engineering principles.
- **Coursework:** DBMS, Data Structures and Algorithms, App Development, Python, Statistics, Mathematics.