Dewashish Lambore

Pune (Maharashtra), India

📞: +91 9307059152 ✉: dewashish.lambore@gmail.com LinkedIn Github

Education

Symbiosis Institute of Technology, Pune **2024-2028**

Bachelor of Technology in Electronics and Telecommunication CGPA (current): 8.40

Experience

AI Intern **July 2025-Present**

Symbiosis Centre of Applied Artificial Intelligence

- Contributed to the design and development of AI-driven solutions in the financial technology sector, focusing on predictive modeling.
- Collaborated with cross-functional teams to integrate AI models into production systems, ensuring compliance with financial industry standards and regulations.
- Worked under strict confidentiality protocols to handle sensitive financial data securely.

Projects

Resilient Multi-Modal Agentic RAG System

An agentic, multi-tool AI system for robust, high-accuracy Q&A across diverse document formats.

- Engineered a multi-modal, agentic RAG system processing **7+ document types** (.pdf, .xlsx, .png), achieving a **90% accuracy** score on complex reasoning tasks—an **8x improvement** over the baseline.
- Implemented a state-of-the-art RAG pipeline featuring a BAAI/bge-large embedding model, **HyDE** query transformation, **Hybrid Search** (Vector + BM25), and a **GPU-accelerated Cross-Encoder Reranker** for high-precision context filtering.
- Reduced initial processing latency for a **400+ page document from over 3 minutes to under 25 seconds** through an offline **pre-processing and caching** strategy, and engineered the FastAPI endpoint to be resilient against adversarial attacks (zip bombs, oversized files) and API rate-limit failures using a **multi-key manager.**

Tech Stack and Methodology: *Python, FastAPI, LangChain, Gemini 1.5, Hugging Face (Transformers, Sentence-Transformers), PyTorch, RAG, Agentic Routing, Multi-Modal Processing, Hybrid Search, Cross-Encoder Reranking, HyDE, Few-Shot Prompting, Caching, Concurrency Management (asyncio), Pandas, Unstructured, ChromaDB, OCR (Tesseract), Robustness Engineering, API Design.*

Unsloth Model Finetuning for Conversational AI

Finetuning a quantized LLM with multi-turn chat capabilities using Unsloth and LoRA

- Finetuned a pre-quantized **4-bit LLM** for multi-turn conversational AI, leveraging **Unsloth and LoRA adapters** to drastically reduce training time and memory overhead on **ChatML-formatted datasets**.
- Implemented an end-to-end training pipeline in **Google Colab** using TRL's SFTTrainer, managing model serialization into multiple deployment formats **(4-bit, 16-bit, GGUF**) for compatibility with inference engines like **VLLM and llama.cpp.**

Tech Stack and Methodology*: Python, Unsloth, Huggingface Transformers, TRL (SFTTrainer), PyTorch, Datasets, ChatML, LoRA, 4-bit Quantization, Google Colab, Model Saving & Deployment Formats.*

Technical Skills

Languages: Python, JavaScript, Java, C

Database: MySQL, MS Excel, MS Access, MongoDB, Power BI

Libraries/Frameworks: Numpy, Pandas, Matplotlib, Scikitlearn, Jupyter Notebook, Seaborn, React, Tailwind, Bootstrap, CSS

AI Tools: ChatGPT, Claude, Gemini, Groq, Copilot, n8n, Leonardo AI , Cursor, Windsurf, Bolt Lovable, Maker

Collaboration and Version control: Git, Kaggle, Jira

Achievements

Runner's up, 2Fast2Hack: Secured second place amongst 600+ applicants creating an AI driven Edtech platform from scratch in 8 hours.

GDSC Top 1% Contributor: recognised amongst the top contributor of my college in GSOC and Hacktoberfest