VIKNESH

📞 +91 7092076809 ✉ jeyaviknesh@gmail.com 🔗 Viknesh ⭘ Viknesh

Education

Vel Tech Engineering College
Bachelor of Technology in Artificial Intelligence and Data Science

08/2020 – 05/2024
Chennai, Tamilnadu, India

Experience

Société Générale
Software Engineer - Full Stack

07/2024 – Present
Bengaluru, India

Global SG Chatbot widget - OKR

- Built a **RAG-powered** chatbot (NavyaBot) with React.js widget + FastAPI backend, enabling plug-and-play embedding across internal apps and adopted by multiple teams
- Delivered a reusable, **configurable widget** (theme, access controls, index selection) with one-line script embed; integration effort reduced to **<1 day per team**.
- Can Improve average query resolution time by **~40%**, minimizing dependency on **L1 support teams** and driving cost savings through SoGPT **pay-per-use model**.
- Consolidated Chat api (Internal) base subscription: **80×€772 → 1×€772 (≈ €61.7k base saving)** + pay-per-use

DQHub And Data Catalogue

- Implemented **API versioning** to enable smooth **consumer migration** without breaking integrations.
- Enhanced **business data** model and automated **synonym ingestion**, improving taxonomy and classification accuracy
- Redesigned **metadata crawler** with multi-threading, reducing runtime by **~86% and stabilizing daily runs across Azure/on-prem**.
- Delivered **bulk dataset** onboarding endpoint (**300+ datasets via Excel**), saving hours of manual effort.
- Built interactive dashboards (Highcharts) to **visualize anomaly metrics** & **KPIs** for business users.
- Strengthened system reliability and compliance through **audit trails, vulnerability fixes, and improved code quality**.

Projects

Multimodal AI Agent for Enterprise Dashboards

- Designed a **multimodal AI agent** that interprets enterprise dashboards by combining **Vision-Language Models** with **LLM reasoning**.
- Enabled screenshot-to-insight workflows, where users upload charts or **KPI snapshots** and receive **automated analysis** + **business recommendations**
- Integrated **LangGraph-based** function calling to dynamically fetch backend data (sales, alerts, performance metrics) for **context-aware responses**
- Delivered actionable intelligence in real time, reducing manual dashboard interpretation effort by **over 70%.**
- Fixed multiple attachment download issues in Documentation tab, improving **usability** and **stability**
- Packaged as a **FastAPI microservice** with **Streamlit UI**, making it deployable as a **plug-and-play** enterprise AI assistant

LLM Efficiency Playground - Comparing Prompting & Fine-Tuning

- Built an interactive **LLM efficiency playground** showcasing the trade-offs between **prompt engineering** and **lightweight fine-tuning**
- Demonstrated advanced **prompting techniques (zero-shot, few-shot, and Chain-of-Thought)** across real-world tasks.
- Implemented **LoRA/QLoRA fine-tuning** on an open-source model to boost task accuracy with minimal compute
- Integrated evaluation workflows to compare **baseline vs. fine-tuned** outputs, highlighting quality improvements.
- Deployed the **playground** via **FastAPI + Jupyter**, enabling reproducible experiments and hands-on exploration.

Skills and Technologies

Languages/Frameworks: Python, Java, FastAPI, Spring Boot, SQL (Postgres/MySQL)
Generative AI / ML: LLMs (GPT, LLaMA, Mistral), Prompt Engineering (Few-shot, CoT), Retrieval-Augmented Generation (RAG), Vision-Language Models (VLMs), Function Calling & Agents, LLM Orchestration
AI Frameworks / Libraries: PyTorch, TensorFlow, Hugging Face Transformers, Scikit-learn, Keras, LangChain,LlamaIndex, AutoGen,CrewAI
MLOps / Infrastructure:CI/CD (GitHub Actions, Jenkins), Docker, Kubernetes (basics), MLflow, Weights & Biases (W&B), OpenAI Evals / RAGAS
Efficiency & Scalability: LoRA / QLoRA (fine-tuning), PEFT, Quantization, PyTorch Lightning, Accelerate
Data & Search: Embeddings & Similarity Search, Vector DBs
Security & Responsible AI: Audit Logging, Data Governance, Bias/Fairness, Hallucination Mitigation, Red-Teaming AI Models