Shubham Yadav

Bangalore, Karnataka 560037
7047749505 sy247792@gmail.com linkedin shubham-yadav github.com/shubshub-bol

Profile

- AI Developer with a strong foundation in Machine Learning, NLP, and Generative AI, specializing in building and deploying end-to-end AI solutions. Skilled in implementing and fine-tuning LLMs, developing intelligent systems using RAG pipelines, prompt engineering, and Agentic AI frameworks. Proficient in Python, LangChain, and cloud-based ML workflows (LLMOps) with a strong foundation in data-driven decision-making. Passionate about transforming innovative ideas into scalable, production-ready AI applications.

Technical Skills

Programming: Python, SQL, R
Generative AI / LLM: LangChain, LlamaIndex, Hugging Face Transformers, OpenAI API, RAG Architecture, Prompt Engineering, Embeddings, Fine-tuning, Agentic AI, MCP
Databases: MySQL, Vector Databases (Chroma, FAISS, Pinecone), PostgreSQL
Machine Learning: Scikit-learn, TensorFlow, PyTorch, XGBoost, LightGBM, Deep Learning, Supervised & Unsupervised Learning, Model Evaluation
MLOps / Deployment: MLflow, LLMOps, Docker, Kubernetes, CI/CD, FastAPI, Flask, Streamlit, AWS SageMaker
Data Handling: Pandas, NumPy, Data Preprocessing, Feature Engineering, Large Dataset Processing
Tools & Collaboration: Git, GitHub, Jupyter Notebook, VS Code, LangFuse, Weights & Biases, Cursor AI, Zerve AI

Experience

Rubixe AI **Apr 2025 – Sep 2025**
AI ENGINEER *Bengaluru, India*

- Developed an AI-powered news research tool for equity analysis using LangChain, OpenAI, and Streamlit.
- Implemented a RAG pipeline utilizing FAISS vector indexes for efficient semantic search and retrieval of financial news.
- Engineered a text data preprocessing workflow to load, chunk, and create embeddings from multiple unstructured sources.
- Integrated OpenAI models to perform model evaluation and generate concise, context-aware summaries for analysts.

Projects

Conversational AI Chatbot with Custom Streamlit UI | *Python, LangChain, Gemini API — Personal Project* **2025**

- Engineered a sophisticated conversational AI chatbot using Python, LangChain, and the Google Gemini API to deliver intelligent, context-aware responses.
- Designed and implemented a fully custom, responsive user interface with Streamlit and advanced CSS, featuring a dynamic "liquid glass" theme to enhance user engagement.
- Diagnosed and resolved complex front-end layout bugs and backend API connectivity errors, ensuring a seamless and stable user experience upon deployment.

Reasoning LLM Fine-Tuning (Llama 3.2) | *Python, PyTorch, Unsloth, Hugging Face, Ollama* **2025**

- Fine-tuned the **Llama-3.2-3B** model using **Unsloth** and **QLoRA** on the ServiceNow R1-Distill dataset, enhancing the model with DeepSeek-R1 style iterative reasoning capabilities.
- Implemented custom "stream-of-consciousness" prompt templates and optimized training using Hugging Face **TRL (SFT-Trainer)** with 4-bit quantization to reduce VRAM usage.
- Deployed the fine-tuned model for local inference by converting adapters to **GGUF** format and integrating with **Ollama**.

Personality Prediction Model | *Python, Scikit-learn, Pandas — Rubixe AI* **2025**

- Performed data preprocessing and feature engineering on behavioral datasets using Pandas and Seaborn.
- Trained and evaluated Logistic Regression and Random Forest models, achieving **99% classification accuracy**.
- Contributed to documentation and reporting of model performance and insights for marketing campaigns.

Ask AI YouTube RAG Extension | *JavaScript, Chrome Manifest V3, Gemini API, Vector Search* **Dec. 2025**

- Architected a **serverless Retrieval-Augmented Generation (RAG)** system entirely within the browser, eliminating backend infrastructure costs by utilizing the user's local memory for vector storage.
- Engineered a client-side **in-memory vector store** using JavaScript to perform cosine similarity searches on video transcripts, achieving real-time query responses (500ms).
- Reverse-engineered YouTube's internal API to intercept `timedtext` and `continuation` network requests, enabling the extraction of hidden comments and transcripts without brittle DOM scraping.
- Integrated **Google Gemini API** for zero-cost natural language inference, processing user queries against dynamic video context with high relevance and low latency.

Generative Art with Diffusion Models | *Python, TensorFlow/Keras — Personal Project* **2025**

- Learned and implemented a basic diffusion model from tutorials to generate simple images
- Understood the core process of adding 'noise' to an image and training a model to reverse the process, creating new images from scratch.
- Used Python and Keras/TensorFlow to build the neural network, and used Matplotlib to visualize the generated images at different stages.

Certifications

NASSCOM - Data Science **2025**
Certificate by *NASSCOM*

IABAC - Data Analysis **2025**
Certificate by *IABAC*

ORACLE CERTIFIED GEN AI PROFESSIONAL **2025**
Certificate by *ORACLE UNIVERSITY*

Education & Training

Bansal Institute of Science & Technology **2018 – 2022**
B.Tech in Electronics & Communication Engineering — CGPA: 8.6 / 10 *Bhopal, M.P*

Datamites Certified Course **Apr 2025 – Sep 2025**
Certified Data Scientist — Grade : A *Bengaluru*