Satyam Sharma

Uttarakhand, India
+91-8218510659 • satyamsharma4949@gmail.com
linkedin.com/in/satyam-sharma-6632a61b4 • github.com/SALEX0R

Objective

Driven by machine learning enthusiasm. Passionate about leveraging algorithms and skills to drive organizational growth. Committed to continuous learning and applying generative AI to solve complex problems.

Work Experience

Rakuten **India**
Consultant AI Engineer *Nov 2023 – Jul 2024*

- Conducted research and development on **multi-modal OCR** systems, including **Duckling, Donut, InternLm, and QwenVl**, advancing document analysis and recognition technologies. **Integrated NLP** techniques to discern variations between documents, **enhancing accuracy and efficiency**, and aligning with cutting-edge projects in Generative AI.
- Streamlined development processes by implementing CI/CD pipelines, **automating testing and deployment of AI services**, and ensuring effective integration of ML solutions for cloud infrastructure and methodologies.
- Interacted with **stakeholders** to demonstrate product features, leading to successful production of deployment and ongoing management.
- Demonstrated expertise in AI/ML with hands-on experience in **LLM models** such as **BLOOM, Imperia, Ollama, BERT, Claude Sonnet, and Mistral**. Utilized **API services** like **Open AI and Azure**, and deployed **agents including Open Hermes**. Proficient in **frameworks** such as **Crew AI and Functionary**, contributing to strategic planning for innovative computer vision solutions.

IBM **India**
Machine Learning Intern *June 2022 – August 2022*

- Developed a machine learning model to categorize music samples into different genres, improving song selection efficiency by 40%.
- Leveraged audio processing techniques and deep learning algorithms to automate music genre classification, reducing manual labor by 60%.
- Implemented a recommendation system based on music classification, enhancing the user experience with an 85% accuracy rate.

Jan Bask **India**
Freelance Consultant *September 2022 – January 2023*

- Conducted customized training sessions to impart extensive knowledge in Data Science.
- Provided aspiring students with expert guidance and consultation on Data Science and Python.
- Stayed updated on industry trends and emerging technologies in Data Science.

Skills

Tech Stack: Python, SQL, No SQL, ML, NLP, LLM, GenAI, Transformers

Frameworks and Libraries: Flask, Crew AI, TensorFlow, Functionary, Hugging Face, LangChain, RAG, Numpy, Pandas, Scikit-learn, PDF-Annotation, easyOCR, Duckling, Donut, QwenVl, InternLm

Game Development: Pygame

Tools: MS Office, MySQL, Git, Adobe Creative Cloud

LLM Models: spaCy, Impira, LayoutLMv3, BERT, BLOOM, Ollama, Open-Llama, Mistral-7B, Claude Sonnet, Open Hermes, Open-AI, Azure AI

Projects

- **API Task Scheduling Agent**

leveraging **Open-AI's GPT-4, Lang-Chain, and an Agentic RAG (Retrieval-Augmented Generation) pipeline** for **multiple API calls** based on user prompts. For example, a user input like "I want to book a meeting at 2 pm" prompts the agent to search for the relevant API, and complete the task seamlessly.

- **AI-Powered SQL Generation Agent**

Developed an advanced agent **leveraging Open-AI's GPT-4, Lang-Chain, and an Agentic RAG (Retrieval-Augmented Generation) pipeline** to revolutionize database interactions. The project features seamless **MySQL integration**, translating natural language queries into SQL for easy data access.

- **Optimized LLM with RAG Pipeline**

Developed a **RAG pipeline using Lang-Chain**, which involved **recursively splitting and chunking text, storing it in a vector database, invoking the chain** to optimize LLM performance. Additionally, implemented **Redis Cache** to store output prompts and compare queries, significantly **reducing redundant processing and optimizing LLM calls**.

- **Spam Detector with Server-less API Deployment**

  GitHub: `https://github.com/SALEXOR/AWS-Spam-Detector-Serverless-API`

Developed a machine learning-based **spam detector and deployed it as a serverless API using AWS Elastic Beanstalk**. Performed regular application versioning and server log analysis, achieving a 20% improvement in overall server performance. Implemented a robust performance monitoring system, reducing downtime by 15% and ensuring optimal ML model deployment.

Education

University of Petroleum and Energy Studies **India**
B.Tech in Computer Science, (4 years of in-depth specialization in Artificial Intelligence and Machine Learning)
CGPA: 7.47/10

Languages

- English (Full Professional Proficiency)
- French (Elementary Proficiency)

Interests

- Gaming
- Music Production
- UI/UX Design
- Football
- Snooker
- Traveling