SHREYAS PACHPUTE

Surat, Gujarat | +91 7228903007 | shreyaspachpute1107@gmail.com | Portfolio | LinkedIn | GitHub

SUMMARY

AI Engineer & AWS ML-Specialist with deep expertise in GenAI/LLMs. Architected and deployed end-to-end AI systems—from fine-tuning open-source models to RAG pipelines and multi-agent orchestration. Automates AI workflows, ships production APIs, and scales cloud-native solutions on AWS.

EXPERIENCE

Commercient LLC · AI/ML Engineer *Jan 2024 – Present*

- **Dynamic RAG Platform:** Built a next-gen Retrieval-Augmented Generation system allowing users to spin up AI bots with a Google Drive/YouTube link; automated data ingestion into Pinecone, with seamless deployment across Slack, Zoom, and web apps.
- **Optimized LLM Inference:** Engineered multi-GPU, multi-server inference pipelines using vLLM, providing 150tokens/sec throughput under concurrent user loads.
- **LLM-Powered AI Agents:**
  - **Sales Agent:** Integrated into the Commercient website to engage prospects with real-time company data, contributing to increased lead generation.
  - **Accounting Agent:** Assisted customers with account balances, invoices, and payment queries.
  - **Support Agent:** Integrated into the company's helpdesk platform to provide instant, automated ticket resolutions.
- **SQL View Generation for Data Sync:** Automated the generation of SQL views essential for robust data synchronization between ERP and CRM systems, streamlining bidirectional integration and business workflows.
- **Model Ops & APIs:** Fine-tuned Llama models for key customers and exposed prediction services through .NET/C# APIs and AWS SageMaker.

CERTIFICATIONS

- **AWS Certified Machine Learning – Specialty** (*Badge*)
- **Amazon ML Summer School 2023** (*Certificate*)

SKILLS

- **Languages:** Python · C# · JavaScript · SQL
- **Backend & Databases:** FastAPI · Flask · .NET APIs · JWT · Redis · PostgreSQL · MySQL · Vector DBs
- **Cloud & DevOps:** AWS · Docker · Kubernetes · GitHub Actions
- **GenAI & LLMs:** Langchain · LangGraph · CrewAI · AutoGen · LlamaIndex · HF Transformers · VLLM
- **GenAI Techniques:** RAG · Prompt Engineering · MCP · Multi-Agent Orchestration · Fine-Tuning · Evaluation
- **AI/ML/NLP/CV:** ML/DL Algorithms · Text/Image Processing · Neural Networks
- **Other:** Git · Linux · System Design

EDUCATION

Sarvajanik College of Engineering & Technology · BE Computer Science
Surat, Gujarat | 2020 – 2024 · CGPA 8.3/10

PROJECTS

CONVERSATIONAL CONTRACT ANALYZER | *GitHub* | *Demo*

- Ingests and OCR-processes legal contract PDFs to extract and index clauses for natural-language querying.
- Achieves 95% accuracy in identifying key entities (parties, obligations, deadlines, penalties) via a custom NER pipeline.

MOVIE RECOMMENDATION SYSTEM | *GitHub* | *Demo*

- Developed an interactive movie recommendation system enabling users to select a movie and receive personalized suggestions for similar titles.
- Implemented a content-based filtering approach by processing movie metadata (e.g., genres, keywords, cast) to calculate similarity scores and rank recommendations.