SHREYAS PACHPUTE

📍 Surat, Gujarat | ☎ +91 7228903007 | ✉ shreyaspachpute1107@gmail.com | 🌐 Portfolio | 🔗 LinkedIn | 💻 GitHub

EXPERIENCE

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

- **Dynamic RAG Platform:** Developed a Retrieval-Augmented Generation system enabling users to create AI bots by simply providing a Google Drive/YouTube link. The system automatically ingests content into Pinecone, making the bot readily usable across platforms like Slack, Zoom, and web interfaces.
- **Optimized LLM Inference:** Engineered multi-GPU, multi-server pipelines using vLLM, achieving throughput of up to 150 tokens/sec under concurrent loads, ensuring efficient and scalable inference.
- **Custom 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.
- **Enterprise ML Solutions:** Designed and deployed ERP/CRM forecasting and churn-prediction models, improving inventory turnover by 15%, enhancing pipeline accuracy by 20% and reducing churn risk by 15%.
- **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*)
- **Kaggle Micro Courses** (*Certificates*)

SKILLS

Programming & Tools:	Python | C# | SQL | JavaScript | Git | Linux
AI/ML & GenAI:	Supervised & Unsupervised Learning | RAG | Prompt Engineering | Fine-Tuning | NLP | CV
Frameworks & Libraries:	Langchain | Hugging Face | Scikit-Learn | TensorFlow | PyTorch | VLLM | OpenCV
Backend & APIs:	FastAPI | Flask | Django (REST Framework) | JWT | Redis Caching | .NET Web APIs
Deployment & MLOps:	Docker | Kubernetes | CI/CD | AWS (SageMaker, Lambda, EC2, CloudWatch)
Data Engineering & EDA:	Pandas | NumPy | Matplotlib | PostgreSQL | Pinecone (Vector DB) | ETL Pipelines

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.

SMART PARKING SYSTEM | *GitHub* | *Demo*

- Architected a real-time Smart Parking System leveraging Computer Vision (Python, OpenCV, YOLO) for precise vehicle detection and parking spot occupancy tracking.
- Delivered application featuring a FastAPI backend for event logging and a responsive web dashboard for live status updates.

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.