MANISH SHARMA

+91-9342533525

G-Mail Github Linkedin Twitter Kaggle

Bangalore, 560068

WORK EXPERIENCE :

Machine Learning Engineer - 2 : **Parspec.io, Bangalore** Website | Linkedin

Jan 2024 – Present

- ***Latency Reduction***: Built Datasheet Recommendation System for model number to datasheet mapping across 5M Documents with **latency reduction by 80%** using **Cache [Aws Dynamo DB]** and Gemini 2.0 Flash LLM from OpenRouter.
- ***Model Fine-Tuning***: **Fine-tuned Llama 3.3 70B Instruct Model** on a custom Alpaca format dataset for attribute extraction on **Modal Labs using H100.**
- ***Accuracy Achievements***: **Achieved 90% accuracy** in matching model numbers across 5 million documents, extracting 10 distinct attributes with **91% overall accuracy.**
- ***Family Name Extraction***: Helped develop a **family name extraction algorithm** with significant improvements by integrating the Gemini 2.5 Flash model via the Gemini SDK, boosting **recall from 89% to 96%.**
- ***Kubernetes Migration***: Migrated the entire **AI-dev Kubernetes workload to AWS EKS** with guidance from the Snapsoft team. Mapped the load balancer to API Gateway and created isolated partitions for staging and production environments.
- ***Order Detection Algorithm***: Designed and deployed an order information detection algorithm for lighting datasheets using T5-base, **achieving 95% accuracy** and automating data extraction.
- ***Algorithm Enhancement:*** Enhanced header and column detection algorithm, increasing capacity from 4 to 8 **columns with 97% accuracy** and reducing manual processing by 30%.
- ***LLM as Judge Pipeline***: Designed and implemented an "**LLM as a Judge**" pipeline to assist human annotation workflows. Leveraged **GPT-4o and Gemini 2.5** Flash in parallel execution to evaluate datapoints for manual annotation. **Decreased human annotation to ~ 74%**
- ***Multimodal RAG Pipeline***: Built a **RAG pipeline supporting** the LLM-as-Judge framework to evaluate retrieved knowledge base chunks. The KB included both images and text, using BGE for text embeddings, CLIP for image embeddings, and **FAISS** as the vector store. **Achieved MRR of 0.94 in retrieval and 0.96 accuracy in generation.**

Machine Learning Scientist : **Docsumo-AI, Bangalore** Website | Linkedin

Dec 2022 – Dec 2023

- Designed **Wireframes** and Implemented Advanced **Document KV + Table Extractor** using **LayoutLM, BROS, YOLO** architectures for both **Fixed and Unstructured Document Categories.** Deployed on product.
- Successfully Integrated these ML & DL architectures into **10+ Custom API's** for our clients having **MRR in range $80K-100K.**
- Built and Integrated **Chat-AI,** a powerful and seamless **LLM** integration of **LangChain** and **PineCone-DB for QA and** other support-tasks in product.
- **Reduced the Annotation Time** from **1 full day to ~2hrs** using **GPT-KV LLM Extractor,** powered by GPT-4. This has saved lots of human efforts.
- Implemented an **Advanced Synthetic Data Generation Pipeline** capable of producing **Duplicate-Data** with exceptionally high correlation similarity for any fixed forms using FP-Tree algos. Through extensive **benchmarking**, our solution has demonstrated **superior performance, outperforming the Production Level APIs** by leveraging only **about 20% of real data.**
- Worked on different Fixed Forms for **KV and Checkbox Extraction** like Insurance [1040, 1120's, W9's], Bank Cheques and so on.

Research Assistant Intern : **Indian Institute of Science, Bangalore**
June 2021 – Sep 2021

- Responsible for Collating and PreProcessing Massive Hindi Datasets for **OCR** and **Speech Recognition task.**
- Leveraging the power of **PyTesseract** and **EasyOCR** for Text Extractions, and **WordLevel** Acc for Evaluating the OCR Model.
- Worked with **Librosa** and **MelSpectrogram** to build and analyze ASR Tasks
- Melinda Gates Foundation under SpireLab

PROJECTS / SIDE BUILDS :

V-Rag [Video Based RAG System] : Qdrant, RAG, Video-Querying Github | Loom

- **Video RAG System**: Developed a system allowing users to query video content via YouTube URL or uploaded videos.
- **Video Chunking & Indexing**: Implemented video chunking and indexing with **Qdrant** for efficient vector search.
- **QA Pipeline Integration**: Integrated a QA pipeline to retrieve **relevant frames, timestamps, and precise answers.**
- **Vector Database Optimization**: Experimented with 3 vector databases to optimize performance.
- **Deployment**: Deployed the solution with Streamlit for an intuitive user interface.

AutoCommit Generator : Mistral, Ollama, Github, LLM Github | Twitter

- **AutoCommit Generator**: Built an AutoCommit Generator using **Ollama and Mistral to automatically generate commit messages for projects locally and quickly.**
- **Local & Privacy-Focused**: Fully **local** solution with no privacy concerns, ensuring secure usage.
- **Simple & Fast**: Developed as a **bash script for quick installation** and seamless terminal integration.
- **Efficient & Lightweight**: Designed with under **100 lines of code** for simplicity and powerful functionality.

Company Scraper [AI-Powered Agent] : Relevance AI, LLMs, Markdown Agent-UI | Twitter

- Built a lightweight agent that generates clean, structured summaries from company URLs (e.g., *pixxel.space*). **automatically generate commit messages for projects locally and quickly.**
- **Auto-extracts**: Overview, Products, Features, Audience, Integrations & more
- Powered by **Relevance AI + custom LLM prompts**
- Outputs are markdown-formatted for easy reading
- **Perfect for:** Due diligence, competitor analysis, interviews, startup research

TECH SKILLS / FRAMEWORKS :

- **Languages & Frameworks**: Python, Cursor, FastAPI, Flask, GitHub, Ollama, OpenRouter, Lovable, HuggingFace, Grok
- **AI & Machine Learning**: ML, DL, NLP, PyTorch, TensorFlow, HuggingFace, Transformers, LLMs (GPT's), Finetuned Models, RAGs, Multi-Agent RAGs, Agents, VectorDB, GenAI Solutions
- **Natural Language Processing (NLP)**: NLTK, Spacy, SciSpacy, MedXN, Librosa, PyTesseract, OCR
- **Databases & Analytics**: MySQL, Neo4j, Tableau, Amplitude Analytics, Hasura-DB, Amazon DynamoDB, S3 Buckets, GCs
- **Cloud & Collaboration**: AWS, GCP, Google Colab, Kubernetes, Docker, EKS, ECS
- **Data Structures & Algorithms**: Strong foundation in problem-solving and optimization