Devesh Surve

(857) 313-1660 | Boston, MA | deveshsurve5@gmail.com | Linkedin | Github | Portfolio | Medium | Kaggle

Summary

Experienced ML Engineer with 3+ YOE building AI Platforms and Products. Developed patented solutions in Gen AI at Fidelity Investments with Research experience at Northeastern University in LLM Hallucinations. Graduate Teaching Assistant for Neural Networks and Gen AI

Professional Experience

Northeastern University **May 2024 – Present**

Graduate Research Assistant (under Prof. Dino Konstantopoulos) *Boston, MA*

- Implemented the paper on Self-Instruct to automatically scale data for finetuning from 100 rows to 100k rows using Llama-3, updating the paper from its original implementation in da-vinci model

Graduate Teaching Assistant (INFO-6106 - Neural Networks under Prof. Konstantopoulos & CSYE-7380 - Gen Ai under Prof Das) *Boston, MA*

- Graded over 200 assignments, ensuring the accurate assessment in topics covering RNN, CNN, and Transformers

Fidelity Investments **Jul 2023 – Dec 2023**

Data Scientist Co-op *Boston, MA*

- Developed a LLM factual-consistency evaluation method using an ensemble of Chain/Tree of Thought prompts, achieving 93% accuracy against BLEU, BARTScore (Patent filed) (Paper Link)
- Improved the accuracy of 401(k) forms' table preprocessing by 38% by replacing the rule-based approach with custom object detection model, using table-transformers (TATR) and Camelot
- Boosted FlanT5 LLM baseline accuracy by 29% to extract values from 401k forms using QLoRA fine-tuning along with creating a tool to generate annotated data automatically on AWS EC2
- Improved a RAG-based customer query resolution system by enhancing document reranking, resulting in a 30% boost in response relevance by integrating Cohere's reranking model with OpenSearch Vector DB

LTIMindTree **Aug 2019 – Jun 2022**

Senior Product Engineer (AI/Machine Learning) *Pune, India*

- Increased ML Platform reliability by 35% by developing an automated testing framework using Jenkins and pytest, enabling QA teams to validate models with 50% faster turnaround times.
- Reduced model drift by 18% by integrating continuous A/B testing into the MLOps pipeline using MLFlow, enabling data scientists to compare deployed performance against a baseline during deployment
- Accelerated deployment timelines by 23% by creating an Auto-Model-Deployment system automatically serializing, containerizing, and deploying pickled ML models using FastAPI and Docker
- Achieved 100% model reproducibility and ensured regulatory compliance by integrating MLflow for model versioning and implementing audit trails, reducing non-compliance risk by 20%.
- Optimized 25+ APIs for interacting with MySQL PostgreSQL databases using ORMs, improving code consistency by 34%

Technical Skills

Languages / Databases :	Python, C, C++, Java, SQL (MySQL, PostgreSQL), NoSQL (MongoDB, DynamoDB), Snowflake, Pinecone, Weaviate
Tools and Frameworks :	REST APIs, Swagger/Flasker, FastAPI, GitHub, CI/CD, MLFlow, TensorFlow, PyTorch, HuggingFace, Keras FlanT5, Phi-2, GPT-3.5/4, Langchain, LLama-Index, Pandas, Numpy, Scikit-learn
Cloud / MLOps :	AWS (S3, EC2, Lambda, SageMaker, EKS), Google Cloud (Vertex AI, GKE), Azure (ML Services, AKS), Prometheus

Projects

Parallel Deep Learning for Image Captioning using PyTorch *Link* **Feb 2024 - May 2024**

- Applied deep learning techniques for image captioning using the COCO dataset, building a ResNet 50 model with PyTorch, processing 20GB of data and 100K images
- Implemented 5 types of parallelism on the Discovery cluster to enhance performance and efficiency of data and model training
- Utilized PyTorch's DataLoader and compared its performance with COCO's lazy loader; implemented preprocessing on the Discovery cluster using multiprocessing techniques like IMAP and APPLY ASYNC, achieving a 1.41x speedup
- Employed PyTorch DDP (Distributed Data Parallel) for model training across 1, 2, 3, and 4 GPUs, recording a 1.31x speedup
- Implemented model parallelism by splitting the encoder-decoder across multi-CUDA cores, achieving a 17% efficiency increase.

Nike Product Search Bot: Multimodal RAG with Virtual Try-On *Link* **April 2024 - June 2024**

- Implemented multimodal search functionality, allowing users to query products using both text+image inputs using CLIP Embeddings
- Engineered a virtual try-on feature leveraging Florence-2 for caption generation and DALL-E-3 for image synthesis
- Developed an AsyncIO-based guardrailing system for LLMs, reducing inappropriate responses by 99% whilst maintaining low latency
- Integrated Azure Computer Vision for background removal and image processing, improving virtual try-on accuracy by 25%

Education

Northeastern University **Sep 2022 - May 2024**

Master of Science in Computer Software Engineering *Boston, MA*

University of Mumbai **Aug 2015 – May 2019**

Bachelor of Engineering in Computer Engineering (1st Rank in CS Dept) *Mumbai, India*

Publications

Detecting Errors through Ensembling Prompts (DEEP): An End-to-End LLM Framework for Detecting Factual Errors — TBD, Vol. , Issue- - Link : Proposes a new approach combining multiple LLM prompts for improved text validation in open-domain QA, aiming to surpass current automatic evaluation system complexities