Shreyas Verma

Google Scholar | shreyas-verma | Website | ✉ shreyas301197@gmail.com | ☎ 404-916-1923

Experience

Simplr AI - An Asurion Company Feb 2024 - Present
Data Scientist - Generative AI *San Francisco, CA*

- **Expert Copilot Agent:** Leveraged closed-source LLMs, Redis-based RAG systems and enhanced prompting techniques to build an agent that helps live experts address product information and product recommendations-based customer queries.
- **DPO-based LLM Alignment Pipeline:** Leveraged the above Copilot pipeline to create a GPT-4 generated preference dataset, which was then used to align Llama-3 8B model using DPO. The aim is to reduce the dependency on GPT-4 calls.
- **LLM-based Tech Troubleshooting Chatbot:** Architected and deployed a production-grade multi-turn conversational AI system, focusing research on multi-agent coordination and effective LLM integration for technical support automation.
  - **Multi-Agent Framework Research:** Engineered a novel multi-agent architecture that leverage tool-use along with RAG to support troubleshooting with high accuracy. Designed a "Personalization Agent" concept for context-aware cross-selling.
  - **Multi-Stage Intent Detection Strategy:** Optimized intent detection by instruction fine-tuning Mistral-7B on user-query data, coupled with a few-shot prompting strategy on GPT-4o as a fallback layer to optimize for cost and latency.
  - **Optimizing for Latency:** Leading research and implementation of techniques like efficient memory management,robust semantic caching and optimizing RAG retrieval pipelines to reduce latency in bot responses. Bought down latency by 40%.

Zillow Group, Search & Discovery AI Team May 2023 - Aug 2023
Applied Science Intern *Seattle, WA*

- **Multimodal Representation Learning:** Developed a Transformer-based architecture to create engagement-based embeddings for Zillow home listings. Improved downstream Similar Homes recommendations by $\sim$ 3% NDCG.
- **Finetuning CLIP for Image Modality:** Fine-tuned the open-sourced CLIP (Vision+Language) model on Zillow listing data to provide image embeddings,using an ensemble of language outputs from models like GPT-3.5, LLaVa-2 and InstructBLIP.
- **Zero-Shot OpenAI Embeddings for Text Modality:** Obtained Zero-shot embeddings for Zillow listing descriptions using the OpenAI API to provide text embeddings as input to the above multimodal architecture.

American Express, Acquisitions Data Science Team Dec 2019 - Jul 2022
Data Scientist *Bangalore, India*

- **Question Generation and Retrieval System:** Used Attention-enhanced Graph Neural Network for generating questions, fine tuned for financial documents - achieved 0.4 BLEU score.
- **NGBoost Proof-of-Concept:** Researched utility of the NGBoost algorithm to quantify the uncertainties in tree-based model predictions in prospect acquisition models by incorporating a Bayesian prior distribution. Reduced run-time by 1.5x.
- **In-house modeling Pipeline:** Leveraged Hadoop MapReduce to develop a distributed modeling pipeline for product-affinity models,utilizing XGBoost, Adjusted Mutual Information and BayesOpt.

Recent Publications

- **Polymath: A challenging multi-modal mathematical reasoning benchmark:** Created a 5,000-image dataset to evaluate multimodal LLMs' cognitive reasoning across 10 categories, including pattern recognition and spatial reasoning.[Link]
- **TarGEN: Targeted Data Generation with Large Language Models (COLM'24):** Designed a multi-step prompting strategy to generate high-quality synthetic datasets, integrating a 'self-correcting' mechanism for label accuracy.[Link]
- **Context-NER : Contextual Phrase Generation at Scale (ENLSP Workshop, NeurIPS'22) :** Introduced a novel NLP task to generate contextual phrases to enhance interpretability of Named Entities in complex financial datasets.[Link]

Projects

- **Reducing LLM Hallucinations using Epistemic Neural Networks :** Experimented with a combination of DoLa and ENN architectures to reduce hallucinations in Large Language Models (worked with Llama-2 7B specifically)[Link]
- **Instruction-tuned Clinical Notes Scoring :** Fine-tuned T5-based Large Language Models for extracting medically relevant phrases from patient notes using Instruction-based learning paradigm.[Link]
- **Question Answering with joint reasoning from Language Models and Knowledge Graphs :** Improved the joint reasoning of LMs and KGs for Question Answering tasks by adding contexts for named entities in the questions.[Link]

Education

Georgia Institute of Technology Atlanta, US
Masters of Science in Analytics - Computational Data Science track *Jul 2022 - Dec 2023*

Birla Institute of Technology & Science, Pilani Pilani, India
Bachelor of Engineering in Electronics & Instrumentation *Jul 2015 - Jun 2019*

Skills & Coursework

Languages & Tools: Python, Apache Spark, Hadoop MapReduce, C/C++, SQL, Bash, JAVA, Hive, Alteryx, Tableau
Frameworks & Libraries: Pytorch, DeepSpeed, FSDP, Huggingface, Spacy, NLTK, Gensim, Pyspark, Keras, Plotly