Shubham Patel

shubhampatel77.github.io
github.com/shubhampatel77 linkedin.com/in/shubham-patel shubham177k@gmail.com

Skills

Core ML/AI: LLMs, Retrieval-Augmented Generation, Computer Vision, Multimodal Learning, Diffusion Models
Frameworks & Libraries: PyTorch, HuggingFace, transformers, FAISS, OpenCV, scikit-learn, LangChain, Ray
Languages & Dev Tools: Python, C++, SQL, Java, Git, Linux, CUDA, React, Node.js
Infrastructure: Docker, Kubernetes, AWS (ECS/EKS), MLflow, Terraform, Spark

Education

The University of Texas at Dallas Aug 2023 - May 2025
M.S. Computer Science; Intelligent Systems specialization *GPA: **4.0/4.0***

Indian Institute of Technology Madras Jul 2019 - May 2023
B.Tech. Mechanical Engineering; Minor in AI *CGPA: **8.41/10.0***

Experience

Retrieval-Augmented Generation Systems Research May 2024 - Present
Research Assistant - Natural Language Processing Lab (Prof. Vincent Ng) *The University of Texas at Dallas*

- **Novel training**: Developed end-to-end fine-tuning framework for decoder-only LLMs in RAG settings, enabling joint optimization of retriever and generator components.
- **Knowledge Organization**: Developed information-theoretic scoring mechanisms to strategically select documents for the datastore, approximating the intractable optimization of jointly determining the optimal LLM parameters and datastore contents. This approach achieved a 30% improvement in accuracy per unit of memory usage.
- **Open Source**: Released ContextFlow library implementing the core mathematical formalism for RAG optimization.

Multimodal Learning Research Aug 2024 - Present
Research Assistant - Computer Vision and Multimodal Computing Lab *The University of Texas at Dallas*

- **Architecture Design**: Created novel cross-attention mechanism aligning temporal audio features with spatial embeddings for active speaker detection.
- **Annotation**: Built automated annotation pipeline for face tracking and python-based tool to label speaker activity.

Single Image Editing using Diffusion Models Research Jan 2024 - May 2024
Student Researcher - Computer Vision and Multimodal Computing Lab *The University of Texas at Dallas*

- **Two-Stage Fine-Tuning**: Developed a setup to learn unique identifier tokens for a specified subject and background from single images each. Utilized those identifiers in prompts to generate images integrating the subject with the given background enabling generative image editing.

Projects

LLM Systems
Technologies: PyTorch, HuggingFace, FAISS, Ray, AWS

- **RAG Framework**: Developed first-of-its-kind probability marginalization for RAG with decoder-only LLMs; open-sourced as ContextFlow with comprehensive documentation and HuggingFace integration.
- **Data Pipeline**: Built intelligent document chunking system with configurable overlaps and dynamic index updates, deployed on AWS ECS with auto-scaling.
- **Retrieval System**: Engineered hybrid BM25/dense retrieval system with custom dataloaders preserving document boundaries; better retains long-range context while achieving 40% faster retrieval.

AI Infrastructure Systems
Technologies: Kubernetes, Docker, Terraform, MLflow

- **Model Training**: Designed distributed RL training infrastructure using Ray and MLflow; reduced training time by 35% through parallel environments.
- **Deployment Pipeline**: Developed CI/CD pipeline for ML models with Docker and Kubernetes; achieved 95% uptime through auto-scaling.