KHUSHANG ZAVERI

kzaveri1@jh.edu | +1 (443) 563-4154 | Linkedin | Github | Portfolio

EDUCATION

JOHNS HOPKINS UNIVERSITY | Master of Science in Engineering in Data Science | Expected: May 2026 | CGPA: 3.73/4
Relevant courses: NLP: Self-supervised Models, Machine Translation, Data Mining, Information Retrieval & Web Agents
MAHINDRA UNIVERSITY | Bachelor of Technology in Computation & Mathematics | August 2020- June 2024 | CGPA: 3.72/4
Relevant courses: Deep Learning, Mathematics in Machine Learning, Design & Analysis of Algorithms, Probability and Statistics

SKILLS

- **Programming Languages**: Python, SQL, R, C, HTML, MATLAB, Bash, C++
- **Data Handling**: NumPy, Pandas, OS, Scikit-learn, SciPy, NLTK, Matplotlib, Seaborn, BeautifulSoup, Requests, ggplot2
- **Model Building and Training**: PyTorch, HuggingFace Transformers, Diffusers, FAISS, MONAI, scikit-learn, TensorFlow
- **NLP & LLM Ecosystem**: LangChain, Ollama, Transformers, spaCy, NLTK, ChromaDB, LangGraph, Hydra, Monai, LLaVA
- **Competencies:** Generative AI, Natural Language Processing, Multimodal AI, Deep Learning, Machine Learning
- **Software:** Git, Docker, Jira, LIWC, Apache Hadoop, Apache Hive, Apache Spark, Microsoft Excel, Git, AWS SageMaker

EXPERIENCE

Graduate Research Intern | Najim Dehak's Lab, CLSP, Johns Hopkins University | August 2025 - Present
- Developed first vocal tic detection system using low latency **dataset pipeline with VAD and speaker diarization**
- Trained **temporal CNN-BiLSTM, TDNN models** on WavLM embeddings **for frame-level tic detection**; achieved 0.73 F1, 0.83 AUC
- Designed baseline experiments using SVM, MMD, cosine, LDA to analyze tic-related transitions across intra-, inter-patient splits

AI Research Intern | St. Jude Children's Research Hospital | May 2025– Present
- Designed experiments to **analyze behavior, generation patterns of stable diffusion models in medical imaging settings**
- Built **DDIM-based inversion pipeline using VAE, Autoencoder, U-Net for latent-to-image reconstruction with spatial denoising**
- Implemented patient verification algorithms using **Siamese nets, logistic thresholds, FAISS hard-negative mining**
- Developed retrieval, verification pipelines to benchmark our contrastive copy-detection model, **achieved SOTA accuracy**
- **Trained DenseNet-121, CXR model for multi-label chest X-ray classification**; reported AUROC, AUPRC, F1 across 14 diseases
- Engineered dataset pipelines with Hydra, PyTorch, MONAI, DICOM preprocessing to ensure reproducibility

AI Intern | **Veltris** | January 2024 - July 2024
- **Engineered Biomedical NER model**, fine-tuned BERT and RoBERTa, achieved **80% accuracy**, enhanced biomedical research
- Leveraged PyTorch to build CNN model from scratch to classify images into streets, forests, and mountains with **90% accuracy**
- Gained expertise in Knowledge Graph, PEFT, RAG, Agentic AI, Azure through training program

AI Intern | **KPMG** | June 2023 - August 2023
- **Devised application with Language Models** such as Flan-T5, GEC-T5_small, Falcon-7B-Instruct
- Enabled auditors to get started with relevant questions, developed the GEC model, improved grammatical accuracy

Research Intern| University of Pennsylvania | Finding Social Norms | January 2023 - January 2024
- **Led creation of largest cross-cultural dataset** from 5,435 movie transcripts to identify harmful social norms
- Engineered an automated Python pipeline (BeautifulSoup, Requests, OS, SQL) to extract, preprocess massive unstructured data
- Applied advanced NLP techniques (LDA, LIWC, agglomerative clustering) to uncover 10,000 unique social norms

KEY PROJECTS

RAG Multimodal Research Paper Retrieval & Summarization Pipeline | Faiss, Ollama, LLaVA, LangChain | **Github.com/RAG**
- **Built multimodal RAG pipeline**, provide state-of-the-art NLP approach for any problem statement, used research paper corpus
- Used relevant papers, LLaVA to extract information from both text & images, enhanced quality of suggested NLP solution
- **Validated retrieval using masked-query filling, BARTScore, and LLM-as-a-judge; improved results via hypothetical querying**

Medical Audio and SOAP Note Generation | OpenAI Whisper, pydub, librosa, blaze999/Medical-NER | **Github.com/SOAP**
- **Implemented Whisper-based pipeline for audio denoising**, transcription, classification; preprocessed using pydub, Librosa
- Fine-tuned Flan-T5 for SOAP note section detection, integrated medical NER to extract clinical entities from transcriptions

PUBLICATION & AWARDS

- **NAACL 2025 Main Conference: "Social Norms in Cinema: A Cross-Cultural Analysis of Shame, Pride, and Prejudice"**
- **COLM 2025**: Paper accepted at MELT Workshop on social norms in cinema
- **IC2S2 2024:** Presented abstract of research paper at conference
- **Received two merit scholarships** for academic year 2021-2022 and 2022-2023 for academic excellence