Vishal Keswani

vkeswani | Webpage | +91 8741-867-137 | keswanivishal1997@gmail.com | LinkedIn | Scholar

WORK EXPERIENCE

VMOCK INDIA PVT LMT | Data Scientist (Jul21-Present)

Multimodal chatbot for customised interview using Agentic AI and RAG

- Created a chatbot that takes in user query, figures out the requirement, & guides to the relevant interview feature (tool)
- Prompt engineered **LLM**s for autonomous tool selection, implemented **RAG pipeline** to ingest product/tool description
- Experimented with **phi-3** and **gpt-3.5** using **Langchain**, **ChromaDB**, **FastEmbed**, **Azure OpenAI** and **streamlit**
- Added speech-to-text via **whisper** and enabled video response with **SadTalker** (lip-sync) & GFPGAN (face restoration)

Automatic Speech Recognition for interview transcription

- Streamlined the 3 step ASR pipeline, removing third party dependency and saving upto **$2k per month** variable costs
- Voice Activity Detection: compared low-latency vad models, used **silero-vad** to find no speech regions for audio splitting
- Audio Transcription: developed in-house **speech-to-text** capability with **Whisper**, testing across sizes and frameworks
- Phoneme Alignment: used **wav2vec2** for generating time-stamps of text over audio for subtitling and targeted feedback
- Reduced word error rate by **51%** on technical transcripts and **29%** overall, reduced latency by **20%** using onnx format

LLM-based Smart Composition for Resume Builder

- Designed a predictive text feature for faster & convenient bullet completion, trained baseline **gpt-neo**-125m from scratch
- Performed parameter efficient fine-tuning of gpt-neo-1.3b with **LoRA** & **QLoRA** on a single 16GB Nvidia Tesla T4 gpu
- Achieved a perplexity reduction of **60%** using rank 8 adapter for all linear layers, improved latency by **12%** using 4-bits

Transformer-based Content Feedback for Technical Transcripts

- Developed and deployed sentence and entity level language models for customized feedback on Elevator Pitch transcripts
- Experimented with transformer (**BERT**, DistilBERT) and RNN (**Bi-LSTM**, LSTM, QRNN with GloVE and Word2vec)
- Achieved macro-F1 score of **82%** on sentence classification and **75%** on NER, further improved by post-processing logics

Question recommendation system for Smart Interviews

- Developed an algorithm for generating interview set based on JD-resume, performed candidate **retrieval** via **opensearch**
- Perfromed candidate **ranking** using modified ES score, affinity propagation clustering, **MPnet** based similarity & logics
- Scaled questions database by 12x, automated **CRUD** operations in MySQL DB & ES index, reduced randomness by **8x**
- Implemented GenAI pipeline using **gpt-4o-mini**, **requests** & **langfuse**, achieved 6x latency gain with in-house algorithm

PUBLICATIONS

Formulating Sentence Ordering as the Asymmetric Travelling Salesman Problem

14th International Conference on Natural Language Generation, INLG 2021

- Classified Sentence-Pairs with BERT, used probabilities as distance input for aymmetric TSP (exact and heuristic)
- Predicted orders overtook baseline by upto **20%** in Perfect Match, **11%** in Kendall Tau, **6%** in Position-wise Accuracy

Hypernym Detection in the Financial Domain via Context-Free and Contextualized Word Embeddings

FinNLP-2020, 2^{nd} *International Workshop on Financial Technology and NLP, IJCAI-PRICAI 2020*

- Used **Word2vec** word-embeddings trained from scratch and pre-trained **BERT** word-embeddings with simple classifiers
- Word2vec with **Naïve Bayes** and BERT with **Logistic Regression** gave best test accuracy of **88%** and mean rank 1.2

Unimodal and Bimodal Sentiment Analysis of Internet Memes

SemEval-2020, 14^{th} *International Workshop on Semantic Evaluation,COLING-2020*

- Implemented Naïve Bayes (text), Combined **CNN** (image) and **Feed-Forward Neural Network** (text) using **SVM**
- Fine-tuned BERT and Multimodal Bitransformer, text-only FFNN with Word2vec gave best macro-F1 **63%**>baseline

INTERNSHIP PROJECTS

MURATA VIOS | Computer Vision Intern (May20-Aug'20)

Computer Vision on Edge devices for the visually impaired using TensorFlow Lite

- Customized **Object detection** (COCO dataset) and **Face Detection** (Open Images dataset) for Raspberry Pi 4
- Trained quantized **MobileNet V2** on LFW dataset and user faces, obtained **Face Recognition** accuracy upto **93%**

NOKIA SOLUTIONS AND NETWORKS | Data Science Intern (May19-Jul19)

Auto-Suggesting inquiry questions to Care Engineer based on client case

- Extracted e-mail bodies (doc to csv), followed by tokenization, removal of stop words, stemming, lower casing in **nltk**
- Clustered client queries using **k-means** (using **tf-idf** scores and **cosine-similarity**), reported 3 most similar questions

RELEVANT SKILLS

Programming: Python, R, C, C++, HTML, CSS, JavaScript | **Others:** SQL, Bash, Linux, Windows, MS Office, Latex
Python: HuggingFace, PyTorch, TensorFlow, Keras, Spacy, MLFlow, Gradio, NumPy, Pandas, Flask, Onnx, DeepSpeed
Tools: Git, Docker, DVC, Devspace, Kubernetes, Celery, Redis, New relic, AWS S3, Sagemaker, Copilot, OpenSearch

ACHIEVEMENTS

- Recognized as **STAR employee** of Capabilities team at VMock for outstanding contribution to Interviews product
- **Ranked 1^{st}** in FinSim task (FinNLP'20) and Memotion Analysis task (SemEval'20) international challenges in NLP
- Secured **AIR 538** in JEE Mains (City Rank 1) and AIR 2191 in JEE Advanced (City Rank 2) among 1.5 million

EDUCATION

Indian Institute of Technology, Kanpur	BS-MS - Economic Sciences Minor - Machine Learning	10/10 (PG) 8.6/10 (UG)	Department Rank 1 in MS batch Graduated with Distinction	2021
Kendriya Vidyalaya 1, Ajmer	CBSE - Senior Secondary	96.4/100	School Rank 1 and Best Student	2015
East Point School, Ajmer	RBSE - Secondary	94.2/100	School Rank 1 and House Captain	2013