MIHIR BANSAL

Email: mihirban@andrew.cmu.edu Contact: +1 4125124820 Webpage: https://mihir86.github.io

EDUCATION

Carnegie Mellon University — Pittsburgh, PA, USA
Master of Science, Data Science | ***GPA:*** *3.92/4.00* — *Aug 2023 - Dec 2024*
Courses: Machine Learning, Advanced NLP (Teaching Assistant), Search Engines, Distributed Systems, Cloud Computing

Birla Institute of Technology and Science, Pilani — Hyderabad, India
Bachelor of Science, Computer Science | ***GPA:*** *9.37/10.00* — *Aug 2018 - Jul 2022*
Courses: Data Structures & Algorithms, Object Oriented Programming, Database Systems, Operating Systems, Computer Networks

TECHNICAL SKILLS

Programming Languages C++, Python, C, Java, C#, R, SQL, MATLAB, Shell Scripting, Go, Scala, JavaScript
Libraries & Frameworks PyTorch, Tensorflow, Keras, NLTK, OpenAI, Transformers, LangChain, Hive, Elasticsearch, Pig, AWS, GCP, Azure, Spark, Hadoop, Kafka, PostgreSQL, Kubernetes, Docker, Tableau, CUDA, A/B Testing

WORK EXPERIENCE

Microsoft — Hyderabad, India
Machine Learning Engineer — *Jul 2022 - Aug 2023*
- Worked in the Microsoft Search team to improve the XGBoost Machine Learning ranking model for acronym search by adding a multigram matching feature with top user AI graph nodes. Achieved an offline precision (MAP) of 92% and improved the online Click-Through Rate by 24%, monitored via A/B experimentation.
- Developed an end-to-end LLM-based summarizer with webpage summary and FAQs for Bing enterprise web search results by using GPT-3 with a scalable online query serving microservice and an offline time-based indexing service using OpenAI and Kubernetes.
- Implemented the data pipeline of a profanity-filtering backend API for Outlook calendar search by using Apache Spark for processing large-scale search logs from a sharded Cassandra database, reducing the search index loading time by 17%.

Seagate Technology — Fremont, CA, USA
Machine Learning Engineer Intern — *May 2024 - Aug 2024*
- Implemented an auto-merging retrieval approach with PostgreSQL logging in the LLM-based Retrieval Augmented Generation pipeline for a chatbot application with Azure AI Search, improving context relevance by 37% and answer relevance by 32%.

JP Morgan Chase & Co. — Mumbai, India
Quantitative Research Analyst Intern — *Jan 2022 - Jun 2022*
- Developed a Time Series Rating Migration model to predict credit ratings with macroeconomic scenarios, with Spark-based ETL to process historical ratings data using SQL, and using LSTM and ARIMA models for prediction, achieving an RMSE of 0.19.

University of Hamburg — Hamburg, Germany
Research Intern — *Aug 2021 - Dec 2021*
- Collaborated with the NLP group on fine-tuning BERT language model with knowledge graph from Wikidata for improving semantics-based question answering using NLTK and Transformers, achieving an improvement of 13% in the GLUE score.

Microsoft — Hyderabad, India
Software Engineer Intern — *May 2021 - Jul 2021*
- Improved the coverage of Outlook Calendar Search by 37% by performing optimized REST API calls and caching to index grouped user entities for person-based Natural Language queries using Elasticsearch and Redis, reducing the overall response time by 60.8%.

Central Electronics Engineering Research Institute — Pilani, India
Research Intern — *May 2020 - Jul 2020*
- Implemented a Genetic Algorithm for feature subset selection using OpenFace and classifying human emotions from facial images with Neural Networks and Random Forest ML classifier models using Tensorflow and PyTorch, achieving an F1 score of 0.91.

PUBLICATIONS

1. Meriem Beloucif, **Mihir Bansal**, Chris Biemann. *Using Wikidata for Enhancing Compositionality in Pretrained Language Models.* Recent Advances in Natural Language Processing. 2023. ACL Anthology. [Paper]

2. B.S.A.S. Rajita, **Mihir Bansal**, Bipin Sai Narwa, Subhrakanta Panda. *Cuckoo search in threshold optimization for better event detection in social networks.* Social Network Analysis and Mining. 2022. Springer. [Paper]

PROJECTS

Improving Search Engines with LLM Embeddings | *Research Assistant, CMU* — *Aug 2023 - Present*
- Working with Prof.Jamie Callan on improving Neural Information Retrieval models by using FAISS and Hypothetical Document Embeddings from GPT-3.5, improving the MAP by 17%, NDCG by 29% and Recall by 18% on the MSMARCO dataset.

Evaluation of Bias in LLMs with Retrieval Augmented Generation | *Research Assistant, CMU* *Jan 2024 - Present*

- Working with Prof.Graham Neubig on developing a framework for detecting bias in Mistral-7b and LLaMA-7b LLMs using HuggingFace for Question Answering using Retrieval Augmented Generation on the American Election Study 2020 dataset.

Twitter Recommendations System with Apache Spark | *CMU* *Jan 2024 - May 2024*

- Developed a highly scalable end-to-end Twitter topics recommendations service by ranking user topics for 10M users by their PageRank scores in a large Twitter social network graph with Azure Databricks, Scala and Spark SQL.

Cab Booking Application with Apache Kafka | *CMU* *Jan 2024 - May 2024*

- Implemented a real-time distributed stream processing cab driver-user matching backend service with an ads recommendations and pricing microservice from user clicks by using Apache Kafka, Samza and Amazon Elastic MapReduce.

Adapting Multilingual Visual LLMs for Code-mixed Visual QA | *CMU* *Aug 2023 - Dec 2023*

- Finetuned the mBLIP LLM model for low-resource Hinglish code-mixed VQA by using LoRA and QLoRA finetuning, achieving an improvement of 34% in accuracy on the MS-COCO dataset as compared to the baseline model.