Ananya Singha:

Email: ananyasingha2000@gmail.com
Linkedin: linkedin.com/in/ananyasingha
Github: github.com/ananyas168
ananyas168.github.io | Google Scholar

Education

- **INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH, BHOPAL** Bhopal, India
  BS - EECS(Electrical Engineering and Computer Science) *August 2018 - June 2022*
  🏆 Batch Topper, Proficiency Gold Medal Awardee, GPA: 9.34
  Courses: *Data Structures, Analysis Of Algorithms, Artificial Intelligence, Machine Learning,Natural Language Processing, Digital Image Processing, Cryptography, Quantum Computer Science, Error Decoding*

Skills Summary

- **Languages**: Python, C, C++, Matlab, R
- **Frameworks**: Scikit, NLTK, SpaCy, TensorFlow, Keras, Qiskit, Dgl, Pytorch, Blazer, MangoDB, FlaskAPI, Streamlit
- **Tools**: Docker, GIT, PostgreSQL, MySQL, SQLite, Oracle
- **Soft Skills**: Leadership, Event Management, Writing, Public Speaking, Time Management

Experience

- **Associate Researcher at Microsoft Research** India
  Advisors*: Manager: Dr. Vu Le, PROSE Team* *May 2024 - ongoing*
  - **Topic - NLP, AI4Code, Language Agents, Agentic Framework**: Working on AI4Code specifically, Table2code, Program Repair and language agent framework for prompt optimization and benchmark data curation including fine-tuning transformer model for different tasks.
    * Developed MetaReflecion workflow for prompt optimization by generating optimal set of rules. The workflow involves actor(task) agent, reflection agent and MetaReflection agent which analyses the actor response log and draws learnings/rules for the prompt. This work resulted in a patent and is published at EMNLP Mains 2024.
    * Working on creating an agentic framework for dataset curation for Program Repair where the seed data points are curated by leveraging the agentic framework on Community Forums posts. Further, there are adversarial actor agents and translation agents which generate more synthetic augmentations using the seed samples, which help generate fine-tuning data for Program Repair.
    * Developed data-analytics pipeline, where we explored ways to generated semantically aligned NL and Code pairs given the table using LLM under the hood. Our work is published at LLM4Code Workshop @ ICSE 2024.
  - **Tools**: Pytorch, AutoGen, Python etc

- **Research Fellow at Microsoft Research** India
  Advisors*: Dr. Austin Henley and Dr. Chris Parnin, PROSE Team* *August 2022 - May 2024*
  - **Topic - NLP, AI4Code and Human Computer Interaction**: Leveraging Deep Learning, NLP for tasks like text to code and building tools to enhance end user experience.
    * Building testing and evaluation framework for LLM based techniques. Developed a self-supervised technique to generate test cases to evaluate LLM understanding over a task.
    * Working on training transformer based models (LLM) for Natural Language to code translation for languages like Power Query and Python in an efficient way. Exploring fine-tuning and building better transformer models for the task.
    * Analyzing formula churns using different statistical models like Survival Analysis and ML algorithms to identify difficult formula features interrupting the formula authoring journey.
    * Exploring use of GPT-3 and codex in code and semantic label generation for unstructured data, like review data, spreadsheet data, etc.
    * Developed end to end pipeline for extracting user concern/suggestion from Microsoft community forums/verbatim and deployed it as a Web Application. Also it supports extraction of concerns from shared screenshots using OCR algorithms. Conducted A/B testing to understand impact. Along with user studies to gather pain-point users used to face with manual verbatim analysis.
  - **Tools**: C#, python, Tensorflow, Pytorch, PowerQueryM, SQL, Pandas, Streamlit, Blazer, etc

- **Research Intern at PSL-ENS Paris** France
  Advisor*: Dr. Alejandrina Cristia, Cognitive Lab* *May 2022 - July 2022*
  - **Topic - Speech Data Processing and Classification**: Worked on Speech Recognition
    * Explored different techniques for preprocessing of Infant (3-15 months) Speech Data. Developed my own pipeline for the preprocessing task which included cleaning , standardizing, structuring of audio file corpus.
    * Used VTC2 for speech classification into Child, Male , Female, Other category. And worked on enhancing performance for multilingual data by pre-trainig VTC2 Model by using samples from different corpus.
  - **Tools**: Pytorch, Pyannote, Pandas , etc

- **Research Intern at IIT Delhi— IISER Bhopal (Collaboration)** India
  Advisor*: Dr.Vaibhav Kumar* *Apr 2022 - May 2022*

- **Topic - 3D Lidar Cloud point**: Worked on development of Damage detection model for historical Building images.
  - Working on labelling of 3D Lidar cloud Data points.
  - Developing 3D Lidar cloud Data classification and segmentation model based on GNN architectures which are best suitable for damage detection in historical building.
- **Tools**: Pytorch, CloudCompare, OpenCV, etc

- **Research Intern at University of Cambridge** UK
  ***Advisor**: Dr. Peter Murray-Rust and Dr.Gitanjali Yadav* *Jan 2022 - Mar 2022*
  - **Topic - Computer Vision and Image Analysis**: Worked on development of Image Analysis tool and statistical modeling of pathway images. Contributed to open source text mining software.
    - Developed statistical model to do analysis on annotated data of pathway images. Used sklearn and scipy statistical tools.
    - Contributed in building robust abbreviation extraction and keywords extraction toolkit using NLP modules and developed Knowledge graphs from the data.
    - Using Tesseract to extract text from pathway images and developing an NLP model to classify images into categories based on extracted text, further to extract Chemical synopsis given Pathway Diagram.
    - I was part of a reputed research project TIGR2ESS an collaboration of University of Cambridge with India to promote research and science.
  - **Tools**: Pytorch, NLP, OpenCV, etc

- **Data Science Intern** New Delhi, India
  ***Company**: AndWeMet* *Nov 2021 - Dec 2021*
  - **Topic - NLP and Computer Vision**: Developed recommendation system
    - Used NLP to developed a robust recommendation system for dating based on profile information.
    - Modeled a face verification system to give feedback to new user and verify uploaded profile pictures by using OpenCV and a pretained Inception-v3 network.
  - **Tools**: Tensorflow, Pytorch, Keras, OpenCV, etc

- **Research Intern at IISC Bangalore** Bangalore, India
  ***Advisor**: Dr. Sundeep Prabhaker Chepuri* *May 2021 - July 2021*
  - **Topic - Representation Learning and Graph Neural Network**: Worked on developing novel representation learning method using Graph Neural Network.
    - Used Graph Neural Network for link prediction and for drug discovery. Used Pytorch geometric and dgl library for generating the data and modeling GNN networks. Developed a Novel approach to solve the entity alignment problem among KGs with an accuracy of 98% on DBPK15 Dataset.
  - **Tools**: Pytorch Geometric, Pytorch, NetworkX, dgl, etc.

- **Research Intern at UTS Sydney** Australia
  ***Advisor**: Dr. Sunny Verma* *Jan 2021 - May 2022*
  - **Topic - Fairness in Deep Learning/ Computer Vision**: Worked on the problem statement: "Identifying and Mitigating the Cause of Bias in Face Recognition system".
    - Proposed an experimental setup to analyze the importance of facial attributes (ex. eyeglasses, earrings, etc.) towards gender classification. Generated training and testing data using OpenCV masking tools. Performed several statistical analysis to detect dataset shift and the results suggested that the most biased attribute is eyeglasses.
  - **Tools**: Tensorflow, Pytorch, Keras, OpenCV, etc

- **Summer Intern at IISER Bhopal** Bhopal, India
  ***Advisor**: Dr. Nirmal Ganguli* *May2019 - July 2019*
  - **Topic - Mathematical approach of machine learning and AI**: Applied Deep Convolution Neural Networks to classify MNIST dataset.Solved complex equations using Deep neural networks along with identification of kinds of perturbation in the given set of Schrodinger equations
  - **Tools**: Tensorflow, Keras, Scikit, etc

Projects

- **Alzheimer Disease Prediction using Graph Neural Network. (Graph Representation Learning, DL)** :
  Advisor: Dr. Sujoy Bhore (BS Project Advisor)
  Developed AD's prediction model using Heirarichal Graph neural network and customized loss functions to overcome data imbalance. Used FSL Software and Pytorch geometric to generate brain network using ADNI fmri Dataset .
  Tools: Pytorch etc (May '22)
- **Autonomous driving. (Robotics and Reinforcement Learning)** :
  Advisor: Dr. Sujit Pedda Baliyarasimhuni
  Developing an autonomous driving system which uses Reinforcement Learning component for intelligent decision making.
  Tools: Pytorch etc (march '22)

- **Financial sentiment analysis using NLP. (Machine Learning, NLP)**:
  Advisor: Dr. Tanmay Basu
  To extract sentiments I used different data preprocessing and cleaning techniques to get the structured data and used it on Bert sentence encoder to compute embedding and capture sentiment and performed classification using logistic regression, RF, SVM.
  Tools: Python,NLP toolkit, Wornet etc (October '21)
- **Unsupervised Opinion Mining. (Machine Learning, NLP)**:
  Used POS tagger to extract nouns, adjectives, and adverbs as they describe the qualities better. Added extra vocabulary for similar words using Wordnet. Used a Customized clustering algorithm to cluster the major opinions/Qualities and to give the desired outcome. The customized algorithm used embedding similarity for each word and the semantic synonyms to give better results than K - means clustering.
  Tools: Python,NLP toolkit, Wornet etc (November '21)
- **Analysis of covid data using Machine Learning algorithms. (Machine Learning, Statistical Learning)**: **Advisor**: Dr. Parthiban Srinivasan
  Used Chembl data to detect covid protein based on bioactivity value using RDkit tool
  Tools: Python,RDkit, scikit, TensorFlow (October '20)

Publications

- **METAREFLECTION: Learning Instructions for Language Agents using Past Reflections**:
  Priyanshu Gupta*, Shashank Kirtania*, Ananya Singha*, Sumit Gulwani, Arjun Radhakrishna, Sherry Shi, Gustavo Soares
  EMNLP Mains 2024, * equal contribution
- **Semantically Aligned Question and Code Generation for Automated Insight Generation**:
  Ananya Singha, Bhavya Chopra, Anirudh Khatry, Sumit Gulwani, Austin Henley, Vu Le, Chris Parnin, Mukul Singh, Gust Verbruggen
  LLM4Code Workshop @ ICSE 2024. (🏆 Best Paper)
- **Tabular Representation, Noisy Operators, and Impacts on Table Structure Understanding Tasks in LLMs**:
  Ananya Singha, José Cambronero, Sumit Gulwani, Vu Le, Chris Parnin
  TRL Workshop @ NeurIPS 2023. (🏆 Runner-Up Best Paper, Spotlight Paper)
- **TSTR: Target Similarity Tuning Meets the Real World**:
  Anirudh Khatry, Sumit Gulwani, Priyanshu Gupta, Vu Le, Ananya Singha, Mukul Singh, Gust Verbruggen
  EMNLP Findings 2023
- **Conversational Challenges in AI-Powered Data Science: Obstacles, Needs, and Design Opportunities**:
  Bhavya Chopra, Ananya Singha, Anna Fariha, Sumit Gulwani, Chris Parnin, Ashish Tiwari, Austin Z Henley
  CHI 2023 (Submitted)
- **Deep Learning Applications in Medical Image Analysis**:
  Ananya Singha, Rini Smita Thakur, Tushar Patel
  Book Chapter: Biomedical Data Mining for Information Retrieval, Published in August'21 by Wiley Publications.

Honors and Awards

- Proficiency Medal Winner, Department Topper Batch'22
  Ranked 1st in EECS Department
- CERN Technical Studentship 2022 in Geneva, Switzerland
  Selected among top 90 candidates globally (declined)
- Indian Academy of Science Summer Fellowship - May, 2021
  Student Researcher under FAST-SF
- Runner's Up at Inter IISER ML Hackathon - August, 2019

Elective, Summer School and MOOCs

- **Electives**:
  Digital Image Processing, NLP, Reinforcement Learning, Quantum Machine Learning, Fairness in Machine learning
- **Summer School:**
  MLSS 2021, Taipei
- **MOOCs**:
  Artifical Intelligence Data Fairness and Bias, Machine Learning courses on Coursera, Graph Machine Learning by Stanford Courses.