Avik Dutta

Research Fellow, PROSE Team, Microsoft, Bengaluru, India
avikdutta772000@gmail.com — +91 8521946682 — Linkedin — Google Scholar — avikdutta772000.github.io

RESEARCH INTERESTS

Machine Learning, Natural Language Processing, Information Retrieval, AI for Code, Language Model Agents

EDUCATION

Indian Institute of Technology, Kharagpur, India   July 2019 — May 2023
Bachelors in Electronics and Electrical Communication Engineering   Cumulative GPA: 9.40/10.00
Minor in Computer Science and Engineering   Department Rank: 8/84
Micro Specialization in Artificial Intelligence and Applications

Narbheram Hansraj English School, Jamshedpur, India   April 2017 — March 2019
Indian School Certificate Examinations (Computer Science)   Percentage: 97.00/100

Carmel Junior College, Jamshedpur, India   April 2015 — March 2017
Indian Certificate of Secondary Education (Computer Science)   Percentage: 97.00/100

PUBLICATIONS   [C=Conference, P=Preprint]

C.1 RAR: Retrieval-augmented retrieval for code generation in low resource languages 📄
    Avik Dutta, Mukul Singh, Gust Verbruggen, Sumit Gulwani, Vu Le; *EMNLP 2024 Main (Long Paper)*
C.2 DistALANER: Distantly Supervised Active Learning Augmented Named Entity Recognition in the Open Source Software Ecosystem 📄
    Somnath Banerjee, Avik Dutta, Aaditya Agrawal, Rima Hazra, Animesh Mukherjee; *ECML-PKDD 2024 (ADS Track)*
C.3 Context Matters: Pushing the Boundaries of Open-Ended Answer Generation with Graph-Structured Knowledge Context 📄
    Somnath Banerjee, Amruit Sahoo, Sayan Layek, Avik Dutta, Rima Hazra, Animesh Mukherjee; *EMNLP 2024 Industry Track*
P.1 Redefining Developer Assistance: Through Large Language Models in Software Ecosystem 📄
    Somnath Banerjee, Avik Dutta, Sayan Layek, Amruit Sahoo, Sam Conrad Joyce, Rima Hazra; *Preprint*

Awards & Achievements

- Won **Gold Medal** as a contingent member of 159 students at Inter IIT Tech Meet 11.0 hosted by IIT Kanpur.
- Won **Silver Medal** in Robosoccer Fresher's Challenge conducted by Technology Robotix Society, IIT Kharagpur.
- Achieved **10.00 CGPA** in the first semester, ranking among the top 23 out of 1,600 undergraduates admitted.
- Obtained an All India Rank of **881**, among 245,000 other candidates (top 0.36%) in JEE Advanced 2019.
- Felicitated with **Sir JC Bose** award for being top scorer (among 250 students) at Narbheram Hansraj English School.
- Received **All-Rounder of the Batch** award and scored **100%** in Maths, CSE and Science in Carmel Junior College.

EXPERIENCE

Microsoft   Bengaluru, India (*remote*)
Research Fellow @PROSE team   November 2023 — Present

- Deployed **conditional formatting formulas** along rows and tables in **Excel Copilot**, which rolled out into production.
- Refactored the Copilot codebase during model migration from GPT4PPO to GPT4Turbo and GPT4o. Achieved $< 1\%$ regressions across all features and T1 languages. Successfully rolled out to production, impacting Excel users globally.
- Designed a framework for context retrieval from spreadsheets to aid Copilot to answer ambiguous and multi-turn queries.

Complex Networks Research Group (CNeRG)   IIT Kharagpur, India
Undergraduate Student Researcher   July 2022 — October 2023

- Devised a **distantly supervised annotation** framework and trained on **CRF** models to tag unseen software entities.
- Implemented a **GNN**-based QA retrieval strategy for grounding LLMs for community question answering tasks.
- Pre-trained BERT-based models using **MLM** technique on software texts extracted from StackOverflow, Launchpad etc. to improve downstream applications on NER and RE. Yielded an average perplexity of **18.35** on roberta-base.
- **Fine-tuned Llama** on software texts to demonstrate superiority over other domain-specific models (Vicuna and Alpaca).

Piramal Finance   Bengaluru, India
Graduate Engineering Trainee   July 2023 — October 2023

- Applied data-driven methods as a business analyst to understand customer attrition and suggested ways to reduce it.
- Reported portfolio profits through interactive charts for different cohorts in the monthly reports at TownHall meetings.
- Used **OCR** to extract text from cheque images. Designed a classifier using empirical heuristics achieving **87%** accuracy.
- Used **R**etrieval-based-**V**oice-**C**onversion and **T**ext-**t**o-**S**peech tools to produce automated calling service for customers.

Computer Graphics Society   IIT Kharagpur, India
Governor & Head of Game Development   February 2020 — April 2023

- Used **Unity Game Engine** along with **C#** language to develop Android games – Ricksy Run, Bubble Meow't.
- Organised and led workshops, talks and hands-on tutorials teaching game developement to students.
- Built an AR-based EduTech game for school students under Prof. **Nian Shing Chen** in **Yuntech University, Taiwan**.

PROJECTS

CoDABench: Agentic Benchmarking and Evaluation for Advanced Data Analysis   Microsoft
Excel AI [Confidential]   July 2024 — Present

- Developed a multi-agent framework for curating question-answer pairs from published articles for benchmark creation.
- Designed a user-proxy for automated and large-scale model evaluation simulating focused interaction similar to a human.
- Reported an evaluation framework to measure answer accuracy and conversation quality on multiple rounds of interaction.

Multi-Agent Framework for Advanced Data Analysis   Microsoft
Fix Hack Learn Hackathon [Slides]   May 2024

- Developed a chat assistant which uses multiple agents, *code_interpreter* and external function calls to solve complex data analytics tasks. Better than ChatGPT-ADA in terms of answer correctness, data wrangling and quality user-interaction.
- Defined a state-machine to streamline the order of interaction happening internally between other agents and reviewer.
- Implemented a File Management System that allows extending to other external tools which cannot access remote files.

Graph-based Policy Network Design for Psuedo-Relevance Feedback   IIT Kharagpur
Advisor: Prof. Plaban Kumar Bhowmick [Report]   February 2022 — November 2022

- Used **Pseudo-Relevance Feedback** for retrieving documents from corpus with functionalities of *Indri* search engine.
- Designed a **GNN**-based policy network architecture on which the **REINFORCE** algorithm was applied for training.
- Used a temporal gain of **M**ean **A**verage **P**recision of extracted documents as reward function for training the RL framework. Our design improved MAP by $\approx$**5.10%** over the neural-based RML baseline on **TREC678** dataset.

Explainable Bayesian Machine Learning   IIT Kharagpur
Advisor: Prof. Pabitra Mitra [Report]   July 2022 — November 2022

- Compared explain-ability of a CNN by subjecting it to adversarial attacks under deterministic and Bayesian Inferencing.
- Used **Variational Inference** to estimate the posterior distribution by assuming prior to have a spike-and-slab function.
- Designed the probabilistic models using tensorflow-probability and used **Lime** to obtain explainability masks of images.
- Demonstrated better explanation through Bayesian Inferencing, in terms of robustness, by studying the intersection, union and weighted average of masks sampled from a distribution rather than choosing the maximum-a-posteriori.

SELECTED COURSES

Computer Science: Algorithms-I^L, Computer Architecture & Operating Systems, Machine Learning^L, Deep Learning^L, Artificial Intelligence^L, Advanced Machine Learning, Big Data Processing, Machine Learning with Graphs*, Natural Language Processing*   (* indicates MOOC, ^{L} includes a lab component)
Mathematics: Vector Algebra, Differential Equations, Matrix Algebra, Probability and Stochastic Processes

TEACHING

Creative Tanima Academy Conducted online classes to teach game-development to 18 students in a 2 week crash course.
Inter-Hall Data Analytics 2023 Taught Machine Learning and mentored a team of 15 juniors during the competition.
Computer Graphics Society Organized 2 annual workshops and directed hands-on tutorials to teach game development.

TECHNICAL SKILLS

Programming: Python, C, C++, Java, MATLAB, R, SQL, C#, Typescript, Spark, LaTeX
Libraries: Numpy, Pandas, SkLearn, Matplotlib, Tensorflow, Pytorch, Tensorflow-Probability, Deep Graph Library
Software: Visual Studio, Indri (Lemur Projects), Unity Game Engine, Arduino, LT Spice, Power Bi, Adobe Illustrator