Avik Dutta

Research Fellow, Microsoft Research, Bengaluru, India
avikdutta772000@gmail.com — +91 8521946682 — Linkedin — Google Scholar — Website

RESEARCH INTERESTS

Natural Language Processing, Machine Learning, Information Retrieval, LLM for Code, AI for Software Engineering

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
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.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) [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) [paper]

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 [paper]

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 [paper]

EXPERIENCE

Microsoft India (R&D) Pvt. Ltd. — Bengaluru, India
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) — Kharagpur, India
Undergraduate Student Researcher — July 2022 — October 2023

- Designed 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 Capital & Housing Finance Ltd. — 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 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

Agentic Benchmarking and Interactive Evaluation of LMs for Advanced Data Analysis Microsoft Research
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 Research
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 ≈**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.

Automating NPC Behaviour in a simulated environment IIT Kharagpur
Term project for Artificial Intelligence course [Presentation] October 2021 — November 2021

- Implemented an A* heuristic search algorithm to decide the active state of a **N**on-**P**layable-**C**haracter in response to the player's movement in a game environment. Used player distance, orientation and barrier position for defining heuristics.
- Designed the Finite-State-Machine with arrows defining heuristics to attain a state that keeps the health value high.

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

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), Unity Game Engine, Arduino, LT Spice, Power Bi, Adobe Illustrator

VOLUNTEER WORK

- **The Stray Army Charitable Trust:** Tending to injured and stray dogs and support the trust with monthly donations.
- **Creative Tanima Academy:** Conducted classes to teach game-development to 18 students in a 2 week crash course.
- **National Cadet Corps:** Military training for 2 years and volunteered in cleanliness drives, blood donation camps, etc.