Avik Dutta

Research Fellow, Microsoft Research, Bengaluru, India
avikdutta772000@gmail.com — +91 8521946682 — www.linkedin.com/in/AvikDutta

RESEARCH INTERESTS

Natural Language Processing, Large Language Models, Program Synthesis, Graph Neural Networks, Reinforcement Learning

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

- **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.
  Accepted at ECML-PKDD 2024 (Long Paper). (https://arxiv.org/abs/2402.16159)

- **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.
  Preprint (https://arxiv.org/abs/2401.12671)

- **Redefining Developer Assistance: Through Large Language Models in Software Ecosystem** Somnath Banerjee, **Avik Dutta**, Sayan Layek, Amruit Sahoo, Sam Conrad Joyce, Rima Hazra.
  Preprint (https://arxiv.org/abs/2312.05626)

EXPERIENCE

Microsoft India (R&D) Pvt. Ltd. — Bengaluru, India
Research Fellow @PROSE team — November 2023 — Present

- Designed **RAR**, a novel retrieval framework that leverages grammar and examples in the prompt for low-resource code generation. Achieved **26.14%** improvement over baseline for grammar retrieval and **3.56%** for example retrieval.
- Deployed Formula based conditional formatting feature in Excel Copilot, which can be applied to rows and tables, in addition to columns. This feature addressed a common user request and received positive results from internal testers.
- Designed an evaluation framework to assess OfficeScripts and built an ExcelJsonParser tool for extracting relevant data from a workbook for context aware LLM tasks. These tools are proactively used by other members from Excel AI.
- Responsible for making changes to the product's code base during model migration for Commanding scenarios in Excel.

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

- 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.
- Designed a distantly supervised annotation framework and trained on CRF models to tag unseen software entities.
- Implemented a GNN-based QA retrieval strategy to provide context to LLMs for solving bugs posted in online platforms.

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. Devised strategies to improve portfolio-based profits on different cohorts on a monthly basis.
- Used OCR to read scanned cheque images, and classified them as either Balance/non-Balance Transfer based on heuristics applied on the extracted text. The model had a design accuracy of 67%.

- Used **R**etrieval-based-**V**oice-**C**onversion (RVC) and **T**ext-**t**o-**S**peech tools to produce automated call facilities, where the customized message typed in English was converted to the MD's voice and relayed as calls to the customers.
- Used various ML and time-series models to estimate future foreclosures and proposed necessary steps to prevent them.

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. The former has an overall rating of 3.8 in google play store with over 35k downloads.
- Organized an international level seminar and workshop event dedicated to teaching basics of game development.
- Built an AR-based EduTech game for school students under Prof. Nian Shing Chen in Yuntech University, Taiwan.

PROJECTS

Multi-Agent Advanced Data Analytics tool Microsoft Research
Fix Hack Learn Hackathon 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 the state-machine for agents, to channelize the order of interaction that should happen between the agents.
- Implemented a File Management System to sync the local data with the version in OpenAI server after every agent call.

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

Bachelor's Courses

- Algorithms-I (Theory+Lab)
- Computer Architecture & Operating Systems
- Probability and Stochastic Processes
- Machine Learning (Theory+Lab)
- Deep Learning (Theory+Lab)
- Artificial Intelligence (Theory+Lab)
- Advanced Machine Learning
- Big Data Processing

MOOC Courses

- Applied Data Science with Python
- Machine Learning with Graphs (CS224W)
- Natural Language Processing (CS224N)

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, LT Spice, Power Bi, Adobe Illustrator, DaVinci Resolve

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.