Anwoy Chatterjee

Email : anwoychatterjee@gmail.com

Google PhD Fellow, IIT Delhi

🌐: https://c-anwoy.github.io/
in: https://www.linkedin.com/in/anwoy-chatterjee/
: https://github.com/C-anwoy

Education

- **Indian Institute of Technology, Delhi** New Delhi, India
  Doctor of Philosophy *in* ***Electrical Engineering****; GPA: 10/10* *July 2023 - Present*
  - ○ **Areas of Specialization**: Natural Language Processing, Deep Learning, Large Language Models
  - ○ **Prospective Thesis**: *"Towards Robust Post-Training Adaptation of Language Models"*
  - ○ Supported by **Google PhD Fellowship**
  - ○ **Courses Credited**: Deep Learning for NLP, Cloud Computing
- **Indian Institute of Technology, Delhi** New Delhi, India
  Master of Technology *in* ***Machine Intelligence and Data Science****; GPA: 8.808/10* *July 2022 - June 2023*
  - ○ Transitioned to PhD
  - ○ **Ranked 1st in the program on the basis of GPA**
  - ○ **Courses Credited**: Artificial Intelligence, Machine Learning, Data Mining, Mathematical Foundations of AI, Computer Vision, Stochastic Control and Reinforcement Learning, AI for Earth Observation, Ethical Considerations in AI
- **Indian Institute of Technology (BHU), Varanasi** Varanasi, India
  Bachelor of Technology *in* ***Computer Science and Engineering****; GPA: 9.72/10* *July 2018 - June 2022*
  - ○ B.Tech Thesis: "Trajectory prediction of dynamic agents around an autonomous vehicle using GNNs"
  - ○ **Ranked 2nd (among 75 graduates) in the department on the basis of GPA**

Skills Summary

- **Programming Languages**: Python, C++, C, SQL, Unix scripting
- **Libraries & Frameworks**: PyTorch, HuggingFace, PyTorch Geometric, TensorFlow, Keras, Scikit-Learn, Numpy, Pandas
- **Tools**: Springboot, Git, MySQL

Publications

C: Conference, **J**: Journal, **P**: Preprint

[J1] **Anwoy Chatterjee**, H S V N S Kowndinya Renduchintala, Sumit Bhatia, Tanmoy Chakraborty,"*On the Effect of Instruction Tuning Loss on Generalization*", **Transactions of the Association for Computational Linguistics (TACL)**, arXiv:2507.07817, July 2025.

[C3] Eshaan Tanwar, **Anwoy Chatterjee**, Michael Saxon, Alon Albalak, William Yang Wang, Tanmoy Chakraborty "*Do You Know About My Nation? Investigating Multilingual Language Models' Cultural Literacy Through Factual Knowledge*", **EMNLP 2025**, August 2025.

[C2] **Anwoy Chatterjee**, H S V N S Kowndinya Renduchintala, Sumit Bhatia, Tanmoy Chakraborty,"*POSIX: A Prompt Sensitivity Index For Large Language Models*", **EMNLP 2024 (Findings)**, arXiv:2410.02185, September 2024.

[C1] **Anwoy Chatterjee**, Eshaan Tanwar, Subhabrata Dutta, Tanmoy Chakraborty,"*Language Models can Exploit Cross-Task In-context Learning for Data-Scarce Novel Tasks*", **ACL 2024**, arXiv:2405.10548, May 2024.

[P1] **Anwoy Chatterjee**, Yash Goel, Tanmoy Chakraborty,"*HIDE and Seek: Detecting Hallucinations in Language Models via Decoupled Representations*", **Preprint** *(Under Review)*, arXiv:2506.17748, July 2025.

Experience

- **Research Intern** Adobe Inc.
  Media and Data Science Research Lab, Adobe India; ***Mentor*** *- Dr. Sumit Bhatia* *Jan 2025 - July 2025*
  - ○ Worked on developing robust and efficient post-training strategies for enhancing the instruction-following and reasoning abilities of LLMs.
  - ○ **A part of the work done during the internship is published at TACL'25**.
- **Research Intern** Adobe Inc.
  Media and Data Science Research Lab, Adobe India; ***Mentor*** *- Dr. Sumit Bhatia* *May 2024 - Aug 2024*
  - ○ Worked on analyzing and quantifying the sensitivity of LLMs to alterations in the input prompt.
  - ○ **A part of the work done during the internship was published at EMNLP'24**.

Selected Research Projects

- **Studying the Effect of Instruction Tuning Loss on Generalization** Adobe & IIT Delhi
  Ongoing PhD Project; ***Supervisors*** *- Prof. Tanmoy Chakraborty and Dr. Sumit Bhatia* *Nov 2024 - Present*
  - We observe that the standard instruction tuning loss often yields suboptimal performance across benchmarks and limited robustness to input prompt variations.
  - We proposed Weighted Instruction Tuning (WIT) as a better alternative to conventional instruction tuning and observed that assigning a low-to-moderate weight to prompt tokens coupled with a moderately high weight to response tokens yields best-performing models across various settings, achieving an average gain of $\sim 6.55\%$ over the conventional loss across five models, three training datasets and four benchmarks. This work is now accepted to **TACL'25**.
  - We are currently working on developing a novel instruction tuning loss with dynamic token weighting to enhance both generalization and robustness of the language models.

- **Detecting Hallucinations in LLMs** IIT Delhi
  PhD Project; ***Supervisor*** *- Prof. Tanmoy Chakraborty* *December 2024 - May 2025*
  - We observed that hallucinations often stem from a statistical decoupling between the hidden states corresponding to input and output tokens in LLMs.
  - We proposed `HIDE` as a statistically-inspired white-box method for detection hallucinations effectively in long-form generations by LLMs. Our proposed method is single-pass, proving to be both more effective and efficient compared to the current state-of-the-art detection methods which are primarily multi-pass methods.
  - The paper on this work is currently *under review*.

- **Quantifying the Sensitivity of LLMs towards Prompt Perturbations** Adobe & IIT Delhi
  PhD Project; ***Supervisors*** *- Prof. Tanmoy Chakraborty and Dr. Sumit Bhatia* *Jan 2024 - Sept 2024*
  - LLMs are observed to generate varied outputs on slightly changing the input prompt. This is often concerning for end users, as finding the optimal prompt is non-trivial for naive users.
  - We developed POSIX, an index to quantify the prompt sensitivity of an LLM on a benchmark. POSIX is agnostic to the accuracy or performance of LLMs on the benchmark. The work was published at **EMNLP'24**.

- **Cross-Task In-Context Learning in LLMs to Solve Data-Scarce Novel Tasks** IIT Delhi
  PhD Project; ***Supervisor*** *- Prof. Tanmoy Chakraborty* *Aug 2023 - Feb 2025*
  - We first identified the possibility of cross-task in-context learning (ICL) in LLMs – the work was published at **ACL'24**.
  - We also developed a method to effectively learn task representations, which can be utilized for selecting source tasks to facilitate effective cross-task information transfer in ICL. The work is currently *under submission*.

Services

- **Conference Reviewer:** ACL Rolling Review (Feb'25, May'25)
- Served as a **Teaching Assistant** for the **Introduction to Large Language Models** course offered jointly by IIT Delhi and IIT Bombay on **NPTEL**.
- Served as a **Graduate Teaching Assistant** for the following courses at IIT Delhi: AIL861/ELL8299 (Advances in Large Language Models), ELL884 (Deep Learning for Natural Language Processing), AIL821/ELL881 (Large Language Models: Introduction and Recent Advances), MTL101 (Linear Algebra and Differential Equations).

Honors and Awards

- Awarded with the **Google PhD Fellowship**, 2024 in the area of **Natural Language Processing**.
- Selected as a finalist of **Qualcomm Innovation Fellowship India**, 2024 (39 finalists out of 122 proposals).
- Selected to attend the *Google Research Week*, 2024 and *Google DeepMind Research Symposium*, 2025.
- Selected to present my research work at *Amazon Research Days*, 2024.
- Ranked 2^{nd} in the Department of Computer Science and Engineering, IIT (BHU), among the graduating batch of 2022.
- Qualified among the top 2% of the students (about 160,000) appearing for *JEE-Advanced*, 2018.
- Selected for the prestigious *KVPY Fellowship* by Government of India (while studying in Class 11) in 2016.
- Ranked 9^{th} in West Bengal in *NTSE (National Talent Search Examination)*, 2015.