Statement of Research Interests

Subhabrata Dutta

My research interests generally fall under the broad umbrella of machine learning applications on textual and network data. In this statement, I outline my past research experiences that are part of my current ongoing work and doctoral thesis, followed by the possible research directions that I am interested to explore in future.

Past Research Interests

Large Language Models (LLMs). Exploring the emergent abilities of LLMs and their practical applications is my primary research interest. In one of my recent works [11], I have developed a framework towards eliciting robust reasoning capabilities in LLMs by separating the solver component from the reasoning decomposer component. A relatively small model is finetuned as an agent that interacts with another (possibly larger) LLM. While the latter solves the reasoning task via chain-of-thought, the former verifies the solution and guides the solver by asking subproblems that constitute the original task. Additionally, I have been working on the nuances of in-context learning in low-resource settings. In my recent work with multilingual LLMs [12], I have shown that cross-lingual in-context learning can be elicited in such models by aligning the source language examples and the target language inputs using semantic similarity and task descriptions. I have worked on aligning pretrained LMs with unsupervised finetuning towards superior argument understanding [6]. Earlier, I have explored the possibilities of building compute-efficient Transformer architectures from the perspective of dynamical systems [4].

Social discussion mining. In my doctoral research, I have worked on predictive modelling of user engagement in online platforms. Exogenous and endogenous influences play a major role in modulating content popularity online; I have developed frameworks for modelling the temporal dynamics of influence-modulated popularity for two different modes of online engagement: discussion forums [3, 1] and microblogging sites [7]. Another key research question explored in my doctoral thesis is the interplay between user opinion and the dynamics of the engagement network formed by them; my past explorations resulted in a semi-supervised framework for tweet stance detection leveraging follow networks [5], characterization of online conflict [2] and how hate-spreaders organize themselves in online echo chambers [9].

Future Research Interests

The advent of LLM ubiquity has posed multiple challenges in front of us, both theoretical as well as application-specific. A considerable portion of AI research in general (that is, beyond straightforward NLP tasks) are now considering LLMs as backbones. On the other hand, we lack any robust theoretical framework to link *what they do* to *how they do*. In this context, following are the key research areas that I am eager to explore in future.

Modular LLMs beyond traditional learning. In my past research, I have explored training LLMs beyond traditional next token prediction setup, primarily using RL to guide the model to use feedback from blackbox environments like another LLM‘[11] or deterministic tools. I seek to explore this direction further in future, given in my opinion, *intelligence*, in any definition, emerges from multiple, heterogeneous interactions with the world around itself.

Mechanistic Interpretability. This is an emerging research direction, aimed towards explaining Transformer-based models by reverse engineering and identifying sparse components that implement algorithms [8, 13]. The current development being still in a nascent state, it is a promising direction towards in-depth understanding of how LLMs function. I am currently working on interpreting mult-istep abstract reasoning and very much interested in continuing in this direction.

Memorization, hallucination, and training dynamics. Most applications of LLMs as AI assistants assume a knowledge-base implicitly present in the parametric representations. Yet, little do we know about how the knowledge is actually stored in the humongous parameter configuration. As a result, a hallucinating LLM is almost beyond repair except a few extrinsic strategies like retrieval augmentation. Recent research [10] shows evidence of many peculiar phenomena like superposition, data double descent, etc. that are intertwined with the feature memorization of a trained neural network. Very closely linked to mechanistic interpretability, studying training dynamics is likely to explain different emergent abilities of LLMs while prescribing robust solutions to the existing limitations. I very much look forward to exploring this area of research in my future engagements.

References

[1] Subhabrata Dutta, Dipankar Das, and Tanmoy Chakraborty. Modeling engagement dynamics of online discussions using relativistic gravitational theory. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 180–189. IEEE, 2019.

[2] Subhabrata Dutta, Dipankar Das, Gunkirat Kaur, Shreyans Mongia, Arpan Mukherjee, and Tanmoy Chakraborty. Into the battlefield: Quantifying and modeling intra-community conflicts in online discussion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1271–1280, 2019.

[3] Subhabrata Dutta, Sarah Masud, Soumen Chakrabarti, and Tanmoy Chakraborty. Deep exogenous and endogenous influence combination for social chatter intensity prediction. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1999–2008, 2020.

[4] Subhabrata Dutta, Tanya Gautam, Soumen Chakrabarti, and Tanmoy Chakraborty. Redesigning the transformer architecture with insights from multi-particle dynamical systems. *Advances in Neural Information Processing Systems*, 34:5531–5544, 2021.

[5] Subhabrata Dutta, Samiya Caur, Soumen Chakrabarti, and Tanmoy Chakraborty. Semi-supervised stance detection of tweets via distant network supervision. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, pages 241–251, 2022.

[6] Subhabrata Dutta, Jeevesh Juneja, Dipankar Das, and Tanmoy Chakraborty. Can unsupervised knowledge transfer from social discussions help argument mining? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, May 2022.

[7] Subhabrata Dutta, Shravika Mittal, Dipankar Das, Soumen Chakrabarti, and Tanmoy Chakraborty. Incomplete gamma integrals for deep cascade prediction using content, network, and exogenous signals. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[8] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1, 2021.

[9] Vasu Goel, Dhruv Sahnan, Subhabrata Dutta, Anil Bandhakavi, and Tanmoy Chakraborty. Hatemongers ride on echo chambers to escalate hate speech diffusion. *PNAS nexus*, 2(3):pgad041, 2023.

[10] Tom Henighan, Shan Carter, Tristan Hume, Nelson Elhage, Robert Lasenby, Stanislav Fort, Nicholas Schiefer, and Christopher Olah. Superposition, memorization, and double descent. *Transformer Circuits Thread*, 2023.

[11] Gurusha Juneja, Subhabrata Dutta, Soumen Chakrabarti, Sunny Manchanda, and Tanmoy Chakraborty. Small language models fine-tuned to coordinate larger language models improve complex reasoning. *arXiv preprint arXiv:2310.18338*, 2023.

[12] Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, July 2023. Association for Computational Linguistics.

[13] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.