Research Statement

Mohna Chakraborty

My current post-doctoral research focuses on advancing the practical applications of Generative AI (GenAI) with an emphasis on enabling large language models (LLMs) to exhibit human-like social behavior in controlled experimental settings. My work specifically aims to enhance the personalization, explainability, interpretability, and robustness of LLMs in social contexts, with an eye toward real-world applicability in varied social interactions. During my doctoral studies, I concentrated on the domain of data mining, natural language processing, and machine learning, with a specific focus on low-cost information extraction using minimal supervision. My research goal was to *obtain quality-aware annotations for diverse data types with a limited budget and minimal human effort* for various natural language processing tasks. My research tackled the scarcity of labeled textual data by developing methods that significantly streamline the annotation process, minimizing human input while ensuring cost-effective, accessible solutions compatible with everyday systems. This focus on efficient, budget-friendly approaches has been central to my research philosophy. My research benefits from and facilitates interdisciplinary collaboration within and outside the university and with industry research labs.

I have contributed several key methods in review analysis and crowd-sourcing. I have published over **fourteen research papers** [3, 5, 2, 1, 6, 4] in top-tier international conferences and workshops, including conferences in natural language processing, artificial intelligence, data mining, and software engineering (ACL, UAI, KDD, PAKDD, SIAM, ESEC/FSE), and workshops in natural language processing and machine learning (ICML, ICLR, WWW, RANLP, PAKDD, ML Reproducibility Challenge). Among them, I am the first author for one paper accepted at ESEC/FSE and co-first author for six papers, including one in ACL (the most reputed conference in natural language processing) and in KDD, PAKDD (the most reputed conferences in data mining). I have several submitted and ongoing works where I am the first or the co-major author. I have worked on the following research directions in data extraction and analysis during my doctoral journey:

1 Quality Aware Annotations with Limited Budget

Data labeling is essential to infer useful information from data and to train quality machine learning models. When the budget is limited, the feasible approaches for data labeling are crowd-sourcing or using pre-trained language models. Both these approaches may result in annotations of low quality. My research aims to improve the quality of annotations obtained by employing these approaches with a limited budget.

1. **Bias and noise in weakly labeled data:** One of the limitations of large language models is their sensitivity to bias and noise in the training data. When the budget is limited, and ground truth labels are unavailable, this LLM sensitivity can result in poor data labeling quality. In my paper published in **KDD'22** [2], I propose a novel double-layer span extraction framework to tackle the challenge of noise and bias in weakly labeled data for the task of review analysis.

2. **Perturbation sensitivity to prompts:** Other limitations of large language models is their sensitivity to the perturbations of prompts when used in zero-shot setting. Furthermore, it is also challenging to find quality prompts for a given LLM in a zero-shot setting. To address these challenges, in my paper published in **ACL'23** [3], the proposed approach automatically generates multiple prompts similar to a reference base prompt and ranks them using a novel ranking metric in a zero-shot setting.

3. **Label dependencies to improve data labeling quality:** Several labeling tasks have implicit or explicit label dependencies between samples in the unlabeled corpus. These label dependencies can help propagate labeling information to dependent samples and enlarge the impact of labels from crowd workers. In my paper published in **UAI'23** [5], I take advantage of label dependencies to improve data labeling accuracy for node classification tasks in citation networks.

2 Future Plan

Building upon my doctoral research and ongoing work, my future goal is to work on proposing cost-effective approaches to enhance the reliability of LLMs for extracting valuable insights and generating high-quality labels from overwhelmingly large and complex data. Specifically, I aim to design strategies that will enable LLMs to function as dependable proxies by improving their decision-making transparency, which will help clarify the reasoning processes underlying the models' outputs. This will involve investigating the rationale behind label assignments by LLMs, focusing on developing frameworks that ensure the explainability, interpretability, and replicability of these models. In doing so, I intend to address critical challenges associated with the current limitations of LLMs.

Also, I plan to explore ways to steer LLMs in more effectively representing the rich diversity of human personalities, voices, and experiences, with particular attention to marginalized communities and individual perspectives. My work will focus on fostering the development of inclusive and trustworthy language models that accurately reflect a broad spectrum of human viewpoints, encompassing both majority and minority perspectives. By doing so, I seek to promote the alignment of LLM reasoning with human thought processes, enhancing their capacity to understand and reflect the complexity of human experience.

2.1 Improving the reliability of LLMs

LLMs often exhibit undesirable behaviors like hallucination, sycophancy, and others that confine their reliability in sensitive domains like fact-checking, medical decision-making, and legal analysis. A hallucination occurs when the models generate factually incorrect or fabricated content, while sycophantic behavior involves producing overly agreeable or flattering responses. These challenges arise partly due to the limited reasoning capabilities of LLMs. These limitations hinder the models' ability to produce accurate, trustworthy outputs, primarily when tasked with nuanced or sensitive applications. One promising direction is to decompose complex tasks into more manageable subtasks either in a few-shot or zero-shot setting. This will allow LLMs to leverage their reasoning capabilities effectively by focusing on solving simpler components individually, which can reduce hallucinations and improve output quality. Task decomposition also provides an opportunity to mitigate sycophantic behavior by encouraging the model to produce more balanced, neutral responses. My research aims to develop dynamic frameworks for task decomposition that improve LLM reasoning, minimize inaccuracies, and enhance the reliability and transparency of these models for real-world applications.

2.2 Aligning LLM with human reasoning

LLMs have shown strong capabilities in text generation but often struggle with tasks requiring human-aligned reasoning, particularly in subjective contexts such as ethical decision-making and

societal discourse. To address this, I aim to develop a hybrid framework that combines theoretical concepts, multi-LLM reasoning aggregation, and role-based prompting to generate coherent, human-aligned reasoning. This framework will identify core reasoning aspects from theoretical concepts, aggregate outputs from diverse LLMs, and dynamically select roles relevant to specific tasks and values. The effectiveness of this approach will be evaluated using metrics for aspect coverage, coherence, and label alignment to ensure the outputs closely align with human perspectives. Additionally, for LLMs to perform tasks requiring human-like presence, they must reflect distinct, realistic human characteristics. My research will focus on methods for personalizing LLMs to emulate individual human traits, internalizing morals, values, and ideologies, and adapting to nuanced social contexts. This includes developing mechanisms to represent diverse backgrounds authentically and introducing evaluation metrics to assess the deep alignment between LLM outputs and human responses. Ultimately, the goal is to create LLMs capable of dynamic and contextually appropriate personalization, enhancing their applicability in complex, subjective domains.

References

[1] Chakraborty, M. Does reusing pre-trained nlp model propagate bugs? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (New York, NY, USA, 2021), ESEC/FSE 2021, Association for Computing Machinery, p. 1686–1688.

[2] Chakraborty, M., Kulkarni, A., and Li, Q. Open-domain aspect-opinion co-mining with double-layer span extraction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2022), pp. 66–75.

[3] Chakraborty, M., Kulkarni, A., and Li, Q. Zero-shot approach to overcome perturbation sensitivity of prompts. 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, pp. 5698–5711.

[4] JILES, R. D. [re] domain generalization using causal matching. In *ML Reproducibility Challenge 2021 (Fall Edition)* (2022).

[5] Kulkarni, A., Chakraborty, M., Xie, S., and Li, Q. Optimal budget allocation for crowdsourcing labels for graphs. In *Uncertainty in Artificial Intelligence* (2023), PMLR, pp. 1154–1163.

[6] Mishra, A. K., and Chakraborty, M. Does local pruning offer task-specific models to learn effectively ? In *Proceedings of the Student Research Workshop Associated with RANLP 2021* (Online, Sept. 2021), INCOMA Ltd., pp. 118–125.