Recent advances in Natural Language Processing (NLP), particularly in large language models (LLMs), have revolutionized human-AI interaction [1, 2]. However, these innovations often overlook key aspects of user engagement in various environments, such as the effective integration of real-time human feedback, which is crucial for creating intuitive and accessible interfaces. This gap hinders the adaptability and trustworthiness of interactive NLP systems in diverse human contexts. In my PhD, I aim to address these challenges by designing new communicative system architectures and analytics methods. These will tackle the core limitations of real-time situated human-AI interaction, ensuring adaptability and trustworthiness in AI design.

Adaptive and Trustworthy NLP. Last year, guided by Professor Sheng Wang, I focused my research on addressing neural model issues from a user perspective. The two main challenges are: 1) Data noise sensitivity: Minor errors can significantly alter model outcomes, necessitating more forgiving models. 2) Output explainability: Modern models may generate biased results or compromise user privacy, requiring a user-centric approach in Machine Learning (ML) and NLP. Addressing these involves developing evaluation protocols and analytical methods prioritizing human knowledge and feedback. The goal is to create a more harmonious interaction between humans and generative models, ensuring reliable, understandable outputs. This research trajectory is vital for advancing ML/NLP, making it more adaptable and safe for everyday users.

Electronic medical records (EMRs), usually stored in relational databases, require Structured Query Language (SQL) to retrieve relevant information. Efficiently making such queries can be challenging for medical professionals without specific software skills. Recently, large-scale pre-trained language models (PLMs) have shown promise in Text2SQL tasks. This piqued interest in whether Text2SQL PLMs can capture domain-specific knowledge. To investigate, we created a large-scale biomedical Criteria2SQL dataset with 4000 pairs of clinical trial criteria and SQL. We then did empirical analysis on the performance of OpenAI Codex [3] on the Crit2SQL task and classified the mistake into seven major types. We notice that incomplete generation was the most common error type, accounting for 58.9% of errors. Further analysis showed 55.5% of Codex-generated SQL queries were under 25 tokens compared to 31.5% of human-annotated queries. This data highlights Codex's tendency for shorter outputs and oversensitivity to input specifics, reflecting the need for models resilient to input variability for robust, complete outputs.

We selected structured criteria as our target text for analysis due to its divisibility. However, this approach raises questions about handling complex language forms. The performance of language models with ambiguous, comparative, or explanatory language, or language influenced by cultural contexts, largely depends on their training and contextual understanding. A model trained on a diverse dataset is more adept at interpreting varied language types. Despite this, current state-of-the-art (SOTA) models [4] have limitations. They are sensitive to hyper-parameter settings and random seed choices, lack nuanced understanding of domain-specific knowledge, and may poorly generalize on unseen data. Such limitations can lead to poor generalization on unseen data, posing risks in practical applications and emphasizing the need for robust model design.

Motivated by these challenges, we introduced BioTranslator [5], a novel multilingual and multimodal translation method, transforming biological data analysis and annotation. Traditional methods [6] rely heavily on controlled vocabularies (CVs) and struggle with new data. BioTranslator overcomes these limitations by converting user-written text into biological data like gene expression, using various biomedical ontologies. This allows data exploration beyond existing CVs. Unlike standard models like ProTranslator [4] that are limited to bilingual translation, BioTranslator's multilingual framework handles various "languages"—genes, drugs, phenotypes, pathways—without paired data, enabling new class identification without pre-existing instances. BioTranslator achieves an average AUROC of 0.90 in cell subtype classification based on textual descriptions alone, eliminating the need for annotated cells or markers. A unique feature of BioTranslator is its ability to translate between two biological modalities via a third one, avoiding direct paired data. This approach simulates real-world scenarios, like analyzing new drugs identified by shorthand codes without textual descriptions, demonstrating its potential in discovering and understanding new languages or contexts for a broader audience.

Looking ahead, the widespread societal applications of ML models, particularly LMs, underscore the necessity of enhancing their adaptability and trustworthiness. My work with BioTranslator has involved generalizing PLMs to out-of-distribution (OOD) ontologies. In the next step, I aim to design MLMs that align with human rationales, avoid spurious correlations, and generalize to OOD data. I am particularly drawn to recent developments in explainable NLP, especially free-text rationales (FTRs), focusing on evaluating FTRs and distilling reasoning knowledge from them to improve LMs' transparency and reliability. It is also imperative to refine models to adapt to new information from the latest facts as well as nuances and cultural variations in human communication, enhancing LMs' adaptability, trustworthiness, and fairness. I am eager to explore further in these directions and beyond.

Human-AI Interaction. The recent advancements in natural language processing have opened up numerous possibilities for building human-compatible AI. Language, being a natural interface for human-AI interaction, can be leveraged to develop agents that can learn useful behaviors, communicate, and explain their processes.

In my work on ICT4D, I focused on designing responsible technologies that aid people's lives in developing regions. I joined the eKichabi project [7] in the ICTD Lab, aiming to benefit farmers and businesses in low-income areas such as rural Tanzania through digital technologies. Under Richard Anderson's supervision, I led the design and implementation of new features, including a user authentication system and a user action logging system. I have tested and re-implemented the android app over 32 iterations, achieving remarkable improvements in the app's user interface design, runtime, memory usage, storage usage, privacy protection, and reliability. The new version has been released on the Google Play Store for further testing with a massive group of farmers and businesses from over 300 villages across 6 districts in Tanzania. This experience allowed me to investigate user interactions with the app and their association with users' background and culture, sparking my interest in enabling LLMs to understand and learn through user interactions.

At Cornell, I'm working on projects to design language models for specific communities, such as LGBTQ+ individuals with vocal impairments, elderly people with voice disabilities, language learners, and fashion designers. This work is part of my effort to create AI systems that cater to the diverse goals and needs of different human communities. Besides, my internship at a wellness company allows me to develop interactive systems, enhancing my experience in creating effective AI solutions that resonate with real-world users. These experiences guide my research towards developing AI systems that are responsive, adaptable, and empathetic to the complexities of human interaction and communication.

Driven by these experiences, I would like to continue my research on the development of interactive NLP systems that learn and evolve through both direct human communication and implicit feedback. This research direction could lead to designing and building NLP systems that can infer user intention and preferences, allow for fine-grained control, and adapt to natural language feedback. In particular, I am interested in designing systems that are adept at understanding user intentions and preferences and capable of fine-tuning their responses based on the subtle nuances captured through implicit feedback signals. I also admire a user-centric research approach, which is to conduct an in-depth study on real users, design interactive NLP systems from a user-centric perspective, and deploy to test. Following this approach, I aim to capture users' essential needs in human-AI interaction and enhance interactive NLP systems with diverse user groups and settings.

Goal and Future Outlook. My career aspiration is to become a professor in academia. This choice is particularly derived from my positive experiences in both teaching and mentoring. As a TA for the AI course, I led a discussion section and held office hours, while my duties as a mentor in the COM2 Big/Little Mentorship Program involved advising first-year and transfer students on exploring research projects on BioNLP; I found that I enjoyed both teaching and research advising. Attending a PhD program will allow me to continue my research, while also gaining further experience mentoring team members to turn research ideas into products.

Past experiences have led to my interests in adaptability, trustworthiness, and human-AI interaction, yet I am also open to exploring other important problems in NLP and HCI. I hope my next frontier can be the launchpad for my journey in further exploring and contributing to NLP and HCI research.

References

[1] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

[2] Tongshuang Sherry Wu, Michael Terry, and Carrie J. Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2021.

[3] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, David W. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, S. Arun Balaji, Shantanu Jain, Andrew Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374, 2021.

[4] Hanwen Xu and Sheng Wang. Protranslator: Zero-shot protein function prediction using textual description. In *Annual International Conference on Research in Computational Molecular Biology*, 2022.

[5] Hanwen Xu, Addie Woicik, Hoifung Poon, Russ B. Altman, and Sheng Wang. Multilingual translation for zero-shot biomedical classification using biotranslator. *Nature Communications*, 14, 2023.

[6] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather L. Butler, J. Michael Cherry, Allan Peter Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna E. Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[7] Galen Cassebeer Weld, Trevor Perrier, Jenny Aker, Joshua Evan Blumenstock, Brian Dillon, Adalbertus Kamanzi, Editha Kokushubira, Jennifer Webster, and Richard J. Anderson. ekichabi: Information access through basic mobile phones in rural tanzania. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.