Human-Centered Evaluation of Language Technologies

Su Lin Blodgett[1], **Jackie Chi Kit Cheung**[2], **Q. Vera Liao**[1], **Ziang Xiao**[3]
[1]Microsoft Research, Canada
[2]McGill University, Canada
[3]Johns Hopkins University, USA
sulin.blodgett@microsoft.com jackie.cheung@mcgill.ca,
veraliao@microsoft.com, ziang.xiao@jhu.edu

Abstract

Evaluation is a cornerstone topic in NLP. However, many criticisms have been raised about the community's evaluation practices, including a lack of human-centered considerations about people's needs for language technologies and technologies' actual impact on people. This "evaluation crisis" is exacerbated by the recent development of large generative models with diverse and uncertain capabilities. This tutorial aims to inspire more human-centered evaluation in NLP by introducing perspectives and methodologies from the social sciences and human-computer interaction (HCI), a field concerned primarily with the design and evaluation of technologies. The tutorial will start with an overview of current NLP evaluation practices and their limitations, then introduce complementary perspectives from the social sciences and a "toolbox of evaluation methods" from HCI, accompanied by discussions of considerations such as what to evaluate for, how generalizable the results are to the real-world contexts, and pragmatic costs of conducting the evaluation. The tutorial will also encourage reflection on how these HCI perspectives and methodologies can complement NLP evaluation through Q&A discussions and a hands-on exercise.

Type of Tutorial: Introductory

1 Tutorial Description

Designing effective evaluation methods for natural language processing (NLP) has long been challenging due to the complex nature of language, open-endedness of tasks, and multifaceted and context-dependent definitions of language quality. This challenge is exacerbated as "general" capability models (e.g., large language models) become more capable and prevalent. Not only must they be evaluated across a diverse range of tasks and domains, which can be difficult to define and validate, but their wide range of potential capabilities, including those potentially unanticipated by model developers (Ganguli et al., 2022), may also render evaluation results ungeneralizable to and unreliable in real-world contexts where the model is to be used.

Researchers have pointed out shortcomings of popular NLP benchmarks, metrics, and human evaluation methods (e.g., human ratings), such as their inability to capture nuanced meanings, their lack of validity, their perpetuation of biases and potential harm, and a lack of standardization and reproducibility (Howcroft et al., 2020; Clark et al., 2021; Jacobs and Wallach, 2021; Gehrmann et al., 2023). Ultimately, NLP models are to be incorporated into real-world applications, interacted with by people, and can have a profound impact on people's lives. Evaluation methods must take on a human-centered perspective that centers around people's needs, values, and interaction behaviors in order to produce results that can realistically reflect real-world performance and possible impacts.

These kinds of human-centered considerations are at the forefront of evaluation practices in social science where the validity of measurements is a key focus, as well as in human-computer interaction (HCI), a field primarily focusing on how to design technologies and evaluate the designs. In the past half-decade, HCI researchers have developed a "toolbox of methods" as different "ways of knowing" (Olson and Kellogg, 2014) people's needs, usage, and interaction outcomes with technologies. This tutorial aims to provide an introduction to these HCI perspectives and evaluation methods to inspire more human-centered evaluation methods in NLP, and to facilitate collaboration between the HCI and NLP communities.

This 3-hour tutorial will include 110 minutes of instructors' presentations followed by Q&A and a hands-on exercise. The presentations will start with a brief overview of current evaluation practices in NLP, including automatic evaluation and human evaluation. In this part, we will review common goals and assumptions that are built into existing

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 39–43
November 12–16, 2024.

evaluation practices. We also aim to highlight concerns and limitations—e.g., lack of reliability, realism, and standardization—which may lead to an overall lack of validity in the evaluation outcomes.

With these concerns and limitations of NLP evaluation in mind, we will introduce complementary perspectives in social sciences and HCI. We will introduce measurement modeling—a framework that disentangles what is measured (i.e., theoretical, frequently unobservable constructs) from how it is measured (operationalizations) and offers a rich vocabulary via *validity* and *reliability* to assess measurements (Jacobs and Wallach, 2021). We will further illustrate how these concepts can be applied to better assess NLP evaluation approaches (e.g., Xiao et al., 2023; Liu et al., 2024).

We will then provide an overview of common HCI evaluation methods, from human-subjects studies and surveys to analytical and simulated evaluations, and discuss the benefits and drawbacks of each. By comparing these different methods, we will particularly highlight the consideration of realism (McGrath, 1995; Schmuckler, 2001; Liao and Xiao, 2023)—designing evaluations in a way that the conclusion can be generalized to the real-world contexts where the technology will be used, and pragmatic costs to conduct the evaluation. Our goal is to inspire NLP researchers to explore diverse evaluation methods as alternatives to benchmarks and automated metrics, and develop human-centered evaluation methods with downstream human needs and lower adoption barriers (for people who should be doing evaluation, such as model developers) in mind. To further ground the introduction to HCI evaluation, we will present examples of HCI works conducting evaluations for language technologies such as chatbots (Langevin et al., 2021; Xiao et al., 2020) and writing support (Jakesch et al., 2019; Wu et al., 2019).

Lastly, the hands-on exercise will ask participants to work in groups to choose an evaluation method and design the details for a given use case. The exercise is designed to encourage participants to explore and compare different evaluation methods they learn from the tutorial, and facilitate further reflections and discussions.

2 Tutorial Content

2.1 Introduction and Background (10 min)

This section will motivate the importance of human-centered evaluation for language technologies, and why we believe valuable lessons can be learned from the field of HCI, which has a primary focus on evaluating and understanding human interactions with and impact from technologies.

2.2 Evaluation in NLP (30 min)

This section will review typical evaluation practices in NLP, and discuss how they may fail to inform real-world performance and usefulness because of a lack of human-centered focus. The goal of this section is not to be comprehensive about the wide range of metrics, datasets, and benchmarks in NLP, but to illustrate common assumptions in their design and application.

We will present examples of evaluation techniques, and ways to distinguish them (e.g., automatic vs. manual, or intrinsic vs. extrinsic). We will examine common motivations behind the development of new evaluations (e.g., to reduce costs or to evaluate a targeted type of model behavior).

We will present measurement modeling and the related concept of validity, and discuss ways in which measurements from the application of current evaluations can fail to exhibit validity, thus yielding unsupported conclusions.

2.3 Evaluation in HCI

2.3.1 Overview of HCI Evaluation Methods (40 min)

HCI researchers have developed and relied on a "toolbox of methods" to conduct evaluations of technologies. In this section, we will give an overview of common HCI evaluation methods (Barkhuus and Rode, 2007; Olson and Kellogg, 2014)—field studies, lab studies, surveys, and simulated evaluations—and discuss their benefits and drawbacks. We will highlight important considerations when making choices from the toolbox, such as quantitative v.s. qualitative, empirical v.s. analytical, and tradeoffs between realism and evaluation costs, which may depend on the types of claimed research contribution, technology development stage, and so on.

We will also include an orthogonal discussion about evaluation criteria commonly used in HCI research (MacDonald and Atwood, 2013; Hornbæk, 2006), including effectiveness, efficiency, user satisfaction, and other experiential and affective dimensions such as engagement and autonomy. Our tutorial will include a list of references for established scales and/or study procedures to evaluate

these criteria. We will also touch on or provide references for practical considerations for evaluation studies such as human-subjects recruitment, analyses of results, and study design best practices as well as ethical considerations.

2.3.2 Case Studies (20 min)

After mapping the landscape of HCI methods, we will walk through two case studies of how language technologies are evaluated in HCI research, such as decades of work on chatbots and more recent work on writing support using LLMs.

2.4 Reflection and Open Questions (10 min)

In this section, we will reflect on current NLP evaluation practices through the lenses employed in HCI research regarding how to assess and select from different evaluation methods. We will discuss how the evaluation practices in HCI and NLP communities can complement and learn from each other. We will also pose open questions and suggest future directions for the community to work towards human-centered evaluation.

2.5 Q&A and Hands-on Exercise (20+50 min)

We will leave Q&A time for audience to directly engage with the instructors. In the last 50 minutes, we will ask participants to form groups and work on a hands-on exercise. The exercise will present participants with choices of case studies, which may include a type of language technology and an "effect of interest" of the technology on people. Participants will work in groups to choose an appropriate evaluation method and design the details. In the end, we will ask the groups to share their evaluation design and encourage collective reflection on common threads and challenges.

3 Expected Outcome

We plan to make the tutorial presentation materials public and the videos accessible to a wide population. With participants' consent, we may also share notes from the Q&A session and discussions in the hands-on exercise.

Expected audience size: We expect to have more than 100 in-person attendees, based on the audience size of a NAACL 2022 tutorial on human-centered evaluation focusing on explanation (Boyd-Graber et al., 2022), and the recent popularity of the topic of model evaluation.

Target audience and prerequisite background: As an introductory tutorial, our presentation will not assume any prior familiarity with HCI evaluation methods or the HCI literature more generally. We expect the audience to have some familiarity with common NLP tasks but not necessarily expert knowledge of NLP evaluation.

Technical requirements: We do not expect technical support beyond regular presentations. To encourage group discussions during the Q&A and the hands-on group exercise, we would like to request roundtables for participants.

Preferred venue: Due to the personal leave schedule of one of the instructors, we have a strong preference for this tutorial to be held later in the year at EMNLP 2024.

4 Diversity Considerations

Instructors: The instructors consist of researchers across NLP, HCI, and psychology at varying career stages, spanning both industry and academia, with equal gender balance.

Diversifying audience participation: The tutorial format is designed to encourage broad participation from researchers and practitioners across industry and academia; no prior familiarity with HCI methods is expected, and the presentation materials will be made publicly available.

5 Presenter Biographies

Su Lin Blodgett is a researcher at Microsoft Research Montréal. Her work has examined measurement and evaluation in NLP, and she has co-organized three editions of the HCI+NLP Workshop, a CHI panel on responsible language technologies, and a FAccT tutorial on measurement and NLP.

Jackie Chi Kit Cheung is an associate professor at McGill University and at the Mila Quebec AI Institute. His work has involved developing new evaluation methods and datasets for a range of NLP tasks including common sense reasoning, automatic summarization, and authorship attribution.

Q. Vera Liao is a principal researcher at Microsoft Research. She is an HCI researcher by training and recently works on human-AI interaction, explainable AI, and responsible AI. She taught tutorials at NAACL 2022, NeurIPS 2022, CHI 2023, CHI 2020, as well as various seminars internationally. She is frequently involved in organizing events (e.g. panels, workshops) that connect the AI and HCI communities.

Ziang Xiao is an assistant professor in the Department of Computer Science. His work lies in the intersection of human-computer interaction, natural language processing, and social psychology. Ziang is on the organizing committee and an associate chair for multiple HCI venues (CHI, CSCW, IUI). He co-organized the 3rd HCI+NLP workshop at NAACL 2024. He co-organized the first workshop on Human-centered Evaluation and Auditing of Language Models at CHI 2024.

6 Ethics Statement

We hope that our tutorial will inspire human-centered evaluation practices that may help alleviate potential harm and ethical concerns brought about by language technologies. As many of the evaluation methods we will present involve human participants, we will also address ethical considerations emerging from their application, e.g., risks and best practices surrounding human-subjects recruitment and study design.

References

Louise Barkhuus and Jennifer A Rode. 2007. From mice to men-24 years of evaluation in chi. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, volume 10. ACM New York, NY.

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-centered evaluation of explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 26–32.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human'is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Kasper Hornbæk. 2006. Current practice in measuring usability: Challenges to usability studies and research. *International journal of human-computer studies*, 64(2):79–102.

David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, pages 169–182. Association for Computational Linguistics.

Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Conference on Fairness, Accountability, and Transparency (FAccT)*.

Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. Ai-mediated communication: How the perception that profile text was written by ai affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Yu Lu Liu, Su Lin Blodgett, Jackie Cheung, Q. Vera Liao, Alexandra Olteanu, and Ziang Xiao. 2024. ECBD: Evidence-centered benchmark design for NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16349–16365, Bangkok, Thailand. Association for Computational Linguistics.

Craig M MacDonald and Michael E Atwood. 2013. Changing perspectives on evaluation in hci: past, present, and future. In *CHI'13 extended abstracts on human factors in computing systems*, pages 1969–1978.

Joseph E McGrath. 1995. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human–Computer Interaction*, pages 152–169. Elsevier.

Judith S Olson and Wendy A Kellogg. 2014. *Ways of Knowing in HCI*, volume 2. Springer.

Mark A Schmuckler. 2001. What is ecological validity? a dimensional analysis. *Infancy*, 2(4):419–436.

Shaomei Wu, Lindsay Reynolds, Xian Li, and Francisco Guzmán. 2019. Design and evaluation of a social media writing support tool for people with dyslexia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore. Association for Computational Linguistics.

Ziang Xiao, Michelle X Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 2020. If i hear you correctly: Building and evaluating interview chatbots with active listening skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.