Designing, Learning from, and Evaluating Human-AI Interactions

Tongshuang Wu[†] **Diyi Yang**[‡] **Sebastin Santy**[±]
[†] Carnegie Mellon University [‡] Stanford University [±] University of Washington

1 Introduction

With the rapid advancement of natural language processing (NLP) research, there are numerous applications across a wide range of domains that require models to interact with humans — for example, chatbots responding to human inquiries (Thoppilan et al., 2022), machine translation systems aiding human translators (Santy et al., 2021), designers prompting Large Language Models for co-creation (Gero et al., 2022) or prototyping AI-infused applications (Park et al., 2022). In each of these cases, (timely) human interaction has been the key to the success; and any potential misconceptions or differences introduced to this interaction process might lead to error cascades at later stages (Sambasivan et al., 2021). Such interaction involves a lot of design choices *around* models — the sensitivity of interfaces (Amershi et al., 2019) and modalities (Ravichander et al., 2021), the impact of questions during human evaluation (Clark et al., 2021), or incorporating steer-ability in models (Dathathri et al., 2019).

These choices are equally (if not more) important compared to the algorithms or datasets, but they are often undervalued and sometimes even considered a trivial part of the equation. In fact, while many of these topics have been extensively investigated in Human-Computer Interaction (HCI), they have only recently gained sufficient attention in NLP. NLP researchers entering the interaction world typically have to go through a steep learning curve before they can fully utilize the best practices from HCI, resulting in some unintentional decisions that have adversely affected the reproducibility of earlier work (Clark et al., 2021).

In this tutorial, we aim to provide a systematic and up-to-date overview of key considerations and effective approaches for studying human-NLP model interactions. Interactions can take various forms depending on the stage of model development and the human involved; For example, NLP researchers and developers may interactively debug models during development, crowdworkers may participate in data annotation, etc.Our tutorial will focus specifically on the scenario where *end users* — lay people and domain experts who try to use and benefit from NLP models — interact or collaborate with deployed models (Wu & Bansal et al., 2021).

Throughout the tutorial, we will use four case studies (on model-assisted decision making, machine-aided translation, dialog systems, and prompting) to cover three major themes: (1) how to conduct usability ***evaluations*** to ensure that models are capable of interacting with humans; (2) how to ***design*** user interfaces (UIs) and interaction mechanisms that allow end users to easily access NLP models; (3) how to ***learn from*** and improve NLP models through human interaction. We will ground our discussion in HCI best practices, highlighting current challenges and future directions.

2 Tutorial Outline

This will be a **three-hour tutorial** devoted to the **cutting-edge topic** of *Designing, Learning from, and Evaluating Human-AI Interaction*. Each theme will take 35 mins, followed by 10 mins for Q&A and 10 mins for a break. Each part includes an overview of the corresponding topics, widely used methods, and a deep dive into a set of representative NLP and HCI work. In the last 15 minutes, we will conclude our tutorial by highlighting challenges and research opportunities in the field.

2.1 Walkthrough Case Studies

For consistency, we will use four case studies throughout the tutorial. They demonstrate how humans and models would play different roles, sometimes working together, sometimes supporting one another. We use them to discuss interaction initiation, usability priorities, etc.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 13–18
December 6-10, 2023

Model-assisted decision making. NLP Models are quite often used when making decisions such in clinical settings. In this setup, humans and AI collaborate towards a common goal, with the hope that each makes a decisions that they are best suited to make. It is an example of how standard evaluation may not translate to model usability in an interaction setup, because accurate models may not be complementary to human strengths (Bansal et al., 2021). Meanwhile, numerous studies have explored how humans would interact with classifiers making recommendations given various visual representations of model outputs and various forms of model explanations (e.g., Wu & Bansal et al., 2021; Gonzalez et al., 2020). This more mature and well-researched scenario will be used to give an overarching introduction on evaluation (§2.2) and interaction design (§2.3).

Machine-aided Translation. This instead illustrates a situation where humans take the initiative while the model provides assistance. With humans making the final judge on model usefulness, various evaluation dimensions are affected. For example, humans would deem a model useful even if they are partially correct (Green et al., 2014b) and different user groups get different benefits (Santy et al., 2021). Meanwhile, users' needs and perceptions on the model also affect their use patterns, e.g., they may only use models for keyword translation if model outputs are not fluent (Green et al., 2013, 2014a), which in turn points to future model improvement. We will use this case study to review the importance of human understanding and tracking in evaluation and learning (§2.4).

Dialog systems. Chatbot/dialog system is another early adoption of NLP techniques that also fall under models supporting humans. It represents the use case where evaluation is dynamic (1) the model performance is easily swayed by human responses and can hardly be measured on benchmark datasets (Li et al., 2021), (2) the model has to balance multiple criteria like interestingness, informativeness, etc. which could be subjective for different user groups (Thoppilan et al., 2022), (3) it is essential to implement fallback options (e.g., responses like "sorry I didn't understand" that's built around the model at the UI level) when the model does not behave as expected or safety modules when there is potential for controversiality (Kim et al., 2022). These properties also make dialog systems an ideal testbed for discussing UI designs (§2.3) and personalization (§2.4).

Prompting Large Lanuge Models Recently, large Language Models has made NLP models more accessible to end users, and has led to the emergence of a brand new interaction mechanism — prompting. Prompting perhaps represents a rare case where humans are "supporting" the model, i.e., they try to search for optimal instructions that maximize model performances on certain tasks. We will review various recent papers on prompting strategies (e.g., chaining (Wu et al., 2022), defining shareable prompt templates (Dang et al., 2022), inducing personas from LLMs (Reynolds and McDonell, 2021)), with an emphasis on the trade-off of expressiveness and learning curve (Jiang et al., 2022), and the potential of learning from user feedback (e.g., InstructGPT (Ouyang et al., 2022)). We will also emphasize on the differences between LLMs (which can respond to arbitrary human input text) and other modeling structures (which make more assumptions on possible text inputs).

2.2 Theme 1: Evaluate Model Usability

The first part of our tutorial will focus on evaluating NLP model usability. As mentioned in §2.1, NLP models that interact with (make suggestions to, have conversations with) humans need to go beyond accuracy (Ribeiro et al., 2020; Bhatt et al., 2021). User interaction experiences are affected by human-centered metrics such as safety, latency, faithfulness, responsiveness, etc. We refer to these dimensions as *usability evaluation*. In most cases, these evaluations are conducted on human subjects. Users would interact with both a target (experiment) model and a baseline (control) model, and compare them on effectiveness, usefulness, etc. through self-rating. The usability evaluations determine whether a model is ready for actual use. Unfortunately, their results are often easily swayed by arbitrary design choices (e.g., the survey question, the task instruction) (Roopa and Rani, 2012), making them unreliable.

This tutorial will guide the participants to design rigorous usability evaluations. Following the evaluation categorization in HCI (Kuniavsky, 2003), we will cover (1) survey design, (2) think-aloud protocol, (3) cognitive walkthrough, and (4) Experimentation and A/B testing. We will also discuss useful qualitative (e.g., Likert Scale results) and quantitative metrics (e.g., retrieving interaction speed from user clickstream (Lee et al., 2022)), best use sce-

nario and typical design pitfalls for each approach (e.g., leading questions in *survey design*).

Besides methodologies, this tutorial will also discuss the user group selection (Olsen Jr, 2007): (1) the potential impact of running studies on crowdsourcing platforms (where motivating participants is challenging and denoising is essential), in the lab (where graduate students are frequently used but can only represent a biased distribution), and in the actual deployment environment (which is costly); (2) the importance of identifying the targeted user group and achieving good coverage.

2.3 Theme 2: Interaction Design

Usability evaluation can help judge whether a model is usable, but user interfaces are still needed to make it user friendly. This part concerns the interface and interaction design, with two focuses:

(1) *Communication*, i.e., what inputs the model should take from humans and how to present the results. We will present different modes of human input (e.g., Natural Language input vs. traditional WIMP interfaces) and discuss their trade-offs (Wang et al., 2022). Additionally, we will discuss the desiderata for visualizing NLP model training information, their predictions, uncertainties, and (where applicable) explanations, as well as the impact of design choice (Khadpe et al., 2020). In addition, we will discuss how NLP models can have a design bias that make them difficult for people from different demographics (culture, language, age, gender), and how interactions may rectify the issues to some extent.

(2) *Initiation*, i.e., how the NLP model and the human can take the leading roles interchangeably. We will ground our discussion on the mixed-initiative interaction mechanism (Avula et al., 2022) — a flexible interaction strategy in which each agent contributes what it is best suited at the most appropriate time — and discuss how model initiations impact the perceived model usefulness (Avula et al., 2022; Santy et al., 2019), and how human initiations may be used as not only a driving force on achieving human goals (Oh et al., 2018), but also a fallback option when the model does not behave as expected (Lee et al., 2022).

2.4 Theme 3: Learn from Interactions

As users interact with NLP models, they generate rich signals that reveal model incorrectness and point to future model improvements (Krishna et al., 2022). For example, users may submit explicit feedback (e.g., users flagging a translation as incorrect) (Cabrera et al., 2021; Stiennon et al., 2020), or their clickstream may implicitly reflect their expectations on a model (e.g., when they revise a model-generated text after accepting the suggestion (Lee et al., 2022)).

Here, we review different types of human feedback that can be naturally retrieved from human interactions, as well as different modeling approaches to incorporate human feedback. Building on the survey from our presenter team (Wang et al., 2021), we will review recent studies that incorporate human feedback with respect to their goals, human interactions, and feedback learning methods, with a focus on example-based feedback (Wallace et al., 2019, e.g.,) and reinforce learning (Ouyang et al., 2022; Stiennon et al., 2020). In particular, we will also re-emphasize how the feedback can be retrieved through the methods introduced in §2.3. Additionally, to help researchers make practical use of these methods, we will discuss the potential trade-offs between intuitiveness vs. expressiveness (e.g., labeling functions in weak supervision (Ratner et al., 2016) might be more scalable but more difficult than labeling a single counterexample (Wallace et al., 2019)).

2.5 Breadth

While we will give pointers to dozens of relevant papers, we plan to cover around 7-8 research papers in close detail. Only 1-2 of the "deep dive" papers will come from the presenter team.

2.6 A Comparison with Relevant Tutorials

Given the rising awareness of human-centered NLP (a special theme at NAACL 2022), it is not surprising that some tutorials have already touched on some relevant topics. To the best of our knowledge, two tutorials that are closest to ours are: (1) *Case Studies in Benchmark Data Collection* at EMNLP 2021[1]) which uses six case studies to present a wide variety of data collection crowdsourcing methods and principles; and (2) *Human-centered Evaluations of Explanations* at NAACL 2022[2], which contributes a taxonomy of human-centered evaluation of explanations. Both tutorials have some topical overlaps with our tutorial: data labeling is a particular form of interaction, crowdsourcing-based interaction will be covered in *Evaluate Model Us-*

[1] https://nlp-crowdsourcing.github.io/
[2] https://xai-hcee.github.io/

ability, and explanations presentation will be covered in *Interaction Design*. However, we believe the overlap is not substantial, as we only instruct these elements as "parameters" in human-model interaction. Instead, we hope our tutorial will be complementary to the previous ones.

Additionally, workshops like CHAI, NLP+HCI, and DADC (NAACL 2022) has gathered researchers in the field to explore the frontiers of relevant topics, whereas our tutorial will do a systematic reflection correspondingly.

3 Diversity Considerations

Our chosen tutorial topic inherently touches on *human user distribution*. As mentioned before, we will discuss the importance of high coverage of user groups, and the impact of design biases on people from different demographics (e.g., ages, cultures, languages, and gender). As such, we believe our tutorial will be a strong advocate for diversity in the NLP model and interaction designs.

Besides diversity-related topics, our presenter team will also make our tutorial more accessible to different user groups. Specifically, we will share our tutorial with a worldwide audience by promoting it on social media. we will also work with *CL D&I teams, and consult resources such as the BIG directory to diversify our audience participation.

4 Prerequisites & Reading List

The tutorial is targeted toward NLP researchers and practitioners working with humans. The prerequisite includes familiarity with basic knowledge of NLP and language systems. Knowledge of system deployment is a plus. We will also provide a more paced introduction to some materials.

The authors will also release an NLP+HCI playbook as a resource for people interested in getting started in human-centered NLP research. Here are a few papers that lay a foundation for this area:

- Putting Humans in the Natural Language Processing Loop: A Survey (Wang et al., 2021);
- All That's Human Is Not Gold: Evaluating Human Evaluation of Generated Text (Clark et al., 2021);
- Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design (Yang et al., 2020);
- Does the whole exceed its parts? The effect of AI explanations on complementary team performance (Wu & Bansal et al., 2021);
- Principles of mixed-initiative user interfaces (Horvitz, 1999);
- Guidelines for Human-AI Interaction (Amershi et al., 2019);
- Training language models to follow instructions with human feedback (Ouyang et al., 2022);
- Learning to summarize with human feedback (Stiennon et al., 2020)

5 Tutorial Presenters

Sherry Tongshuang Wu (she/her) is an assistant professor at the Human-Computer Interaction Institute, Carnegie Mellon University. Her primary research investigates how humans (AI experts, lay users, domain experts) interact with (debug, audit, and collaborate) AI systems. Sherry has organized two workshops at NLP and HCI conferences: Shared Stories and Lessons Learned workshop at EMNLP 2022 and Trust and Reliance in AI-Human Teams at CHI 2022. She is currently developing a new course on Human-Centered NLP at CMU.

Diyi Yang (she/her) is an assistant professor in the CS Department at Stanford University. Her research focuses on learning with limited and noisy text data, user-centric language generation, and computational social science. Diyi has organized four workshops at NLP conferences: Widening NLP Workshops at NAACL 2018 and ACL 2019, Casual Inference workshop at EMNLP 2021, and NLG Evaluation workshop at EMNLP 2021. She also gave a tutorial at the ACL 2022 on Learning with Limited Data. She has taught courses on natural language processing at Georgia Tech since 2019 and is now developing a new course on Human-Centered NLP at Stanford University.

Sebastin Santy (he/him) is a second-year PhD student at the Paul G. Allen School of CSE, University of Washington. He works on problems in the intersection of HCI and NLP and specifically his research focuses on uncovering design biases in NLP systems. He previously worked on multilinguality and machine translation.

6 Ethics Statement

We do not anticipate any ethical issues related to the tutorial logistics, but we plan to cover ethical considerations in our content, especially when we discuss human-centered evaluation metrics like safety, and when we review the impact of different communication and initiation methods in interaction designs (e.g. leading to confirmation biases).

References

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13.

Sandeep Avula, Bogeum Choi, and Jaime Arguello. 2022. The effects of system initiative during conversational collaborative search. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–30.

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414.

Shaily Bhatt, Rahul Jain, Sandipan Dandapat, and Sunayana Sitaram. 2021. A case study of efficacy and challenges in practical human-in-loop evaluation of nlp systems using checklist. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 120–130.

Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. 2021. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–22.

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. *arXiv preprint arXiv:2107.00061*.

Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to prompt? opportunities and challenges of zero-and few-shot learning for human-ai interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Designing Interactive Systems Conference*, pages 1002–1019.

Ana Valeria Gonzalez, Gagan Bansal, Angela Fan, Robin Jia, Yashar Mehdad, and Srinivasan Iyer. 2020. Human evaluation of spoken vs. visual explanations for open-domain qa. *arXiv preprint arXiv:2012.15075*.

Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D Manning. 2014a. Predictive translation memory: A mixed-initiative system for human language translation. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 177–187.

Spence Green, Jeffrey Heer, and Christopher D Manning. 2013. The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 439–448.

Spence Green, Sida I Wang, Jason Chuang, Jeffrey Heer, Sebastian Schuster, and Christopher D Manning. 2014b. Human effort and machine learnability in computer aided translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1225–1236.

Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166.

Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. Promptmaker: Prompt-based prototyping with large language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–8.

Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–26.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. *arXiv preprint arXiv:2205.12688*.

Ranjay Krishna, Donsuk Lee, Li Fei-Fei, and Michael S Bernstein. 2022. Socially situated artificial intelligence enables learning from human interaction. *Proceedings of the National Academy of Sciences*, 119(39):e2115730119.

Mike Kuniavsky. 2003. *Observing the user experience: a practitioner's guide to user research*. Elsevier.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Huihan Li, Tianyu Gao, Manan Goenka, and Danqi Chen. 2021. Ditch the gold standard: Re-evaluating conversational question answering. *arXiv preprint arXiv:2112.08812*.

Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Dan R Olsen Jr. 2007. Evaluating user interface systems research. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*, pages 251–258.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. NoiseQA: Challenge Set Evaluation for User-Centric Question Answering. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Online.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

S Roopa and MS Rani. 2012. Questionnaire designing for a survey. *Journal of Indian Orthodontic Society*, 46(4_suppl1):273–277.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Sebastin Santy, Kalika Bali, Monojit Choudhury, Sandipan Dandapat, Tanuja Ganu, Anurag Shukla, Jahanvi Shah, and Vivek Seshadri. 2021. Language translation as a socio-technical system: Case-studies of mixed-initiative interactions. In *ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 156–172.

Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. Inmt: Interactive neural machine translation prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 103–108.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Yun Wang, Zhitao Hou, Leixian Shen, Tongshuang Wu, Jiaqi Wang, He Huang, Haidong Zhang, and Dongmei Zhang. 2022. Towards natural language-based visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*.

Zijie J Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. Putting humans in the natural language processing loop: A survey. *arXiv preprint arXiv:2103.04044*.

Tongshuang & Gagan Wu & Bansal, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-ai interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13.