Human-AI Interaction in the Age of LLMs

Diyi Yang[*] **Sherry Tongshuang Wu**[†] **Marti A. Hearst**[◇]
[*]Stanford University [†]Carnegie Mellon University [◇]University of California, Berkeley

1 Introduction

Recently, the development of Large Language Models (LLMs) has revolutionized the capabilities of AI systems. These models possess the ability to comprehend and generate human-like text, enabling them to engage in sophisticated conversations, generate content, and even perform tasks that once seemed beyond the reach of machines. As a result, the way we interact with technology and each other — an established field called "Human-AI Interaction" and have been studied for over a decade — is undergoing a profound transformation.

This tutorial will provide an overview of **the interaction between humans and LLMs**, exploring the challenges, opportunities, and ethical considerations that arise in this dynamic landscape. It will start with a review of the types of AI models we interact with, and a walkthrough of the core concepts in Human-AI Interaction. We will then emphasize the emerging topics shared between HCI and NLP communities in light of LLMs.

2 Tutorial Outline

This will be a **three-hour tutorial** devoted to the **cutting-edge topic** of *Human-AI Interaction in the Age of LLMs*. Each theme will take 35 minutes, followed by 10 minutes for Q&A and 10 minutes for a break. Each part includes an overview of the corresponding topics, and a deep dive into a set of representative studies. We will conclude our tutorial by highlighting challenges and research opportunities in the field.

2.1 Human-AI Interaction up to 2021

Though the interaction between humans and LLMs is still an emergent topic in NLP, it has been studied for more than a decade by other related fields. In this section, we will abstract the AI systems and interactions into taxonomies and desiderata for human-AI interaction that has been established prior to the introduction of LLMs. We plan to cover the following aspects:

Slot	Theme
Session 1: Human-AI Interaction before 2021	
14:00 – 14:10	Tutorial presenters introduction
14:10 – 14:35	Types of human-AI interaction and design thinking
14:35 – 15:15	Mixed-initiative interaction
15:15 – 15:45	Coffee Break
Session 2: Deep-dive into AI types	
15:45 – 15:55	Classic models in human-AI collaboration and case studies
15:55 – 16:10	Large language models as agents
16:10 – 16:30	Comparison with human-human interaction and human-AI interaction
Session 3: Human-LLM Interaction (HLI) and Challenges	
16:30 – 16:45	Paradigms and models (e.g., decomposition, planning) in HLI
17:45 – 17:00	Evaluation metrics and issues
17:00 – 17:15	Conclusion

Table 1: Example tutorial schedule.

- **Types of interaction**: We will enumerate the objective of interaction, including *human-AI collaboration* (the coordinated interaction between humans and AI to achieve certain goals) (Oh et al., 2018), *humans getting assistance from AI-infused applications* (humans using AIs as a tool, not a partner) (Amershi et al., 2019), and *humans analyzing AIs* (humans systematically understand NLP models) (Wu et al., 2019).

- **Design-thinking**: We will review desiderata for designing optimal interactions between AIs and humans. This will include HCI methods like need-finding, user-centered design, etc. (Amershi et al., 2019; Yang et al., 2020; Laban et al., 2021)

- **Goals for the interaction**: we will discuss typical evaluation metrics that represent the success of human-AI interactions, in particular centering around *complemen-*

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts), pages 34–38

tary performance. To achieve better outcomes than either could accomplish alone, by leveraging the strengths of both AI and humans (Wu & Bansal et al., 2021).

Mixed-Initiative Interaction Besides broad discussions on the aforementioned aspects, we will focus on discussing *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 (Horvitz, 1999) — 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., 2022a).

2.2 Deep-dive: Types of AIs, LLM Agents

In this section, we will concretize the theoretical grounding with more specific examples, grouped by how AIs are presented in the context of interaction. We will use research studies and real-world products that involve Human-AI Interaction as case studies, and reflect on their interactions design (e.g., through displaying model suggestions, dialog systems, GUI interactions).

We will first discuss the use of **single-purpose AIs** who take over dedicated tasks through a single form of interaction. This includes, e.g., toxicity detectors making recommendations in decision making tasks like content moderation (Zhang et al., 2023b), language models making autocompletion suggestions in writing tasks (Lee et al., 2022a), etc.

We will then move to the more current advancement of **general purpose AIs**, where the AI plays certain roles in social contexts, and interact with humans in more diverse manners, e.g., intelligent tutors offering multiple types of hints, explanations, followup questions etc. (OpenAI, 2023). This thread of work is becoming more prevalent as the AI systems become more competent in simulating human behaviors, and will ground our discussion on Human-LLM Interaction in §2.3.

LLM agents Among general-purpose AIs, we will particularly emphasize on how these LLMs are usually framed as *agents* (Talebirad and Nadiri, 2023; Wang et al., 2023), and how the interactions with these models follow social norms. Based off research on **human-human interaction**, we will cover how domain knowledge and skills can be operationalized into this process to support an effective workflow, and discuss possible limitations of using an agent (e.g., the introduction of human insights is very likely to trigger cognitive load for users). One example is our current survey comparing human-human pair-programming and human-AI pair-programming (Ma et al., 2023).

We will also compare the human-LLM agent interactions with the recent agent-agent interactions where both subjects of the interaction are LLM-simulated agents, including generative agent simulations where multiple LLM agents simulate a small town similar to The Sim (Park et al., 2023), and red-teaming research where an LLM plays the role of a malicious character for testing the safety of another model (Ganguli et al., 2022).

2.3 Human-LLM Interaction

The design of Human-LLM Interaction Directly leveraging LLMs for complex tasks, especially when it comes to sophisticated tasks that might require different expertise and both humans and LLMs, is non-trivial. Going beyond standard prompting engineering to supporting different aspects of interaction, we will cover a few key subareas under human-LLM interaction, ranging from decomposition to planning, refinement, and interaction (Cai et al., 2023; Li et al., 2023). Concretely, we will cover how prompts are often designed to generate certain outcomes, chain of thought prompting (Wei et al., 2022) to demonstrate desired actions, as well as few-shot learning techniques to tailor the generation. Purely relying on prompting requires substantial expertise and time for design and implementation, and makes it difficult to leverage end-user feedback. Thus, we will discuss how planning and human-in-the-loop (Zhang et al., 2023a) can help boost the workflow via techniques like structured planning, conditional generation (Hsu et al., 2023), and memory mechanisms (Park et al., 2023), for more transparent and collaborative human-LLM collaboration.

Evaluation We will discuss the evaluation of human-LLM interaction (Lee et al., 2022b), ranging from quantitative measures to user-centered evaluation. This will not only cover task-level performances, but also interaction dimensions such as usability, satisfaction, and engagement, as well

as long-term effects on users. Beyond evaluation, we will provide an in-depth summary of existing datasets (Lin et al., 2023), environments, and platforms that support the study of human-LLM interaction and provide guidelines on the pros and cons of different datasets, as well as how practitioners in this space could design innovative interaction paradigms tailored to their interests.

3 Tutorial Presenters

Diyi Yang is an assistant professor in the Computer Science Department at Stanford University. Her research focuses on human-centered natural language processing 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, NLG Evaluation workshop at EMNLP 2021, and Shared Stories and Lessons Learned workshop at EMNLP 2022. She also gave a tutorial at ACL 2022 on Learning with Limited Data, and a tutorial at EACL 2023 on Summarizing Conversations at Scale. Diyi and Sherry have co-developed a new course on Human-Centered NLP that has been offered at both Stanford and CMU.

Sherry Tongshuang Wu 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 and 2023. She will give a tutorial at EMNLP 2023 on Designing, Learning from, and Evaluating Human-AI Interactions.

Marti A. Hearst is a professor and the Interim Dean for the UC Berkeley School of Information. She is both an ACL Fellow and a SIGCHI Academy member, and former ACL President. Her research has long combined HCI and NLP; recent projects include adding interactivity to scholarly documents and creating interactive newspods. She recently gave invited keynote talks at the EACL NLP + HCI workshop, the KDD Workshop on Data Science with a Human in the Loop, and she advised the 2022 NAACL program chairs on the Human-Centered Natural Language Processing special theme. She has taught courses in NLP, HCI, and information visualization for 25 years.

4 Diversity Considerations

The topic of human AI interaction will be inclusive to both NLP and HCI communities. We will make our tutorial materials digitally accessible to all participants. During the tutorial sessions, we will work with student volunteers to encourage open dialogue and promote active listening, allowing participants to share their thoughts and experiences without fear of judgment. After the tutorial, we will actively collect feedback to identify areas for improvement related to diversity and inclusion and share it with future tutorial presenters.

Our presenter team will share our tutorial with a worldwide audience by promoting it on social media, and to diverse research communities. Our presenters include both junior and senior researchers. Thus, we have diversified instructors which will also help encourage diverse audience. Diyi has experience co-organizing Widening NLP Workshops at both NAACL and ACL, and actively works on inviting undergraduate students to research and promoting diversity such as by speaking at AI4ALL and local high-schools at Atlanta. We will work with ACL/NAACL D&I teams, and consult resources such as the BIG directory to diversify our audience participation.

5 Reading List and Prerequisite

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. Here are a few papers that lay a foundation for this area:

- 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);
- Supporting Peer Counselors via AI-Empowered Practice and Feedback (Hsu et al., 2023)

- Evaluating human-language model interaction (Lee et al., 2022b)

Breadth While we will give pointers to dozens of relevant papers over the course of the tutorial, 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.

6 Ethics Statement

Given its strong emphasise on human AI interactions, our tutorial provides insights into the intricate relationship between humans and AIs (e.g., LLMs). In our tutorial, we will provide discussions regarding the capabilities and limitations of LLMs, as well as potential ethical challenges that they might pose, such as around bias, harm and fairness. Our conclusion session will also discuss responsible research design in the space of human-AI interaction, and best practices that can encourage ethical and inclusive uses.

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.

Yuzhe Cai, Shaoguang Mao, Wenshan Wu, Zehua Wang, Yaobo Liang, Tao Ge, Chenfei Wu, Wang You, Ting Song, Yan Xia, et al. 2023. Low-code llm: Visual programming over llms. *arXiv preprint arXiv:2304.08103*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

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

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982*.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph 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 6365–6378.

Mina Lee, Percy Liang, and Qian Yang. 2022a. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022b. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*.

Yunxin Li, Baotian Hu, Xinyu Chen, Lin Ma, and Min Zhang. 2023. Lmeye: An interactive perception network for large language models. *arXiv preprint arXiv:2305.03701*.

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2023. Decision-oriented dialogue for human-ai collaboration. *arXiv preprint arXiv:2305.20076*.

Qianou Ma, Tongshuang Wu, Kenneth Koedinger, et al. 2023. Is ai the better programming partner? human-human pair programming vs. human-ai pair programming. *arXiv preprint arXiv:2306.05153*.

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.

OpenAI. 2023. Gpt-4 technical report.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

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.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended

embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 747–763.

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.

Tianyi Zhang, Isaac Tham, Zhaoyi Hou, Jiaxuan Ren, Liyang Zhou, Hainiu Xu, Li Zhang, Lara J Martin, Rotem Dror, Sha Li, et al. 2023a. Human-in-the-loop schema induction. *arXiv preprint arXiv:2302.13048*.

Yiming Zhang, Sravani Nanduri, Liwei Jiang, Tongshuang Wu, and Maarten Sap. 2023b. Biasx:" thinking slow" in toxic content moderation with explanations of implied social biases. *arXiv preprint arXiv:2305.13589*.