Social Intelligence in the Age of LLMs

Hao Zhu[*], **Bodhisattwa Prasad Majumder**[†], **Dirk Hovy**[◇], **Diyi Yang**[*]
[*]Stanford University [†]Allen Institute for AI [◇]Bocconi University

1 Motivation and Objectives

With the emergence of Large Language Models (LLMs), we now have unprecedented opportunities to incorporate human-like communication and context-aware interactions into artificial systems. But what is the current state of LLMs' capability of social interaction? Can they truly understand social scenarios, perform social reasoning, or interact with humans as socially competent agents? We propose this tutorial as an introduction to and an overview of different aspects of artificial social intelligence and their relationship with LLMs.

In this tutorial, we will explore these questions by introducing scientific methods for evaluating social intelligence in LLMs, highlighting the key challenges, and identifying promising research directions. Participants will not only gain a comprehensive overview of the field's progress, but also acquire technical skills on analysing and developing LLM-based social intelligence.

2 Tutorial Outline

This will be a **three-hour tutorial** devoted to the **cutting-edge topic** of *Social Intelligence and Large Language Models*, divided into four themes. Each theme will take 40 minutes including Q&A. We provide a sample tutorial schedule in Table 1 for a morning tutorial. Each theme 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.

Time	Event	Speaker
09:00-09:05	Opening	Hao Zhu
09:05-09:45	Theme 1	Diyi Yang
09:45-09:50	Break	-
09:50-10:30	Theme 2	Bodhisattwa Prasad Majumder
10:30-10:35	Break	-
10:35-11:15	Theme 3	Dirk Hovy
11:15-11:20	Break	-
11:20-12:00	Theme 4	Hao Zhu

Table 1: Proposed schedule of events.

Theme 1: LLM + Culture and Norm (40 mins)

Social communication takes place within a cultural context, and understanding the influence of culture is important for conflict resolution and decision-making. As LLMs being used to produce vast amounts of data, they reflect and potentially influence social norms and cultural narratives embedded with the data. In this part, we aim to provide a deep dive into how LLMs can facilitate cross-cultural exchange and understanding, as well as how they perpetuate stereotypes and shape societal norms negatively, as a main case study of examining social intelligence in LLMs. We will first share how existing research quantifies cultural awareness (Naous et al., 2023; Ryan et al., 2024) and social norms in and of large language models, and also provide an overview of different multilingual and multicultural evaluations (Bhutani et al., 2024; Myung et al., 2024) that the language technologies communities have worked on. In addition to measuring culture and norm in LLMs, we plan to discuss how to develop more culturally aware and socially aware language technologies by focusing on how one can design culturally grounded objectives (Zhou et al., 2023a), represent culturally relevant concepts and entities as knowledge base (Shi et al., 2024), transfer learning and how diverse cultural and norm preferences are leveraged to align LLMs with such diverse intelligence (Kirk et al., 2024).

Theme 2: Social Reasoning (40 mins)

Social reasoning builds upon the foundations of the theory of mind and beliefs about the world. As LLMs become increasingly anthropomorphized due to their ability to interact like humans, it is crucial to critically investigate their true capabilities. Researchers have been employing various NLP techniques to evaluate LLMs on social reasoning tasks, assessing their ability to infer and interpret observed and perceived beliefs about the social world. We will discuss an array of frame-

Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, pages 51–55

works and benchmarks that aim to evaluate LLM's ability to perform social reasoning. We will investigate methods that tend to improve LLM's intrinsic ability to reason in a social scenario and identify the research gaps that remain in achieving a level of anthropomorphism, tying them to the continuing development of cutting-edge NLP methods to handle diverse social contexts.

Theme 3: Social Demographics (40 mins)

Language is used by people to communicate with other people, and who says something is often as important as what is said (Lynn et al., 2017). Our sociodemographic identity shapes both how we are perceived and how we choose to express ourselves. The meaning of a sentence can change depending on both the speaker's and the listener's sociodemographic profile (Hovy and Yang, 2021). NLP models do so far rarely model sociodemographic aspects of speaker or listener, but there is increasingly research showing the importance of doing so (Hovy, 2015). We will discuss what sociodemographic factors influence language, and in what ways, and how those can be modeled (Lauscher et al., 2022; Soni et al., 2024). We will specifically talk about persona-based prompting in LLMs (Cheng et al., 2023), which allows us now to explicitly incorporate sociodemographics into a model – or does it?

Theme 4: LLM-based Social Agents (40 mins)

Humans' ability to achieve and balance complex, multifaceted social goals in our interactions with others is a crucial part of our social intelligence as a species (Kihlstrom and Cantor, 2000; Tomasello, 2019). In this theme, we will focus on the interactive aspect of social intelligence. We first introduce the various environments where we can ground the social agents in, including text-based simulation environments (Zhou et al., 2023b; Park et al., 2023), virtual environments (Liu et al., 2023), simulated embodied environments (Biswas et al., 2022; team, 2024; Guo et al., 2024), and physical environments (Wang et al., 2024a). We will then introduce various evaluation metrics for social agents in these environments, including qualitative metrics from user surveys, quantitative metrics based on the outcome of the social interaction, and model-based evaluation using large language models. The third part of this theme will introduce various method for improving social intelligence in LLM-based agents, including prompting and in-context learning, *e.g.* Ma et al. (2024), fine-tuning, *e.g.* Wan et al. (2023), and reinforcement learning, *e.g.* Wang et al. (2024b). Finally, we will highlight several applications of social agents, including education (Liu et al., 2024), health (Mukherjee et al., 2024), and policy (Mou et al., 2024), and potential risks of applying social agents in the real world (Matz et al., 2024; Kasirzadeh, 2024). For each of the four parts in this theme, we will deep dive into the 1 to 2 most representative papers, including the technical details, but also offer an overview of the latest progress.

3 Tutorial Presenters

Hao Zhu is a postdoctoral scholar in the Computer Science Department at Stanford University. He got a PhD from Carnegie Mellon University. He focuses on artificial social intelligence, and published relevant papers in various NLP, LLM, ML, CV, Robotics and CogSci conferences. Hao led the organization of the first workshop on theory of mind at ICML 2023.

Bodhisattwa Prasad Majumder Bodhisattwa Prasad Majumder is a Research Scientist at the Allen Institute for AI. One of his key research interests is the development of interactive and communicative agents for social and scientific reasoning. He received his Ph.D. from UC San Diego where he received the CSE Dissertation Award (2024). Bodhisattwa was a keynote speaker at the WordPlay workshop at ACL 2024 and a panelist at the Theory of Mind workshop at ICML 2024. He co-organized the workshop on representation learning and commonsense reasoning at ACL 2022. Bodhisattwa also co-authored a best-selling book on Practical Natural Language Processing that is published in 4 languages and internationally adapted across several universities and organizations.

Dirk Hovy is a Professor in the Computing Sciences Department of Bocconi University. He got a PhD from USC's Information Sciences Institute, and a master's degree in sociolinguistics from Marburg University in Germany. Dirk is interested in what computers can tell us about language, and what language can tell us about society. He has authored over 150 articles on these topics, including 3 best and one outstanding paper awards, and published two textbooks on NLP in Python for social scientists. Dirk has co-founded and organized several workshops (on computational social science,

ethics, cultural aspects, and hate speech in NLP), and was a local organizer for the EMNLP 2017 conference. He leads an ERC Starting Grant project on sociodemographic bias in NLP models.

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.

4 Diversity Considerations

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.

Breadth While we will give pointers to dozens of relevant papers over the course of the tutorial, we plan to cover around 15 research papers in close detail. Only 4 of the "deep dive" papers will come from the presenter team.

1. Having beer after prayer? measuring cultural bias in large language models. (Naous et al., 2023)
2. Unintended impacts of llm alignment on global representation. (Ryan et al., 2024)
3. Seegull multilingual: a dataset of geo-culturally situated stereotypes. (Bhutani et al., 2024)
4. Cross-cultural transfer learning for Chinese offensive language detection. (Zhou et al., 2023a)
5. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. (Kirk et al., 2024)
6. Human centered NLP with user-factor adaptation. (Lynn et al., 2017)
7. Demographic factors improve classification performance. (Hovy, 2015)
8. SocioProbe: What, when, and where language models learn about sociodemographics. (Lauscher et al., 2022)
9. Comparing pretrained human language models: Is it better with human context as groups, individual traits, or both? (Soni et al., 2024)
10. Marked personas: Using natural language prompts to measure stereotypes in language models (Cheng et al., 2023)
11. Becoming human: A theory of ontogeny. (Tomasello, 2019)
12. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. (Zhou et al., 2023b)
13. Computational language acquisition with theory of mind. (Liu et al., 2023)
14. Mosaic: A modular system for assistive and interactive cooking. (Wang et al., 2024a)

6 Tutorial Details

Audience Size We expect the audience size to be around 100 for a physical conference, and around 150 for a virtual conference. Our tutorial will likely bring a similar audience as the NLP+HCI workshop (https://sites.google.com/view/hciandnlp-2021/home?authuser=0).

Open Access We will put the slides, code, and other teaching materials online for public access, as well as consent to adding the video recording of our tutorial in the ACL Anthology.

7 Sharing of Tutorial Material

We plan to make all teaching material available online and agree to allow the publication of slides and video recordings in the ACL anthology.

8 Pedagogical Material

We plan to do some short hands-on exercises to let the audience try to prompt LLMs for the social intelligence tasks introduced in this tutorial.

9 Ethics Statement

When AI is used in social scenarios, the biases and stereotypes might lead to undesirable consequences. The goal of this tutorial is to offer the audience the tools to evaluate these biases and stereotypes and to mitigate them. We would also introduce latest progress of understanding the risks and safety issues of applying LLMs in the real world.

References

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. Seegull multilingual: a dataset of geo-culturally situated stereotypes. *arXiv preprint arXiv:2403.05696*.

Abhijat Biswas, Allan Wang, Gustavo Silvera, Aaron Steinfeld, and Henny Admoni. 2022. Socnavbench: A grounded simulation testing framework for evaluating social navigation. *J. Hum.-Robot Interact.*

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Xudong Guo, Kaixuan Huang, Jiale Liu, Wenhui Fan, Natalia Vélez, Qingyun Wu, Huazheng Wang, Thomas L. Griffiths, and Mengdi Wang. 2024. Embodied llm agents learn to cooperate in organized teams.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Atoosa Kasirzadeh. 2024. Two types of ai existential risk: Decisive and accumulative.

John F Kihlstrom and Nancy Cantor. 2000. Social intelligence. *Handbook of intelligence*, 2:359–379.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. 2024. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Andy Liu, Hao Zhu, Emmy Liu, Yonatan Bisk, and Graham Neubig. 2023. Computational language acquisition with theory of mind. In *The Eleventh International Conference on Learning Representations*.

Jiawen Liu, Yuanyuan Yao, Pengcheng An, and Qi Wang. 2024. Peergpt: Probing the roles of llm-based peer agents as team moderators and participants in children's collaborative learning. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human centered NLP with user-factor adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

Qianou Ma, Hua Shen, Kenneth Koedinger, and Tongshuang Wu. 2024. How to teach programming in the AI era? using LLMs as a teachable agent for debugging. In *International Conference on Artificial Intelligence in Education*.

SC Matz, JD Teeny, Sumer S Vaid, H Peters, GM Harari, and M Cerf. 2024. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1):4692.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation.

Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, Cezanne Bianco, Swapnil Sharma, Rae Lasko, Michelle Voisard, Sanchay Harneja, Darya Filippova, Gerry Meixiong, Kevin Cha, Amir Youssefi, Meyhaa Buvanesh, Howard Weingram, Sebastian Bierman-Lytle, Harpreet Singh Mangat, Kim Parikh, Saad Godil, and Alex Miller. 2024. Polaris: A safety-focused llm constellation architecture for healthcare.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. *arXiv preprint arXiv:2305.14456*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*.

Nikita Soni, Niranjan Balasubramanian, H. Andrew Schwartz, and Dirk Hovy. 2024. Comparing pretrained human language models: Is it better with human context as groups, individual traits, or both? In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 316–328, Bangkok, Thailand. Association for Computational Linguistics.

OpenGenerativeAI team. 2024. Evaluate llms in real time with street fighter iii.

Michael Tomasello. 2019. *Becoming human: A theory of ontogeny*. Harvard University Press.

Yanming Wan, Jiayuan Mao, and Joshua B. Tenenbaum. 2023. Handmethat: Human-robot communication in physical and social environments.

Huaxiaoyue Wang, Kushal Kedia, Juntao Ren, Rahma Abdullah, Atiksh Bhardwaj, Angela Chao, Kelly Y Chen, Nathaniel Chin, Prithwish Dan, Xinyi Fan, et al. 2024a. Mosaic: A modular system for assistive and interactive cooking. In *2nd Workshop on Mobile Manipulation and Embodied Intelligence at ICRA 2024*.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024b. Sotopia-π: Interactive learning of socially intelligent language agents.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023a. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2023b. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents.