Theory of Mind and Application in Educational Context

Effat Farhana[1], Maha Zainab[1], Qiaosi Wang[2], Niloofar Mireshghallah[2], Ramira van der Meulen[3], Max van Duijn[3],
[1]Auburn University, [2]Carnegie Mellon University, [3]Leiden University
{e.farhana, maz0032}@auburn.edu, qiaosiw@andrew.cmu.edu,
niloofar@cmu.edu, {r.van.der.meulen, m.j.van.duijn}@liacs.leidenuniv.nl

Abstract

This tutorial examines the integration of Theory of Mind (ToM) into AI-driven tutoring systems, with a focus on how large language models (LLMs) can represent learners' cognitive and emotional states to enable adaptive, personalized feedback. Participants will learn foundational ToM concepts from cognitive science and psychology and how these ideas can be operationalized in AI systems. We discuss mutual ToM, in which both tutors and learners model each other's mental states, and address challenges including misconception detection, metacognitive modeling, and privacy in data-driven tutoring. The tutorial also includes hands-on demonstrations of machine ToM in programming education using benchmark datasets such as CS1QA and CodeQA. By combining theoretical foundations, empirical insights, and practical exercises, this tutorial will provide an overview of designing human-centered, ethically aware, and cognitively informed AI tutoring systems.

1 Introduction

This tutorial explores the intersection of Theory of Mind (ToM) and AI-driven tutoring systems, focusing on how LLMs can be used to model students' cognitive states. We position this tutorial within the *cutting-edge* category and present a comprehensive overview of *evolution of cognitive modeling* in AI-driven tutoring systems: from rule-based and single-agent systems to reinforcement learning approaches and multi-agent LLM frameworks. Participants will learn about recent advances in this field, along with key limitations including privacy, mutual ToM, and the handling of student misconceptions, as well as practical tools and resources for deploying ready-to-use models in real-world educational NLP applications.

2 Tutorial Outline

The tutorial will be organized into three parts, covering the fundamentals of ToM, reflection, and ToM in educational contexts.

2.1 Part 1: Background (35 min)

Introduction (10 min) We will start with the core principles of ToM from cognitive science and psychology, explaining how humans attribute beliefs, desires, intentions, and emotions to themselves and others (Yeung et al., 2024). We will then discuss the importance of ToM and how ToM can be operationalized in human-centered AI systems (Mutalliyevich et al., 2025) to enhance human-AI interaction (Wang et al., 2024). Finally, we will motivate the integration of ToM into online tutoring systems, highlighting its potential to improve learner engagement, personalization, and adaptive feedback.

Modeling Cognition in AI Tutor (10 min) We will provide an overview of AI tutoring systems and the role of cognitive state modeling in such systems. We will discuss adaptive tutoring systems (Maclellan et al., 2016; MacLellan et al., 2014), emphasizing methods for real-time assessment of student knowledge (Feng et al., 2009), engagement (Baker et al., 2008), and learning difficulties. The discussion will illustrate how AI can provide personalized hints, guidance, and feedback based on learners' cognitive states (Abdelshiheed et al., 2024). We will also examine the challenges in modeling complex learning behaviors (Gao et al., 2023, 2024; Farhana et al., 2022) to give participants a clear understanding of the current capabilities and limitations of existing AI tutoring platforms.

LLM Adaptation in AI Tutoring (15 min) We will focuses on the adaptation of LLMs in educational contexts: examples from programming, mathematics, and other domains where LLMs are used to provide explanations, hints, and interactive feedback. Examples include automated scoring of

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 1–6

math questions using LLM (Morris et al., 2025), automated feedback generation in argumentative writing (Baffour and Crossley, 2024), personalized coach in online learning (Dass et al., 2025). The discussion will highlight the strengths of LLMs in generating contextually appropriate responses while also pointing out limitations in recognizing and responding to students' cognitive states. This sets the stage for exploring research gaps and the integration of ToM with LLM-based systems.

2.2 Part 2: Reflections and Questions (25 min)

In the first **15 minutes**, we will examine the gaps in current LLM-based tutoring systems and opportunities for ToM integration. We will also focus on the limitations of evaluating and measuring ToM in LLM-based systems (Wang et al., 2025; Chen et al., 2025; Ma et al., 2023; Ullman, 2023). Participants will understand implications of this gap and the need for ToM-informed AI in education. We will conclude the session with a **10-minute** reflection and open discussion outlining future research directions for the BEA community to advance human-centered evaluation methods and educational applications.

2.3 Part 3: ToM in Education (120 min)

We will cover core ideas of the tutorial in four interconnected themes: mutual ToM, addressing misconceptions, privacy considerations, and Machine ToM in source code comprehension.

Mutual ToM in AI Tutoring (25 min) We will introduce **mutual ToM**, where both the AI tutor and the learner maintain models of each other's cognitive states (Wang and Goel, 2024; Ashktorab et al., 2025). This reciprocity allows the system to anticipate learner needs while learners understand AI intentions, improving trust, engagement, and learning outcomes. The discussion will cover strategies for implementing mutual ToM in online platforms, including cognitive state modeling, adaptive feedback, and interactive scenarios. Participants will explore how mutual understanding can be operationalized to enhance adaptive learning.

Addressing Misconceptions (25 min) This section examines the identification and correction of learners' misconceptions in AI tutors. We will highlight the importance of integrating ToM principles to ensure that AI tutors do not overgeneralize or misinterpret learner behaviors. Recent research identifies common misconceptions in computational ToM, including the belief that AI should implement ToM as a singular module, that all social interactions require advanced ToM reasoning, that ToM is uniform across individuals, and that current LLMs already possess genuine ToM (van der Meulen et al., 2025). In practice, these insights suggest that LLM-based tutors should selectively apply cognitive modeling, account for variations in learners' backgrounds and knowledge states, and avoid assuming ToM competence based solely on benchmark performance. Participants will learn methods for combining LLM pattern recognition with explicit cognitive modeling to detect misconceptions, deliver targeted feedback, and support students accurate reasoning. The section also highlights how misconception-aware tutoring can enhance learner engagement and outcomes while remaining aligned with human-centered AI principles.

ToM and User Privacy (25 min) Privacy and ToM are deeply intertwined: deciding what to share, with whom, and in what context requires reasoning about others' knowledge, expectations, and interpretation of information. Education is a key domain where this is especially relevant, as learner data is sensitive and shared across multiple stakeholders, though the issue extends beyond education. We organize this section around a progression that mirrors how the field has come to understand the problem: from *whether* models reason about contextual privacy at all, to *what* they should disclose, to *how* users experience disclosures and inferences made on their behalf (Borkar et al., 2025).

We begin with the contextual integrity framing (Mireshghallah et al., 2024a,b), which defines privacy as appropriate information flow rather than secrecy of fixed attributes. ConfAIde (Mireshghallah et al., 2024b) shows that even frontier LLMs leak information in contexts where humans would not, and interventions such as privacy-inducing prompts and chain-of-thought reasoning fail to close the gap. The failure is linked to ToM, as models struggle in tracking knowledge and reasoning mental states of parties involved in an information exchange.

We then turn to persistent memory, which is increasingly central to LLM-based assistants. CIMemories (Mireshghallah et al., 2025) shows that contextual integrity violations compound as memory accumulates and models are deployed across multiple tasks, with GPT-5 violations increasing from 0.1% to 9.6% across tasks and up to 25.1% under repeated prompts. The diagnosed failure mode is a *granularity failure* in which models identify the right information *domain* but cannot

distinguish relevant details from unnecessary disclosures. This naturally raises the next question of what a model *should* disclose, leading to data minimization as an operational principle (Zhou et al., 2025). Recent work shows LLMs are systematically biased toward oversharing and struggle to identify the minimal information they actually need—a capability gap, not just a policy gap.

Finally, we consider the user perspective, focusing on how LLMs not only *disclose* but also *infer* sensitive attributes such as (emotional states, health conditions, life circumstances) from interaction patterns. Recent user-centered (Monteiro et al., 2026) shows that users are more concerned with misrepresentative or unexpectedly shared inferences than with inference itself, and they prefer granular control over how such inferences are generated, stored, and shared. We conclude with best practices for privacy- and ToM-aware systems, particularly in education, emphasizing transparency of inferred states, user control over memory and inference scope, contextual-integrity-aware data flows, and default minimization in prompts/persistent storage.

Q&A and Machine ToM in Source Code Comprehension (45 min) We will present a demonstration of **Machine ToM**, including ToM among AI agents (Zhou et al., 2024) applied to programming and code comprehension tasks (Nikiema et al., 2025). The system models a learner's misunderstanding patterns and selectively intervenes with explanations, or scaffolding (Buçinca et al., 2025). Participants will see how teacher–student LLM architectures simulate human-like tutoring interactions, guiding learners toward better reasoning and problem-solving (Liang et al., 2025). The demo will include examples of adaptive interventions and strategies for evaluating their effectiveness.

3 Specification of the Tutorial

Expected Audience Background The audience is expected to have foundational knowledge of NLP, machine learning, and evaluation methods, including supervised learning, classification, and metrics such as accuracy, precision, and recall. Familiarity with HCI, LLMs, and AI ethics is encouraged for participation in Q&A and discussions. We expect about 70–100 participants, based on growing interest in ToM and attendance at a recent workshop on generative AI and ToM at IJCAI 2025 organized by one of the organizers.

Preferred Venue We strongly prefer ACL 2026 to be held in the U.S. due to visa re-entry uncertainties, as many visa holders—including organizers, favor a U.S. location to reduce travel risks and support broad participation.

Diversity Considerations and Others This tutorial emphasizes in its instructional team, participant access, and content. The instructional team represents multiple career stages, from Ph.D. students to assistant professors, across four institutions on two continents and interdisciplinary backgrounds. To support accessibility, virtual participation will be available via Zoom. Approximately 33% of the session is devoted to related work from the wider research community, as reflected in the schedule and materials. The tutorial requires a standard Python environment with PyTorch. All code, data, and experimental scripts are openly released to facilitate reproducibility `https://github.com/MahaZainab/tom-in-education`.

Reading List To deepen participants' understanding of ToM, we recommend the references cited in this tutorial as well as the readings listed below.

- SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents (Zhou et al., 2024).
- Large language models fail on trivial alterations to theory-of-mind tasks (Ullman, 2023).
- Towards Properly Implementing Theory of Mind in AI: An Account of Four Misconceptions (van der Meulen et al., 2025).
- Towards a Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant (Wang, 2021).
- Framework for a multi-dimensional test of theory of mind for humans and ai systems (Stack et al., 2022).
- Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory (Mireshghallah et al., 2024b).
- CIMemories: A Compositional Benchmark for Contextual Integrity of Persistent Memory in LLMs (Mireshghallah et al., 2025).
- Operationalizing Data Minimization for Privacy-Preserving LLM Prompting (Zhou et al., 2025).
- When Are LLM Inferences Acceptable? User Reactions and Control Preferences for Inferred Personal Information (Monteiro et al., 2026).
- 1-2-3 Check: Enhancing Contextual Privacy in LLM via Multi-Agent Reasoning (Li et al., 2025).

- Mindless Dialogue? A Critical Note on ToM and Communicative Generative Agents (Farhana, 2025).
- Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild (Mireshghallah et al., 2024a).

4 Presenter Biographies

Effat Farhana is Assistant Professor in the Department of Computer Science and Software Engineering at Auburn University. Her work has pioneered designing cog- nitive theory-informed AI systems for personalized learning, designing interpretable ML algorithms, and behavioral data mining within education and autism research. SIn 2021, she was named a Rising Star in Data Science by the University of Chicago for her PhD research. Effat is also an active com- munity builder, serving as co-chair of the AI in Education Track at EAAI/AAAI 2024 and 2025, leading the organization of Theory of Mind work- shop at IJCAI 2025, and serving a Research Fellow at the NSF AI Institute for Adult Learning and On- line Education (ALOE).

Maha Zainab is a Graduate Research Assistant in the Computer Science and Software Engineering Department at Auburn University. She is also a Gavin Graduate Student Fellow. She has participated in multiple international coding competitions and has also won some of them. Her research is in Generative AI, especially LLM evaluation, theory of mind and multi-agent reasoning. Her work is published in an IEEE conference, and her current research focuses on the Theory of Mind in intelligent systems. She has also taught national and international audiences.

Qiaosi Wang (Chelsea) is a Carnegie Bosch Postdoctoral Fellow at Carnegie Mellon University. Her work has proposed and empirically examined the Mutual Theory of Mind interaction paradigm to facilitate communications between online learners and AI agents in various social roles (e.g., teaching assistant, social facilitator) in online education. She has frequently presented her work at AI workshops and led the organization of the Theory of Mind in Human-AI Interaction (ToMinHAI) workshop series at multiple HCI venues (e,g., CHI, CUI).

Niloofar Mireshghallah is a Research Scientist at Meta AI's FAIR Alignment group in San Francisco. Beginning Fall 2026, she will join Carnegie Mellon University's Engineering & Public Policy (EPP) Department and Language Technologies Institute (LTI) as an Assistant Professor. She received her Ph.D. from the CSE department of UC San Diego in 2023. Her research interests are privacy, natural language processing, and the societal implications of ML. She is a recipient of the National Center for Women & IT (NCWIT) Collegiate award in 2020 for her work on privacy-preserving inference, a finalist of the Qualcomm Innovation Fellowship in 2021, and a recipient of the 2022 Rising Star in Adversarial ML award.

Ramira van der Meulen is a Ph.D. student at Leiden University's Institute of Advanced Computer Science. She studies decision-making in Human-Machine Collaboration, with a special interest in its communicative requirements. Her current work focuses on the moment of 'agreement' in scenarios with incomplete information, where humans and AI establish their 'common ground' through varying levels of communication. This work concerns with Theory of Mind: the ability to take the perspective of others' beliefs, intentions and knowledge, and the use of this perspective to make sense of their behaviour and attitudes towards the world.

Max Johannes van Duijn is an Assistant Professor at Leiden University's Institute of Advanced Computer Science (LIACS) and Principal Investigator of the Social Intelligence Modelling [SIM] lab. He is a lecturer in Leiden's Data Science and Artificial Intelligence (DSAI) and Creative Science and Technology (CrIT) education programmes. In 2023 he was elected a member of The Young Academy, where he co-heads the Science and Society track. He is the associate editor for Machine Behaviour and Cognition of *Behaviour*. Research in the SIM lab combines methods from cognitive science, linguistics, and AI to study social intelligence, in particular empathy, perspective-taking, and Theory of Mind. Their work includes modeling these capacities in humans as well as in AI systems such as LLMs. His work is published in English and Dutch, in scholarly as well as popular venues, and has regularly attracted media attention.

5 Ethics Statement

As our tutorial will focus on modeling student cognition and providing personalized feedback, we will address ethical best practices for data collection and algorithmic design. We will emphasize inclusivity and fairness by supporting diverse learners and mitigating biases, aiming to advance safe, transparent, human-centered AI in education.

References

Mark Abdelshiheed, Tiffany Barnes, and Min Chi. 2024. How and when: the impact of metacognitive knowledge instruction and motivation on transfer across intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 34(3):974–1007.

Zahra Ashktorab, Djallel Bouneffouf, Krissy Brimijoin, Rachel Bellamy, Murray Campbell, Arielle Goldberg, Gabriel Enrique Gonzalez, Stephanie Houde, Miao Liu, Dario Andres Silva Moran, and 1 others. 2025. Bridging the gap: Unifying hci & ml perspectives on mutual theory of mind. In *International Joint Conference on Artificial Intelligence*.

Perpetual Baffour and Scott Crossley. 2024. Advances in automating feedback for argumentative writing: Feedback prize as a case study. In *The Routledge international handbook of automated essay evaluation*, pages 303–328. Routledge.

Ryan Baker, Jason Walonoski, Neil Heffernan, Ido Roll, Albert Corbett, and Kenneth Koedinger. 2008. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, 19(2):185–224.

Jaydeep Borkar, Matthew Jagielski, Katherine Lee, Niloofar Mireshghallah, David A Smith, and Christopher A Choquette-Choo. 2025. Privacy ripple effects from adding or removing personal information in language model training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18703–18726.

Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Finale Doshi-Velez, and Krzysztof Z Gajos. 2025. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–25.

Ruirui Chen, Weifeng Jiang, Chengwei Qin, and Cheston Tan. 2025. Theory of mind in large language models: Assessment and enhancement. *arXiv preprint arXiv:2505.00026v1*.

Rahul K Dass, Rochan H Madhusudhana, Erin C Deye, Shashank Verma, Timothy A Bydlon, Grace Brazil, and Ashok K Goel. 2025. Ivy: a hybrid knowledge-based and generative ai coach for explaining procedural skills. In *International Conference on Artificial Intelligence in Education*, pages 233–246. Springer.

Effat Farhana. 2025. Mindless dialogue? a critical note on theory of mind and communicative generative agents. In *Proceedings of the 12th Advances in Cognitive Systems*.

Effat Farhana, Teomara Rutherford, and Collin F Lynch. 2022. Predictive student modelling in an online reading platform. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 12735–12743.

Mingyu Feng, Neil Heffernan, and Kenneth Koedinger. 2009. Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction*, 19(3):243–266.

Ge Gao, Xi Yang, and Min Chi. 2024. Get a head start: On-demand pedagogical policy selection in intelligent tutoring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12136–12144.

Qitong Gao, Ge Gao, Juncheng Dong, Vahid Tarokh, Min Chi, and Miroslav Pajic. 2023. Off-policy evaluation for human feedback. *Advances in Neural Information Processing Systems*, 36:9065–9091.

Wenkai Li, Liwen Sun, Zhenxiang Guan, Xuhui Zhou, and Maarten Sap. 2025. 1-2-3 check: Enhancing contextual privacy in llm via multi-agent reasoning. *arXiv preprint arXiv:2508.07667*.

Keyu Liang, Zhongxin Liu, Chao Liu, Zhiyuan Wan, David Lo, and Xiaohu Yang. 2025. Zero-shot cross-domain code search without fine-tuning. *Proc. ACM Softw. Eng.*, 2(FSE).

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Christopher J Maclellan, Erik Harpstead, Rony Patel, and Kenneth R Koedinger. 2016. The apprentice learner architecture: Closing the loop between learning theory and educational data. *International Educational Data Mining Society*.

Christopher J MacLellan, Kenneth R Koedinger, and Noboru Matsuda. 2014. Authoring tutors with simstudent: An evaluation of efficiency and model quality. In *International conference on intelligent tutoring systems*, pages 551–560. Springer.

Niloofar Mireshghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024a. Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv preprint arXiv:2407.11438*.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. 2024b. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Niloofar Mireshghallah, Neal Mangaokar, Narine Kokhlikyan, Arman Zharmagambetov, Manzil Zaheer, Saeed Mahloujifar, and Kamalika Chaudhuri. 2025. CIMemories: A compositional benchmark for contextual integrity of persistent memory in LLMs. *arXiv preprint arXiv:2511.14937*.

Kyzyl Monteiro, Minjung Park, Alexander Ioffrida, Angelina Sanna, Niloofar Mireshghallah, Yang Wang, and Sauvik Das. 2026. When are LLM inferences acceptable? user reactions and control preferences for inferred personal information. *arXiv preprint arXiv:2605.10013*.

Wesley Morris, Langdon Holmes, Joon Suh Choi, and Scott Crossley. 2025. Automated scoring of constructed response items in math assessment using large language models. *International journal of artificial intelligence in education*, 35(2):559–586.

Dedakhanov Abdumalik Mutalliyevich, S Balapriya, Deepak Dharrao, and Yuldashev Ulugbek Vokhidjon Ugli. 2025. Ai-based intelligent tutoring system for the cognitive development. In *AIP Conference Proceedings*, volume 3306, page 030072. AIP Publishing LLC.

Serge Lionel Nikiema, Jordan Samhi, Abdoul Kader Kaboré, Jacques Klein, and Tegawendé F Bissyandé. 2025. The code barrier: What llms actually understand? *arXiv preprint arXiv:2504.10557*.

Caoimhe Harrington Stack, Effat Farhana, Xinyu Shen, Simeng Zhao, Angela Maliakal, Roxanne Rashedi, Joel Michelson, and Maithilee Kunda. 2022. Framework for a multi-dimensional test of theory of mind for humans and ai systems. In *The Tenth Annual Conference on Advances in Cognitive Systems*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Ramira van der Meulen, Rineke Verbrugge, and Max J. van Duijn. 2025. Towards properly implementing theory of mind in ai: An account of four misconceptions. *arXiv preprint arXiv:2503.16468*.

Qiaosi Wang. 2021. Towards mutual theory of mind in human-ai interaction: How language reflects what students perceive about a virtual teaching assistant.

Qiaosi Wang and Ashok K. Goel. 2024. Mutual theory of mind for human-ai communication. In *Proceedings of the Workshop on Theory of Mind in Human-AI Interaction at CHI 2024*.

Qiaosi Wang, Sarah Walsh, Mei Si, Jeffrey Kephart, Justin D Weisz, and Ashok K Goel. 2024. Theory of mind in human-ai interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Qiaosi Wang, Xuhui Zhou, Maarten Sap, Jodi Forlizzi, and Hong Shen. 2025. Rethinking theory of mind benchmarks for llms: Towards a user-centered perspective. In *Proceedings of the Human-centered Evaluation and Auditing of Language Models Workshop at CHI 2025 (HEAL @ CHI'25 Workshop)*, page 7, New York, NY, USA. ACM.

Elaine Kit Ling Yeung, Ian A Apperly, and Rory T Devine. 2024. Measures of individual differences in adult theory of mind: A systematic review. *Neuroscience & Biobehavioral Reviews*, 157:105481.

Jijie Zhou, Niloofar Mireshghallah, and Tianshi Li. 2025. Operationalizing data minimization for privacy-preserving LLM prompting. *arXiv preprint arXiv:2510.03662*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.