Towards Educational Theory of Mind for Generative AI: A Review of Related Literature and Future Opportunities

Sumit Asthana
University of Michigan
Ann Arbor, USA

Kevyn Collins Thompson
University of Michigan
Ann Arbor, USA

ABSTRACT

Ensuring equitable educational outcomes requires understanding and addressing each student's needs. Expert instructors achieve this by effectively assessing student knowledge gaps through the mistakes that they commit, employing varying degrees of subject-matter, teaching and social knowledge. While wide-spread availability of online courses has greatly expanded equitable access to educational resources internationally, effective student feedback is still limited by instructor availability. With increasing class sizes and higher student-teacher ratios, we need methods to scale up instructor expertise, and also give instructors better training and tools for supporting students at scale. Large Language Models (LLMs) are one tool that show promise in this direction. However, little is known about how effectively LLMs can infer student needs from their behaviors (e.g., test performance, questions asked) the way some human teachers can, especially in arbitrary domains and social situations. We motivate the task of identifying student's knowledge gaps as a form of educational Theory of Mind and provide a literature synthesis of key papers from traditional and more recent educational diagnostics and remediation. We discuss building on this extensive prior work in order to create new, more general effective methods and datasets for assessing and optimizing educational ToM capabilities of generative AI.

ACM Reference Format:
Sumit Asthana and Kevyn Collins Thompson. 2024. Towards Educational Theory of Mind for Generative AI: A Review of Related Literature and Future Opportunities. In *Proceedings of Workshop on Theory of Mind in Human-AI Interaction at CHI 2024 (ToMinHAI at CHI 2024).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Understanding students' knowledge gaps is essential to providing them with individual personalized support in course learning environments [33]. Experienced instructors are able to assess their students' knowledge gaps from their class behaviors (e.g., question asking patterns) and mistakes on test assessments [28]. However, instructors may not have the time or bandwidth to address every individual student's knowledge gaps comprehensively, especially in online environments where class sizes are large, and instruction is remote [24]. With the increasing use of generative AI in education to address the challenge of scale [2], it is important to understand the ability and potential for generative models like LLMs to reliably infer underlying knowledge states and suggest effective remediation actions based on observed student interactions. In social and communicative situations, the ability to infer the intent and mental states of listeners is a key component of people's adaptive behavior and is essential for artificial systems to navigate diverse social contexts [1, 11]. This capability is referred to as Theory of Mind (ToM) in Cognitive Psychology [30].

Instructors in learning environments need to infer their students' latent cognitive states based on their students' observed behaviors (e.g., using a student's questions to better estimate the state of their mental model and reasoning processes). Much of this inference is implicit in their teaching activities. Designing tests for assessment and activities for student discussions all aim to surface student understanding of concepts and their refinement through feedback. On the other hand, artificial agents such as LLMs are not currently built with explicit ToM capabilities [9], but research suggests that LLMs can perform reasonably on at least a few ToM benchmarks [27] even though they may lack higher-order reasoning necessary for complex tasks [14].

In this work, we provide an overview of selected key papers from the rich history of educational research from the lens of Theory of Mind. In this context, we also discuss challenges and opportunities for new forms of benchmark methods that could be used to evaluate and optimize the educational Theory-of-Mind capabilities of LLMs. Such advances would be crucial for key pedagogical tasks such as supporting instructors in creating lesson plans, assessing student preparedness, evaluating student knowledge gaps and progress over time, and providing appropriate adaptive remediation in classroom settings. We also discuss opportunities for leveraging education to develop agents with a social theory of mind.

2 LITERATURE REVIEW

Our current research framework is informed by a rich, decades-long history of work in education, particularly in mathematics education. We provide a short literature review that ties together a broad set of key papers from diverse fields as critical background on which to build future ToM efforts involving generative AI.

A significant fraction of education research has focused on early childhood development. A meta-study by Beaudoin et al. (2020) provides a systematic inventory of existing ToM measures for young children [7], resulting in the synthesis of a new taxonomy of ToM sub-domains they called "Abilities in Theory of Mind Space" (ATOMS). They enumerated seven ToM categories of mental states and social situations: Intentions, Desires, Emotions, Knowledge, Percepts, Beliefs, and mentalistic understanding of non-literal communication, as well as an eighth meta-category "Comprehensive ToM measures" that could include multiple subcategories. While recognizing the potential importance of all of these categories in a learning scenario, work involving educational ToM has tended to focus most on the Knowledge ToM category, in which student

ToMinHAI at CHI 2024, May 12th, Honolulu, Hawaii
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

reasoning errors and misconceptions are diagnosed, understood, and even predicted. We describe that subfield in Sec. 2.3.

2.1 LLMs and theory of mind

In LLMs, exploration of Theory of Mind has relied on benchmarks that explore social scenarios based on the criteria of information asymmetry between parties in a social setting (e.g., a multi-party conversation where some members are privy to more information than others) [35]. LLMs are currently not constructed with sufficient theory of mind capabilities [9] that would allow them to reliably reason about social situations and take effective actions in partnership with humans, including in educational scenarios. Reasoning about the other person's mental state is essential for LLMs as they are being used in social and dyadic interaction scenarios that demand collaboration [50]. One recent example of this trend is an attempt to increase robot navigation effectiveness by including a ToM model [13], which helps the agent plan by reasoning about a user's actions [20]. LLMs, however, when applied to predict human perception of robot behaviors, were found to fail belief tests through simple prompt perturbations [46]. Human-AI interaction scenarios such as virtual assistants for tasks can also benefit from inclusion of ToM models by allowing LLMs to efficiently leverage context [8] and reducing interaction breakdowns [49].

2.2 Pedagogical content knowledge

Subject matter knowledge (SMK) is a necessary requirement for teaching. However, while subject matter knowledge allows us to apply the knowledge correctly, it does not indicate the different ways in which learners with insufficient subject knowledge may approach problems. Understanding how students may incorrectly apply knowledge in tests and real-world scenarios is crucial to providing them with relevant feedback to address those gaps. Shulman et al. described this knowledge as pedagogical content knowledge (PCK) [39]. It allows teachers to connect their teaching experience with subject matter expertise to present their instruction. It was later revised to have more emphasis on knowing and understanding subject and teaching matter as active processes [10]. It encompasses several dimensions of knowledge, such as using the right language for complex concepts [47], ordering concept presentation in the right hierarchy [34], the right difficulty of content for effective recall [6], and managing the cognitive overload by strategizing concept presentation [44]. We also note that Minsky argued that in modeling expertise, we must also consider negative knowledge [31]: in this case, an additional category of learner 'error' in their mental model would be lack of knowledge about what *not* to do. Related to pedagogical knowledge, teacher's beliefs and attitudes also determines the student learning experience. Ernst et. al [15] discusses how teachers belief of mathematics as 'a set of rules and facts' or 'as a dynamic problem-driven field of enqiury' can determine the learning activities they employ, and how they impart knowledge to students. Artificial educational support tools like LLMs need to be aware of such values to better support and align with teaching activities (e.g., creating lesson plans) [26]. To our knowledge, these aspects of LLMs have so far been little explored or evaluated.

2.3 Diagnosis of misconceptions and reasoning errors

Foundational work in mathematics education by Brown, Burton [5] and VanLehn [4] originated Repair Theory, which was an attempt to explain how people learn procedural skills, and especially how and why they make mistakes in acquiring and applying those skills. This view decomposed complex basic math tasks such as subtracting two three-digit numbers into all of its conceptual procedures (subroutines) using a procedural network. Notably, the authors also introduced a chatbot-based game called BUGGY [5] that simulated a student with a particular knowledge 'bug'. BUGGY was used as a training and validation process in which a human teacher had to discover the underlying misconception by providing strategic test problems. While BUGGY used a hand-created procedural network, we believe a compelling and timely update to this idea would be to use recent advances in conversational generative AI to implement a more general advanced version of BUGGY that could extract a more general form of knowledge network and operate in any domain, using a hybrid model that merges LLMs with more traditional student knowledge models. In that scenario, the computer could also serve in an arbiter mode, in which a teacher must describe the student's error in words and also simulate the error by answering the same way a student would. Our research focus also includes a scenario where the roles of the LLM student and human teacher are exchanged so that an AI teacher could attempt to make correct inferences about errors and provide correct guidance for a human student.

The utility of well-defined design methods for diagnostic tests that go beyond just factual correct or incorrect responses has long been recognized. A report by Herman and Webb (1983) on assessing student understanding of science [21] provided a comprehensive landscape of past work in diagnostic assessment. The authors emphasize the need to go beyond summary scores and overall accuracy, with a strong need to identify specific misconceptions (e.g., 'molecular'-level mistakes instead of more 'global' mistakes). They describe the ideal goal of identifying all possible error types. Since a teacher cannot track misconceptions for all students and curriculum areas, let alone tailor instruction, a more realistic goal is to improve the power of diagnostic instruments, 'so they provide more precise but practical information'. A successful diagnostic test should be able to isolate specific errors of reasoning by which consistency could be diagnosed using a rational choice of distractors. In this way, assessments would enable instructors to examine differential patterns of error at different levels of comprehension through orderly structuring of items by the complexity of item content.

Traditionally, diagnostic assessments have started with a comprehensive analysis and definition of the space of possible errors. The work of Brown & Skow (2016) is an excellent representative example that focuses on multiplication but gives a clear, general framework for how to conduct an error analysis [23]. In their words, an error analysis helps a teacher to (1) Identify which steps the student is able to perform correctly (as opposed to simply marking answers either correct or incorrect, something that might mask what it is that the student is doing right); (2) Determine what type(s) of errors a student is making ; (3) Determine whether an error is a

one-time miscalculation or a persistent issue that indicates an important misunderstanding of a mathematics concept or procedure; and (4) Select an effective instructional approach to address the student's misconceptions and to teach the correct concept, strategy, or procedure.

To date, creating new diagnostic assessments of error types in a specific domain has been labor-intensive, with a relatively inflexible result. By retaining domain experts in the loop, but augmenting the assessment generation and selection process using generative AI, we believe there are significant opportunities for efficiently generalizing these traditional educational methodologies to several new domains and populations. By leveraging their ability to synthesize and summarize large amounts of existing literature, LLMs could help systematically explore the error category space in a domain and, in collaboration with an expert, generate an appropriate domain-specific error framework and corresponding diagnostic tests (e.g., appropriate multiple-choice question distractors), for review by experienced teachers. This could be paired with online experimental methods [36] that conduct assessments dynamically for specific students or populations to maximize the information gain from student interaction.

2.4 Social contexts of learning

Another representative point of reference from math education is a study summarizing the reflections of teacher candidates on how they respond to student errors [17]. The authors note the connection with social classroom issues, such as public collective learning, e.g., public discussion of other students' errors, vs. private individual learning. Their work implies that any remediation of student errors, in deciding which action to take and in what manner to address it, must consider not only knowledge-type ToM constructs but also social ToM constructs, especially the context in which the error occurred (public vs private). The authors make a distinction between student-centric actions (peer discussion of erroneous facts) vs teacher-centric (immediate correction). Social ToM constructs within an LLM could help a tutorial system suggest the most effective implementation of a given remediation action predicted by the knowledge-based ToM capabilities of the (same or complementary) LLM system.

2.5 Predicting student mistakes

Predicting common errors is an important aspect of inferring student's knowledge. Recent work in predicting student errors and remediation in mathematical problems [48] suggests that LLMs are still constrained in their ability to diagnose student errors or provide intelligent remediation strategies. Case studies of eliciting common wrong answers in data science [40] and in K-12 mathematics curriculum [19] suggest that even identifying common misconceptions could be a challenging task. Ways to address misconceptions can range from administering multiple-choice questions targeting the misconception to open-ended answers [42] with tradeoffs of expressivity of student response and difficulty to evaluate response. In our opinion this is another research direction for LLM development with significant future potential.

2.6 Augmenting learner-teacher interactions with ToM models

The idea of simulating ToM abilities within a machine learning teaching objective is intended to have the teaching algorithm 'see' the world from the student's point of view. This idea has recently been applied to algorithmic learners, at the level of learner sensory capacities and memory, to improve the 'coaching' usefulness of teacher agents in robotics. Grislain et al. [18] designed a robot maze teaching experiment with a robot teacher guiding a robot learner towards a maze but only having access to the observed behaviors of the learner. They introduced a simple Bayesian 'Theory of Mind' model of the learner's hidden state, which quantified the learner's goals and sensory capacity (maximum distance, vision resolution). In the first phase, a teaching agent observed a learner's behavior (trajectory) in a different maze environment with an assumed known policy in order to estimate the posterior distributions of each learner's ToM model parameters. In the second phase, the teaching agent chose a demonstration from an available set that was tailored to individual learner goals and cognitive (sensory) abilities. The utility of a demo was defined in terms of the learner's goal and sensory capacity (hidden variable estimated by the teacher), and the demo with the maximum expected utility of all available demos was selected to show the learner. As we explain further below, this idea of augmenting teaching models with basic ToM capabilities is a promising area for future research.

3 FUTURE OPPORTUNITIES

In the context of the above related contributions, we summarize a few key areas where we advocate for additional research. This is by no means meant to be a complete list, but is intended to showcase representative problems we believe are high priority for the field.

3.1 New educational ToM assessment benchmarks in arbitrary domains

In addition to tools that can support teachers in identifying common misconceptions, training teachers to identify misconceptions and providing feedback increases access to quality education for more students [16]. Evaluating content knowledge necessary for teaching is essential to provide the right feedback to students when they make mistakes and understand and address common misconceptions they may have [38]. In mathematics, research has investigated measures that not only elicit teachers' understanding of concepts but also whether they can reason about mistakes with the concepts. For example, measures developed by Ball et al. [3] considered whether teachers could come up with examples of mathematical expressions (e.g., mixed fractions), represent them in different ways, or relate them to word problems. Such understanding has been shown to be positively correlated with student learning outcomes [37]. These measures were later condensed into Mathematical Knowledge for Teaching (MKT) measures that capture teacher competencies for effective elementary mathematics instruction [22]. The measures span dimensions such as how teachers represent numbers, how they interpret unusual student answers, assessing material difficulty, and providing individualized feedback to students. We advocate for applying a more general form of such measures to evaluate LLMs for their capability to infer student knowledge gaps,

as one important indicator of their potential applicability to educational courses.

Looking ahead, progress in incorporating educational ToM capabilities broadly in AI systems will require new ways to evaluate different systems on specific educational ToM tasks, in arbitrary domains. With careful use, the same, or related, benchmarks could also be applied to optimize generative AI system performance on educational ToM tasks. We need to build on the deep research contributions described above in areas like mathematics education to leverage LLMs to create dynamic, automated processes for generating such benchmarks.

We advocate for at least two different modalities for these evaluations: new LLM-based approaches that can create a static assessment similar to existing measures such as MKT for math teachers but in any desired domain[1]; and a dynamic assessment anchored in interactive conversation (inspired by BUGGY above) that can capture important pedagogical aspects of learner-teacher interaction, such as the ability to elicit further diagnostic information in a supportive way. Conversation-based benchmarks incorporating social ToM could also assess teacher effectiveness in both public and private teaching scenarios.

3.2 Hybrid LLM-based teaching and learning models

We believe a powerful trend in upcoming educational systems will be the inclusion of educational ToM capabilities in the objective functions used to both assess and optimize underlying AI algorithms. This idea relates to the computer science subfield of machine teaching [51], which has incorporated human cognitive models into machine learning objective functions [32] in order to obtain more effective training schedules for both humans and machines on difficult tasks. Related work in search engines that help people learn has also integrated simplistic cognitive models into machine learning objectives [45]. Following the examples discussed in Sec. 2.6, we expect that hybrid models that join detailed validated cognitive models of student learning with the linguistic capabilities of LLMs could be an effective path toward incorporating both Knowledge- and Social-based ToM capabilities within AI-based educational support systems.

3.3 Curriculum support tools for diverse student backgrounds

Online education provides opportunities for students to follow diverse learning pathways depending on their interests and economic goals. However, traditionally designed curricula may not be the ideal vehicle for supporting such diversity of student backgrounds and goals [41]. For example, while computing (and more recently, AI) are increasingly being used in diverse domains, lack of programming expertise can often be a barrier to learning for individuals from non-CS majors [25]. Designing curricula for students with different backgrounds requires evaluating their understanding of course materials given their prior knowledge [12]. Strategies like explaining core concepts through visualizations and structured programming exercises can reduce the complexity for learners but this may require additional teaching and domain expertise [43]. AI-supported curriculum design workflows can provide instructors insights on what parts of lessons may be difficult for students from different backgrounds and how to restructure them automatically or with instructor guidance. For example, Kross et. al [29] suggest that tools can support designing datasets for data science instruction that contain "correlations, associations, and relationships" in data that are necessary to illustrate the applicability of data science.

[1]Interestingly, the MKT tests contain a mixture of textual and graphical elements, making them an interesting separate benchmark category for vision-oriented tasks.

3.4 Conclusion

While current generative AI systems do not possess comprehensive theory of mind capabilities, it is still useful to evaluate their utility in understanding and predicting student knowledge gaps. In the context of a broad, cross-disciplinary literature review of work in student and teacher assessment in education [3] and computer science, we advocate for new resources and approaches to evaluate and optimize LLMs for their theory of mind capabilities in education – an area we believe with significant future potential to help both teachers and learners, especially at scale.

Acknowledgements

We thank Michael Ion and the anonymous reviewers for their valuable feedback. This research was sponsored in part by a grant from the Michigan Institute for Data Science (MIDAS), with additional support from the University of Michigan School of Information.

REFERENCES

[1] Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.
[2] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* 7, 1 (2023), 52–62.
[3] Deborah Loewenberg Ball. 1990. The mathematical understandings that prospective teachers bring to teacher education. *The elementary school journal* 90, 4 (1990), 449–466.
[4] J.S. Brown and K. VanLehn. 1980. Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science* 4 (1980), 379–426.
[5] John Seely Brown and Richard R. Burton. 1978. Diagnostic Models for Procedural Bugs in Basic Mathematical Skills*. *Cognitive Science* 2, 2 (1978), 155–192. https://doi.org/10.1207/s15516709cog0202_4 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog0202$_4$
[6] Jerome S Bruner. 2009. *The process of education*. Harvard university press.
[7] Beaudoin C, Leblanc É, Gagner C, and Beauchamp MH. 2020. Systematic Review and Inventory of Theory of Mind Measures for Young Children. *Front Psychol.* (2020). https://doi.org/10.3389/fpsyg.2019.02905
[8] Mustafa Mert Çelikok, Tomi Peltola, Pedram Daee, and Samuel Kaski. 2019. Interactive AI with a theory of mind. *arXiv preprint arXiv:1912.05284* (2019).
[9] Yejin Choi. 2022. The curious case of commonsense intelligence. *Daedalus* 151, 2 (2022), 139–155.
[10] Kathryn F Cochran, James A DeRuiter, and Richard A King. 1993. Pedagogical content knowing: An integrative model for teacher preparation. *Journal of teacher Education* 44, 4 (1993), 263–272.
[11] Michael C Corballis and Stephen EG Lea. 1999. *The descent of mind: Psychological perspectives on hominid evolution*. Oxford University Press.
[12] Adrian A. de Freitas and Troy B. Weingart. 2021. I'm Going to Learn What?!? Teaching Artificial Intelligence to Freshmen in an Introductory Computer Science Course. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (Virtual Event, USA) *(SIGCSE '21)*. Association for Computing Machinery, New York, NY, USA, 198–204. https://doi.org/10.1145/3408877.3432530
[13] Sandra Devin and Rachid Alami. 2016. An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 319–326.
[14] Katherine Elkins and Jon Chun. 2020. Can GPT-3 pass a writer's Turing test? *Journal of Cultural Analytics* 5, 2 (2020).
[15] Paul Ernest. 1989. The knowledge, beliefs and attitudes of the mathematics teacher: A model. *Journal of education for teaching* 15, 1 (1989), 13–33.

[16] Beverley Getzlaf, Beth Perry, Greg Toffner, Kimberley Lamarche, and Margaret Edwards. 2009. Effective instructor feedback: Perceptions of online Graduate students. *Journal of Educators Online* 6, 2 (2009), n2.

[17] F Graif, Baldinger E., and Campbell M. [n. d.]. Teacher Candidates' Reflections on Responding to Errors: Exploring Their Vision and Goals. *Math. Educator* 30, 1 ([n. d.]), 3–24. https://files.eric.ed.gov/fulltext/EJ1315098.pdf

[18] Clémence Grislain, Hugo Caselles-Dupré, Olivier Sigaud, and Mohamed Chetouani. 2023. Utility-based Adaptive Teaching Strategies using Bayesian Theory of Mind. arXiv:2309.17275 [cs.LG]

[19] Ashish Gurung, Sami Baral, Kirk P. Vanacore, Andrew A. Mcreynolds, Hilary Kreisberg, Anthony F. Botelho, Stacy T. Shaw, and Neil T. Hefferna. 2023. Identification, Exploration, and Remediation: Can Teachers Predict Common Wrong Answers?. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (, Arlington, TX, USA,) *(LAK2023)*. Association for Computing Machinery, New York, NY, USA, 399–410. https://doi.org/10.1145/3576050.3576109

[20] Maaike Harbers, Karel Van Den Bosch, and John-Jules Meyer. 2009. Modeling agents with a theory of mind. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 2. IEEE, 217–224.

[21] J Herman and N Webb. [n. d.]. Item Structures for Diagnostic Testing: Methodology Project. *Technical Report* ([n. d.]). https://files.eric.ed.gov/fulltext/ED238935.pdf

[22] Heather C Hill, Stephen G Schilling, and Deborah Loewenberg Ball. 2004. Developing measures of teachers' mathematics knowledge for teaching. *The elementary school journal* 105, 1 (2004), 11–30.

[23] Brown J., Skow K., and the IRIS Center. 2016. Mathematics: Identifying and addressing student errors. *Technical Report, IRIS Center* (2016). https://iris.peabody.vanderbilt.edu/wp-content/uploads/pdf_case_studies/ics_matherr.pdf

[24] Mansureh Kebritchi, Angie Lipschuetz, and Lilia Santiague. 2017. Issues and challenges for teaching successful online courses in higher education: A literature review. *Journal of Educational Technology Systems* 46, 1 (2017), 4–29.

[25] Siu-Cheung Kong, William Man-Yin Cheung, and Guo Zhang. 2021. Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds. *Computers and Education: Artificial Intelligence* 2 (2021), 100026.

[26] Osama Koraishi. 2023. Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. *Language Education and Technology* 3, 1 (2023).

[27] Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083* (2023).

[28] Matthew A Kraft, David Blazar, and Dylan Hogan. 2018. The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of educational research* 88, 4 (2018), 547–588.

[29] Sean Kross and Philip J. Guo. 2019. Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300493

[30] Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. 2022. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review. *Frontiers in Artificial Intelligence* 5 (2022), 62.

[31] M. Minsky. 1994. Negative Expertise. *International Journal of Expert Systems* 7, 1 (1994), 13–19. https://web.media.mit.edu/~minsky/papers/NegExp.mss.txt

[32] Kaustubh R Patil, Jerry Zhu, Ł ukasz Kopeć, and Bradley C Love. 2014. Optimal Teaching for Limited-Capacity Human Learners. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/84438b7aae55a0638073ef798e50b4ef-Paper.pdf

[33] Susan Patrick, Maria Worthen, Dale Frost, and Susan Gentz. 2016. Meeting the Every Student Succeeds Act's Promise: State Policy to Support Personalized Learning. *iNACOL* (2016).

[34] Jean Piaget. 1971. The theory of stages in cognitive development. (1971).

[35] François Quesque and Yves Rossetti. 2020. What do theory-of-mind tasks actually measure? Theory and practice. *Perspectives on Psychological Science* 15, 2 (2020), 384–396.

[36] Anna Rafferty, Huiji Ying, and Joseph Williams. 2019. Statistical Consequences of using Multi-armed Bandits to Conduct Adaptive Educational Experiments. *Journal of Educational Data Mining* 11, 1 (Jun. 2019), 47–79. https://doi.org/10.5281/zenodo.3554749

[37] Brian Rowan, Fang-Shen Chiang, and Robert J Miller. 1997. Using research on employees' performance to study the effects of teachers on students' achievement. *Sociology of education* (1997), 256–284.

[38] Meghan Shaughnessy, Rosalie DeFino, Erin Pfaff, and Merrie Blunk. 2021. I think I made a mistake: How do prospective teachers elicit the thinking of a student who has made a mistake? *Journal of Mathematics Teacher Education* 24 (2021), 335–359.

[39] Lee S Shulman. 1986. Those who understand: Knowledge growth in teaching. *Educational researcher* 15, 2 (1986), 4–14.

[40] James Skripchuk, Yang Shi, and Thomas Price. 2022. Identifying Common Errors in Open-Ended Machine Learning Projects. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education - Volume 1* (Providence, RI, USA) *(SIGCSE 2022)*. Association for Computing Machinery, New York, NY, USA, 216–222. https://doi.org/10.1145/3478431.3499397

[41] David H. Smith, Qiang Hao, Filip Jagodzinski, Yan Liu, and Vishal Gupta. 2019. Quantifying the Effects of Prior Knowledge in Entry-Level Programming Courses. In *Proceedings of the ACM Conference on Global Computing Education* (Chengdu,Sichuan, China) *(CompEd '19)*. Association for Computing Machinery, New York, NY, USA, 30–36. https://doi.org/10.1145/3300115.3309503

[42] Soeharto Soeharto, Benő Csapó, Eri Sarimanah, FI Dewi, and Tahmid Sabri. 2019. A review of students' common misconceptions in science and their diagnostic assessment tools. *Jurnal Pendidikan IPA Indonesia* 8, 2 (2019), 247–266.

[43] Elisabeth Sulmont, Elizabeth Patitsas, and Jeremy R. Cooperstock. 2019. Can You Teach Me To Machine Learn?. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education* (Minneapolis, MN, USA) *(SIGCSE '19)*. Association for Computing Machinery, New York, NY, USA, 948–954. https://doi.org/10.1145/3287324.3287392

[44] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.

[45] Rohail Syed and Kevyn Collins-Thompson. 2017. Retrieval Algorithms Optimized for Human Learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 555–564. https://doi.org/10.1145/3077136.3080835

[46] Mudit Verma, Siddhant Bhambri, and Subbarao Kambhampati. 2024. Theory of Mind abilities of Large Language Models in Human-Robot Interaction: An Illusion? *arXiv preprint arXiv:2401.05302* (2024).

[47] Lev S Vygotsky. 2012. *Thought and language*. MIT press.

[48] Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023. Step-by-Step Remediation of Students' Mathematical Mistakes. *arXiv preprint arXiv:2310.10648* (2023).

[49] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[50] Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2023. I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 11136–11155. https://doi.org/10.18653/v1/2023.acl-long.624

[51] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. 2018. An Overview of Machine Teaching. *CoRR* abs/1801.05927 (2018). arXiv:1801.05927 http://arxiv.org/abs/1801.05927