Large Language Models for Education: Understanding the Needs of Stakeholders, Current Capabilities and the Path Forward

Sankalan Pal Chowdhury[1], Nico Daheim[2], Ekaterina Kochmar[3], Jakub Macina[1], Donya Rooein[4], Mrinmaya Sachan[1], Shashank Sonkar[5]
[1]ETH Zurich, [2]TU Darmstadt, [3]MBZUAI, [4]Bocconi University, [5]Rice University
Alphabetical order of presenters, **Correspondence to:** mrinmaya.sachan@inf.ethz.ch

Motivation and Objectives: Recent advancements in Large Language Models (LLMs) have opened unprecedented opportunities in education but the current development goals of LLMs stand in contrast to the requirements of educational applications. This tutorial aims to bridge the gap between two major communities: Natural Language Processing (NLP) researchers and Artificial Intelligence in Education (AIED) practitioners. Our objectives are: (1) to help NLP researchers understand the requirements and challenges of education, enabling them to develop LLMs that align with educational needs, and (2) to enable educators and AIED practitioners to gain a deeper understanding of the capabilities and limitations of current NLP technologies , fostering effective integration of LLMs in educational contexts. By facilitating cross-disciplinary dialog, we aim to uncover the potential of LLMs in education.

First, we identify several critical challenges: *LLMs must be aligned to complement established pedagogical theories and educational practices*, incorporating principles such as scaffolding (Macina et al., 2023b; Sonkar et al., 2024a) or Socratic questioning (Shridhar et al., 2022), effective feedback mechanisms (Daheim et al., 2024), and cognitive load management (Settles and Meeder, 2016). This ensures that AI systems enhance rather than undermine learning processes. We emphasize that LLMs *need to be integrated with existing AIED technologies*, including knowledge tracing models and intelligent tutoring systems (ITS). As highlighted by UNESCO (Miao and Cukurova, 2024), we also need to explore *human-AI collaboration* to preserve human agency while leveraging the benefits of LLMs. The use of LLMs also raises ethical *concerns about data privacy and security* and *fairness* for students, necessitating robust safeguards. Finally, *AI literacy* among educators, students, and policymakers is important for ensuring that stakeholders understand their potential and limitations.

1 Tutorial Overview and Structure

1. LLMs meet AIED (60 min)
Intro to LLMs (20 min)
Learning science, AIED foundations (20 min)
Misalignment b/w LLMs & AIED (20 min)

2. Case Studies & Coffee Break (120 min)
Intelligent Tutoring Systems (30 min)
Coffee break (30 min)
Automated feedback & assessment (20 min)
Content (e.g. problem) generation (20 min)
Student modeling and adaptivity (20 min)

3. Closing Discussion (30 min)
LLM development for education
Human, ethical and societal aspects
Closing remarks

We will begin with an introduction of key LLM technologies and AIED usecases, focussing on the needs of stakeholders in education, such as pedagogy, and opportunities to harness LLMs for education applications. Then, we will outline how these needs stand in contrast with current LLM development which instead focusses on solving correctness. Afterwards, we will delve into a series of case studies that highlight how LLMs can be adapted for: (1) robust, personalized, and scalable conversational tutoring systems; (2) adaptive and personalized content generation of educational material, lesson plans, and assessments; (3) grading and delivery of detailed and personalized feedback on student work. We will examine the current capabilities of LLMs in these domains, discussing recent research findings and practical applications. The tutorial will interweave the applications with critical challenges such as pedagogical alignment, ethical considerations, and human factors in AI-assisted education. We finally conclude with a discussion of LLM development for education that emphasizes human, ethical, and societal aspects.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1–10

2 LLMs Meet AIED

LLM Training & AIED Requirements LLMs offer significant potential in education but require careful tuning to align with pedagogical goals. For instance, LLMs tend to provide direct answers instead of scaffolding learning which can hinder learning (Macina et al., 2023b; Sonkar et al., 2024a). We will first discuss how LLMs are trained using supervised fine-Tuning (SFT) (Wei et al., 2022), instruction tuning, and reinforcement-learning-based optimization methods (Ziegler et al., 2019; Rafailov et al., 2023). Connected to this, we also highlight the shortcomings of current benchmarks (Hendrycks et al., 2020; Cobbe et al., 2021; Hendrycks et al., 2021) that are used to evaluate LLMs, mainly for solving accuracy. Evaluation of AIED systems is different from this, as pedagogical factors play a large role and have dominated the development of educational systems (Graesser et al., 2005). We highlight these educational needs from different perspectives and show how LLM development goals do not align to them. For example, students require space to think and learn, also by making mistakes (Macina et al., 2023a; Sonkar et al., 2024a), and teachers require flexible student simulations (Markel et al., 2023).

Human Factors & Ethical Considerations: Integrating LLMs into educational contexts brings several human-centered challenges that must be addressed to ensure effective and ethical use. For example, teachers are often not included in the development loop (Shankar et al., 2024), but gaining their trust, also through model explainability (Cortez et al., 2024) is important. We will discuss how instructors can be included effectively, for example, to decide, when and which NLP models to use or which inputs to give to the models. We will also discuss how they can modify the generated outcomes as needed (Lu et al., 2023) and prompt architectures to provide responses to MCQs based on student simulations (Lu and Wang, 2024).

The application of LLMs in schools also raises ethical considerations related to attribution, plagiarism, and the potential for AI-generated content to be presented as original work. To address these issues, universities and educational authorities must strengthen and enforce academic integrity policies while educating students about responsible AI use. Promoting awareness and developing guidelines is essential in maintaining the integrity of academic work in the age of GenAI (Okaiyeto et al., 2023).

3 LLMs for Educational Applications

3.1 Intelligent Tutoring Systems (ITSs)

ITSs have long been the focus of AIED developments including systems such as AutoTutor-based (Nye et al., 2014), example-tracing tutors (Aleven et al., 2009) or Cognitive tutor (Anderson et al., 1997). However, they require extensive human authoring. While LLMs hold great promise to overcome this and enable applications like student tutoring (Chen et al., 2024) or teacher training (Gregorcic et al., 2024; Markel et al., 2023). Yet, they still face limitations, such as generating factually incorrect responses or not offering sufficient pedagogy (Sonkar et al., 2023).

In this tutorial, we will cover a range of works that attempt to alleviate these shortcomings, for example, such that use LLMs within structured dialogs (Schmucker et al., 2024; Pal Chowdhury et al., 2024), data-driven approaches to adding scaffolding capabilities (Macina et al., 2023a; Sonkar et al., 2023; Jurenka and et al., 2024), and mitigating hallucinations by adding intermediate reasoning steps for prompted LLMs (Wang et al., 2024b; Daheim et al., 2024). As large amounts of dialog tutoring data can be hard to collect, we will also discuss synthetic data creation methods (Wang et al., 2024a; Chevalier et al., 2024).

Finally, we will touch upon evaluation protocols that, ideally, should include relevant stakeholders and evaluate learning effectiveness. Such studies include using LLMs in real classrooms, for example, for computer science (Nie et al., 2024) or math education (Cheng et al., 2024), or using LLMs as student simulations to evaluate the effectiveness of automatic dialog tutors (Macina et al., 2023a). Such student simulations can also be effective for teacher training (Gregorcic et al., 2024; Wang and Demszky, 2023) and training teaching assistants (Markel et al., 2023).

3.2 Automated Feedback and Assessment

Hint and Feedback mechanisms play an important role in determining learning outcomes. We will discuss studies that show both the potential and limitations of LLMs in generating quality feedback. (McNichols et al., 2024) show fine-tuned LLMs have limited generalization capabilities. Contrarily, (Dai et al., 2024) find GPT-4 outperforms human instructors in important aspects of effective feedback dimensions such as feeding-up, feeding-forward, and process level. However, student dynamics are

complex; (Nazaretsky et al., 2024) highlights a preference for human-generated feedback when students know its source. We will discuss solutions to overcome these challenges such as reinforcement learning (Scarlatos et al., 2024) and LLM-based student simulation models (Phung et al., 2024).

Another important aspect of feedback is its emotional and motivational impact on students. We will discuss the importance of affective feedback (Li et al., 2024a; Baral et al., 2023). We will also explore how LLMs can be used to provide not just cognitive but also emotional support, offering praise (Thomas et al., 2023) and addressing negative self-talk (Thomas et al., 2024). Additionally, we will touch on ongoing efforts to integrate AI-driven emotional assessment in educational settings (Vistorte et al., 2024) to create empathetic learning environments.

Finally, we'll shift our focus to **automatic assessment**. We will review their performance in Automated Short/Long Answer Grading (Kortemeyer, 2023a; Sonkar et al., 2024b) and Automated Essay Grading (AEG) (Mizumoto and Eguchi, 2023), referencing open-source benchmarks (Ruseti et al., 2024; Dzikovska et al., 2013; Blanchard et al., 2013) for these tasks. Next we will summarize some findings on the real-world deployment of LLMs for grading, which show promise despite certain limitations. We will start with studies on math grading (Morris et al., 2024; Gandolfi, 2024) including those which involve handwritten recognition (Liu et al., 2024a). We will also expand the analysis to other subjects like physics (Kortemeyer, 2023b), computer science (Nilsson and Tuvstedt, 2023), and biology (Mackey et al., 2023) to highlight their capabilities and limitation across domains. We will also explore hybrid grading strategies that incorporate human oversight to enhance reliability (Kaya and Cicekli, 2024).

3.3 Educational Content Generation

LLM-generated content serves teachers (e.g., for curating lessons and exercises) and students (e.g., for writing essays and problem-solving). We will examine studies that use controllable generation to adapt LLMs to diverse learners based on difficulty, grade level, and readability score (Rooein et al., 2023; Kew et al., 2023). We will also discuss LLMs in controlled content generation, focusing on readability scores (Imperial and Tayyar Madabushi, 2023) and novel prompting techniques for difficulty assessment (Rooein et al., 2024).

We will also explore strategies to control and align generated questions with students' abilities, expert requirements, and question taxonomies like Bloom's (Elkins et al., 2024; Hwang et al., 2023). We will mention studies on improving adaptability in question generation (Scaria et al., 2024; Wang et al., 2022) and cover methods like PFQS (Li and Zhang, 2024) for improved control by generating answer outlines before question generation. Evaluation of generated educational questions typically involves expert assessments (Scaria et al., 2024; Biancini et al., 2024), while tools like SQUET (Moore et al., 2024) offer automated quality evaluation. However, challenges remain, as studies show GPT models underperforming in evaluating the pedagogical quality of generated questions (Bulathwela et al., 2023).

Finally, we will also discuss multimodal and multilingual LLMs in education – research has demonstrated the effectiveness of multimodal learning in enhancing educational outcomes, e.g., in science (Bewersdorff et al., 2024). These findings are supported by learning theories emphasizing the cognitive benefits of integrating multiple modes of information, such as combining multimodal representations like text and images (Mayer, 2024).

3.4 Adaptivity and Personalization

In this section, we will discuss personalized learning's potential to address diverse student needs, based on educational theories emphasizing tailored learning experiences. We discuss knowledge space theory (Doignon and Falmagne, 1985), Vygotsky's Zone of Proximal Development (Vygotsky, 1978), and Ebbinghaus's memory model (Ebbinghaus, 1913), which have influenced applications like Duolingo's spaced repetition (Settles and Meeder, 2016) and ETS's assessments (Carlson and von Davier, 2017). We then introduce Knowledge Tracing (KT) techniques, from basic Rasch models (Rasch, 1960) and Item Response Theory (IRT) (Lord, 1980) to advanced Bayesian Knowledge Tracing (Corbett and Anderson, 1994) and Deep Knowledge Tracing (Piech et al., 2015).

Traditionally, KT models have focused on question IDs rather than textual content due to dataset limitations. However, the attention mechanism is well-suited for sequence modeling tasks like knowledge tracing. We will cover models such as MCQStudentBert (Parsa Neshaei et al., 2024), AKT (Ghosh et al., 2020), SAKT (Pandey and Karypis, 2019), Dtransformer (Yin et al., 2023), and SAINT

(Choi et al., 2020), which leverage attention mechanisms to capture complex relationships between knowledge components and student interactions. The emergence of datasets with auxiliary information, like XES3G5M (Liu et al., 2024b), has facilitated the application of pre-trained LLMs in KT, as explored in works like (Lee et al., 2024).

LLMs have also expanded the scope of KT by enabling adaptive exercise generation (Cui and Sachan, 2023; Srivastava and Goodman, 2021) and domain-specific modifications to transformer architecture, e.g. SparseKT (Huang et al., 2023) which models student behaviors like forgetting (Im et al., 2023). LLMs have also been used in student simulation models like OKT (Liu et al., 2022), which predicts actual student textual responses. Despite these advances, challenges remain, such as LLMs' limited context windows which hinder capturing long-range learning trajectories (Li et al., 2024b).

4 Vision and Path Forward

AI in education offers significant opportunities but requires careful technical, ethical, regulatory, and pedagogical consideration. Requirements include balancing technology with human agency, inclusion, and diversity (Miao and Cukurova, 2024), addressing privacy (Baraniuk, 2024; Leitner et al., 2019; O'Hara and Straus, 2022) and transparency (Holmes et al., 2022), promoting AI literacy (Su et al., 2023; Su and Yang, 2023), but also developing LLMs that meet pedagogical goals. We aim to build a common ground between various stakeholders, namely policymakers, educators, developers, and researchers, which can form a basis for human-centered AI development in education.

5 Diversity & Inclusion considerations

Our tutorial aims to bring together NLP, LS and AIED researchers as well as practitioners. The tutorial is designed to be understandable to an audience with a range of backgrounds. Our group of presenters is made up of diverse backgrounds, seniority-levels, genders, and affiliations.

6 About the Speakers

Sankalan Pal Chowdhury is a second year PhD student in the ETH-EPFL Joint Doctoral Program for Learning Science, advised by Mrinmaya Sachan and Tanja Käser. His research focuses on improving tutoring abilities of LLMs. His work has been published in EMNLP, TACL and L@S.

Nico Daheim is a third year ELLIS PhD student advised by Iryna Gurevych and Mrinmaya Sachan. He works on making LLMs equitable dialog tutors that provide studentes with personalized opportunities to learn. His works have been published at EMNLP, NAACL, EACL, ICLR and ICML.

Ekaterina Kochmar is an Assistant Professor at the NLP Department at MBZUAI, where she conducts research at the intersection of AI, NLP, and ITSs. She is the current President of SIGEDU and has been involved in organizing BEA since 2013.

Jakub Macina is a fourth year PhD at ETH advised by Mrinmaya Sachan and Manu Kapur. His research focuses on understanding and improving generative models' reasoning and pedagogical capabilities. His work has been published in venues such as ACL, EMNLP, and RecSys.

Donya Rooein is a Postdoc at Bocconi University; her work revolves around leveraging NLP for Education. She explores the synergy between machine learning, linguistics, and practitioner insights to enhance education systems. Her work has been published in different ML, NLP, and AIED venues, including NAACL, WWW, and EdMedia.

Mrinmaya Sachan is an Assistant Professor at ETH Zurich, focusing on NLP and its interface with Education. His group has published relevant research on the challenges of Pedagogy and LLMs, Educational Chatbots and Tutors, Student Modeling and Assessment across various NLP and Education-focused venues.

Shashank Sonkar is a final-year PhD student at Rice University advised by Richard G. Baraniuk. His work focuses on pedagogical alignment of LLMs, learner modeling, and intelligent assessment. His work has been published in EMNLP, COLING, AIED, EDM, and LAK.

7 Type of Tutorial & Target Audience

The tutorial will be **introductory** and present research from the fields of NLP, AIED and learning sciences. We will discuss seminal as well as recent papers to build a common ground for participants. Therefore, we welcome participants from any of these backgrounds. While it is helpful to have knowledge of either NLP / ML or learning sciences, it is not a requirement. The tutorial will be self-contained and welcomes an estimated 50-100 attendees based on recent BEA iterations. We will recommend the attendees a small reading list comprising of papers listed in the appendix.

References

Vincent Aleven, Bruce M McLaren, Jonathan Sewall, and Kenneth R Koedinger. 2009. Example-tracing tutors: A new paradigm for intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 19(2):105–154.

John R Anderson, Albert T Corbett, Kenneth R Koedinger, and Ray Pelletier. 1997. Cognitive tutors: Lessons learned. *The journal of the learning sciences*, 4(2):167–207.

Sami Baral, Anthony F Botelho, Abhishek Santhanam, Ashish Gurung, John Erickson, and Neil T Heffernan. 2023. Investigating patterns of tone and sentiment in teacher written feedback messages. In *International Conference on Artificial Intelligence in Education*, pages 341–346. Springer.

Richard G. Baraniuk. 2024. Mid-scale ri-2: Safeinsights: A national research infrastructure for large-scale learning science and engineering. NSF Award Number 2153481.

Arne Bewersdorff, Christian Hartmann, Marie Hornberger, Kathrin Seßler, Maria Bannert, Enkelejda Kasneci, Gjergji Kasneci, Xiaoming Zhai, and Claudia Nerdel. 2024. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *arXiv preprint arXiv:2401.00832*.

Giorgio Biancini, Alessio Ferrato, and Carla Limongelli. 2024. Multiple-choice question generation using large language models: Methodology and educator insights. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 584–590.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Sahan Bulathwela, Hamze Muse, and Emine Yilmaz. 2023. Scalable educational question generation with pre-trained language models. In *International Conference on Artificial Intelligence in Education*, pages 327–339. Springer.

James E Carlson and Matthias von Davier. 2017. Item response theory. *Advancing human assessment: The methodological, psychological and policy contributions of ETS*, pages 133–178.

Eason Chen, Jia-En Lee, Jionghao Lin, and Kenneth Koedinger. 2024. Gptutor: Great personalized tutor with large language models for personalized learning content generation. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 539–541.

Li Cheng, Ethan Croteau, Sami Baral, Cristina Heffernan, and Neil Heffernan. 2024. Facilitating student learning with a chatbot in an online math learning platform. *Journal of Educational Computing Research*, 62(4):907–937.

Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Toni Annala, Max Jameson Aragon, Arturo Rodríguez Fanlo, Simon Frieder, Simon Machado, et al. 2024. Language models as science tutors. *arXiv preprint arXiv:2402.11111*.

Youngduck Choi, Youngnam Lee, Junghyun Cho, Jineon Baek, Byungsoo Kim, Yeongmin Cha, Dongmin Shin, Chan Bae, and Jaewe Heo. 2020. Towards an appropriate query, key, and value computation for knowledge tracing. In *Proceedings of the seventh ACM conference on learning@ scale*, pages 341–344.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. In *User modeling and user-adapted interaction*, volume 4, pages 253–278. Springer.

S Magalí López Cortez, Mark Josef Norris, and Steve Duman. 2024. Gmeg-exp: A dataset of human-and llm-generated explanations of grammatical and fluency edits. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7785–7800.

Peng Cui and Mrinmaya Sachan. 2023. Adaptive and personalized exercise generation for online language learning. *arXiv preprint arXiv:2306.02457*.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise verification and remediation of student reasoning errors with large language model tutors. *Preprint*, arXiv:2407.09136.

Wei Dai, Yi-Shan Tsai, Jionghao Lin, Ahmad Aldino, Hua Jin, Tongguang Li, Dragan Gašević, and Guanliang Chen. 2024. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, page 100299.

Jean-Paul Doignon and Jean-Claude Falmagne. 1985. Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, 23(2):175–196.

Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, volume 2, pages 263–274.

Hermann Ebbinghaus. 1913. *Memory: A Contribution to Experimental Psychology*. Teachers College, Columbia University, New York.

Sabina Elkins, Ekaterina Kochmar, Jackie CK Cheung, and Iulian Serban. 2024. How teachers can use large language models and bloom's taxonomy to create educational quizzes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23084–23091.

Alberto Gandolfi. 2024. Gpt-4 in education: Evaluating aptness, reliability, and loss of coherence in solving calculus problems and grading submissions. *International Journal of Artificial Intelligence in Education*, pages 1–31.

Aritra Ghosh, Neil Heffernan, and Andrew S Lan. 2020. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2330–2339.

Arthur C Graesser, Patrick Chipman, Brian C Haynes, and Andrew Olney. 2005. Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48(4):612–618.

Bor Gregorcic, Giulia Polverini, and Andreja Sarlah. 2024. Chatgpt as a tool for honing teachers' socratic dialogue skills. *Physics Education*, 59(4):045005.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C Santos, Mercedes T Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, et al. 2022. Ethics of ai in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, pages 1–23.

Shuyan Huang, Zitao Liu, Xiangyu Zhao, Weiqi Luo, and Jian Weng. 2023. Towards robust knowledge tracing models via k-sparse attention. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2441–2445.

Kevin Hwang, Sai Challagundla, Maryam Alomair, Lujie Karen Chen, and Fow-Sen Choa. 2023. Towards ai-assisted multiple choice question generation and quality evaluation at scale: Aligning with bloom's taxonomy. In *Workshop on Generative AI for Education*.

Yoonjin Im, Eunseong Choi, Heejin Kook, and Jongwuk Lee. 2023. Forgetting-aware linear bias for attentive knowledge tracing. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3958–3962.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Irina Jurenka and et al. 2024. Towards responsible development of generative ai for education: An evaluation-driven approach. *Preprint*, arXiv:2407.12687.

Mustafa Kaya and Ilyas Cicekli. 2024. A hybrid approach for automated short answer grading. *IEEE Access*.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. Bless: Benchmarking large language models on sentence simplification. *arXiv preprint arXiv:2310.15773*.

Gerd Kortemeyer. 2023a. Performance of the pre-trained large language model gpt-4 on automated short answer grading. *arXiv preprint arXiv:2309.09338*.

Gerd Kortemeyer. 2023b. Toward ai grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research*, 19(2):020163.

Unggi Lee, Jiyeong Bae, Dohee Kim, Sookbun Lee, Jaekwon Park, Taekyung Ahn, Gunho Lee, Damji Stratton, and Hyeoncheol Kim. 2024. Language model can do knowledge tracing: Simple but effective method to integrate language model and knowledge tracing task. *arXiv preprint arXiv:2406.02893*.

Philipp Leitner, Markus Ebner, and Martin Ebner. 2019. Learning analytics challenges to overcome in higher education institutions. *Utilizing learning analytics to support study success*, pages 91–104.

Hai Li, Wanli Xing, Chenglu Li, Wangda Zhu, and Neil Heffernan. 2024a. Positive affective feedback mechanisms in an online mathematics learning platform. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 371–375.

Kunze Li and Yu Zhang. 2024. Planning first, question second: An llm-guided method for controllable question generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4715–4729.

Xueyi Li, Youheng Bai, Teng Guo, Zitao Liu, Yaying Huang, Xiangyu Zhao, Feng Xia, Weiqi Luo, and Jian Weng. 2024b. Enhancing length generalization for attention based knowledge tracing models with

linear biases. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 5918–5926. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Naiming Liu, Zichao Wang, Richard G Baraniuk, and Andrew Lan. 2022. Gpt-based open-ended knowledge tracing. *arXiv preprint arXiv:2203.03716*.

Tianyi Liu, Julia Chatain, Laura Kobel-Keller, Gerd Kortemeyer, Thomas Willwacher, and Mrinmaya Sachan. 2024a. Ai-assisted automated short answer grading of handwritten university level mathematics exams. *arXiv preprint arXiv:2408.11728*.

Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. 2024b. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. *Advances in Neural Information Processing Systems*, 36.

Frederic M Lord. 1980. *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. Readingquizmaker: a human-nlp collaborative system that supports instructors to design high-quality reading quiz questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 16–27.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023a. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Lingzhi Wang, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023b. Opportunities and challenges in neural dialog tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2357–2372, Dubrovnik, Croatia. Association for Computational Linguistics.

Brendan P Mackey, Razmig Garabet, Laura Maule, Abay Tadesse, James Cross, and Michael Weingarten. 2023. Evaluating chatgpt-4 in medical education: an assessment of subject exam performance reveals limitations in clinical curriculum support for students.

Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.

Richard E Mayer. 2024. The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychology Review*, 36(1):8.

Hunter McNichols, Jaewook Lee, Stephen Fancsali, Steve Ritter, and Andrew Lan. 2024. Can large language models replicate its feedback on open-ended math questions? *arXiv preprint arXiv:2405.06414*.

Fengchun Miao and Mutlu Cukurova. 2024. *AI competency framework for teachers*. UNESCO, Paris. Foreword by Stefania Giannini, UNESCO Assistant Director-General for Education. Includes bibliography.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Steven Moore, Eamon Costello, Huy A Nguyen, and John Stamper. 2024. An automatic question usability evaluation toolkit. In *International Conference on Artificial Intelligence in Education*, pages 31–46. Springer.

Wesley Morris, Langdon Holmes, Joon Suh Choi, and Scott Crossley. 2024. Automated scoring of constructed response items in math assessment using large language models. *International Journal of Artificial Intelligence in Education*, pages 1–28.

Tanya Nazaretsky, Paola Mejia-Domenzain, Vinitra Swamy, Jibril Frej, and K Käser. 2024. Ai or human? evaluating student feedback perceptions in higher education. In *European Conference on Technology Enhanced Learning, ECTEL 2024*.

Allen Nie, Yash Chandak, Miroslav Suzara, Ali Malik, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech. 2024. The gpt surprise: Offering large language model chat in a massive coding class reduced engagement but increased adopters' exam performances. Technical report, Center for Open Science.

Filippa Nilsson and Jonatan Tuvstedt. 2023. Gpt-4 as an automatic grader: The accuracy of grades set by gpt-4 on introductory programming assignments.

Benjamin D Nye, Arthur C Graesser, and Xiangen Hu. 2014. Autotutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24:427–469.

Amy O'Hara and Stephanie Straus. 2022. Privacy preserving technologies in us education. *International Journal of Population Data Science*, 7(3).

Samuel Ariyo Okaiyeto, Junwen Bai, and Hongwei Xiao. 2023. Generative ai in education: To embrace it or not? *International Journal of Agricultural and Biological Engineering*, 16(3):285–286.

Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15.

Shalini Pandey and George Karypis. 2019. A self-attentive model for knowledge tracing. *arXiv preprint arXiv:1907.06837*.

Seyed Parsa Neshaei, Richard Lee Davis, Adam Hazimeh, Bojan Lazarevski, Pierre Dillenbourg, and Tanja Käser. 2024. Towards modeling learner performance with large language models. *arXiv e-prints*, pages arXiv–2403.

Tung Phung, Victor-Alexandru Pădurean, Anjali Singh, Christopher Brooks, José Cambronero, Sumit Gulwani, Adish Singla, and Gustavo Soares. 2024. Automating human tutor-style programming feedback: Leveraging gpt-4 tutor model for hint generation and gpt-3.5 student model for hint validation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 12–23.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, pages 505–513.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Georg Rasch. 1960. *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, Copenhagen.

Donya Rooein, Amanda Cercas Curry, and Dirk Hovy. 2023. Know your audience: Do llms adapt to different age and education levels? *arXiv preprint arXiv:2312.02065*.

Donya Rooein, Paul Röttger, Anastassia Shaitarova, and Dirk Hovy. 2024. Beyond flesch-kincaid: Prompt-based metrics improve difficulty classification of educational texts. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 54–67, Mexico City, Mexico. Association for Computational Linguistics.

Stefan Ruseti, Ionut Paraschiv, Mihai Dascalu, and Danielle S McNamara. 2024. Automated pipeline for multi-lingual automated essay scoring with readerbench. *International Journal of Artificial Intelligence in Education*, pages 1–22.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation. In *International Conference on Artificial Intelligence in Education*, pages 165–179. Springer.

Alexander Scarlatos, Digory Smith, Simon Woodhead, and Andrew Lan. 2024. Improving the validity of automatically generated feedback via reinforcement learning. In *International Conference on Artificial Intelligence in Education*, pages 280–294. Springer.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle &riley: Insights from designing and evaluating a large language model-based conversational tutoring system. In *International Conference on Artificial Intelligence in Education*, pages 75–90. Springer.

Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics*, volume 1, pages 1848–1858.

Shashi Kant Shankar, Gayathri Pothancheri, Deepu Sasi, and Shitanshu Mishra. 2024. Bringing teachers in the loop: Exploring perspectives on integrating generative ai in technology-enhanced learning. *International Journal of Artificial Intelligence in Education*, pages 1–26.

Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4136–4149, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A Design Framework for Building Intelligent Tutoring Systems Based on Learning Science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961.

Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard G Baraniuk. 2024a. Pedagogical alignment of large language models. *arXiv preprint arXiv:2402.05000*.

Shashank Sonkar, Kangqi Ni, Lesa Tran Lu, Kristi Kincaid, John S Hutchinson, and Richard G Baraniuk. 2024b. Automated long answer grading with ricechem dataset. In *International Conference on Artificial Intelligence in Education*, pages 163–176. Springer.

Megha Srivastava and Noah Goodman. 2021. Question generation for adaptive education. *arXiv preprint arXiv:2106.04262*.

Jiahong Su, Davy Tsz Kit Ng, and Samuel Kai Wah Chu. 2023. Artificial intelligence (ai) literacy in early childhood education: The challenges and opportunities. *Computers and Education: Artificial Intelligence*, 4:100124.

Jiahong Su and Weipeng Yang. 2023. Artificial intelligence (ai) literacy in early childhood education: An

intervention study in hong kong. *Interactive Learning Environments*, pages 1–15.

Danielle Thomas, Xinyu Yang, Shivang Gupta, Adetunji Adeniran, Elizabeth Mclaughlin, and Kenneth Koedinger. 2023. When the Tutor Becomes the Student: Design and Evaluation of Efficient Scenario-Based Lessons for Tutors. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, LAK2023, page 250–261, New York, NY, USA. Association for Computing Machinery.

Danielle R Thomas, Jionghao Lin, Shambhavi Bhushan, Ralph Abboud, Erin Gatz, Shivang Gupta, and Kenneth R Koedinger. 2024. Learning and ai evaluation of tutors responding to students engaging in negative self-talk. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 481–485.

Angel Olider Rojas Vistorte, Angel Deroncele-Acosta, Juan Luis Martín Ayala, Angel Barrasa, Caridad López-Granero, and Mariacarla Martí-González. 2024. Integrating artificial intelligence to assess emotions in learning environments: a systematic literature review. *Frontiers in Psychology*, 15:1387089.

Lev Semyonovich Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024a. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9707–9731, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024b. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico. Association for Computational Linguistics.

Rose E Wang and Dorottya Demszky. 2023. Is chatgpt a good teacher coach? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. *arXiv preprint arXiv:2306.03090*.

Xu Wang, Simin Fan, Jessica Houghton, and Lu Wang. 2022. Towards process-oriented, modular, and versatile question generation that meets educational needs. *arXiv preprint arXiv:2205.00355*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Yu Yin, Le Dai, Zhenya Huang, Shuanghong Shen, Fei Wang, Qi Liu, Enhong Chen, and Xin Li. 2023. Tracing knowledge instead of patterns: Stable knowledge tracing with diagnostic transformer. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 855–864, New York, NY, USA. Association for Computing Machinery.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Reading List

We plan to share the following list of representative papers spanning the topics covered in our tutorial to the attendees to generate interest. However, we do not plan to have this as a requirement for attendees:

1. Wei Dai, Yi-Shan Tsai, Jionghao Lin, Ahmad Aldino, Hua Jin, Tongguang Li, Dragan Gašević, and Guanliang Chen. 2024. Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, page 100299

2. Sabina Elkins, Ekaterina Kochmar, Jackie CK Cheung, and Iulian Serban. 2024. How teachers can use large language models and bloom's taxonomy to create educational quizzes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23084–23091

3. Alberto Gandolfi. 2024. Gpt-4 in education: Evaluating aptness, reliability, and loss of coherence in solving calculus problems and grading submissions. *International Journal of Artificial Intelligence in Education*, pages 1–31

4. Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C Santos, Mercedes T Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, et al. 2022. Ethics of ai in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, pages 1–23

5. Irina Jurenka and et al. 2024. Towards responsible development of generative ai for education: An evaluation-driven approach. *Preprint*, arXiv:2407.12687

6. Shashank Sonkar, Kangqi Ni, Sapana Chaudhary, and Richard G Baraniuk. 2024a. Pedagogical alignment of large language models. *arXiv preprint arXiv:2402.05000*

7. Allen Nie, Yash Chandak, Miroslav Suzara, Ali Malik, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech. 2024. The gpt surprise: Offering large language model chat in a massive coding class reduced engagement but increased adopters' exam performances. Technical report, Center for Open Science

8. Xinyi Lu, Simin Fan, Jessica Houghton, Lu Wang, and Xu Wang. 2023. Readingquizmaker: a human-nlp collaborative system that supports instructors to design high-quality reading quiz questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18

9. Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, pages 505–513

10. Fengchun Miao and Mutlu Cukurova. 2024. *AI competency framework for teachers*. UNESCO, Paris. Foreword by Stefania Giannini, UNESCO Assistant Director-General for Education. Includes bibliography