Rethinking AI Evaluation through TEACH-AI: A Human-Centered Benchmark and Toolkit for Evaluating AI Assistants in Education

Shi Ding
Expressive Machinery Lab
Georgia Institute of Technology
Atlanta, GA 30332
sding84@gatech.edu

Brian Magerko
Expressive Machinery Lab
Georgia Institute of Technology
Atlanta, GA 30332
magerko@gatech.edu

Abstract

As generative artificial intelligence (AI) continues to transform education, most existing AI evaluations rely primarily on technical metrics such as BLEU and ROUGE, focusing on accuracy and speed while overlooking human identity, agency, contextual learning processes, and ethical considerations. In this paper, we present *TEACH-AI (Trustworthy and Effective AI Classroom Heuristics)*—a domain-independent, pedagogically grounded, and stakeholder-aligned benchmark framework with measurable indicators and a practical toolkit for guiding the design, development, and evaluation of generative AI systems in educational contexts. Built on an extensive literature review and synthesis, the ten-component assessment framework and toolkit checklist provide a foundation for scalable, value-aligned AI evaluation in education. TEACH-AI rethinks "evaluation" through sociotechnical, educational, theoretical, and applied lenses, engaging designers, developers, researchers, and policymakers across AI and education. Our work invites the community to reconsider what constructs "effective" AI in education and to design model evaluation approaches that promote co-creation, inclusivity, and long-term human, social, and educational impact.

Keywords: Generative AI Assistant, AI evaluation, Benchmark framework, Toolkit

1 Introduction and Related Work

As generative AI systems increasingly become intelligent assistant in learning environments, they challenge the traditional roles of teachers, tools, and humans [1, 2]. AI education interventions are often designed for and evaluated based on the efficacy of the AI technology in terms of its behavior, sensing capabilities, and reasoning [3, 4] centered on agent-human interactions. Rarely do these works involve the broader learning context of their designs and evaluations [5, 6, 7, 8]. Therefore, rather than benchmarking against the status quo or competing models, this article attempts to enable researchers to evaluate how well their AI-based interventions work with a multi-faceted framework and a practical toolkit that captures both the top-down and bottom-up factors related to human success with generative AI tools. We aim to investigate our research question: *What criteria define effective, value-aligned human–AI collaboration in educational settings, and how might these criteria guide the development of the TEACH-AI benchmark and practical toolkit for evaluating generative AI assisting tools?*

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: ResponsibleFM.

1.1 Generative AI in Educational Contexts

Recent research highlights generative AI (GenAI)'s potential to support multiple areas of education, including: a) educational administration by reducing teachers' workload (e.g., auto-grading and instructional content generation)[9]; b) personalized learning[10], c) digital literacy development [11, 12], and d) with growing interest in its role in developing higher order thinking skills[13, 14, 15]. AI-supported learning aligns with theories like Zone of Proximal Development(ZPD)[16] and Constructionism [17], which emphasize contextualized support and human agency. However, GenAI's effectiveness in K–12 and higher education remains underexplored due to ongoing concerns about over-reliance on Generative AI, socio-emotional and creative limitations in AI-generated feedback, and broader ethical risks including bias, factual inaccuracies, trust, equity, and accountability [5, 18].

Addressing these issues requires design and evaluation approaches aligned with human identities, values, and community norms [10, 19, 20]. In response, we propose a stakeholder-aligned, human-centered benchmark framework that emphasizes explainability, value alignment, co-adaptive refinement, and iterative assessment[21, 7, 22]. We consider stakeholder aligned human-centered design and evaluation as an approach that centers educational design around humans' needs, engagement, and cognitive growth in this paper[6, 23, 7].

1.2 Foundational UX and Human-AI Evaluation Frameworks

User experience (UX) evaluation plays a crucial role in assessing the effectiveness and acceptability of educational AI systems, particularly in human-centered contexts. Frameworks typically combine core evaluation criteria such as accuracy, clarity, feedback usefulness or engagement potential [24, 25, 26]. Human evaluation remains critical[27, 28]. However, studies show that AI-generated feedback, although often perceived as an immediate assistant, also poses unique challenges related to explainability, bias, and ethical use [29, 30, 31]; Trust, fairness, academic integrity, and need of AI literacy among educators and humans[32, 33, 34]. A notable limitation in existing UX evaluation research is its emphasis on domain specific systems, such as tutoring systems for STEAM like code.org[15]or writing tasks[35]. Our work addresses this gap by proposing a domain independent UX evaluation framework that can generalize across creative, interdisciplinary learning environments such as Scratch, a block-based programming platform that supports storytelling, games[36]; Teachable Machine, train machine learning models through images, sounds[37, 38], or Earsketch, expressive programming learning platform that support teaching both music and coding [39]. These platforms support diverse, cross-disciplinary learning, but lack standardized frameworks to assess outcomes like adaptability, ethical awareness, human values, and stakeholder alignment. A flexible, domain independent benchmark is needed to capture the broader educational impact of these tools across varied contexts [40, 41]. In this paper, we define "Domain Independent Evaluation" as evaluating AI across multiple subject areas, requiring generalizable, content-neutral metrics[42, 40].

1.3 Benchmarking and Evaluation of Intelligent Tutoring Systems

For decades, traditional Intelligent Tutoring Systems (ITS) have aimed to deliver individualized instruction by modeling student knowledge and guiding problem-solving (inner loop) and instructional sequencing (outer loop) through predefined rule-based decision trees. These systems have shown positive learning outcomes through features such as immediate feedback and adaptivity [4, 43]. However, their reliance on rules limits responsiveness to dynamic learning scenarios and diverse student behaviors. In contrast, generative AI tutors are emerging to address these limitations by using adaptive techniques such as retrieval augmented generation for producing context-aware and coherent response[44, 45], reinforcement learning to optimize teaching strategies based on students feedback [46], deep knowledge tracing for modeling and predicting student understand over time[47], and long-term retention and self-regulated learning over immediate correctness to support deeper learning [25, 48].

Benchmarks are critical for evaluating educational AI systems, offering standardized tasks, datasets, and metrics to assess performance. In this context, we adopt the definition of benchmarking as "a combination of task, dataset, and metric" used to evaluate how AI systems support learning [42, 25]. But most existing benchmarks for large language models (LLMs) focus on general reasoning or factual recall [40], with few targeting pedagogical efficacy in real-world learning contexts [49]. This gap raises concerns about the lack of stakeholder validation and limited alignment with teaching

and learning needs and context [42]. As Shute and Ventura emphasize, educational evaluation must move beyond correctness to include formative, contextual, and human-centered outcomes [25, 28]. Anderson et al. [50] further demonstrated how AI tutors can be benchmarked to support procedural knowledge through structured feedback. Building on this, our work addresses the need for pedagogically grounded, stakeholder-aligned benchmarks that reflect how generative AI supports learning in authentic, situated contexts [6, 7].

Our contribution to this paper is to address these research gaps through proposing the TEACH-AI (Trustworthy and Effective AI Classroom Heuristics) Benchmark Framework—a domain independent, value-aligned human-centered conceptual benchmark, along with a practical toolkit for evaluating generative AI tutors. Informed by a synthesis of over 126 publications, our framework serve as a start point to guide the design and evaluation of pedagogically meaningful AI-driven learning experiences.

2 Methodology

We conducted a scoping review following Arksey and O'Malley's framework [51, 52] to examine how AI agents are evaluated in educational environments. Guided by the question "How are AI agents evaluated in educational environments?", we performed targeted searches across major venues (e.g., CHI, NeurIPS, IDC, AIED) and Google Scholar. In total, we reviewed 126 relevant sources, including **27 conference papers**, **78 journal articles**, and **21 books and gray literature**. These were categorized into three thematic phases: the **pre-LLM era** (pre-2017, focused on early ITS and HCI) with 37 papers, the **transformer era** (2017-2022, marked by the rise of XAI and AI literacy)[53, 34] with 43 papers, and the **generative AI phase** (2023-present, emphasizing co-design and agent collaboration)[54, 55], with 36 papers.

Through iterative coding and synthesis, we identified ten recurring components relevant to human-centered evaluation, including explainability, adaptivity, usability, ethical use, and accessibility. These insights informed a practical toolkit of reflective prompts [56, 7] and a simplified scoring structure inspired by Meadows' leverage points [57]. Regular weekly meetings with a senior faculty advisor facilitated thematic validation and iterative refinement of interpretations, ensuring conceptual rigor and alignment with human-centered AI evaluation principles in both the TEACH-AI benchmark and toolkit design.

3 TEACH-AI Benchmark Framework and Early Design Implications

In this section, we revisit our research question and present an initial benchmark framework along with a practical toolkit, drawing on existing literature, to address what evaluation components construct effective, value-aligned human–AI collaboration in educational domains. We define each component in detail and synthesize these findings into preliminary design implications to inform future benchmark development for generative AI tutoring agents. This benchmark framework adopts a value-sensitive human-centered perspective and structures the analysis to address the gap in existing evaluation approaches by strengthening the focus across cognitive and sociotechnical arguments and offers a foundation for iterative refinement through future research.

3.1 TEACH-AI Evaluation Components

To address the first part of the research question: What criteria define effective, value-aligned human–AI collaboration in education? We first define ten core components that form the basis of our evaluation framework (see Table 1): explainability, helpfulness, adaptivity, consistency, creative exploration, system usability, ethical responsibility, accessibility, workflow, and refinement. We then provide a detailed table outlining sub-components with indicators or metrics, and relevant key references.

Explainability: The agent's ability to present its reasoning and decision making in clear, contextual meaningful, and human-understandable terms[58, 59, 60].

Helpfulness: The extent to which the agent supports educational stakeholders such as teachers and humans' needs in achieving their goals through actionable, pedagogically appropriate assistance [61, 62, 16].

Adaptivity: The system's responsiveness to human preferences, contexts, and needs through personalization and dynamic guidance. This includes flexible exploration to foster humans autonomy and confidence [63, 64, 65, 26].

Consistency: The stability and trustworthy of system outputs under similar conditions and alignment of behavior, languages, and situations across tasks [58, 26].

Learning Exploration: The agent's capacity to foster curiosity, support diverse solution paths, and encourage reflective, open-ended inquiry, long-term human autonomy [6, 66, 16, 67, 26].

System Usability: The effectiveness and ease of interaction that support efficient, intuitive, role-shifting, and error-resistant interactions between users and AI systems [24, 26, 68].

Responsiblity and Ethics: The system's ability to act in alignment with human values, legal, ethical, and educational norms, and cultural sensitivities, even under adversarial conditions. It requires agents to avoid harm, ensure fairness, protect privacy and safeguard student data and voice [69, 70, 31, 71, 40].

Accessibility: The extent to which the system is usable and equitable access by humans with diverse abilities, including those using assistive technologies [72, 73, 55, 74, 75].

Workflow Integration & Stakeholder Coordination: The agent's ability to support multi-steps, human-AI collaboration between teachers, students, and other stakeholders, while maintaining adaptability in a dynamic learning context [55, 24].

Refinement: The system's ability to support iterative improvement through a) the user correcting AI errors, b) user adjustment of vague or biased feedback, and c) ethical traceable revisions [76, 69, 26].

Overall, TEACH-AI benchmark framework address three critical interconnected arguments in evaluation: (1) the agent's capacity for explainability, adaptability, helpfulness, and consistency, including interpretable, context-aware justifications [6, 63], dynamic adaptation to human needs [62, 77, 78, 79], and stable, reliable outputs scross similar conditions [26, 80]; (2) the extent to which the agent fosters creative exploration, emotional engagement, and deep thinking, by scaffolding open-ended problem-solving, supporting divergent approaches, encouraging productive struggle, and enabling transferable learning process cross domains[24, 43, 81, 63, 26]; and (3) the degree to which the agent operates responsibly, accessibly, and is open to refinement, including ethical behavior under adversarial conditions [69, 31, 40, 82], equitable access for diverse humans, and support for iterative improvement through feedback, error recovery, and coordination in multi-step, multi-stakeholder workflows [26, 83].

3.2 TEACH-AI Benchmark: Early Implications

To illustrate how TEACH-AI can inform early-stage evaluation, we outline how the TEACH-AI framework could be applied to evaluate domain-independent generative AI assistants in educational settings [110]. The framework's ten components can be selectively applied depending on research goals, stakeholder roles, and contextual factors. For instance, studies involving a single agent may emphasize components such as helpfulness or explainability, whereas multi-agent settings may prioritize coordination or workflow support. Similarly, accessibility considerations should be adapted based on the characteristics and needs of the target user population.

More broadly, TEACH-AI encourages researchers and designers to reflect on how generative AI systems support education, creativity, values, and human agency. By applying the framework iteratively, practitioners can identify where the system meets expectations and where further refinement is needed, guiding more thoughtful and contextually grounded algorithmic design decisions.

3.3 Practical Toolkit

3.3.1 TEACH-AI Toolkit Development: Checklist Example

We also introduce a preliminary toolkit intended to help practitioners apply TEACH-AI in practice. The toolkit offers a set of reflective questions aligned with each framework component, supporting structured evaluation across different educational and design contexts. Rather than serving as a prescriptive checklist, these prompts help users identify strengths, gaps, and opportunities for improvement in an AI system's behavior and alignment with human-centered values. The goal of

Table 1: TEACH-AI Benchmark Framework: Evaluation Components for Generative AI Assistants

Component	Subcomponents	Indicators / Metrics	Key References
1. Explainability	Reasoning Clarity Traceability Fidelity Interpretability	XAI metrics, trust ratings, task agreement rate, XAI question bank	[58, 84, 85, 86, 40, 63] [87, 24, 88, 29]
2. Helpfulness	Goal support human-aligned pedagogy	Task success rate, hint relevance (knowledge tracing), human modeling accuracy, user ratings (e.g., CAS-UX)	[43, 62, 77, 78, 61]
3. Adaptivity	Personalization Context awareness Controllability	Adaptation rate, System Usability Scale (SUS scores), User Experience Questionnaire (UEQ scores), NASA-TLX, controllability metrics, human modeling accuracy	[79, 64, 89, 80, 40] [24, 74]
4. Consistency	Appropriate determinism Implementation invariance Cross-evaluator reliability	Output stability across runs, inter-coder agreement, threshold tuning precision, value map stability index	[58, 90, 91, 80, 26]
5. Learning Exploration	Creativity Metacognition Transfer Affective & social engagement	Creativity support index (CSI) scores, human modeling accuracy, transfer task performance, LX scale, Self-Determination Theory (SDT) indicators (autonomy, competence, relatedness)	[6, 66, 16, 67, 81] [92, 93, 94, 26, 55]
6. System Usability	Usability Interface quality Co-regulation support	CSI scores, usability heuristics checklist, interface clarity, feedback visibility, co-regulation cues rating(e.g., tutoring role-switching affordances), cognitive walkthrough analysis	[81, 68, 24, 95, 26, 96, 97]
7. Responsiblity & Ethics	Fairness Transparency Privacy compliance	Fairness stress tests, adversarial prompt handling, stakeholder alignment, traceability, privacy compliance audit, group fairness metrics	[69, 98, 40, 70, 31, 99]
8. Accessibility	Functional adaptation Assistive technology integration multi-modal UX compatibility	Text-to-audio adaptation rate, comprehension scores, error resolution rate, contextual navigability (e.g., keyboard and curriculum switching), XAI question bank, group fairness metrics	[73, 100, 74, 101, 99] [102, 72, 82]
9. Workflow & Coordination	Multi-agent coordination Multi-role flow control (teachers, students, stakeholders)	Workflow coherence score, task decomposition rate, planning cost analysis, human–agent alignment	[103, 104, 41, 83]
10. Refinement	Iterative feedback Error correction Ethical traceability	Error detection rate for revisions, refinement trace logs (keystroke logs, edit history), cross-agent coherence, time-to-refine, value alignment over revisions, longitudinal user satisfaction	[105, 69, 76, 106] [107, 108, 109]

this tool is to guide consistent reflection and comparison across contexts, whether in classroom use, design, or model development reviews, or early research prototyping. Future iterations will refine these prompts and explore ways to support broader, scalable evaluation workflows. This checklist can be used by educators, researchers, and designers to assess human-centered AI alignment.

Table 2: TEACH-AI Toolkit Checklist Example (Index + Tech-Eval depth forthcoming).

Framework	Checklist Questions
1. Explainability	a. Does the AI explain its decisions clearly? b. How would you verify the explanation?
2. Helpfulness	a. Does the AI support the task goal you set?
3. Adaptivity	a. How quickly does the AI provide adaptive feedback (e.g., latency)?
4. Consistency	a. Does the system behave consistently across different conditions, prompts, or contexts?
5. Learning Exploration	a. Does the AI foster creativity, learning transfer, and critical thinking? b. Does the AI foster emotional connection, motivation, and social reasoning?
6. System Usability	a. Was the AI easy to use and navigate?
7. Responsibility & Ethics	a. Does the agent support diverse users equitably? b. Has a safety or privacy audit been conducted?
8. Accessibility	a. Does the system adapt effectively to my accessibility needs?
9. Workflow & Coordination	a. Were roles and responsibilities clear throughout the collaboration process?
10. Refinement	a. Does the feedback clearly identify areas that need correction?

The checklist is intended to support reflective practice rather than function as a prescriptive to-do list. It translates abstract values (e.g., explainability) into actionable criteria that can be applied across technical design, policymaking, training, and research contexts [20]. In classroom settings, including those using tools like ChatGPT, the checklist guide scalable evaluation by allowing raters to assess each criterion using either a simple Yes/No option or a progressive scale [57]. This approach

provides a clear foundation for assessing an AI system's alignment with human values, contextual demands, and broad AI development principles such as transparency and safety across diverse educational environments. The resulting *TEACH-AI index* can support both reflective classroom practice and quantitative research analysis, supporting consistent comparisons and guiding ongoing model evaluation efforts in the educational domain.

4 Conclusion and future work

In summary, we introduce TEACH-AI, a ten-component, human-centered benchmark framework and toolkit for evaluating generative AI systems in education. While the current version is primarily conceptual, it highlights the need for evaluation approaches that align with emerging educational needs, ethical design principles, and human values. Importantly, TEACH-AI bridges human-generated feedback and LLM-generated feedback by providing a unified structure that supports both human evaluators and LLM-as-judge methods.

Moving forward, our work will involve co-design with diverse stakeholders and iterative refinement of the framework across different educational contexts. We also plan to explore technical development, such as integrating TEACH-AI into a scalable digital prototype for large-scale benchmarking. This direction aligns with broader trends in human-centered AI evaluation, for example, the use of LLM-as-judge methods for automated assessment, and research on Reinforcement Learning from AI Feedback (RLAIF), which highlights the growing emphasis on reliable feedback signals in AI behavior. Our long-term goal is to support the development of accessible, responsible, and pedagogically aligned AI evaluation ecosystems that drive meaningful impact in real educational settings.

5 References

References

[1] Shamini Shetye. An evaluation of khanmigo, a generative ai tool, as a computer-assisted language learning app. *Studies in Applied Linguistics and TESOL*, 24(1), 2024.

[2] Danielle R Thomas, Erin Gatz, Shivang Gupta, Jionghao Lin, Cindy Tipper, and Kenneth R Koedinger. Using generative ai to provide feedback to adult tutors in training and assess real-life performance. In *The Learning Ideas Conference*, pages 204–214. Springer, 2024.

[3] Fabian Sperrle, Mennatallah El-Assady, Grace Guo, Rita Borgo, D Horng Chau, Alex Endert, and Daniel Keim. A survey of human-centered evaluations in human-centered machine learning. In *Computer Graphics Forum*, volume 40, pages 543–568. Wiley Online Library, 2021.

[4] Kurt VanLehn. The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265, 2006.

[5] Mitchel Resnick. Generative ai and creative learning: Concerns, opportunities, and choices. 2024.

[6] David Williamson Shaffer and Mitchel Resnick. " thick" authenticity: New media and authentic learning. *Journal of interactive learning research*, 10(2):195–216, 1999.

[7] Batya Friedman, Peter Kahn, and Alan Borning. Value sensitive design: Theory and methods. *University of Washington technical report*, 2(8):1–8, 2002.

[8] Tom McKlin, Brian Magerko, Taneisha Lee, Dana Wanzer, Doug Edwards, and Jason Freeman. Authenticity and personal creativity: How earsketch affects student persistence. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pages 987–992, 2018.

[9] Safinah Ali, Hae Won Park, and Cynthia Breazeal. A social robot's influence on children's figural creativity during gameplay. *International Journal of Child-Computer Interaction*, 28:100234, 2021.

[10] Stéphan Vincent-Lancrin and Reyer Van der Vlies. Trustworthy artificial intelligence (ai) in education: Promises and challenges. *OECD education working papers*, (218):0_1–17, 2020.

[11] Yoram Eshet. Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of educational multimedia and hypermedia*, 13(1):93–106, 2004.

[12] Alan Bundy. Preparing for the future of artificial intelligence, 2017.

[13] John Baer. The case for domain specificity of creativity. *Creativity research journal*, 11(2):173–177, 1998.

[14] Teresa M Amabile. *Creativity in context: Update to the social psychology of creativity*. Routledge, 2018.

[15] Shuchi Grover and Roy Pea. Computational thinking in k–12: A review of the state of the field. *Educational researcher*, 42(1):38–43, 2013.

[16] Lev Semenovich Vygotsky and Michael Cole. *Mind in society: Development of higher psychological processes*. Harvard university press, 1978.

[17] Seymour Papert and Idit Harel. Situating constructionism. *constructionism*, 36(2):1–11, 1991.

[18] Eman A Alasadi and Carlos R Baiz. Generative ai in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education*, 100(8):2965–2971, 2023.

[19] Katherine Ash and Madelyn Rahn. Reimagining workforce policy in the age of disruption: A state guide for preparing the future workforce now. *National Governors Association*, 2020.

[20] Batya Friedman and David G Hendry. *Value sensitive design: Shaping technology with moral imagination*. Mit Press, 2019.

[21] Ariel Han and Zhenyao Cai. Design implications of generative ai systems for visual storytelling for young learners. In *Proceedings of the 22nd annual ACM interaction design and children conference*, pages 470–474, 2023.

[22] Uday Mittal, Siva Sai, Vinay Chamola, et al. A comprehensive review on generative ai for education. *IEEE Access*, 2024.

[23] Ben Shneiderman. *Human-centered AI*. Oxford University Press, 2022.

[24] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[25] Valerie Shute and Matthew Ventura. *Stealth assessment: Measuring and supporting learning in video games*. The mit press, 2013.

[26] Jakob Nielsen. Ten usability heuristics for user interface design. Online; accessed July X, 2025, 1995.

[27] Lucille Alice Suchman. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press, 2007.

[28] Danielle R Thomas, Jionghao Lin, Erin Gatz, Ashish Gurung, Shivang Gupta, Kole Norberg, Stephen E Fancsali, Vincent Aleven, Lee Branstetter, Emma Brunskill, et al. Improving student learning with hybrid human-ai tutoring: A three-study quasi-experimental investigation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 404–415, 2024.

[29] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15, 2020.

[30] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14, 2018.

[31] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[32] Blerta Abazi Chaushi, Besnik Selimi, Agron Chaushi, and Marika Apostolova. Explainable artificial intelligence in education: A comprehensive review. In *World Conference on Explainable Artificial Intelligence*, pages 48–71. Springer, 2023.

[33] Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in education and teaching international*, 61(2):228–239, 2024.

[34] Duri Long and Brian Magerko. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–16, 2020.

[35] Ruth ONeill and Alex Russell. Stop! grammar time: University students' perceptions of the automated feedback program grammarly. *Australasian Journal of Educational Technology*, 35(1), 2019.

[36] Mitchel Resnick and David Siegel. A different approach to coding. *International Journal of People-Oriented Programming*, 4(1):1–4, 2015.

[37] Abigail Zimmermann-Niefield, Makenna Turner, Bridget Murphy, Shaun K Kane, and R Benjamin Shapiro. Youth learning machine learning through building models of athletic moves. In *Proceedings of the 18th ACM international conference on interaction design and children*, pages 121–132, 2019.

[38] David Touretzky, Fred Martin, Deborah Seehorn, Cynthia Breazeal, and Tess Posner. Special session: Ai for k-12 guidelines initiative. In *Proceedings of the 50th ACM technical symposium on computer science education*, pages 492–493, 2019.

[39] Brian Magerko, Jason Freeman, Tom Mcklin, Mike Reilly, Elise Livingston, Scott Mccoid, and Andrea Crews-Brown. Earsketch: A steam-based approach for underrepresented populations in high school computer science education. *ACM Transactions on Computing Education (TOCE)*, 16(4):1–25, 2016.

[40] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[41] Frank F Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, et al. Theagentcompany: benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*, 2024.

[42] Anka Reuel-Lamparth, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813, 2024.

[43] Kenneth R Koedinger, John R Anderson, William H Hadley, and Mary A Mark. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8:30–43, 1997.

[44] Suqing Liu, Zezhu Yu, Feiran Huang, Yousef Bulbulia, Andreas Bergen, and Michael Liut. Can small language models with retrieval-augmented generation replace large language models when learning computer science? In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*, pages 388–393. 2024.

[45] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[46] Min Chi, Kurt VanLehn, Diane Litman, and Pamela Jordan. An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *International Journal of Artificial Intelligence in Education*, 21(1-2):83–113, 2011.

[47] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.

[48] Ido Roll and Philip H Winne. Understanding, evaluating, and supporting self-regulated learning using learning analytics. *Journal of Learning Analytics*, 2(1):7–12, 2015.

[49] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

[50] John R Anderson, C Franklin Boyle, Albert T Corbett, and Matthew W Lewis. Cognitive modeling and intelligent tutoring. *Artificial intelligence*, 42(1):7–49, 1990.

[51] Hilary Arksey and Lisa O'malley. Scoping studies: towards a methodological framework. *International journal of social research methodology*, 8(1):19–32, 2005.

[52] Danielle Levac, Heather Colquhoun, and Kelly K O'brien. Scoping studies: advancing the methodology. *Implementation science*, 5:1–9, 2010.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[54] Paul Denny, Sumit Gulwani, Neil T Heffernan, Tanja Käser, Steven Moore, Anna N Rafferty, and Adish Singla. Generative ai for education (gaied): Advances, opportunities, and challenges. *arXiv preprint arXiv:2402.01580*, 2024.

[55] Prajish Prasad, Rishabh Balse, and Dhwani Balchandani. Exploring multimodal generative ai for education through co-design workshops with students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2025.

[56] Richmond Y Wong, Michael A Madaio, and Nick Merrill. Seeing like a toolkit: How toolkits envision the work of ai ethics. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–27, 2023.

[57] Donella Meadows. Leverage points. *Places to Intervene in a System*, 19:28, 1999.

[58] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[59] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[60] Riccardo Guidotti and Salvatore Ruggieri. On the stability of interpretable models. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2019.

[61] Lawal Ibrahim Dutsinma Faruk, Debajyoti Pal, Suree Funilkul, Thinagaran Perumal, and Pornchai Mongkolnam. Introducing casux: A standardized scale for measuring the user experience of artificial intelligence based conversational agents. *International Journal of Human–Computer Interaction*, 41(9):5274–5298, 2025.

[62] Xavier Puig, Tianmin Shu, Shuang Li, Zilin Wang, Yuan-Hong Liao, Joshua B Tenenbaum, Sanja Fidler, and Antonio Torralba. Watch-and-help: A challenge for social perception and human-ai collaboration. *arXiv preprint arXiv:2010.09890*, 2020.

[63] John D Bransford, Ann L Brown, Rodney R Cocking, et al. *How people learn*, volume 11. Washington, DC: National academy press, 2000.

[64] Ilie Gligorea, Marius Cioca, Romana Oancea, Andra-Teodora Gorski, Hortensia Gorski, and Paul Tudorache. Adaptive learning using artificial intelligence in e-learning: A literature review. *Education Sciences*, 13(12):1216, 2023.

[65] Catherine Mulwa, Seamus Lawless, Mary Sharp, and Vincent Wade. The evaluation of adaptive and personalised information retrieval systems: a review. *International Journal of Knowledge and Web Intelligence*, 2(2-3):138–156, 2011.

[66] Seymour Papert. *The children's machine: Rethinking school in the age of the computer*. Basic Books, Inc., 1993.

[67] Manu Kapur. Productive failure. *Cognition and instruction*, 26(3):379–424, 2008.

[68] Sherry Turkle. *Life on the Screen*. Simon and Schuster, 2011.

[69] Han Wang, An Zhang, Nguyen Duy Tai, Jun Sun, Tat-Seng Chua, et al. Ali-agent: Assessing llms' alignment with human values via agent-based evaluation. *Advances in Neural Information Processing Systems*, 37:99040–99088, 2024.

[70] Wayne Holmes. Artificial intelligence in education. In *Encyclopedia of education and information technologies*, pages 88–103. Springer, 2020.

[71] Prem Lata. Beyond algorithms: Humanizing artificial intelligence for personalized and adaptive learning. *International Journal of Innovative Research in Engineering and Management*, 11(5):10–55524, 2024.

[72] World Wide Web Consortium et al. W3c web content accessibility guidelines (wcag 2.0). *Internet]. World Wide Web Consortium. Accessed*, 22, 2012.

[73] Diana Ruiz and Tom Duenas. Towards inclusive ai: Developing a w3c-inspired accessibility benchmark for large language models. *Research Gate*, 2024.

[74] Emma Goldenthal, Jennifer Park, Sunny X Liu, Hannah Mieczkowski, and Jeffrey T Hancock. Not all ai are equal: Exploring the accessibility of ai-mediated communication technology. *Computers in Human Behavior*, 125:106975, 2021.

[75] Shi Ding, Jason Brent Smith, Stephen Garrett, and Brian Magerko. Redesigning earsketch for inclusive cs education: A participatory design approach. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*, pages 720–724, 2024.

[76] Jiayi Pan, Yichi Zhang, Nicholas Tomlin, Yifei Zhou, Sergey Levine, and Alane Suhr. Autonomous evaluation and refinement of digital agents. *arXiv preprint arXiv:2404.06474*, 2024.

[77] Anaïs Tack and Chris Piech. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *arXiv preprint arXiv:2205.07540*, 2022.

[78] Vedant Bahel, Harshinee Sriram, and Cristina Conati. Personalizing explanations of ai-driven hints to users' cognitive abilities: an empirical evaluation. *arXiv preprint arXiv:2403.04035*, 2024.

[79] Peter Brusilovsky, Charalampos Karagiannidis, and Demetrios Sampson. Layered evaluation of adaptive learning systems. *International Journal of Continuing Engineering Education and Life Long Learning*, 14(4-5):402–421, 2004.

[80] Vero Vanden Abeele, Erik Hauters, and Bieke Zaman. Increasing the reliability and validity of quantitative laddering data with ladderux. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 2057–2062. 2012.

[81] Erin Cherry and Celine Latulipe. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(4):1–25, 2014.

[82] Shi Ding, Jason Brent Smith, and Brian Magerko. Considering large language model integration in expressive computer science learning environments for blind and visually impaired learners through co-design. In *International Conference on Artificial Intelligence in Education*, pages 472–480. Springer, 2025.

[83] Omar Elnaggar and Roselina Arelhi. Quantification of knowledge exchange within classrooms: an ai-based approach. In *The European Conference on Education*, pages 1–11, 2021.

[84] Avi Rosenfeld. Better metrics for evaluating explainable artificial intelligence. In *Proceedings of the 20th international conference on autonomous agents and multiagent systems*, pages 45–50, 2021.

[85] Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction*, 39(7):1390–1404, 2023.

[86] Mireia Ribera and Agata Lapedriza. Can we do better explanations? a proposal of user-centered explainable ai. CEUR Workshop Proceedings, 2019.

[87] Abdallah MH Abbas, Khairil Imran Ghauth, and Choo-Yee Ting. User experience design using machine learning: a systematic review. *IEEE Access*, 10:51501–51514, 2022.

[88] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[89] Sebastian AC Perrig, Lena Fanya Aeschbach, Nicolas Scharowski, Nick von Felten, Klaus Opwis, and Florian Brühlmann. Measurement practices in user experience (ux) research: A systematic quantitative literature review. *Frontiers in Computer Science*, 6:1368860, 2024.

[90] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[91] Marko Robnik-Šikonja and Marko Bohanec. Perturbation-based explanations of prediction models. In *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pages 159–175. Springer, 2018.

[92] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. Co-designing a real-time classroom orchestration tool to support teacher-ai complementarity. *Grantee Submission*, 2019.

[93] Yassine Safsouf, Khalifa Mansouri, and Franck Poirier. Design of a new scale to measure the learner experience in e-learning systems. In *Proceedings of the International Conference on E-Learning*, pages 301–304. ERIC, 2019.

[94] Richard M Ryan and Edward L Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist*, 55(1):68, 2000.

[95] Nigel Bevan. Human-computer interaction standards. In *Advances in human factors/ergonomics*, volume 20, pages 885–890. Elsevier, 1995.

[96] Allyson Fiona Hadwin, Sanna Järvelä, and Mariel Miller. Self-regulated, co-regulated, and socially shared regulation of learning. *Handbook of self-regulation of learning and performance*, 30:65–84, 2011.

[97] Peter G Polson, Clayton Lewis, John Rieman, and Cathleen Wharton. Cognitive walkthroughs: a method for theory-based evaluation of user interfaces. *International Journal of man-machine studies*, 36(5):741–773, 1992.

[98] Selçuk Kılınç. Comprehensive ai assessment framework: Enhancing educational evaluation with ethical ai integration. *Journal of Educational Technology and Online Learning*, 7(4-ICETOL 2024 Special Issue):521–540, 2024.

[99] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[100] Meredith Ringel Morris. Ai and accessibility. *Communications of the ACM*, 63(6):35–37, 2020.

[101] Normala Mohamad, Nor Laily Hashim, Husna Mad Baguri, Hazirah Abdul Pisal, Cik Fazilah Hibadullah, and Nur Hani Zulkifli Abai. Ux metrics of mobile learning for deaf children using fuzzy delphi method. In *2021 IEEE International Conference on Computing (ICOCO)*, pages 309–314. IEEE, 2021.

[102] Selwyn Goldsmith. *Universal design*. Routledge, 2007.

[103] Xiangyuan Xue, Zeyu Lu, Di Huang, Zidong Wang, Wanli Ouyang, and Lei Bai. Comfybench: Benchmarking llm-based agents in comfyui for autonomously designing collaborative ai systems. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24614–24624, 2025.

[104] Shuofei Qiao, Runnan Fang, Zhisong Qiu, Xiaobin Wang, Ningyu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Benchmarking agentic workflow generation. *arXiv preprint arXiv:2410.07869*, 2024.

[105] Shuchen Guo, Ehsan Latif, Yifan Zhou, Xuan Huang, and Xiaoming Zhai. Using generative ai and multi-agents to provide automatic feedback. *arXiv preprint arXiv:2411.07407*, 2024.

[106] Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q Vera Liao, and Jennifer Wortman Vaughan. Generation probabilities are not enough: Improving error highlighting for ai code suggestions. In *HCAI Workshop at NeurIPS*, 2022.

[107] CJ Bryant, Mariano Felice, and Edward Briscoe. Automatic annotation and evaluation of error types for grammatical error correction. Association for Computational Linguistics, 2017.

[108] Shengxin Hong, Chang Cai, Sixuan Du, Haiyue Feng, Siyuan Liu, and Xiuyi Fan. " my grade is wrong!": A contestable ai framework for interactive feedback in evaluating student essays. *arXiv preprint arXiv:2409.07453*, 2024.

[109] Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, et al. Towards responsible development of generative ai for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*, 2024.

[110] Anthony Dunne and Fiona Raby. *Speculative Everything, With a new preface by the authors: Design, Fiction, and Social Dreaming*. MIT press, 2024.