Ceci n'est pas une explication: Evaluating Explanation Failures as Explainability Pitfalls in Language Learning Systems

Ben Knight*
Oxford University Press
Oxford, UK
elt-bench@oup.com

Wm. Matthew Kennedy*
Oxford Internet Institute
University of Oxford
Oxford, UK

Danielle Carvalho
Oxford University Press
Oxford, UK

Isaac Pattis
Oxford University Press
Oxford, UK

James Edgell
Oxford University Press
Oxford, UK

Abstract

AI-powered language learning tools increasingly provide instant, personalised feedback to millions of learners worldwide. However, this feedback can fail in ways that are difficult for learners—and even teachers—to detect, potentially reinforcing misconceptions and eroding learning outcomes over extended use. We present a portion of L2-Bench, a benchmark for evaluating AI systems in language education that includes (but is not limited to) six critical dimensions of effective feedback—diagnostic accuracy, awareness of appropriacy, causes of error, prioritisation, guidance for improvement, and supporting self-regulation. We analyse how AI systems can fail with respect to these dimensions. These failures, which we argue are conducive to "explainability pitfalls", are AI-generated explanations that appear helpful on the surface but are fundamentally flawed, increasing the risk of attainment, human-AI interaction, and socioaffective harms. We discuss how the specific context of language learning amplifies these risks and outline open questions we believe merit more attention when designing evaluation frameworks specifically. Our analysis aims to expand the community's understanding of both the typology of explainability pitfalls and the contextual dynamics in which they may occur in order to encourage AI developers to better design safe, trustworthy, and effective AI explanations.

Keywords

Explanations, AI for Education, Language Learning, Human-AI Interaction, Evaluations

1 Introduction

Feedback in the form of explanations is a critical part of any learning system. In AI systems design, explanations have been proposed as a key mechanism through which more explainable, transparent, and ultimately accountable AI systems can be developed. If an AI system's actions can be explained to a user, that user may, in theory, be more empowered to steer the system towards more beneficial outcomes [37]. Elsewhere, in educational spaces, explanations provided as feedback in learning contexts supports both knowledge and cognitive process dimensions of individual learning [3, 38]. Clear feedback improves learners' ability to identify areas of weakness and understand target behaviours [18, 36, 42]. As it is typically delivered by trusted instructors or peers, it also facilitates social processes that are conducive to knowledge formation [4, 6]. In this vein, feedback is more than information transmission; it engages and motivates learners, provided it is constructive, specific, and timely [40].

These human-computer interaction and pedagogical considerations intersect in the development of digital learning systems. Digital learning system design is aware of the value and function of feedback in digital pedagogy [32, 39, 44]. Some of the first Intelligent Tutoring Systems viewed appropriate feedback as the core function of the system itself [30]. Developers of new digital learning systems—those powered by generative AI technologies—maintain this emphasis on feedback as essential to effective AI pedagogy [19]. Many indeed perceive a new opportunity: recent AI systems may be especially good at delivering constructive, specific, and timely feedback, given their reliance on powerful large language models or multimodal models that perform well on tasks critical to human interaction, such as natural language processing or image recognition [15, 19, 25, 26, 31, 43]. Although little evidence on the efficacy of such systems exists, many efforts to understand the real-world capabilities of AI learning systems are ongoing [27].

At the same time that we await more evidence that AI powered learning systems are indeed capable of supporting learning, and, therefore, providing effective feedback in educational contexts, several researchers and educational stakeholders have warned that AI systems may be prone to failures that can have substantial effects on digital learning [17]. Specifically, although AI tools can generate instant and apparently personalized feedback at scale, the interaction design that structures how feedback is delivered can mask flaws in its quality. These flaws are not the result any intent on the part of system designers to deceive learners (or other educational stakeholders); They occur despite considerable system post-training and fine-tuning for pedagogically appropriate subject matter, interaction design (e.g. optimizing for long conversations), and instructional goals [22]. Given that many of these failures present themselves most clearly in feedback—for example, in conversational turns that follow student completion of a learning task—they can also be exceedingly difficult to detect as those most directly exposed are the least likely to be sufficiently expert to perceive them [16, 22]. Failures can therefore compound over time.

For all these reasons, we believe they may best be approached as "explainability pitfalls" [9]. Problematizing these kinds of failures in AI for education systems in this way has implications both for

*Both authors contributed equally

Accepted to Misleading Impacts Resulting from AI Generated Explanations (MIRAGE) Workshop @ IUI 2026

system designers hoping to improve their systems through monitoring usage data and for learners and instructors who may find that such failures mistakenly reinforce misunderstandings or weaken trust between students and instructors. Understanding when these kinds of failures may occur, with what variety, and to what effect is crucial to advancing any AI for education system design. This need is particularly pronounced in AI for language learning [36].

To this end, our paper discusses our efforts to design L2-Bench [7], an evaluation benchmark to assess the efficacy of AI systems in language learning design, which decomposes the tasks that underpin a 'learning experience designer' into a hierarchical taxonomy of competencies . In this workshop paper, we focus specifically on the portion of our recently proposed taxonomy of competencies that concern the application of "giving feedback" effectively. In the first section of the paper, for each of these six areas, we briefly outline the core criteria for good feedback in language learning and comment on how AI systems may fail in each dimension. In the second section, we enumerate how AI failures discussed above may produce conditions conducive to explainability pitfalls. We also note how the specific context of language learning can create especially pronounced attainment, human-AI interaction, and other socioaffective risks that are specific to learning contexts and particularly intensified in language-learning contexts [28]. Ultimately, we aim to expand our understanding of both the typology of explainability pitfalls, and the contextual dynamics in which they may occur, lessons we believe can be generally useful to all AI system designers who are interested in safe, trustworthy, and effective AI explanations.

2 Giving Effective Feedback: a Review of Core Properties

Research on effective feedback identifies several key properties that distinguish pedagogically sound feedback from superficial or harmful responses [14, 36]. Within L2-Bench, we operationalise the competency of "giving feedback" through a structured evaluation framework examining six critical dimensions. These extend beyond simple error correction to encompass the full range of considerations essential to effective language teaching feedback.

2.1 Diagnostic Accuracy

Does the AI correctly identify the type and locus of a problem (grammar vs. lexis vs. discourse) and infer intended meaning?

Effective feedback begins with accurate diagnosis—understanding not just that an error occurred, but precisely where and what kind [36]. Hallucinations and overconfidence from AI systems are well-documented and are particularly problematic when the nature of an error is ambiguous. For instance, using the wrong past form of a verb could stem from not knowing verb forms, misunderstanding contextual significance, or overlooking subject-verb agreement. Each diagnostic hypothesis suggests different pedagogical interventions, and misdiagnosis can reinforce rather than correct misconceptions.

2.2 Awareness of Appropriacy

Does feedback evaluate appropriateness (formality, stance, audience, genre conventions, cultural pragmatics) rather than just correctness?

Language learning extends beyond grammatical accuracy to encompass using language appropriately for different contexts and purposes [16, 18, 26]. This is particularly salient in the communicative approach to language teaching, where priority is placed on conveying meaning in social situations [36]. A job application may be grammatically flawless yet employ informal language that creates a negative impression due to its inappropriacy for the professional context. Feedback focusing solely on correctness while ignoring appropriacy fails to prepare learners for authentic communication.

2.3 Causes of Error

Does the feedback hypothesize why the error occurred (e.g., transfer from L1 word order, article systems, collocation patterns)?

Understanding the underlying cause—rather than simply identifying the surface manifestation—is a hallmark of expert teaching [36]. A teacher experienced with students of a particular mother tongue can predict where negative transfer is likely to occur. AI systems lacking either sufficiently robust internal representations of a given problem space or causal reasoning capabilities may provide feedback that addresses errors superficially without helping learners understand why those errors occurred, limiting transfer of learning to new contexts.

2.4 Prioritization of Feedback

Does the feedback effectively prioritise what is most useful to the learner?

More feedback is not necessarily better feedback [14, 36]. When learners are overwhelmed with corrections, they struggle to identify priorities. A vital aspect of effective feedback is identifying which areas will be most useful, drawing upon understanding of what is developmentally appropriate for the learner's proficiency level and current learning goals. AI systems that indiscriminately correct every error risk inducing interaction harms such as anxiety and, paradoxically, reducing effectiveness by failing to prioritise [22].

2.5 Guidance for Improvement

Does the feedback help the learner to do better next time?

The fundamental purpose of feedback is to support future improvement [14, 35]. Effective feedback must strike a delicate balance: offering sufficient information without simply providing answers [36]. This aligns with "offering strategies rather than solutions"—an approach that develops learner autonomy rather than dependence [19, 22]. For example, rather than correcting all errors, effective feedback might indicate that a paragraph contains three subject-verb agreement errors and invite the learner to locate and correct them. AI systems providing solutions without strategic guidance (a high likelihood unless some sort of memory is implemented) may short-circuit the learning process [5]. Furthermore, an instructor's own familiarity with AI systems and their outputs (including explanations) may vary. This may subsequently lead them to use AI system explanations differently to how system designers intended, with adverse consequences, as Ehsan et al [8] document in other domains.

2.6 Supporting Self-regulation

The development of self-regulatory skills is central to long-term learning success [36, 42]. Does the AI encourage self-regulated learning (goal setting, monitoring, strategy choice) or foster passive reliance?

A commonly observed phenomenon with AI tools is that learners become dependent on them without developing their own metacognitive capabilities [12, 34]. Such dependencies may be intensified by design choices that promote an anthropomorphization of an instructional character [1, 22]. Feedback must be designed to avoid this dependency trap, ensuring learners do not receive answers without actively engaging with the learning process. This corresponds to the highest level in Absolum's feedback framework—"provocative prompts" that encourage deeper reflection even when success criteria have been met [36]. AI systems that fail in this dimension may produce short-term performance gains while undermining lifelong learning capabilities [29].

3 The Anatomy of an Explainability Pitfall in AI for Language Learning

The six dimensions of effective feedback outlined above are not merely pedagogical ideals—they can function as risk models. They can be used to structure efforts to identify failure modes in which AI systems can generate feedback that appears helpful on the surface but is fundamentally flawed; either failing to support learning or actively undermining it. These failures constitute what we might view as "explainability pitfalls" in language learning contexts. Context is a critical factor in producing explainability pitfalls as, in the words of Weidinger et al [41], context "co-determines" the harms and benefits AI systems (or, for that matter, any sociotechnical system) produces. This is especially true in educational spaces, whose norms and values differ in important ways from those operating in society in general [21].

Drawing on this and our preliminary evaluation of frontier AI models on L2-Bench tasks for the "giving feedback" competency, we observe several characteristic failure patterns that remain difficult for learners (and often teachers) to detect.

3.1 Over-Correction

If the AI system corrects every minor error, learners may become anxious or focus excessively on form rather than meaning, reducing fluency [14, 36]. This failure in prioritisation manifests as an explainability pitfall when feedback overwhelms rather than guides, leaving learners unable to distinguish critical issues from minor ones.

3.2 Ambiguous or Unhelpful Feedback

Feedback like "Incorrect" without explanation doesn't teach. Learners need actionable guidance (e.g., why it's wrong and what to do next) [35, 36]. AI systems that provide corrected versions without explaining reasoning foster dependency and short-circuit learning—resolving the immediate problem without building capacity to solve similar problems independently.

3.3 Ignoring Communicative Competence

Systems that only target grammar accuracy may neglect pragmatic and discourse-level skills, leading to unnatural language use [18]. This represents a fundamental explainability failure: the AI cannot adequately explain 'why' certain language choices are problematic because it lacks robust models of social context and communicative purpose. Although some systems are improving in-context learning capabilities via expanded context windows, negative verification tasks remain challenging to most LLMs [10].

3.4 Cultural and Contextual Misalignment

AI systems trained primarily on certain varieties of English may provide feedback reflecting hidden biases about "correct" usage [23] or may cluster around a central tendency, producing a homogenizing effect [2, 16]. Feedback failing to account for learners' cultural backgrounds or learning goals represents an explainability failure at the level of contextual assumptions—the system cannot adequately explain 'for whom' and 'in what contexts' its guidance is intended [16], which can result in irrelevant or even harmful feedback.

3.5 Bias and Inconsistency

Poorly designed systems might give inconsistent feedback or reinforce incorrect patterns in uncertainty, harming learning progress. When AI systems provide definitive diagnoses without expressing appropriate uncertainty—even when multiple interpretations are plausible—misdiagnosis can reinforce rather than correct misconceptions. In part, this results from methods used to train AI models–methods that OpenAI researchers argue "reward guessing over acknowledging uncertainty" [20]. This is also a data preparation problem–a common step in preparing either training data or data for evaluation datasets is to remove samples with ambiguous properties.

4 Future work

By participating in this workshop, we aim to engage participants to aid our continued exploration of concrete examples of explainability pitfalls in AI-generated feedback for language learning. We are particularly interested in understanding where the general failure modes described in section one and their specific presentation in language learning contexts described in section two can usefully generalize to broader research on explainability pitfalls in general. From our point of view, the development of L2-Bench's feedback evaluation framework raises three critical questions that we welcome perspectives on from the broader community working on AI explanations and educational systems that may inform AI explanations (and risks entailed in generating them) more broadly:

4.1 Cross-Cultural and Contextual Validity

Language learning feedback practices vary significantly across cultural and educational contexts. How can AI systems be evaluated for their ability to adapt feedback styles appropriately to diverse learner populations, particularly in global English language teaching contexts where learners' first languages, educational backgrounds, and cultural expectations of teacher-student interaction differ substantially? Some preliminary investigations from the field of adversarial evaluation exist [24], but more work clearly needs to be done here.

4.2 Explainability in Uncertain Pedagogical Situations

Many feedback scenarios in language learning lack clear "correct" answers—for instance, when multiple formulations are acceptable, when appropriacy depends on unstated contextual factors [11], or when diagnostic uncertainty is high. How should AI systems communicate uncertainty in such cases, and what evaluation criteria can distinguish pedagogically appropriate expressions of uncertainty from evasiveness or overconfidence?

4.3 Multi-turn Interaction

A major limitation of our current evaluation benchmark is that it does not yet account for multi-turn interaction. Although some harms do present themselves within the "foundational unit" of conversation—single turn adjacency pairs [33]—many do not [13]. Indeed this is the key mechanic driving low detection rates of compounding errors in accuracy but also broader sociocultural and contextual misalignment. Yet no AI for education development or evaluation team appears to have made much progress on this front as accounting for multi-turn interaction is an essential condition for evaluating pedagogical efficacy as well as socioaffective risk [22, 26].

5 Conclusion

The unique role that feedback plays in language learning—extending beyond binary correctness judgments to encompass appropriacy, learner development, and socioaffective dimensions—makes it both a critical and challenging domain for AI evaluation. Assessing feedback failures therefore requires a different approach than simply computing an accuracy metric. Instead, we find utility in approaching these types of failures as inadequacies in human-AI interaction design for educational contexts. We first review core principles of feedback as practiced in language learning contexts and discuss where we hypothesize AI systems may perform inadequately. We then discuss how such failures may produce explainability pitfalls and document initial efforts to formalize a portion of a new AI for language-learning evaluation benchmark (L2-Bench) drawing upon that concept. We conclude by observing that several critical challenges remain both for AI for education system designers and evaluation developers. We seek the advice of the broader XAI community about how best to make progress on these important questions, and, in so doing, advance the responsible development of AI systems that genuinely support language learning and boost learning outcomes.

References

[1] Gavin Abercrombie, Alice Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages: On Anthropomorphism in Dialogue Systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 4776–4790.

[2] Disha Agarwal, Mor Naaman, and Aditya Vashistha. 2024. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. arXiv:2409.11360.

[3] Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Raths, and Merlin C. Wittrock. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, NY.

[4] Albert Bandura. 1977. *Social Learning Theory*. Prentice Hall, Englewood Cliffs, NJ.

[5] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haoyang Ge, Ozge Kabakci, and Rani Mariman. 2024. Generative AI Can Harm Learning. SSRN Working Paper, DOI: 10.2139/ssrn.4895486.

[6] Susan M. Brookhart. 2008. *How to Give Effective Feedback to Your Students*. ASCD, Alexandria, VA.

[7] James Edgell, Wm. Matthew Kennedy, Isaac Pattis, Ben Knight, Danielle Carvalho, and Elizabeth Wonnacott. 2026. Beyond Accuracy: Towards a Robust Evaluation Methodology for AI Systems for Language Education. arXiv:2603.20088 [cs.CY] https://arxiv.org/abs/2603.20088

[8] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O. Riedl. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, Article 316. doi:10.1145/3613904.3642474

[9] Upol Ehsan and Mark O. Riedl. 2021. Explainability Pitfalls: Beyond Dark Patterns in Explainable AI. In *Human-Centered AI Workshop at NeurIPS*. arXiv:2109.12480.

[10] Harvey Yiyun Fu, Aryan Shrivastava, Jared Moore, Peter West, Chenhao Tan, and Ari Holtzman. 2025. AbsenceBench: Language Models Can't Tell What's Missing. arXiv:2506.11440.

[11] Iason Gabriel, Andrea Manzini, Geoffrey Keeling, Lisa Anne Hendricks, Verena Rieser, Haroon Iqbal, Nenad Tomašev, Irina Ktena, Zachary Kenton, Manuel Rodriguez, Sam El-Sayed, Sarah Brown, Cansu Akbulut, Andrew Trask, Edward Hughes, Adam S. Bergman, Renee Shelby, Naomi Marchal, Casey Griffin, Juan Mateos-Garcia, Laura Weidinger, William Street, Benjamin Lange, Adam Ingerman, Adam Lentz, Rune Enger, Brandon Andrew, Victoria Krakovna, Joshua O. Siy, Zeb Kurth-Nelson, Angus McCroskery, Valerio Bolina, Hugh Law, Murray Shanahan, Lucas Alberts, Borja Balle, Stephan de Haas, Yemi Ibitoye, Allan Dafoe, Bennett Goldberg, Stephen Krier, Andrew Reese, Scott Witherspoon, Will Hawkins, Maximilian Rauh, David Wallace, Michael Franklin, Jared A. Goldstein, Joel Lehman, Matthew Klenk, Shannon Vallor, Cole Biles, Meredith Ringel Morris, Hilary King, Blaise Agüera y Arcas, William Isaac, and James Manyika. 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244v2.

[12] Michael Gerlich. 2025. AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* 15, 1 (2025), 6. doi:10.3390/soc15010006

[13] Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. 1995. Collaborative Dialogue Patterns in Naturalistic One-to-One Tutoring. *Applied Cognitive Psychology* 9, 6 (1995), 495–522.

[14] John Hattie and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77, 1 (2007), 81–112.

[15] Wayne Holmes. 2024. AIED—Coming of Age? *International Journal of Artificial Intelligence in Education* 34, 1 (2024), 1–11. doi:10.1007/s40593-023-00352-3

[16] Wayne Holmes and Fengchun Miao. 2023. *Guidance for Generative AI for Education and Research*. UNESCO.

[17] Wayne Holmes and Ilkka Tuomi. 2022. State of the Art and Practice in AI in Education. *European Journal of Education* 57, 4 (2022), 542–570.

[18] Fiona Hyland. 1998. The Impact of Teacher Written Feedback on Individual Writers. *Journal of Second Language Writing* 7, 3 (1998), 255–286.

[19] Ivan Jurenka, Matthias Kunesch, Kyle R. McKee, Daniel Gillick, et al. 2024. Towards Responsible Development of Generative AI for Education: An Evaluation-Driven Approach. arXiv:2407.12687.

[20] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why Language Models Hallucinate. arXiv:2509.04664.

[21] Wm. Matthew Kennedy and Daniel Vargas Campos. 2024. Vernacularizing Taxonomies of Harm Is Essential for Operationalizing Holistic AI Safety. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 698–710.

[22] Wm. Matthew Kennedy and Daniel Vargas Campos. 2026. A Vernacularized Taxonomy of Harms for AI in Education. In *Handbook of Critical Studies in AI for Education*, Wayne Holmes and Caroline Pelletier (Eds.). Edward Elgar. Forthcoming.

[23] Val Klenowski. 2009. Assessment for Learning Revisited: An Asia-Pacific Perspective. *Assessment in Education: Principles, Policy & Practice* 16, 3 (2009), 263–268.

[24] Akash Kundu, Adrianna Tan, Theodora Skeadas, Rumman Chowdhury, and Sarah Amos. 2025. Red Teaming for Trust: Evaluating Multicultural and Multilingual AI Systems in Asia-Pacific. In *Building Trust Workshop at the International Conference on Learning Representations*.

[25] LearnLM Team. 2024. LearnLM: Improving Gemini for Learning. arXiv:2412.16429.

[26] LearnLM Team and Google. 2025. Evaluating Gemini in an Arena for Learning. arXiv:2505.24477v1 [cs.CY].

[27] LearnLM Team, Google, and Eedi. 2025. AI Tutoring Can Safely and Effectively Support Students: An Exploratory RCT in UK Classrooms. Technical report.

[28] Danielle S. McNamara, Laura K. Allen, Matthew E. Jacovina, and Aaron D. Likens. 2023. Leveraging Large Language Models for Language Learning. *Journal of Learning Analytics* 10, 3 (2023), 1–15.

[29] Allen Nie, Yash Chandak, Miroslav Suzara, Ali Malik, Juliette Woodrow, Matt Peng, Mehran Sahami, Emma Brunskill, and Chris Piech. 2025. The GPT Surprise: Offering Large Language Model Chat in a Massive Coding Class Reduced Engagement but May Increase Adopters' Exam Performance. In *Proceedings of*

the Twelfth ACM Conference on Learning @ Scale (L@S '25). Association for Computing Machinery, New York, NY, USA, 376–380. doi:10.1145/3698205.3733960

[30] Seymour Papert. 1980. *Mindstorms: Children, Computers, and Powerful Ideas*. Basic Books, New York, NY.

[31] Alison Pease, Anna Zamansky, and Sarah Wiseman. 2023. Pedagogical Implications of Large Language Models: Challenges and Opportunities. *AI & Society* (2023), 1–14. doi:10.1007/s00146-023-01753-4

[32] Chris Piech, Mehran Sahami, Daphne Koller, Steve Cooper, and Paulo Blikstein. 2015. Modeling How Students Learn to Program. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. ACM, 153–158. doi:10.1145/2676723.2677308

[33] Emanuel A. Schegloff and Harvey Sacks. 1973. Opening Up Closings. *Semiotica* 8, 4 (1973), 289–327.

[34] Neil Selwyn. 2019. *Should Robots Replace Teachers? AI and the Future of Education*. Polity Press, Cambridge, UK.

[35] Valerie J. Shute. 2008. Focus on Formative Feedback. *Review of Educational Research* 78, 1 (2008), 153–189.

[36] Gordon Stobart, Elaine Boyd, Anthony Green, and Therese N. Hopfenbeck. 2019. *Effective Feedback: The Key to Successful Assessment for Learning*. Oxford University Press.

[37] Stefano Teso, Oznur Alkan, Wolfgang Stammer, and Elizabeth Daly. 2023. Leveraging Explanations in Interactive Machine Learning: An Overview. *Frontiers in Artificial Intelligence* 6 (2023). doi:10.3389/frai.2023.1066049

[38] Kelsey Urgo, Jaime Arguello, and Robert Capra. 2019. Anderson and Krathwohl's Two-Dimensional Taxonomy Applied to Task Creation and Learning Assessment. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. ACM, 117–124. doi:10.1145/3341981.3344226

[39] Kurt VanLehn. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist* 46, 4 (2011), 197–221. doi:10.1080/00461520.2011.611369

[40] Lev S. Vygotsky. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA.

[41] Laura Weidinger, Maximilian Rauh, Naomi Marchal, Andrea Manzini, Lisa Anne Hendricks, et al. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986.

[42] Dylan Wiliam. 2011. *Embedded Formative Assessment*. Solution Tree Press, Bloomington, IN.

[43] Simon Woodhead, Simon Blatchford, and Michael Webb. 2023. Can AI Tutors Improve Learning Outcomes at Scale? Results from a Randomized Controlled Trial. In *Proceedings of the International Conference on Learning Analytics and Knowledge*. ACM, 489–495.

[44] Beverly Park Woolf. 2010. *Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing E-Learning*. Morgan Kaufmann, Burlington, MA.