AIBAT: AI Behavior Analysis Tool for Teacher-Driven Contextual Evaluation of Language Models in Education

Shamya Karumbaiah[1][0000−0002−7920−4510], Yaxuan Yin[1][0009−0006−6661−2183], and Aayush Bharadwaj[1][0009−0008−2664−5583]

University of Wisconsin–Madison, Madison, WI 53706, USA
{shamya.karumbaiah, yaxuan.yin, abbharadwaj}@wisc.edu

Abstract. With the increasing reliance of AIED on opaque, black-box scaffolds such as large language models to support student learning, there is a growing concern about their limitations when used in diverse pedagogical contexts. This opacity often undermines educators' trust and shapes their perceptions, contributing to resistance toward the adoption of AI scaffolds in schools. To address these challenges, we developed AIBAT, a workflow and system designed to support educators in auditing and critically evaluating the potential benefits and harms of AI systems within their specific pedagogical contexts (e.g., subject matter, grade level, English proficiency). With AIBAT, teachers can specify expected behaviors — i.e., what they anticipate the AI scaffold should do — and test the system against those expectations. We conducted an exploratory user study with 14 teachers to examine how AIBAT facilitates the identification and sensemaking of AI-related risks, while enabling educators to use evidence to calibrate their trust in AI scaffolds. Our findings reveal that teachers valued the ability to engage with AI decisions rather than passively accepting them, describing the process as a "conversation" that enhanced transparency, trust, and a sense of control. We identify key opportunities to foster meaningful user engagement in AI auditing processes and discuss broader implications for promoting responsible and effective teacher participation in the evaluation and deployment of AI systems in educational settings.

Keywords: Large Language models · Evaluation · Teachers · Context · Behavior analysis · Scaffolds · Responsible AI in education

1 Introduction

An equitable and synergistic use of AI for scaffolding in classrooms must acknowledge differences in pedagogical contexts. In a classroom, scaffolding is often distributed across materials, peers, technology, and the teacher [17]. AI systems that scaffold student learning can free up teachers' time, allowing them to focus on students who need support the most [23, 24]. Hence, AI systems designed to work synergistically with teacher practices are more effective [25]. Since pedagogical contexts vary, so too do scaffolding needs and priorities — as well as teachers' definitions of system failure.

However, AI evaluation typically assumes universal deployment, ignoring differences in contexts. Indeed, the dominant approach for AI evaluation (i.e., generalization estimation on test data) often tends to be an overestimation of real-world performance [30, 21]. In most cases, the test data used to evaluate generalization poorly represent the range of real-world scenarios and often contain the same biases as the training data. Even when tested with data from the deployment context, the population distribution may shift over time [18].

Recent research in natural language processing (NLP) — a kind of AI system involving computational models of text — has shown that engaging stakeholders to analyze NLP model behaviors (in conditions relevant to their context) was effective in identifying failures that are likely to go unnoticed in tests for generalization [9, 8]. Stakeholders with limited prior experience in AI were able to identify failures in key capabilities such as fairness (e.g., biases against linguistic minorities), robustness to perturbations (e.g., spelling errors), and domain-specific vocabulary when evaluating NLP systems used for language translation, content moderation, and question-answering.

Although analyzing AI system behavior in context offers a promising avenue for enabling stakeholders to interrogate AI black boxes, we still don't know: (1) what value this approach holds for stakeholders of educational AI such as teachers; (2) how we could align it better with teachers' contextual needs; (3) how to build teacher expertise and agency in identifying beneficial and harmful scaffolding behaviors; and (4) how it impacts teacher trust and practices with AI scaffolds. Several forms of AI systems already exist for distributed scaffolding (e.g., NLP systems for writing feedback, conversational agents, AI tutors) and now with the advent of generative AI, teachers are faced with the difficult choice of trusting these advanced technologies to take advantage of them. Trusting such systems is especially difficult as they become less transparent and raise equity concerns for minoritized students. As learning sciences raises critical questions about student and teacher agency with technology [1], we ask: How do we equip educational stakeholders with tools that build their expertise and agency in trustworthy and equitable AI use in classrooms? Hence, the central research question in our study is:*How do teachers analyze AI system behaviors to identify the benefits and harms of using AI scaffolds with their students?*

2 Conceptual Framing

We first frame AI systems as scaffolds to support student learning and argue that human mediation is necessary for the appropriate use of AI in classrooms. Then, we show how system failures differ by pedagogical context in ways not captured by generalization estimates. Last, we explain how allowing stakeholders such as teachers to evaluate AI system behaviors in their pedagogical context fosters trust and agency for an equitable and synergistic human–AI scaffolding.

2.1 AI Systems as Scaffolds to Support Student Learning

Scaffolding is closely related to the sociocultural perspective of learning proposed by Vygotsky [2], wherein learning occurs first at the social level in interactions with others and later at an individual level. Although scaffolding was originally described as the one-on-one support an adult offers to a child in their zone of proximal development (ZPD), practical limitations of present-day classrooms pose constraints on one teacher responsible for multiple ZPDs at the same time. Hence, scaffolding is often synergistically distributed across instructional materials, peers, technology, and the teacher [23]. AI systems, designed as scaffolds [31], could anticipate students' changing needs in their ZPD and support them to reach their potential [6]. Well-designed scaffolds are sensitive to students' ZPD [17]. For example, over the past 40 years, research on AI tutors has developed models to assess students' current skill levels and adapt instructional support in the moment [13], allowing teachers to attend to students who need individualized support [19].

2.2 Human–AI Synergy in Distributed Scaffolding

Although AI systems as scaffolds may be narrowly defined to support a specific task [31], scaffolding, as a dialogic process, is theoretically grounded in the sociocultural approach [17]. Human mediation, involving interpersonal interactions within the learning environment, is key for material scaffolds or tools to be used appropriately [34]. In classrooms, social structures facilitating dialogues with peers and teachers serve as social scaffolds that enable learning with tools (e.g., [32]). Moreover, teachers play a critical role in adapting and continually adjusting scaffolding in the moment based on their students' changing needs [20], making sensitive, conscious decisions to complement the support provided by the tool. Teachers are also responsible for responsively fading scaffolds to promote student independence. Hence, a system of scaffolding (i.e., distributed scaffolding [17]), involving tools and social scaffolds, must function synergistically to effectively support students in classrooms. For instance, past research with AI tutors shows that even when technology is designed to work with minimal human intervention, teacher practices matter to student learning [7]. Moreover, adaptive experiences are jointly facilitated by both teachers and technology [15]. Hence, in contrast to an automation-first approach to AI, a more beneficial approach for AI use in classrooms could be intelligence augmentation [22, 27], where AI builds on teachers' pedagogical knowledge.

2.3 The Need to Go Beyond Generalization Estimates for Equitable AI

Despite significant differences in the pedagogical contexts in which educational AI systems are used, AI evaluation is often limited to generalization estimates (e.g., overall accuracy, mean squared error) that inherently assume universal deployment [33]. Moreover, AI systems are known to systematically fail on rare groups not obvious in aggregate evaluation [28], such as minoritized student populations. Past research demonstrates how ignoring learner context in the design of AI tutors could introduce harmful biases in them [3]. Despite a recent spike in efforts to identify and mitigate bias in educational AI [10], significant challenges remain. A common approach to fairness is ensuring that the system performs well for student subgroups. In addition to oversimplified or politically influenced categorizations of student demographics, these approaches fall short in considering the myriad of ways in which student identities intersect [16]. Moreover, technical conceptions of bias are often vague, lack normative grounding, and diverge from how bias is socially understood [29]. Hence, answers to the question of "what kinds of system behaviors are harmful, in what ways, to whom, and why?" [26] need to be contextually grounded in the lived experiences of the stakeholders.

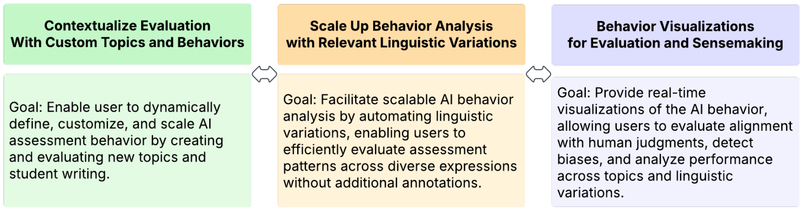
2.4 Stakeholder-Driven Contextual Evaluation of Behaviors for Human–AI Trust

There has been an increasing call for contextual evaluation [12] that recognizes differences in the lived experiences of stakeholders, which in turn define system failures differently [14]. Specifying desired system behavior serves an important role in transparency and trust, opening the AI system for stakeholder scrutiny [8, **?**]. Prior work has formalized human–AI trust as contractual, i.e., trust is built on explicit, context-specific contracts that stakeholders specify based on the expected behaviors of the AI system [5]. System *behaviors* defined through testing can act as components of such contracts — particularly when they incorporate stakeholders' contextual expertise to translate implicit expectations (e.g., fairness) into explicit, testable criteria [11]. In addition to better-informed trust, contextual evaluation of AI scaffold behaviors can improve AI transparency [4]. With improved awareness of both the benefits and limitations of using AI scaffolds in their context — as well as the associated equity concerns — stakeholders can more effectively situate AI scaffolds within their broader system of scaffolding distributed across tools and social supports. This may involve adjusting scaffolding practices to amplify the benefits and mitigate the limitations of the AI scaffold.

3 AIBAT System Design

With AIBAT, teachers specify *behaviors* (i.e., what they expect the AI scaffold to do) and test scaffolds such as large language models (LLMs) against those expectations. For example, to analyze how fairly an LLM treats bilingual students, teachers can define a target LLM behavior (e.g., tag incorrect student response for feedback) and test whether that behavior stays invariant based on a specified condition (e.g., no change in assessment when a relevant Spanish idiom is added to the response). Although AI scaffolds may be expected to perform various tasks such as assessment, feedback, and question answering, we focus on assessment as an illustrative example in this study because it is often the first step in diagnosing students' scaffolding needs. In this section, we first describe the infrastructure and iterative workflow of AIBAT in Section 3.1. We then introduce three integrated design features from Sections 3.2 to 3.4, along with the interface shown in Figure 2, aimed at helping teachers understand the harms and benefits of AI performance and trust.

3.1 System Overview



Contextualize Evaluation With Custom Topics and Behaviors

Goal: Enable user to dynamically define, customize, and scale AI assessment behavior by creating and evaluating new topics and student writing.

Scale Up Behavior Analysis with Relevant Linguistic Variations

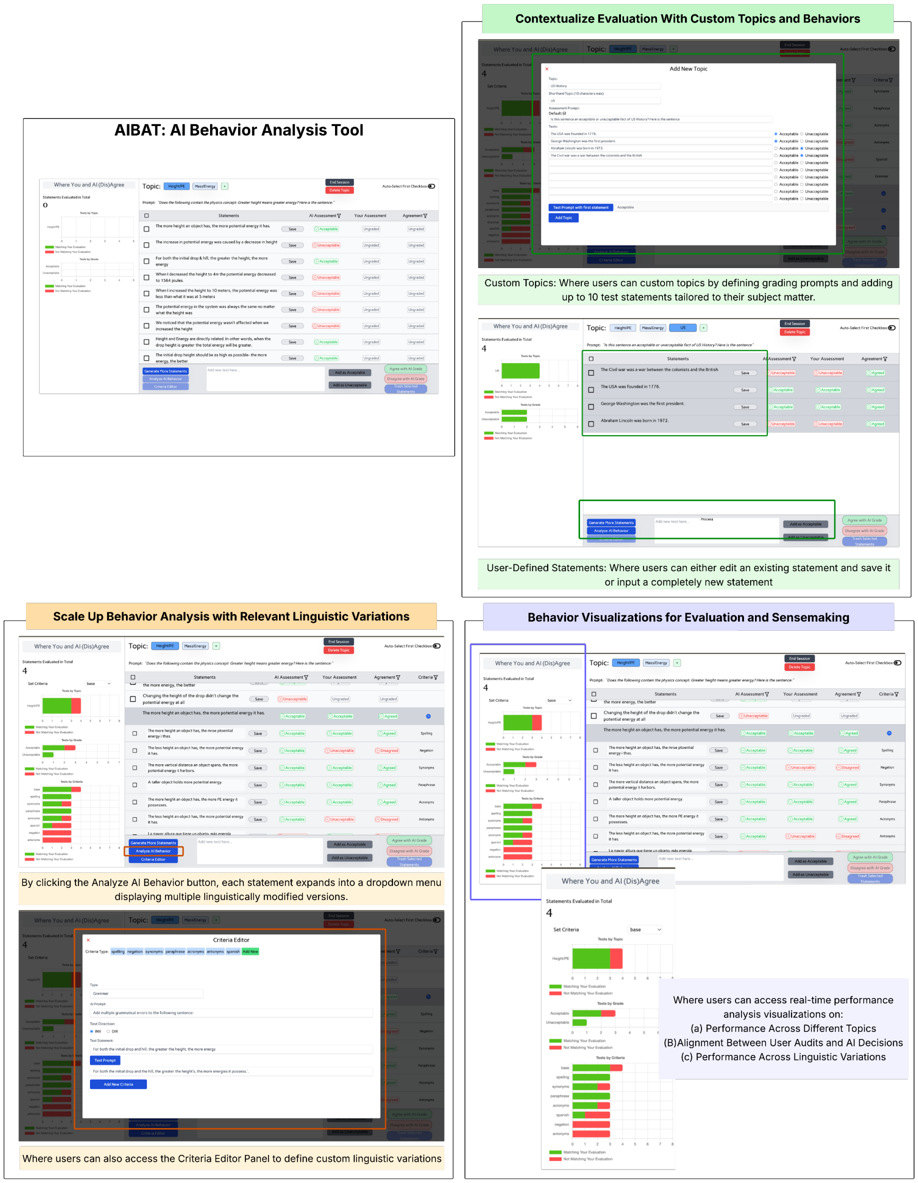
Goal: Facilitate scalable AI behavior analysis by automating linguistic variations, enabling users to efficiently evaluate assessment patterns across diverse expressions without additional annotations.

Behavior Visualizations for Evaluation and Sensemaking

Goal: Provide real-time visualizations of the AI behavior, allowing users to evaluate alignment with human judgments, detect biases, and analyze performance across topics and linguistic variations.

Fig. 1. AIBAT's workflow consists of Contextualized Evaluation with Custom Topics and Behaviors, Scale Up Behavior Analysis with Relevant Linguistic Variations, and Behavior Visualizations for Evaluation and Sensemaking to facilitate the sensemaking of the benefits and harms of LLM assessments..

As shown in Figure 1, there are three key design features in AIBAT: 1) Contextualize Evaluation With Custom Topics and Behaviors, 2) Scale Up Behavior Analysis with Relevant Linguistic Variations, and 3) Behavior Visualizations for Evaluation and Sensemaking. First, teachers define expected AI behavior by providing test cases (e.g., student responses) and expected assessments (e.g., acceptable versus unacceptable). Teachers can contextualize the evaluation process by creating new topics relevant to their subject matter and adding test cases that reflect their students' writing (see Section 3.2). Second, the platform offers full test case management, enabling users to edit, delete, or add test cases as needed. After reviewing the AI's assessment, teachers can indicate whether they agree or disagree with it, and AIBAT visualizes these results through interactive graphs. These visualizations provide insights into AI assessment performance, potential biases, and overall consistency (see Section 3.4). Third, AIBAT allows teachers to introduce linguistic variations (see Section 3.3), such as misspellings, missing punctuation, translations, and paraphrased versions of test statements, to evaluate how the AI assessment handles the diverse ways in which their students write. AIBAT is built using Django and Next.js and hosted on Lambda Cloud GPU instances.



AIBAT: AI Behavior Analysis Tool

Contextualize Evaluation With Custom Topics and Behaviors

Custom Topics: Where users can custom topics by defining grading prompts and adding up to 10 test statements tailored to their subject matter.

User-Defined Statements: Where users can either edit an existing statement and save it or input a completely new statement

Scale Up Behavior Analysis with Relevant Linguistic Variations

By clicking the Analyze AI Behavior button, each statement expands into a dropdown menu displaying multiple linguistically modified versions.

Where users can also access the Criteria Editor Panel to define custom linguistic variations

Behavior Visualizations for Evaluation and Sensemaking

Where users can access real-time performance analysis visualizations on:
(a) Performance Across Different Topics
(B)Alignment Between User Audits and AI Decisions
(c) Performance Across Linguistic Variations

Fig. 2. AIBAT's Interface and Key Features. The interface includes default status (top left), Contextualize Evaluation with Custom Topics and Behaviors (top right), Scale Up Behavior Analysis with Relevant Linguistic Variations (bottom left), and Behavior Visualizations for Evaluation and Sensemaking (bottom right).

3.2 Contextualize Evaluation With Custom Topics and Behaviors

By default, the platform includes two sample middle-school science topics — how Potential Energy (PE) varies with height and how mass affects total energy — along with a set of predefined student responses. For each topic, users are presented with 20 predefined statements on the main panel, each accompanied by an AI-generated assessment indicating whether the statement is deemed acceptable or unacceptable. Users can review these assessments and provide feedback by agreeing or disagreeing with the AI's decision with a single click. To provide adaptability to different subject areas and grading criteria, AIBAT offers a Custom Topic Definition feature, allowing users to define and evaluate new topics dynamically.

This feature comprises two configurable options to support varied evaluation needs. First, Custom Topics allow teachers to create their own topics by defining assessment prompts and adding up to 10 test statements tailored to their subject matter. These user-defined topics function as prompts that help AIBAT generate a bookmarkable topic page, similar to the default Height/PE and Mass/Energy topics, allowing users to switch between them. If teachers wish to augment their evaluation with additional, auto-generated test statements, they can select "Generate More Statements", prompting AIBAT to produce new statements under the defined topic.

Second, User-Defined Statements allow users to manually input specific statements under an existing topic for AI evaluation. Users can either edit an existing statement and save it, prompting the LLM to regrade the modified statement, or input a completely new statement instead of modifying a predefined one, indicating whether they consider it acceptable or unacceptable. Once submitted, the statement appears in the panel under the selected topic, accompanied by the LLM-generated assessment (acceptable/unacceptable) and an indication of whether the AI's assessment aligns with the teacher's decision (agree/disagree). Together, these options allow educators to tailor AIBAT's evaluation system to a range of grading needs, ensuring flexibility beyond its default topics and statements.

3.3 Scale Up Behavior Analysis with Relevant Linguistic Variations

AIBAT incorporates a Linguistic Variation feature to account for the diverse ways a given statement can be expressed. When teachers enable this feature by clicking the Analyze AI Behavior button, each statement expands into a dropdown menu displaying multiple linguistically modified versions. These variations include adjustments such as spelling modifications, negation, synonyms, paraphrasing, acronyms, antonyms, and translations into Spanish by default. Additionally, users can access the Criteria Editor Panel to define custom linguistic variations, providing key information that helps fine-tune the model's responses. Taken holistically, this feature aims to enhance scalability by allowing teachers to assess AI-graded patterns across multiple linguistic forms with their previous decisions populated automatically, rather than manually reviewing each case in isolation. In real-world classroom settings, student responses naturally vary in wording, grammar, and phrasing, yet traditional AI evaluation methods often rely on fixed expressions, limiting the scope of assessment. By automating the generation of systematic linguistic variations, AIBAT enables educators to efficiently examine how AI models handle diverse inputs at scale.

3.4 Behavior Visualizations for Evaluation and Sensemaking

AIBAT includes a real-time AI performance analysis panel that provides dynamic insights into the AI model's evaluation accuracy through bar charts. The visualization is structured around three key dimensions: (a) Performance Across Different Topics — The system evaluates AI behaviors across various subject areas, enabling users to assess whether the model generalizes effectively or struggles with certain topics. This feature helps identify subject-specific biases and inconsistencies that may impact grading fairness. (b) Alignment Between User Audits and AI Decisions — This metric measures the extent to which users' auditing decisions align with the AI's assessment. It distinguishes between true positives (correct acceptances), true negatives (correct rejections), false positives (incorrect acceptances), and false negatives (incorrect rejections). This allows users to identify patterns in the AI's errors and assess its reliability in different grading scenarios. (c) Performance Across Linguistic Variations — AIBAT also analyzes how the AI responds to different linguistic variations, such as changes in grammar, spelling, and negation. By tracking performance across these variations, users can detect potential biases or weaknesses in the model's ability to handle diverse linguistic expressions.

3.5 LLM Model Specifications

AIBAT leverages various LLMs throughout the application to perform different tasks effectively. For grading the default topics, the platform utilizes a fine-tuned RoBERTa classification model, specifically trained to evaluate responses related to the default subjects. When it comes to grading new, user-defined topics, AIBAT switches to a general-purpose, pre-trained Llama 3 model, which provides flexibility for a broader range of grading tasks. For generating new test cases, AIBAT relies on a Mistral model, which helps create diverse and relevant test scenarios. Linguistic perturbations are generated using Mistral. Through these carefully selected models, AIBAT ensures accurate grading, efficient test case generation, and robust perturbation handling for various educational needs.

4 Method

4.1 Participant Recruitment

We recruited 14 participants (3 male, 11 female) through institutional networks to participate in this study, which was conducted under Institutional Review

Board (IRB) approval. All participants had prior teaching experience, ranging from K–12 to the college level, with an average of 16 years of teaching experience. Their disciplinary backgrounds spanned various fields, including arts, history, social studies, technology, and entrepreneurship. While nine participants self-reported familiarity with LLM tools such as ChatGPT, the remaining five had no prior experience with them. We conducted the think-aloud sessions in person, with each session lasting between 58 and 110 minutes.

4.2 Procedure and Data Analysis

To evaluate the usability and effectiveness of AIBAT, we conducted a user study in which participants explored the tool's features while providing real-time feedback. We began by introducing the goal of AIBAT, explaining its intended application scenarios to ensure participants understood its purpose and relevance. Following this introduction, researchers guided participants through each feature, demonstrating its functionality and encouraging hands-on interaction. Participants were then given the opportunity to explore AIBAT independently while being instructed to think aloud, verbalizing their thoughts, impressions, and any challenges they encountered. To capture their real-time interactions, participants were asked to share their screens throughout the session. This enabled researchers to closely observe navigation patterns, engagement levels, and any usability issues that emerged during exploration. At the conclusion of each session, researchers conducted a brief debriefing with participants to reflect on their experiences, gather feedback, and identify areas for improvement.

To analyze the study data, we reviewed all video recordings, transcribed the discussions, and extracted relevant quotes and actions that illustrated how participants approached creating example sets and tests, as well as how they reasoned about and reflected on the model. Using affinity diagrams, we annotated and categorized these insights into thematic groups. We iteratively analyzed the transcripts, grouped interpretation notes, and identified emerging themes from the data, allowing us to discern key patterns and trends in participant interactions and reasoning.

5 Findings

5.1 Evaluation Mechanism Facilitates Human-AI Trust

Our findings indicate that AIBAT's evaluation mechanism, which allows users to audit AI-generated assessments, plays a critical role in fostering user autonomy and trust in LLM-based assessment. Specifically, participants expressed that the ability to engage with AI decisions — rather than passively accepting them — contributed to a greater sense of control over the system's outputs. As P13 noted: *"Some other tools are just like, well, I put this in, this came out, and we can only hope it's right or we have to verify it ourselves."* One participant highlighted how the ability to see AI assessments directly within the interface and compare them against their own judgments provided a clearer understanding of the model's decision-making process, stating, *"Looking at it, like we had talked about with how we know the information is accurate... now we can actually see it right there."* . Moreover, P5 described this evaluation as a "conversation" with the AI, stating, *"I generally love this tool...It kind of shows that it's a conversation almost — like a back and forth. Transparency can also be super critical in getting anybody else to accept it."* Overall, our results suggest that the ability to evaluate AI assessments empowered users to critically assess the model's outputs, aligning with the idea that AI should be a tool for collaboration rather than a source of unquestioned authority.

Additionally, participants emphasized that the ability to interactively evaluate the model increased their trust in the system. P4 explained, *"Being given the tools to kind of test it and move it and kind of train it and understand a little bit about what it's looking for gives me that stronger trust, and I would be more likely to use it."* Similarly, P11 noted *"It kind of helps you look through these lenses, if you will, to get a more accurate tool and move towards trust and reliability. If I was given the tool on my own and just told to use it, my trust would be really low."* Participants also emphasized how the evaluation process encouraged active engagement and critical thinking beyond just grading. As one P7 explained: *"One of the things I keep thinking of is how important it always has been, but even more so now, we need to teach students critical thinking. If we do not start now, they will not know what to trust and what not to trust. The evaluation in this tool is a good starting point."*

5.2 Structured Evaluation Through Custom Topics and Linguistic Variation

Participants emphasized that AIBAT's structured evaluation approach, particularly the ability to assess performance across different linguistic variations, provided a critical framework for systematic evaluation. Having a well-defined structure allowed users to effectively test and refine the tool, as P1 noted: *"It tries to come up with different ways to say the same thing so that you don't have to create that exhaustive list of what would be official."* This suggests that without structured guidance, users may struggle to generate diverse and meaningful examples, limiting their ability to thoroughly assess AI outputs. Moreover, participants agreed that AIBAT's ability to automatically generate a variety of statements significantly reduced the cognitive burden associated with manually creating test cases. As P9 observed: *"Just the way that middle school students write — it varies so much that it's hard to capture all the possibilities. Like all the different ways they could say the same thing or you know, even they can say the same thing with commas."* Similarly, P3 stated that *"Not having to use my brain power to come up with all of those statements."* Overall, participants emphasized that AIBAT's ability to systematically capture linguistic variation strengthened human-AI collaboration by reducing the manual effort required for grading while still allowing educators to apply their expertise in evaluating AI behavior.

5.3 Evaluating AI Grading Across Linguistic Variations for Inclusive and Context-Aware Assessment

The Linguistic Variation feature in AIBAT provides educators with a structured way to assess AI grading across diverse linguistic expressions, allowing for a more nuanced evaluation of student responses. Our results show that linguistic variation may serve as a parameter for aligning grading practices with pedagogical goals. Specifically, it allows educators to ensure that students are not unfairly penalized for deviations in language use when they successfully convey core concepts. As P12 described, the feature makes it possible to adjust AI grading to be *"a little bit more permissive"* under certain conditions, ensuring that rigid language rules do not obscure a student's understanding of the material. P4 also emphasized this challenge, particularly for students who struggle with writing mechanics: *"If students really struggle with writing, I can understand their reading — or their writing, rather — and know that they grasp the main idea... even though they might not be using all the correct vocabulary, they might not be using correct punctuation or any punctuation at all."* Moreover, P3 recognized the need for flexible grading criteria based on student skill levels: *"when you're dealing with certain [students'] ages, there's a more level playing field. But in middle school, you know, I have students with probably a third-grade reading and writing level all the way up to like a high, you know, middle of high school writing level...I would have different grading criteria in different grade levels."* This perspective highlights a key challenge in AI-driven assessment — balancing linguistic accuracy with instructional goals. AIBAT's scalability through linguistic variation enables teachers to see where AI applies "stricter" or more "lenient" grading, allowing them to adjust their decisions accordingly.

5.4 Interpreting Model Bias Through Visualization to Enhance Transparency and Decision-Making

Our results show that Behavior Visualizations for Evaluation and Sensemaking provided users with an intuitive means to recognize model biases, facilitating sensemaking and encouraging deeper reflection on AI decision-making processes. Beyond detecting biases, the visualization actively engaged users in deliberating their role in addressing systemic inconsistencies. As P2 reflected: *"That visualization makes me think — Do I need to do more work with the tool to get consistency across the board, or... just accept that I'm not going to track Spanish, for instance, because it's already doing a good job?"* By surfacing these discrepancies in an accessible format, AIBAT's visualization feature supported a more nuanced approach to AI trust and oversight. Furthermore, P4 recommended reorganizing the data to prioritize areas of greatest disagreement, suggesting that results be sorted *"from least agreed to most disagreed"*, which underscores the practical need for structured prioritization, enabling users to quickly identify and address problematic classifications, streamlining the auditing and decision-making process.

6 Discussion and Future Directions

We introduce AIBAT, an iterative workflow and tool designed for teacher-driven contextual evaluation of language models in education. Unlike traditional model evaluation approaches that emphasize generalization performance and quantitative benchmarks, AIBAT shifts the focus toward contextualizing model harms in the pedagogical context and instructional goals.

Our think-aloud study findings highlight the role of AIBAT's evaluation mechanism in fostering user autonomy and trust in AI-generated assessments. Participants valued the ability to engage with AI decisions rather than passively accepting them, describing the process as a "conversation" that enhanced transparency and control. This interactive approach empowered users to critically assess AI outputs, reinforcing the idea that AI should serve as a collaborative tool rather than an unquestioned authority. Additionally, AIBAT's structured evaluation framework, particularly its ability to assess linguistic variations, enabled systematic and efficient grading assessment. By generating diverse student responses automatically, the tool reduced cognitive effort while ensuring comprehensive evaluation. Participants also emphasized the importance of linguistic variation in grading fairness, noting that it helped avoid penalizing students for language differences while aligning assessments with pedagogical goals. The ability to adjust AI grading criteria based on student proficiency levels allowed for more inclusive and context-aware evaluations. Furthermore, visual representations of model behavior provided an intuitive way to detect AI biases, facilitating informed decision-making and strengthening trust. Some participants suggested prioritizing areas of greatest disagreement to streamline the auditing process. Overall, AIBAT's evaluation, linguistic flexibility, and bias visualization features enhanced human-AI collaboration, improving transparency, trust, and usability.

AIBAT also inherits limitations that we have identified as areas for future iterations. One challenge is expanding beyond single-sentence test cases to support context-aware evaluation. Currently, many AI evaluation methods assess responses in isolation, which may not fully capture how model outputs align with broader pedagogical structures, discourse patterns, and curricular themes. Another key direction is deepening AIBAT's interactive feedback mechanisms. In its current form, AIBAT enables educators to provide constructive evaluations of AI-generated responses, but future work can strengthen its iterative feedback loop, allowing teachers not only to identify issues but also receive explanations of why the model produced a particular response and make adjustments dynamically to improve alignment with pedagogical expectations.

In conclusion, we argue that current AI evaluation frameworks offer limited support for teachers to critically assess AI models based on the specific knowledge, skills, and values emphasized in their curriculum. AIBAT addresses this gap by offering a scaffolded evaluation process that enables teachers to customize evaluation topics, surface critical examples where model outputs misalign with pedagogical expectations, and engage in deeper sensemaking. This approach empowers teachers to be active participants in AI development, rather than positioning them as passive adopters.

Acknowledgments. This work was supported by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin–Madison, with funding from the Wisconsin Alumni Research Foundation. We thank Nora Tseng, Lucas Anderson, Pranav Poddutoori, Aditya Anand, and Daryush Ghadiali. We also appreciate Dana Gnesdilow, Ananya Ganesh, Nate Smith, Kaycie Barron, Sandra Taylor-Marshall, Kiyo White, and all the think-aloud participants for their valuable feedback.

References

1. Vakil, S., McKinney de Royston, M.: Youth as philosophers of technology. Mind Cult. Act. **29**(4), 336–355 (2022).
2. Vygotsky, L.S.: Mind in Society: The Development of Higher Psychological Processes. Harvard University Press (1978).
3. Karumbaiah, S., Ocumpaugh, J., Baker, R.S.: Context matters: Differing implications of motivation and help-seeking in educational technology. Int. J. Artif. Intell. Educ. **1**, 1–40 (2021).
4. Bommasani, R., Liang, P., Lee, T.: Holistic evaluation of language models. Ann. N.Y. Acad. Sci. (2023).
5. Jacovi, A., Marasović, A., Miller, T., Goldberg, Y.: Formalizing trust in artificial intelligence: Prerequisites, causes, and goals of human trust in AI. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, pp. 624–635 (2021).
6. Reiser, B.J.: Scaffolding complex learning: The mechanisms of structuring and problematizing student work. J. Learn. Sci. **13**(3), 273–304 (2004).
7. Holstein, K., Aleven, V., Rummel, N.: A conceptual framework for human–AI hybrid adaptivity in education. Artif. Intell. Educ. **12163**, 240–254 (2020).
8. Suresh, H., Shanmugam, D., Chen, T., Bryan, A.G., D'Amour, A., Guttag, J., Satyanarayan, A.: Kaleidoscope: Semantically-grounded, context-specific ML model evaluation. ACM Hum. Factors Comput. Syst. **1**, 1–13 (2023).
9. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: Behavioral testing of NLP models with CheckList. arXiv preprint arXiv:2005.04118 (2020).
10. Baker, R.S., Hawn, A.: Algorithmic bias in education. Int. J. Artif. Intell. Educ. **1**, 1–41 (2021).
11. Yin, Y., Karumbaiah, S., Acquaye, S.: Responsible AI in Education: Understanding Teachers' Priorities and Contextual Challenges. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), Athens, Greece, June 23–26, 2025, pp. 1–21. ACM, New York (2025). https://doi.org/10.1145/3715275.3732176
12. Raji, D., Denton, E., Bender, E.M., Hanna, A., Paullada, A.: AI and the Everything in the Whole Wide World Benchmark. In: Neural Information Processing Systems (2021).
13. Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Model. User-Adapt. Interact. **4**, 253–278 (1994).
14. D'Ignazio, C., Klein, L.F.: Data Feminism. MIT Press (2020).
15. Dillenbourg, P.: The evolution of research on digital education. Int. J. Artif. Intell. Educ. **26**, 544–560 (2016).
16. Crenshaw, K.W.: On Intersectionality: Essential Writings. The New Press (2017).
17. Puntambekar, S.: Distributed scaffolding: scaffolding students in classroom environments. Educ. Psychol. Rev. **34**(1), 451–472 (2022).

18. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D. (eds.): Dataset Shift in Machine Learning. MIT Press (2008).
19. Schofield, J.W., Eurich-Fulcer, R., Britt, C.L.: Teachers, computer tutors, and teaching: The artificially intelligent tutor as an agent for classroom change. Am. Educ. Res. J. **31**(3), 579–607 (1994).
20. van de Pol, J., Volman, M., Oort, F., Beishuizen, J.: Teacher scaffolding in small-group work: An intervention study. J. Learn. Sci. **23**(4), 600–650 (2014).
21. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do ImageNet classifiers generalize to ImageNet? In: International Conference on Machine Learning, pp. 5389–5400. PMLR (2019).
22. Engelbart, D.C., English, W.K.: A research center for augmenting human intellect. In: Proceedings of the December 9–11, 1968 Fall Joint Computer Conference, Part I, pp. 395–410 (1968).
23. Tabak, I.: Synergy: A complement to emerging patterns of distributed scaffolding. J. Learn. Sci. **13**(3), 305–335 (2004).
24. Koedinger, K.R., Corbett, A.T.: Cognitive Tutors: Technology bringing learning science to the classroom. In: Sawyer, K. (ed.) The Cambridge Handbook of the Learning Sciences, pp. 61–78. Cambridge University Press (2006).
25. Karumbaiah, S., Borchers, C., Shou, T., Falhs, A.C., Liu, P., Nagashima, T., Rummel, N., Aleven, V.: A spatiotemporal analysis of teacher practices in supporting student learning and engagement in an AI-enabled classroom. Artif. Intell. Educ. **450**, 450–462 (2023).
26. Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of "bias" in NLP. arXiv preprint arXiv:2005.14050 (2020).
27. Shneiderman, B.: Human-Centered AI. Oxford University Press (2022).
28. Rajani, N., Liang, W., Chen, L., Mitchell, M., Zou, J.: SEAL: Interactive tool for systematic error analysis and labeling. arXiv preprint arXiv:2210.05839 (2022).
29. Birhane, A.: Algorithmic injustice: A relational ethics approach. Patterns **2**(2) (2021).
30. Patel, K., Fogarty, J., Landay, J.A., Harrison, B.: Investigating statistical machine learning as a tool for software development. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 667–676. ACM, New York (2008).
31. Pea, R.D.: The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. J. Learn. Sci. **13**(3), 423–451 (2004).
32. Martin, N.D., Tissenbaum, C.D., Gnesdilow, D., Puntambekar, S.: Fading distributed scaffolds: The importance of complementarity between teacher and material scaffolds. Instr. Sci. **47**(1), 69–98 (2019).
33. Paquette, L., Ocumpaugh, J., Li, Z., Andres, A., Baker, R.: Who's learning? Using demographics in EDM research. J. Educ. Data Min. **12**(3), 1–30 (2020).
34. Kozulin, A.: Psychological tools and mediated learning. In: Kozulin, A., Gindis, B., Ageyev, V.S., Miller, S.M. (eds.) Vygotsky's Educational Theory in Cultural Context, pp. 15–38. Cambridge University Press (2003).