Rethinking NLP Evaluation for LLM Assistants in Education: A Human-Centered Evaluation Framework

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly adopted as educational assistants and feedback providers; however, dominant NLP evaluation practices continue to emphasize technical metrics such as accuracy, fluency, and automated judgment. We argue that these approaches are misaligned with educational contexts because they overlook human values, learner agency, and pedagogical goals. This paper presents an argument-driven critique of NLP evaluation practices as applied to educational LLMs and introduces a literature-informed, human-centered, and sociotechnical evaluation framework with guiding questions to support its use. We highlight criteria such as explainability, consistency, and refinement, which are critical to human-AI interactive instructional effectiveness but absent from existing NLP benchmarks, and argue for more human-aligned, co-creative, and long-term evaluation practices for educational LLM systems.

1 Introduction

Large language models (LLMs) are being rapidly adopted in educational settings, yet their evaluation remains governed by general-purpose NLP benchmarks. While these benchmarks support scalable and standardized comparisons, they are often misaligned with the pedagogical, contextual, and ethical requirements of educational use. This paper asks which evaluative criteria are overlooked by dominant NLP benchmarks when educational language models are used as collaborative, teachable, and value-aligned agents. We critically examine current NLP evaluation practices in the context of educational LLMs and argue that existing metrics insufficiently capture learning processes, learner agency, educational context, and stakeholder values. As a starting point for future work, we introduce a human-centered, value-aligned evaluation framework and distill early takeaways to inform the design, benchmarking, and assessment of educational language models in future NLP research.

2 NLP Evaluation critiques

2.1 Dominant NLP Evaluation Practices

NLP evaluation has traditionally emphasized task correctness through metrics such as accuracy, precision, recall, and F-score measures (Hutchinson et al., 2022). While these metrics capture specific error types, most of them ignore contextual factors. Widely used automated metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are well-suited to short, constrained tasks with a single correct answer, but break down for longer, open-ended generations where multiple valid responses may exist. Even more semantically informed metrics, such as BERTScore improve alignment at the embedding level, yet still struggle to capture reasoning quality, domain-specific correctness, and human preferences. This highlights a fundamental mismatch between dominant NLP evaluation metrics and the evaluative demands of complex generative tasks (Zhang et al., 2019).

LLM-as-a-Judge frameworks have recently adopted prominence due to their low cost, ease of use, and reported correlations with human stylistic preferences. However, prior work demonstrates that LLM judges exhibit systematic biases that can distort evaluation outcomes, often relying on output accuracy or text similarity (Li et al., 2025; Szymanski et al., 2025). As a result, LLM-as-a-Judge evaluations risk presenting an illusion of holistic assessment while remaining grounded in opaque evaluative criteria. Benchmarks such as MT-Bench (Zheng et al., 2023), and Arena-Hard-Auto (Li et al., 2024) exemplify this trend, emphasizing conversational ability but offering limited insight into model behavior, or contextual appropriateness. Benchmark suites such as GLUE and SuperGLUE(Wang et al., 2018) further exemplify dom-

inant evaluation practices in NLP by aggregating diverse natural language understanding tasks and emphasizing sample-efficient learning and cross-task transfer (Wang et al., 2018, 2019). However, uneven data availability across tasks, genre mismatches between training and test sets, and reliance on privately held test data limit transparency, raising concerns about whether benchmark performance reliably translates to real-world deployment settings.

Overall, evaluating whether an NLP model is appropriate for real-world use requires more than benchmark scores and accuracy, it requires attention to context and responsibility. Prior work highlights the need to avoid the "portability trap", the assumption that what works well in one context will generalize across others by accounting for how models interact with specific human, social, and technical environments (Selbst et al., 2019; Martin Jr et al., 2020). Such evaluations should incorporate multiple metrics, distributional analyses, and qualitative assessments of impact (Mitchell et al., 2019). They should also be reflexive about social and intentional factors in dataset construction and model development (Scheuerman et al., 2021; Miceli et al., 2021; Mantelero, 2018; McGregor et al., 2019), explicitly document data provenance and known biases (Andrus et al., 2021), and consider broader obligations such as ethical alignment, institutional accountability, and potential privacy leakage (Mantelero, 2018; McGregor et al., 2019; Yeom et al., 2018; Raji et al., 2021). Together, these considerations highlight that evaluating NLP systems for responsible deployment requires assessing not only performance, but also harms, benefits, and obligations within the application ecosystem (Hutchinson et al., 2022; Raji et al., 2021).

2.2 NLP Evaluation in Educational Contexts

Despite the growing adoption of large language models (LLMs) in educational settings, including reducing instructional workload through automated grading and content generation (Ali et al., 2021), supporting personalized learning interactions (Vincent-Lancrin and Van der Vlies, 2020), supporting digital and AI literacy (Eshet, 2004; Bundy, 2017), and fostering higher-order cognitive skills such as reasoning, creativity, and computational thinking (Baer, 1998; Amabile, 2018; Grover and Pea, 2013), educational contexts have received limited attention in dominant NLP evaluation research, which have primarily been developed for general-purpose language tasks. As a result, general NLP evaluation approaches often fail in the educational context because they prioritize surface-level correctness and fluency over deep-level learning processes. Accurate or well-formed language does not guarantee conceptual understanding, appropriate scaffolding (Vygotsky and Cole, 1978; Papert and Harel, 1991), or productive struggle (Kapur, 2008), nor does it adequately reflect learners' needs, engagement, or cognitive growth (Shaffer and Resnick, 1999; Shneiderman, 2022; Friedman et al., 2002). These limitations also lead to education-specific language challenges, such as domain-specific terms, ambiguity, sarcasm, and imbalanced data, which make human feedback and learning signals more difficult to interpret. As emphasized in prior work on educational measurement, meaningful evaluation must consider incorporating formative, contextual, and human-centered outcomes that capture how AI systems support learning in authentic, situated environments (Shute and Ventura, 2013; Thomas et al., 2024; Anderson et al., 1990; Shaffer and Resnick, 1999; Friedman et al., 2002).

2.3 Conceptual Human-centered Evaluation Framework

In this section, we revisit our research question and propose an initial NLP evaluation framework, along with a set of guiding questions grounded in prior literature. The framework targets observable linguistic behaviors and focuses on how instructional qualities, particularly those that emerge when educational scenarios act as stress tests, such as explanation transparency, behavioral consistency, and iterative refinement, are reflected in generated language across interaction turns. By defining each component and synthesizing the analysis into preliminary implications for future evaluation and development of NLP-based generative tutoring agents, we highlight limitations of automated evaluation practices, including LLM-as-a-judge approaches, as well as sociotechnical considerations that are often overlooked by dominant NLP benchmarks. This framing provides a starting point for iterative refinement through future research.

2.3.1 Evaluation Framework Components for NLP

We focus on a subset of our framework's components that most directly surface misalignments between current NLP benchmarks and the evaluation

needs of educational LLMs. These components emphasize model behaviors that are critical to instructional effectiveness.

Explainability Explainability refers to an agent's ability to present its reasoning and decision-making processes in clear, contextually meaningful, and human-understandable terms (Nauta et al., 2023; Arrieta et al., 2020; Guidotti and Ruggieri, 2019). In educational contexts, explanations are core instructional acts rather than auxiliary outputs. However, current NLP evaluations rarely assess whether explanations are faithful, interpretable, or supportive of human understanding. Relevant evaluation sub-components include reasoning clarity, traceability between inputs and outputs, and alignment between explanations and task outcomes (Rosenfeld, 2021; Silva et al., 2023; Ribera and Lapedriza, 2019; Bommasani et al., 2021; Bransford et al., 2000; Abbas et al., 2022; Amershi et al., 2019; Guidotti et al., 2018; Liao et al., 2020).

Consistency Consistency captures the stability and trustworthiness of system behavior under similar conditions, including alignment across prompts, languages, situations, and evaluators (Nauta et al., 2023; Nielsen, 1995). While educational trust depends on predictable feedback, most benchmarks report only aggregate scores, masking behavioral variance. Indicators of consistency include output stability across runs, sensitivity to prompt perturbations, where minor rephrasings of semantically equivalent prompts lead to divergent outputs, and cross-evaluator agreement (Carvalho et al., 2019; Robnik-Šikonja and Bohanec, 2018; Vanden Abeele et al., 2012).

Refinement Refinement illustrated an agent's capacity to support iterative human–AI interaction, including user-guided error correction, clarification of vague or biased feedback, and traceable revisions over time (Pan et al., 2024; Wang et al., 2024; Nielsen, 1995). Existing NLP evaluations primarily evaluate static, single-turn outputs, implicitly assuming quality is fixed at generation time. In contrast, educational use is inherently iterative. Evaluating refinement behaviors such as responsiveness to feedback, coherence across revisions, time-to-refine, value alignment over revisions, these metrics or indicators are largely absent from current evaluation components(Guo et al., 2024; Wang et al., 2024; Pan et al., 2024; Vasconcelos et al., 2022; Bryant et al., 2017; Hong et al., 2024; Jurenka et al., 2024).

2.3.2 Reflecting Evaluation Through Questions

We re-frame the evaluation of NLP systems in educational contexts through a small set of guiding questions grounded in a human-centered framework(Jacovi and Goldberg, 2020; Shneiderman, 2022). These questions present how evaluators can move beyond outcome-focused judgments toward more transparent, process-oriented evaluation. Practitioners may select, adapt, or extend these questions based on their roles, contexts, and decision-making needs.

Explainability

- Does the AI explain its decisions or recommendations through its generated language in a clear and contextually meaningful way?
- Are sources, assumptions, or reasoning made visible to the user?
- To what extent would a user trust the system's explanation, and why?
- How could a user verify the explanation if needed?

Consistency

- Does the system produce linguistically consistent outputs across similar prompts, conditions, or learning contexts?
- Are evaluations stable across sessions or human evaluators?

Refinement

- Does the AI support iterative revision of its generated responses across multiple attempts rather than providing a single final answer?
- Do users understand why a change or refinement was suggested?
- Does the refinement process remain aligned with the learner's context and goals?

These questions are intended as starting points for reflective evaluation, helping surface pedagogical, social, and human-centered considerations that support discussion, comparison, and reflection across stakeholders.

2.3.3 Illustrative Application: Evaluating an LLM assistant Response

To show how our framework presents differences missed by current NLP metrics, consider an LLM programming assistant responding to the prompt: "Why does the `fitMedia` call in this Python code produce no output?" One response provides a direct answer by identifying a misspelled sound name, while another also explains how fixing errors allows the loop to run and connects the fix to curriculum-aligned debugging. Standard automated metrics such as BLEU, embedding-based similarity, and LLM-as-a-judge score, rate both responses similarly as fluent and correct. In contrast, our human-centered evaluation framework distinguishes the second response by its stronger explainability and closer alignment with instructional debugging practices in educational stress-test scenarios.

3 Implications for NLP Evaluation Research

Our analysis highlights key limitations in dominant NLP evaluation practices, particularly the common assumptions about correctness, fluency, and automated judgment when applied to educational LLMs. We distill two early implications that clarify why current evaluation practices are insufficient for human-centered, instructional language models.

3.1 Implication 1: Limitation of LLM-as-a-Judge for Evaluation

LLM-as-a-Judge frameworks offer scalability and convenience and have shown promise in correlating human preference judgments. However, our analysis identifies important limitations in their applicability to pedagogical evaluation. Educational quality depends on domain expertise, sensitivity to learner context, and value-based judgment, which cannot be reliably inferred from surface-level textual comparison alone. In instructional settings, LLM judges tend to prioritize fluent responses while overlooking misleading explanations, inappropriate levels of assistance, or misalignment with learning objectives. These patterns indicate that LLM-based judgment captures only a partial view of educational quality and highlights the need for hybrid evaluation approaches that combine automated judgment with human expertise.

3.2 Implication 2: Evaluation as a Sociotechnical Process

Dominant NLP evaluation practices focus on single-turn correctness and surface-level similarity metrics, such as BLEU, ROUGE, Exact Match, and embedding-based overlap scores. These metrics work well for short tasks with one correct answer, but they do not capture how educational interactions involve over time through dialogue and adaptive scaffolding. In educational settings, correctness alone does not indicate whether a model helps learners understand concepts, supports productive struggle, or aligns with instructional goals. These limitations highlight evaluation as a sociotechnical process shaped by human values, institutional expectations, and interactional context. Framing evaluation around stakeholder-aligned questions makes these assumptions explicit and surfaces concrete aspects of model behavior, positioning evaluation as an integral practice in how educational language models are assessed and governed.

Together, these implications suggest that future NLP benchmarks and evaluation methods should incorporate interaction-level metrics, multi-turn consistency analyses, and hybrid human–AI evaluation approaches, rather than relying solely on single-turn automated judgments.

4 Limitations

This work presents a conceptual analysis and does not involve human subjects, data collection, or system deployment at the current stage. The proposed framework has not yet been empirically validated or evaluated with relevant stakeholders, and the current discussion focuses on LLMs in educational contexts.

5 Conclusion and Future Work

This paper argues that current NLP evaluation practices exhibit gaps when applied to educational LLMs and emphasizes human-centered criteria such as explainability, consistency, and refinement in NLP evaluation. At this stage, the work presents a conceptual analysis without empirical validation, and the discussion is limited to LLMs in educational contexts. Looking ahead, we plan to iteratively refine the framework through feedback from interdisciplinary stakeholders in education and to conduct empirical and cross-domain studies to examine how its evaluation principles generalize within and beyond educational domains.

5.1 References

Acknowledgments

References

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

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

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

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

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

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 249–260.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, and 1 others. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

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

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

John D Bransford, Ann L Brown, Rodney R Cocking, and 1 others. 2000. *How people learn*, volume 11. Washington, DC: National academy press.

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

Alan Bundy. 2017. Preparing for the future of artificial intelligence.

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

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

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

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

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

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

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

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

Ben Hutchinson, Negar Rostamzadeh, Christina Greer, Katherine Heller, and Vinodkumar Prabhakaran. 2022. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1859–1876.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, and 1 others. 2024. Towards responsible development of generative ai for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*.

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

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. 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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Alessandro Mantelero. 2018. Ai and big data: A blueprint for a human rights, social and ethical impact assessment. *Computer Law & Security Review*, 34(4):754–772.

Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Extending the machine learning abstraction boundary: A complex systems approach to incorporate societal context. *arXiv preprint arXiv:2006.09663*.

Lorna McGregor, Daragh Murray, and Vivian Ng. 2019. International human rights law as a framework for algorithmic accountability. *International & Comparative Law Quarterly*, 68(2):309–343.

Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting computer vision datasets: An invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 161–172.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

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

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

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

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

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

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

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

Morgan Klaus Scheuerman, Alex Hanna, and Remi Denton. 2021. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68.

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

Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.

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

Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. 2023. 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.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966.

Danielle R Thomas, Jionghao Lin, Erin Gatz, Ashish Gurung, Shivang Gupta, Kole Norberg, Stephen E Fancsali, Vincent Aleven, Lee Branstetter, Emma Brunskill, and 1 others. 2024. 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.

Vero Vanden Abeele, Erik Hauters, and Bieke Zaman. 2012. 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.

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

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

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.

Han Wang, An Zhang, Nguyen Duy Tai, Jun Sun, Tat-Seng Chua, and 1 others. 2024. Ali-agent: Assessing llms' alignment with human values via agent-based evaluation. *Advances in Neural Information Processing Systems*, 37:99040–99088.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Example Appendix

This is an appendix.