Evaluating Multi-turn Human-AI Interaction

Shi Ding
Georgia Institute of Technology
Atlanta, Georgia, USA
sding84@gatech.edu

Sijian Tan
Georgia Institute of Technology
Atlanta, Georgia, USA
stan43@gatech.edu

Abstract

Large language models (LLMs) are increasingly used as collaborative assistants, yet dominant NLP evaluation practices remain centered on aggregate metrics such as accuracy and fluency. These approaches often overlook behaviors that are critical in human-facing settings (e.g., consistency across multiple turns and iterative refinement). In this paper, we examine limitations of current NLP evaluation practices and introduce TCR, a structured framework for evaluating human–AI interaction using educational LLM assistants as an illustrative example. TCR emphasizes dimensions such as transparency, consistency, and refinement. We further present structured evaluation prompts and illustrative interaction examples demonstrating how structured evaluation can complement aggregate metrics and LLM-as-a-judge approaches. Our work highlights the need for more human-centered evaluation practices for interactive LLM systems.

1 Introduction

Large language models (LLMs) are increasingly adopted in human-facing settings, yet their evaluation remains largely governed by general-purpose NLP benchmarks. While these benchmarks support scalable and standardized comparison, they often provide limited insight into interactive behaviors that emerge in real-world use(Ding and Magerko, 2026). This paper examines gaps in current LLM evaluation practices and introduces **TCR (Transparency, Consistency, and Refinement)**, a structured framework for analyzing model behaviors in multi-turn human–AI interaction. Using educational settings as an illustrative example, we discuss how structured evaluation may reveal limitations often overlooked by dominant NLP evaluation approaches and outline implications for the design and assessment of interactive LLM systems. We take the position that aggregate-centric evaluation may overlook critical behaviors in multi-turn interaction, highlighting the need for more interpretable behavioral evaluation. We make two contributions in this position paper. First, we identify gaps between dominant NLP evaluation practices and real-world interactive AI systems. Second, we introduce TCR, a structured framework that supports interaction-level evaluation beyond aggregate correctness metrics for interactive LLM systems.

2 Current Model Evaluation

2.1 Aggregate Evaluation

Dominant NLP evaluation practices have traditionally emphasized correctness-oriented metrics such as accuracy, precision, recall, and F-score measures (Hutchinson et al., 2022). These approaches primarily rely on reference-based metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), as well as reference-free approaches including LLM-as-a-judge frameworks. While these methods support scalable and standardized comparison, they remain limited for evaluating open-ended and human-facing generation tasks(Kashid and Bhattacharyya).

LLM-as-a-judge frameworks have recently gained prominence due to their scalability(Wang et al., 2018, 2019). However, prior work shows that LLM judges exhibit systematic biases that can distort evaluation outcomes (Li et al., 2025; Szymanski et al., 2025). In many cases, these approaches continue to prioritize surface-level fluency and preference signals while providing limited insight into how interaction quality evolves across multi-turn and human-facing contexts. Benchmarks such as MT-Bench (Zheng et al., 2023) and Arena-Hard-Auto (Li et al., 2024) exemplify this trend by emphasizing conversational performance while providing limited insight into model behavior across real-world interaction contexts.

While single-turn evaluation has long been dominant, recent work shows that static benchmarks

can obscure demographic biases and failures in AI-mediated interaction(Star, 2026), Recent work has increasingly explored multi-turn and long-horizon evaluation settings beyond static single-turn benchmarks. but often remain centered on preference modeling, or aggregate outcome optimization rather than broader human-centered interaction qualities (Chen et al., 2026; Abdulhai et al., 2026). Prior work further emphasizes that meaningful evaluation should account for broader contextual and human-centered considerations beyond benchmark performance (Selbst et al., 2019; Mitchell et al., 2019; Raji et al., 2021). Together, these considerations highlight that meaningful evaluation should account not only for performance outcomes, but also for the broader assumptions, behaviors, and consequences that emerge when language models are deployed in human-facing settings.

2.2 Human-Facing Evaluation

Prior research argues that evaluating machine learning systems requires attention not only to model outputs, but also to human behavior, institutional practices, and deployment contexts (Ji et al., 2023; Hendrycks et al., 2020; Shen et al., 2024). Interactive benchmarks such as Chatbot Arena (Chiang et al., 2024) introduce more dynamic evaluation settings, but often remain outcome-centric and rely heavily on aggregate success metrics. Human-facing LLM systems frequently require contextual guidance, iterative interaction, and adaptation across multiple turns. However, dominant evaluation approaches still emphasize correctness and fluency metrics, providing limited insight into interaction quality and evolving user context. Educational settings provide an illustrative example of these limitations because effective instructional support depends on iterative interaction, contextual guidance, and learner-centered communication (Vygotsky and Cole, 1978; Kapur, 2008; Ding et al., 2025). Correctness alone provides limited insight into whether AI systems support understanding, refinement, or productive interaction over time (Shneiderman, 2022; Friedman et al., 2002). Recent work explores multi-turn tutoring evaluation and personalized interaction modeling in educational AI. However, these approaches often remain centered on aggregate optimization and reward-based evaluation rather than structured analysis of interaction quality (Abdulhai et al., 2026; Wan et al., 2026). Educational settings therefore illustrate why human-facing multi-turn interaction may require more interpretable and process-oriented evaluation approaches.

Existing evaluation approaches for human-facing LLM systems are often either aggregate-centric or too coarse to capture model behaviors in authentic multi-turn settings. Unlike prior work that evaluates isolated usability or explainability factors, TCR focuses specifically on interaction-level behaviors emerging across multi-turn human–AI interaction. This position paper builds on a prior scoping review examining how AI agents are evaluated in human-facing environments (Levac et al., 2010). The review analyzed 126 sources across HCI, AI, and education venues and identified recurring evaluation themes(Ding and Magerko, 2025). Rather than proposing a broad evaluation taxonomy, this paper focuses on interaction quality in human-facing multi-turn LLM systems using education as an illustrative example.

3 TCR Evaluation Framework

3.1 Framework Overview

We introduce TCR as a lightweight framework for evaluating model behaviors in multi-turn human-facing AI systems. Rather than focusing only on traditional aggregate correctness or fluency, TCR emphasizes how models support users through iterative interaction, contextual guidance, and refinement across turns. The framework focuses on behaviors that are often overlooked in dominant NLP evaluation practices, particularly in settings where users rely on AI systems for explanation, clarification, and ongoing support. TCR also includes structured evaluation prompts designed to support more interpretable and process-oriented assessment practices. As shown in Table 1, TCR emphasizes dynamic multi-turn behaviors that are often overlooked by traditional evaluation approaches.

Traditional Evaluation	TCR
Correctness	**Transparency**
Preference	**Consistency**
Task Completion	**Refinement**

Table 1: Comparison between traditional evaluation approaches and the dynamic multi-turn behaviors emphasized by TCR.

3.1.1 Evaluation Dimensions

TCR focuses on three behavioral dimensions that commonly emerge in human-facing multi-turn interaction: transparency, consistency, and refinement. These dimensions are presented as an intentionally scoped framework rather than an exhaustive taxonomy, while remaining extensible across additional domains and interaction contexts.

Transparency Transparency refers to whether an agent communicates reasoning in clear and interpretable ways (Nauta et al., 2023; Arrieta et al., 2020; Guidotti and Ruggieri, 2019). In educational settings, explanations are part of the learning process rather than auxiliary outputs. Relevant indicators include reasoning clarity, traceability, and alignment between explanations and task outcomes (Silva et al., 2023; Bommasani et al., 2021; Bransford et al., 2000; Amershi et al., 2019; Radanliev, 2025; Kim et al., 2026).

Consistency Consistency refers to the stability of system behavior across similar conditions (Nauta et al., 2023; Nielsen, 1995). While benchmarks often report aggregate scores, they may overlook behavioral variance across prompts, runs, or evaluators. Relevant indicators include output stability and sensitivity to prompt perturbations(Carvalho et al., 2019; Robnik-Šikonja and Bohanec, 2018; Vanden Abeele et al., 2012; Lu and Song, 2025; Shen et al., 2025).

Refinement Refinement refers to an agent's ability to support iterative interaction through clarification, revision, and feedback over time (Wang et al., 2024; Pan et al., 2024). Unlike static single-turn evaluation, educational interaction is inherently iterative. Relevant indicators include responsiveness to feedback and coherence across revisions(Guo et al., 2024; Wang et al., 2024; Hong et al., 2024; Nam et al., 2026; Cai et al., 2025).

While these dimensions are defined conceptually above, Table 2 summarizes illustrative indicators for how they may appear in practice. The framework focuses on educational contexts as an example of human-facing multi-turn interaction.

While TCR is presented conceptually in this work, the framework may also support structured evaluation through lightweight annotation practices. These dimensions may further enable hybrid human–AI evaluation workflows and analyses such as agreement comparison across evaluators.

Dimension	Example Indicators
T Transparency	Step-by-step reasoning; source references
C Consistency	Output stability under prompt paraphrases; variance across runs
R Refinement	Responsiveness to feedback; clarity of revision rationale

Table 2: Illustrative indicators associated with TCR dimensions.

3.1.2 Structured Evaluation Prompts

TCR may also support reflective evaluation through lightweight structured prompts (Jacovi and Goldberg, 2020; Shneiderman, 2022). These prompts illustrate how evaluators may move beyond outcome-focused judgments toward more process-oriented assessment. Table 3 presents example prompts for each dimension.

Dimension	Example Prompts
Transparency	Does the AI explain its reasoning in a clear and understandable way? Are reasoning steps or supporting information visible to the user?
Consistency	Does the system remain consistent across similar prompts or interaction contexts? Are outputs stable across sessions or evaluators?
Refinement	Does the AI support iterative revision and improvement over multiple turns? Does the refinement process remain aligned with the user's goals and context?

Table 3: Illustrative reflective prompts associated with TCR dimensions.

3.1.3 Evaluation Example

To illustrate behaviors that may be overlooked by aggregate evaluation, Table 4 presents an example multi-turn interaction with an LLM programming assistant. While both responses may appear technically correct under aggregate or LLM-as-a-judge evaluation, the example highlights differences in transparency, contextual continuity, and refinement-oriented support across turns.

4 Implications for Model Evaluation

Our analysis highlights key limitations in dominant model evaluation practices, particularly the common assumptions around correctness, fluency, and

Turn	Interaction
1	**User:** Why does my `fitMedia` code not play any sound?
	Model A: This issue may be related to how the audio variable is referenced in your program. In programming, variables need to remain consistent across function calls and media operations.
	Model B: Check whether the sound variable name inside `fitMedia()` matches the variable you created earlier. For example, if you created `AK_UNDOG_808_1`, the same name should also appear inside `fitMedia()`.
2	**User:** I do not understand how to fix it.
	Model A: Variable consistency is important because programming functions depend on correct parameter references and object assignment.
	Model B: First, find the line where you created the sound variable. Then check the name used inside `fitMedia()`. The two names should match exactly. Even one missing letter in `AK_UNDOG_808_1` can stop the sound from playing.
Evaluation Summary: Aggregate or LLM-as-a-judge evaluation may rate both responses as generally correct, while TCR highlights differences in how the models support novice users through concrete guidance, step-by-step clarification, and iterative interaction.	

Table 4: Illustrative multi-turn interaction and evaluation comparison.

automated judgment in the evaluation of human-facing LLM systems. We distill two implications that clarify why current evaluation practices remain insufficient for interactive language models.

4.1 Implication 1: Limitations in LLM-as-a-Judge Systems

LLM-as-a-judge frameworks offer scalability and efficiency for automated evaluation. However, our analysis identifies limitations in their applicability to human-facing evaluation settings. Evaluating interaction quality often requires sensitivity to context, user goals, and iterative interaction patterns that cannot be reliably inferred through fluency- and similarity-based evaluation alone. In our illustrative example, LLM judges may prioritize technically correct responses while overlooking contextual misalignment, inconsistent behavior, or ineffective refinement support. These patterns suggest that LLM-as-a-Judge Systems provide only a partial view of interaction quality and highlight the need for hybrid evaluation approaches that combine automated auditing with structured human-centered assessment.

4.2 Implication 2: Evaluating Multi-Turn Interaction

Current evaluation approaches often collapse multi-turn interaction into aggregate correctness or preference signals, masking behaviors that emerge through refinement, clarification, and contextual adaptation over time. While these approaches support scalable comparison, they provide limited insight into how interactions evolve through dialogue, revision, and iterative refinement. These limitations highlight the sociotechnical nature of evaluation, which is shaped by human values, institutional expectations, and contextual interaction. As illustrated in our educational example, structure evaluation through structured evaluation prompts helps surface assumptions that are often hidden by aggregate evaluation practices and makes model behavior more transparent in interactive contexts.

Together, these implications suggest that future model evaluation methods should incorporate contextual evaluation, multi-turn consistency analysis, and hybrid human–AI assessment approaches.

5 Limitations and Future Work

While TCR is intentionally scoped in this work, future work may further validate and extend the framework through large-scale empirical studies and more accessible evaluation practices for diverse users and interaction contexts.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2026. Consistently simulating human personas with multi-turn reinforcement learning. *Advances in Neural Information Processing Systems*, 38:52920–52957.

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.

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.

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.

Yufan Cai, Zhe Hou, David Sanán, Xiaokun Luan, Yun Lin, Jun Sun, and Jin Song Dong. 2025. Automated program refinement: Guide and verify code large language model with refinement calculus. *Proceedings of the ACM on Programming Languages*, 9(POPL):2057–2089.

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

Boyuan Chen, Donghai Hong, Jiaming Ji, Jiacheng Zheng, Bowen Dong, Jiayi Zhou, Kaile Wang, Juntao Dai, Xuyao Wang, Qirui Zheng, and 1 others. 2026. Intermt: Multi-turn interleaved preference alignment with human feedback. *Advances in Neural Information Processing Systems*, 38.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Shi Ding and Brian Magerko. 2025. Rethinking ai evaluation in education: The teach-ai framework and benchmark for generative ai assistants. *arXiv preprint arXiv:2512.04107*.

Shi Ding and Brian Magerko. 2026. Bridging responsible ai and ai literacy: The teach-rai framework and toolkit for education, design, and research. In *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V. 2*, pages 1295–1296.

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

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

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*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

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*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

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

Harshvivek Kashid and Pushpak Bhattacharyya. Large language models for nlp evaluation: A survey.

Chanwoo Kim, Soham U Gadgil, and Su-In Lee. 2026. Transparency of medical artificial intelligence systems. *Nature Reviews Bioengineering*, 4(1):11–29.

Danielle Levac, Heather Colquhoun, and Kelly K O'brien. 2010. Scoping studies: advancing the methodology. *Implementation science*, 5(1):69.

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*.

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

Cheng Lu and Yang Song. 2025. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pages 50611–50649.

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.

Jaehyun Nam, Jinsung Yoon, Jiefeng Chen, Jinwoo Shin, Sercan Arik, and Tomas Pfister. 2026. Mle-star: Machine learning engineering agent via search and targeted refinement. *Advances in Neural Information Processing Systems*, 38:116692–116712.

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*.

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.

Petar Radanliev. 2025. Privacy, ethics, transparency, and accountability in ai systems for wearable devices. *Frontiers in Digital Health*, 7:1431246.

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*.

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.

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.

Fei Shen, Hu Ye, Sibo Liu, Jun Zhang, Cong Wang, Xiao Han, and Yang Wei. 2025. Boosting consistency in story visualization with rich-contextual conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6785–6794.

Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, and 1 others. 2024. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2406:1–56.

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

Michelle Star. 2026. Evaluating supportive llm behavior over multiple turns across demographics. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–4.

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.

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.

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

Yanming Wan, Jiaxing Wu, Marwa Abdulhai, Lior Shani, and Natasha Jaques. 2026. Enhancing personalized multi-turn dialogue with curiosity reward. *Advances in Neural Information Processing Systems*, 38:155857–155894.

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.

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.