AERA Chat: An Interactive Platform for
Automated Explainable Student Answer Assessment

Jiazheng Li[1*], **Artem Bobrov**[1*], **David West**[2], **Cesare Aloisi**[2], **Yulan He**[1,3]

[1]King's College London [2]AQA [3]The Alan Turing Institute

{jiazheng.li, artem.bobrov, yulan.he}@kcl.ac.uk {caloisi, dwest}@aqa.org.uk

Abstract

Generating rationales that justify scoring decisions has emerged as a promising approach to enhance explainability in the development of automated scoring systems. However, the scarcity of publicly available rationale data and the high cost of annotation have resulted in existing methods typically relying on noisy rationales generated by large language models (LLMs). To address these challenges, we have developed AERA Chat, an interactive platform, to provide visually explained assessment of student answers and streamline the verification of rationales. Users can input questions and student answers to obtain automated, explainable assessment results from LLMs. The platform's innovative visualization features and robust evaluation tools make it useful for educators to assist their marking process, and for researchers to evaluate assessment performance and quality of rationales generated by different LLMs, or as a tool for efficient annotation. We evaluated three rationale generation approaches on our platform to demonstrate its capability.

1 Introduction

Automated student answer scoring (ASAS) systems are vital educational NLP applications that streamline the manual grading process, offering a swift and consistent evaluation of student performance (Larkey, 1998; Alikaniotis et al., 2016; Dong et al., 2017). Traditional ASAS systems typically utilize unexplainable text classifiers built on pre-trained language models (Mayfield and Black, 2020; Xie et al., 2022), which take question, key answer elements, marking rubric, and student responses as input to derive marks.

The lack of transparency in traditional ASAS systems has prompted concerns about their trustworthiness in real-world assessments. Various approaches have been developed to address this issue by interpreting the model marking processes.

These include feature analysis (Tornqvist et al., 2023; Vanga et al., 2023) and visualizations of internal mechanisms, such as weights and attention (Alikaniotis et al., 2016; Yang et al., 2020). However, these interpretations often require a substantial understanding of NLP, which becomes a barrier for educators without technical backgrounds.

The recent development of large language models (LLMs) has introduced a new approach that leverages in-context learning and reasoning capabilities (Brown et al., 2020; Wei et al., 2022) to generate natural language rationales that justify model decisions (Camburu et al., 2018; Marasovic et al., 2022; Gurrapu et al., 2023). This approach improves explainability, making it more accessible to both educators and students. However, acquiring rationale annotations for ASAS datasets is costly. Most ground truth labels used for training rationale generation models are based on noisy rationales prompted from LLMs without human verification (Li et al., 2023a, 2024). *Establishing a platform for comparing and evaluating LLM-generated rationales is crucial for researchers to understand the models' capability and enhance the rationale quality in future developments.*

At the same time, many researchers have developed applications powered by LLM for educational purposes (Wang et al., 2024), such as student feedback platform (Matelsky et al., 2023; Tobler, 2024), student learning assistant systems (Park and Kulkarni, 2024; Kabir and Lin, 2023; Schmucker et al., 2024), or exercise generation frameworks (Xiao et al., 2023; Cui and Sachan, 2023). However, while existing educational interface developments primarily focus on applications that enhance teaching or learning, they often overlook the potential of using LLMs in the assessment stage. *Developing a visualization platform that grades student answers and generates rationales for assessments could enable educators to engage more effectively with the latest rationale generation technologies and sup-*
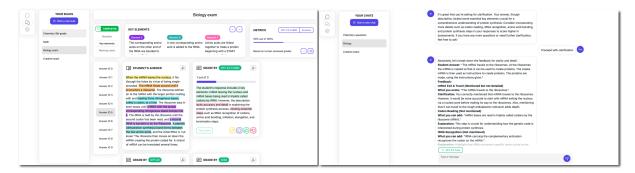
---

[*]Equal contribution.

Figure 1: Two main interfaces of the AERA Chat platform: *(1) Bulk Marking Interface (left side):* Users can enter question details and perform automated assessment on student answers in batches. Our platform supports using multiple LLMs concurrently, with the ability of highlighting key components in both the answers and the rationales provided. *(2) Chat Interface (right side):* Users can select an initialized question along with its assessed rationales to query LLMs for more detailed explanation of the marking decisions or reflections on any incorrect rationales.

port practical use in evaluating student work.

To *facilitate access to methods for generating assessment rationales* and *simplify the rationale evaluation and annotation process*, we introduce an interactive **A**utomated **E**xplainable student **R**esponse **A**ssessment platform: **AERA Chat**. Our platform features a newly designed interactive user interface, leveraging multiple LLMs as the backend for automated assessment and rationale generation. AERA Chat is designed to promote the practical application of explainable assessment methods in real-world education scenarios and to boost research by providing a unified platform for comparisons. To the best of our knowledge, AERA Chat is the first open-source interactive platform explicitly designed to utilize LLMs in explainable student answer scoring.

Our contributions are as follows:

- We developed **a novel interactive platform to perform simultaneous student answer assessment and rationale generation using multiple LLMs**, accessible through public API or local deployments.

- Our **visualization interface is innovatively designed to highlight key components in student answers and assessment rationales**, enhancing users' ability to focus on key parts of the context.

- Our system provides a comparative display of rationales generated by different LLMs and offers **a streamlined rationale annotation/verification environment through preference selection or direct annotation**.

- We showcase the efficacy of our platform's evaluation capability through automated metrics and human preference evaluations for rationale quality, using three LLMs on four datasets.

A demonstration video is available at https://www.youtube.com/watch?v=WIK2OM99Hb8.

2 Overview of AERA Chat

The AERA Chat platform, intended for both educators and AI researchers, offers a comprehensive environment for providing explainable and transparent assessment decisions in automated student answer scoring. The main features of our platform are §2.1 bulk explainable student answer assessment and rationale generation, with the flexibility to §2.2 interactively chat with LLMs for detailed explanation or reflection on assessment rationales.

2.1 Bulk Marking Interface

As shown on the left-hand side of Figure 1, the bulk marking interface allows users to set up a question and a batch of student answers they want to assess. Once those inputs are submitted to our system, the backend will concurrently process the student answers and query all user-selected LLMs for a scoring decision and a rationale that explains the assessment decision. Upon completion, our platform can highlight the critical components in student answers or rationales to draw users' attention to the most relevant parts of the context. Furthermore, our platform can automatically evaluate the LLMs' assessment performance and display the metric results in histograms. Moreover, the interactive platform serves as a convenient interface for annotating or verifying the rationales.

Bulk Initialization During the bulk initialization step, users can set up the question, key answer

(a) Key Component Highlighting   (b) Interactive Rationale Annotation   (c) Automated Assessment Evaluation
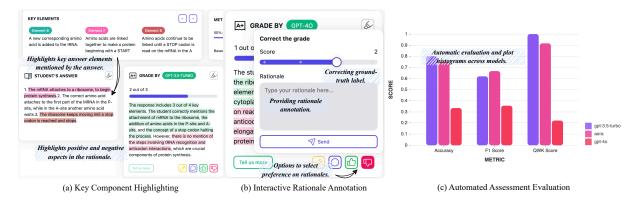
Figure 2: Key feature functionalities provided in the Bulk Marking Interface.

elements (the essential key phrases for a correct response), and a point-based marking rubric (marking criteria based on key elements) in the system. Our system automatically compiles the question information into a template prompt, denoted as Q_{info}, which is then provided to LLMs for assessment. Users can upload a batch of student answers they want to assess in a file, which may or may not include the ground-truth marks. We denote the set of user-uploaded student answers and their corresponding ground-truth marks, if available, as $D = (x_i, y_i)_{i=1}^{N}$, where N is the total number of student answers for the same question.

Automated Scoring and Rationale Generation Once the question information and student answers have been uploaded to the system, our platform can automatically assess student answers and generate rationales with the user-selected LLMs. For demonstration, we used two OpenAI models: GPT-3.5-turbo and GPT-4o, as well as our developed rationale generation model – the AERA model (Li et al., 2023a). Our system can be easily extended to include other models for student answer assessment and rationale generation.

To fulfil the requirement for simultaneously assessing scores and generating free-form text rationales, we instruct the LLM LLM$_\theta$ to format the output in JSON format (Li et al., 2023b). This generation process can be represented as:

$$(\hat{y}_i, \hat{r}_i) = \text{LLM}_\theta(x_i, Q_{\text{info}}), \qquad (1)$$

where $\hat{y}_i$ represents the predicted mark of the student answer and $\hat{r}_i$ is a textual rationale that justifies the marking decision.

Our database system will automatically monitor the assessment status. As shown on the left-hand side of Figure 1, once the assessment is completed, our platform will display rationales generated by

each LLM in parallel, presented as a card view. The score assessed by each LLM will be displayed at the top of its respective card, allowing users to compare the scoring decisions across LLMs easily.

Key Component Highlighting Verifying the faithfulness of the assessment rationales with respect to a student's answer is a complex task that requires a detailed understanding of the context. Our platform aims to enhance the users' reading experience by providing a clear, high-contrast indication of the relevant contexts using colours within the student answers and assessment rationales. Unlike the mechanism visualizations discussed in previous research (Alikaniotis et al., 2016; Yang et al., 2020), our highlighting feature acts as an independent semantic comparison tool, aiding in the comprehension of the LLM's decision-making process.

As shown in Figure 2 (a), users can choose to visualize the key answer elements mentioned in the student answers with the same colour assigned or to visualize the positive aspects (reasons for awarding points) and negative aspects (reasons for deducting points) in the LLM-generated rationales. This feature is implemented by automatically querying GPT-4o to provide word-level tagging and highlight the context using contrasting colours.

Annotation Toolkit Current evaluations of rationale quality have typically been done manually. To provide tools for rationale evaluation and annotation, we have implemented three functionalities, as demonstrated in Figure 2 (b):

(1) Ground-truth label correction: As noted in (Li et al., 2023a), some ground-truth labels provided by publicly available ASAS dataset (Hamner et al., 2012) could be wrong. Therefore, we offer users with an option to correct these ground-truth labels. This feature could be benefited from the multi-LLM assessment functionality; when the

3

LLMs agree on a particular score, users can recheck the original label if it appears to be incorrect.

(2) Rationale preference selection: Evaluating the quality of rationales presents challenges due to a lack of evaluation metrics; existing works often model the qualitative evaluation task as binary preferences (Li et al., 2023a, 2024), where the factually correct and more detailed rationale is preferred. Consequently, our platform includes options for users to select "preferred" or "not preferred" rationales. These selections are automatically recorded in the system's database, allowing researchers to construct preference data that can be used to leverage the latest reinforcement learning from human feedback techniques (Rafailov et al., 2023; Ouyang et al., 2022) in training rationale generation LLMs.

(3) Rationale annotation: If none of the candidate rationales generated by LLMs is correct, our platform allows users to alternatively submit their assessment rationales annotation as annotated data for supervised fine-tuning.

Assessment Performance Evaluation If the user has uploaded ground-truth labels for student answers to our interface, as shown in Figure 2 (c), our platform will automatically evaluate the performance of assessment decisions across selected LLMs. This evaluation is visualized in a histogram displaying the Accuracy, macro F1 Score, and QWK (Quadratic Weighted Kappa) score. We adopt the Accuracy and F1 Score implementations from the Sci-kit Learn package[1], and implemented the QWK score using the following equation (Reighns, 2020):

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} O_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} E_{ij}} \qquad (2)$$

where k is the number of score categories, w is the weighted matrix, calculates as: $w_{i,j} = \frac{(i-j)^2}{(k-1)^2}$. O is a $k \times k$ observed agreement matrix, and E is the $k \times k$ expected agreement matrix.

2.2 Chat Interface

To leverage the extensive chat capabilities of LLMs and harness their potential to assist with explainable student answer scoring, as shown on the right-hand side of Figure 1, our platform includes a chat interface to allow users to bring question information and rationales from the bulk marking system to interact with LLMs.

Educators can use these functions to ask LLMs to provide more detailed explanations for ambiguous assessment rationales. Conversely, researchers can utilize LLMs to reflect on their incorrect assessment rationales and regenerate assessment decisions. The Chat Interface is equipped with various LLM choices; if a user is not satisfied with the response from the current LLM, they can choose to integrate with a more powerful LLM (e.g., GPT-4o) or a more specialized model (e.g., AERA). These chats will be automatically recorded in our system's database so that researchers can utilise them as potential training data for improving the rationale generated by LLMs.
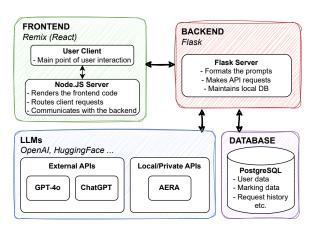

Figure 3: System Architecture of AERA Chat.

3 Implementation Details

AERA Chat is designed using a microservices architecture that integrates multiple LLMs through a unified web interface. We use Docker[2] for the modularization of services, which enhances modularity and improves the system's scalability and maintainability. This modularization strategy enables flexible deployment of AERA Chat using individual or distributed across multiple servers.

As depicted in Figure 3, AERA Chat's system architecture includes four main components: the §3.1 Frontend, §3.2 Aggregation Backend Layer, §3.3 LLM services, and §3.4 Database.

3.1 Frontend

The frontend provides a responsive web interface tailored for educators and researchers to engage with the system. It is designed with optimal usability in mind, featuring functionalities such as login, registration, chat interactions, bulk assessments, and comprehensive feedback mechanisms.

---

[1] https://scikit-learn.org/

[2] https://www.docker.com/

4

The frontend is developed using the Remix[3] framework, which builds on React. To achieve both peak performance and a contemporary aesthetic, we use isomorphic TypeScript, primarily for server-side rendering, along with a nested routing approach. Our frontend design efficiently shifts a significant processing load from the client to the server side.

3.2 Aggregation Backend Layer

The backend layer serves as the backbone of the AERA Chat platform, integrating all other layers. It operates based on a REST API, handling requests over HTTP connections without maintaining the state of each connection. This layer is implemented using the Flask[4] framework. In addition to handling standard HTTP requests, which are inherently blocking, we have integrated WebSocket technology. This allows for a persistent connection between the client and server, enabling the client to immediately receive updates from the server without the need for repeated refresh requests.

3.3 LLM Services

Student answer assessment is a pivotal task in educational settings. Accordingly, users of our system can opt to employ publicly available API-based LLMs, such as GPT-4 (OpenAI et al., 2024) or Gemini (Team et al., 2024), or they may choose to develop and utilize privately trained, customized LLMs to enhance question-specific performance or ensure better data privacy. In our implementation, we integrated the OpenAI API service as an example of external LLM resources. Additionally, we developed a local API module that allows users to deploy any privately trained models using the HuggingFace[5] package as their private API services. To demonstrate this capability, we used the publicly available AERA model from (Li et al., 2023a), serving as a specialized private LLM configuration.

3.4 Database

This layer utilizes a PostgreSQL[6] relational database to manage a variety of data, including user profiles, assessment records (both for active tracking and background processing), and chat histories. It is interconnected with the backend layer and accessed via SQLAlchemy, an object-relational

mapping tool, to ensure optimal security and operational speed. Our architecture enhances system robustness and modularity by hosting the database as an isolated service, preventing platform failure.

4 Evaluation

In this section, we present an evaluation that utilizes AERA Chat's evaluation capability for automated assessment, as well as its interactive annotation interface for rationale verification.

Datasets Similar to prior work (Li et al., 2023a, 2024), we validate the evaluation capability for LLMs assessment performance on four sub-datasets covering subjects like science and biology from the ASAP-AES dataset (Hamner et al., 2012). Table 2 presents the test set statistic. We used the same prompts as those provided in the appendix of (Li et al., 2023a).

Models As discussed in Section 2.1, we evaluated three model settings using our platform: external API-based models, specifically `gpt-3.5-turbo-0125` and `gpt-4o-2024-05-13` from the OpenAI API, and the AERA model, for which we used our publicly released checkpoint[7].

Both OpenAI models are queried in a zero-shot query manner, with a temperature of `0.7`. On the contrary, the AERA model is specifically trained for automated assessment rationale generation on the ASAP-AES dataset. Instead of training and testing on single questions as presented in (Li et al., 2023a), the model we used was trained with student answers and ChatGPT-generated rationales across all questions.

4.1 Evaluation of Assessment Performance

As shown in Table 1, our platform provides a unified evaluation for those three aforementioned model settings. The AERA model achieved the highest test results across the four datasets. This is likely because the model has been specifically trained to perform assessment and rationale generation on these four datasets and is familiar with the prompt format, making this an in-domain testing scenario. In comparison, models like GPT-3.5-turbo and GPT-4o, which were trained for more generalized tasks, show less accurate assessment performance in zero-shot inference.

---

[3]https://remix.run/
[4]https://flask.palletsprojects.com/
[5]https://huggingface.co/
[6]https://www.postgresql.org/

[7]https://huggingface.co/jiazhengli/
long-t5-tglobal-large-AERA

Dataset (Subject)	#1 (Science)			#2 (Science)			#5 (Biology)			#6 (Biology)		
Model	Acc	F1	QWK	Acc	F1	QWK	Acc	F1	QWK	Acc	F1	QWK
gpt-3.5-turbo	0.300	0.227	0.040	0.207	0.144	0.012	0.585	0.345	0.276	0.579	0.252	0.084
gpt-4o	0.312	0.292	0.083	0.293	0.214	0.013	0.741	0.395	0.362	0.621	0.250	0.080
AERA	0.654	0.658	<u>0.765</u>	0.547	0.550	<u>0.734</u>	0.861	0.567	<u>0.818</u>	0.888	0.579	<u>0.802</u>

Table 1: Evaluation of assessment performance on three models over four datasets.

Datasets	ASAP #1	ASAP #2	ASAP #5	ASAP #6
Test	554	426	598	599
Score Range	0-3	0-3	0-3	0-3

Table 2: Test set statistics.

Models	ChatGPT	GPT-4o	AERA
Correctness	<u>45%</u>	40%	40%
Preference	20%	25%	<u>35%</u>

Table 3: Human evaluation of rationale quality.

Although various evaluations have been conducted to validate LLMs' capability to perform exams (OpenAI et al., 2024), the ability to assess student answers in exams has been rarely evaluated. As demonstrated in Table 1, we found that GPT-4o generally show slightly better performance across all evaluation metrics compared to GPT-3.5-turbo. Since the OpenAI models' predicted distributions have a quite significant gap from the ground-truth labels, the QWK score penalizes more heavily when the predicted score gap is larger. Therefore, while GPT-4o demonstrates a certain level of accuracy in the assessment of Questions 5 and 6, the wrongly predicted labels have large gaps, which causes lower macro F1 and QWK scores.

When examining the subject dimension, we can see that LLMs generally perform worse on science questions compared to biology questions. This could be due to the complex question setup for Questions 1 and 2, which involves providing a lengthy context that requires students to reflect on how to improve experimental designs.

4.2 Human Evaluation of Rationale Quality

To evaluate the quality of the generated rationales, we conducted a human evaluation using the AERA Chat, following an evaluation methodology similar to that proposed by (Li et al., 2023a, 2024). We randomly selected five student responses from each dataset to undergo human evaluation. As shown in Table 3, annotators used the AERA Chat's built-in annotation function to evaluate the rationales in two distinct modes: (1) Correctness of the rationales and (2) User preferences among rationales produced by different LLMs.

Evaluation of Rationale Correctness In this evaluation, annotators assessed each rationale's correctness based on its alignment with key elements of the answers and its adherence to the marking rubrics. Annotators are requested to dislike the rationale if any mistakes appear in the assessment. Notably, ChatGPT achieved the highest correctness score, particularly when evaluating responses that received zero points. In contrast, GPT-4o and AERA were more capable at correctly assessing responses that received higher point values.

Evaluation of User Preferences In the second evaluation task, annotators are asked to choose their preferred rationales from those generated by various LLMs, one at a time. If none of the rationales were satisfactory, annotators were instructed not to select any. AERA, which had been fine-tuned on four distinct datasets, received the highest preference score, followed by GPT-4o. Despite producing more detailed rationales, GPT-4o and ChatGPT occasionally deviated from the assessment context, leading to incorrect assessment decisions, which are less preferred by annotators.

5 Conclusion

In conclusion, we have developed an interactive platform, AERA Chat, for explainable student answer assessment via rationale generation. Our platform supports public and private LLM assessment, uses highlighting techniques for better visualization, and includes annotation functions to help educators and researchers generate more accurate data for developing trustworthy automated assessment LLMs. We evaluated AERA Chat's assessment and rationale generation capabilities using built-in automated evaluation metrics and manual rationale evaluation via the preference selection tool.

Acknowledgements

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*.

Peng Cui and Mrinmaya Sachan. 2023. Adaptive and personalized exercise generation for online language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*.

Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, and Feras A. Batarseh. 2023. Rationalization for explainable nlp: a survey. *Frontiers in Artificial Intelligence*.

Ben Hamner, Jaison Morgan, Mark Shermis Lynnvandev, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring.

Md Rayhan Kabir and Fuhua Oscar Lin. 2023. An llm-powered adaptive practicing system. In *LLM@AIED*.

Leah S. Larkey. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98. Association for Computing Machinery.

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023a. Distilling Chat-GPT for explainable automated student answer assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Jiazheng Li, Hainiu Xu, Zhaoyue Sun, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2024. Calibrating llms with preference optimization on thought trees for generating rationale in science question scoring. *Preprint*, arXiv:2406.19949.

Jiazheng Li, Runcong Zhao, Yongxin Yang, Yulan He, and Lin Gui. 2023b. Overprompt: Enhancing chat-GPT through efficient in-context learning. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics.

Jordan K. Matelsky, Felipe Parodi, Tony Liu, Richard D. Lange, and Konrad P. Kording. 2023. A large language model-assisted education tool to provide feedback on open-ended responses. *Preprint*, arXiv:2308.02439.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Soya Park and Chinmay Kulkarni. 2024. Thinking assistants: Llm-based conversational assistants that help users think by asking rather than answering. *Preprint*, arXiv:2312.06024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Reighns. 2020. Understanding the quadratic weighted kappa.

Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle &riley: Insights from designing and evaluating a large language model-based conversational tutoring system. In *Artificial Intelligence in Education*. Springer Nature Switzerland.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Samuel Tobler. 2024. Smart grading: A generative ai-based tool for knowledge-grounded answer evaluation in educational assessments. *MethodsX*.

Maximilian Tornqvist, Mosleh Mahamud, Erick Mendez Guzman, and Alexandra Farazouli. 2023. ExASAG: Explainable framework for automatic short answer grading. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics.

Roopchand Reddy Vanga, C. Sindhu, M. S. Bharath, T. Charandeep Reddy, and Meghana Kanneganti. 2023. Autograder: A feature-based quantitative essay grading system using bert. In *ICT Infrastructure and Computing*.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *ArXiv*, abs/2403.18105.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by LLMs: A showcase of ChatGPT in education applications. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.