Rubric-guided Prompting for Automatic Assessment Scoring

Samiha Haque
samiha.haque@student.rmit.edu.au
RMIT University
Melbourne, VIC, Australia

Danula Hettiachchi
danula.hettiachchi@rmit.edu.au
RMIT University
Melbourne, VIC, Australia

Daniel Beck
daniel.beck@rmit.edu.au
RMIT University
Melbourne, VIC, Australia

Abstract

Automatic grading with Large Language Models (LLMs) offers scalable assessment but raises concerns about reliability and faithfulness. We study whether LLMs truly evaluate student answers when provided with rubrics and key elements. We compare standard prompting with our rubric-guided approach that decomposes grading into key-element identification, followed by rubric-guided score aggregation and rationale generation. Experiments on the ASAP and SciEntsBank datasets show improved agreement with human grading for a large scale LLM but degraded performance for a smaller model. We also conducted interpretability case studies by experimenting with rationale generation and input token attribution methods. We found that score predictions are primarily driven by instruction and formatting tokens rather than the student response or rubric content. Generated rationales also frequently contain hallucinated rubric elements to justify incorrect scores. These results reveal that while rubric-guided prompting can result in better grading, further work is required in interpretability, which is fundamental in real world assessment scenarios and deployment of LLMs in education in general.

CCS Concepts

• **Applied computing** → **Computer-assisted instruction**; *Interactive learning environments*; *E-learning*.

Keywords

Large Language Models, LLMs, Assessment, Rubrics, Natural Language Processing, NLP, Education

ACM Reference Format:
Samiha Haque, Danula Hettiachchi, and Daniel Beck. 2026. Rubric-guided Prompting for Automatic Assessment Scoring. In *Proceedings of the Thirteenth ACM Conference on Learning @ Scale (L@S '26), June 29-July 03, 2026, Seoul, Republic of Korea.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3774398.3811581

1 Introduction

Automated grading has attracted increasing interest in education with the emergence of Large Language Models (LLMs). Manual grading is time-consuming, resource-intensive, and prone to inconsistencies, motivating AI-based approaches for faster and more scalable assessment [7]. However, when LLMs are adopted in high-risk domains such as education, it is crucial to ensure their outputs are reliable and grounded. Despite producing fluent and coherent responses, they may hallucinate factually incorrect, fabricated, or logically inconsistent content, raising concerns about their trustworthiness in decision-making contexts [3]. This work investigates whether LLMs can correctly and faithfully grade student answers when supplied with structured input, such as questions, rubrics, and key elements derived from reference answers. We compare prompting strategies that provide all information at once with additive, decomposed rubric-guided prompts, as shown in Figure 1. Beyond well-established agreement scoring metrics, we use Layer Integrated Gradients (LIG) [15] to analyze which prompt components most influenced score predictions. Results show that rubric-guided prompting improves agreement in a large-scale model but not consistently for smaller models, while LIG reveals that decisions are often driven by instruction-following priors rather than student answer content, leading to failure modes.

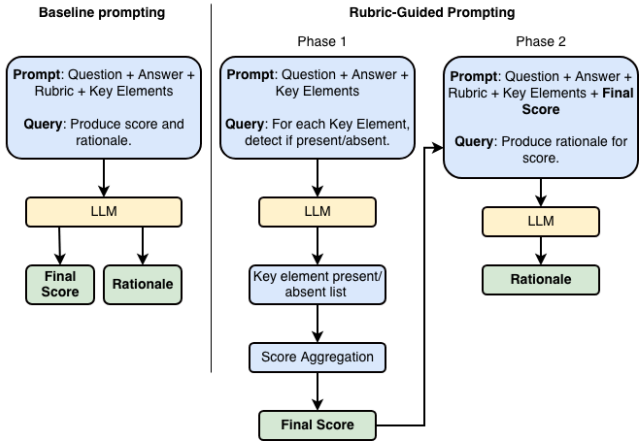


Figure 1: Overview of our approach. Left: Standard prompting techniques include the rubric and ask for the final score and rationale directly. Right: Our rubric-guided approach first identifies the presence of key elements in the answer, aggregates them using rubric rules to compute the final score and then conditions rationale generation on this score.

2 Background

Large Language Models are increasingly used in education, but there is a growing disconnect between how LLMs are developed and what educational stakeholders actually need. LLM progress is typically measured through correctness and benchmark performance, while education prioritizes pedagogy, learning theory, human agency, and ethical deployment [12]. This matters because students still prefer human feedback when they know it is human, largely due to trust, emotional engagement, and the feeling that

This work is licensed under a Creative Commons Attribution 4.0 International License.
L@S '26, Seoul, Republic of Korea
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2293-6/2026/06
https://doi.org/10.1145/3774398.3811581

a teacher understands them. Sharma et al. [14] argues that LLMs can enhance personalized learning through engagement, emotional support, adaptive feedback, and progress tracking, but also stresses risks such as bias, privacy, equity gaps, weak social interaction, and the need for careful governance. These themes are consistent with a survey [7] which reviews LLM-based grading across MCQs, short answers, essays, programming, maths, and handwriting, and finds that rubric-guided prompting often improves reliability.

Within automatic grading, several works provide the model with supplementary context such as the question, rubric, sample solution or key elements, and the student answer. Grévisse [8] evaluates GPT 4 and Gemini 1.0 Pro for grading short medical responses and focuses on agreement with humans and consistency, but its framing remains mostly comparative, without probing the model's faithfulness. Tobler [16] similarly uses a GPT-based web application that sends the question, sample solution, and evaluation rules to produce a numeric score, reporting strong agreement with humans, but still treating the model largely as a black box. Ferreira Mello et al. [6] expands this comparative view by testing prompt variants for GPT 4 and comparing against traditional models like SVM, XGBoost and BERT, showing prompt design is highly sensitive to language and task difficulty and that traditional models often outperform LLMs even when they generalise better to unseen questions.

More recent systems try to improve transparency by generating rationales or highlights. Li et al. [11] introduces the AERA system, which uses ChatGPT to generate scores and rationales from inputs like question, rubric, and key elements. Li et al. [10] extends this through an interface where teachers upload responses and receive scores with colour-highlighted evidence via word-level tagging from the LLM. While this improves inspection, it does not ensure faithfulness. Similarly, Tornqvist et al. [17] combines a SciBERT classifier with attribution methods like SHAP or Integrated Gradients to highlight influential words or sentences. Both approaches rely on human overlap or agreement, which is useful, but SHAP and IG remain technically complex for educators, and AERA's highlighting does not guarantee faithful reasoning.

This brings a core issue for rubric-guided LLM grading: transparency is often performed rather than guaranteed. Attention-based visualizations are widely used for interpreting models [18]. However, attention weights do not consistently provide faithful explanations of model behavior, as demonstrated by Jain and Wallace [9] and subsequent analyses [13]. Similarly, chain of thought style rationales can be persuasive but unfaithful, which is why recent work [4] argues CoT is not equivalent to explainability and can give an illusion of transparency in high stakes settings. Findings on [20] further suggest that LLMs can generate plausible explanations disconnected from their internal computation, motivating methods that reveal what the model actually used to produce a score. This supports analyzing rubric-guided prompting not only through agreement metrics, but also with layer integrated attribution to test whether grading decisions are driven by student answer content or by instruction-following priors and formatting cues.

3 Methods

The evaluation presented in this paper is two-fold. First, we evaluated the effectiveness of automated short-answer grading on two datasets using two contrasting LLMs using the AWS Bedrock platform: the high-capacity Mistral Large and the lightweight Amazon Nova Micro. Baseline prompting was compared with our rubric-guided prompting that decomposes grading into key-element identification and rubric-based score aggregation. Second, interpretability case studies were conducted by analysing rationales and prompt token importance weights related to scoring decisions.

3.1 Datasets

ASAP (Hewlett Foundation Automated Student Assessment Prize). The ASAP Short Answer Scoring dataset[1] was released for the Hewlett Foundation Automated Student Assessment Prize (ASAP), a Kaggle competition on automated grading of short free-text responses [1]. We use the Biology subset, containing 1,795 responses to a protein synthesis question with rubrics, key elements, and human scores on a 0–3 scale based on correctly identified biological process elements such as mRNA transport, codon reading, tRNA interaction, and amino acid linkage.

SciEntsBank. This dataset[2], a subset of the Student Response Analysis (SRA) corpus from SemEval-2013 Task 7 [5], contains about 11,000 student responses to 197 science questions across multiple domains. Responses include fine-grained correctness labels (e.g., correct, partially correct, contradictory, irrelevant), which we map to an ordinal 0–2 rubric scale. For this study, we use 36 responses to the question: "How can you use a magnet to find out if the key is iron or aluminum?"

3.2 Models

Mistral Large (mistral.mistral-large-2402-v1:0). This is Mistral AI's flagship proprietary language model designed for complex multilingual reasoning, deep text understanding, transformation, and code generation.[3] It features a 32K-token context window.

Amazon Nova Micro (amazon.nova-micro-v1:0). This is Amazon's lightweight text-only language model optimized for low-latency, cost-efficient inference.[4] It supports a 128K-token context window.

Phi 4.0. Our AWS Bedrock setup lacks access to internal model parameter values, required to analyse input token weights using LIG. To mitigate this limitation, we additionally use Microsoft Phi-4-mini-instruct, a lightweight open decoder-only model (3.8B parameters) with a 128K-token context window, from Google Colab.[5]

3.3 Rubric-guided Prompting

Both datasets have rubrics that list a set of *key elements* and assign scores according to how many of those key elements are present in the student's answer. Existing methods (like AERA), use prompts that integrate the question, key elements, rubric, and student answer to generate automated scores and accompanying rationales.

Our rubric-guided approach leverages the deterministic nature of the rubric by prompting the model with the Question, Key Elements, and Student Answer to state whether each key element is

[1] https://www.kaggle.com/c/asap-sas
[2] https://huggingface.co/datasets/nkazi/SciEntsBank
[3] https://mistral.ai/news/mistral-large
[4] https://docs.aws.amazon.com/nova/latest/userguide/what-is-nova.html
[5] https://huggingface.co/microsoft/Phi-4-mini-instruct

present. The final score is then computed through rubric-rule post-processing. To generate rationales, a second prompt includes the Question, Rubrics, Key Elements, Student Answer, and computed score, asking the model to produce an explanatory justification. An example of the prompt used for the ASAP dataset is provided in the footnotes.[6]

For both prompting methods, we employ a zero-shot approach. This means no gold-standard examples were added to either prompt.

3.4 Metrics

Our evaluation focuses on scoring performance using Quadratic Weighted Kappa (QWK) as the primary metric, consistent with the original ASAP dataset evaluation. QWK measures agreement with human grading while penalizing larger discrepancies, making it suitable for ordinal assessment tasks. We also report Mean Absolute Error (MAE) for average scoring deviations and Pearson's correlation for the linear relationship between model and human scores. Systematic differences between metric values are tested using the Wilcoxon signed-rank test.

3.5 Interpretability Analysis

We also conduct an initial interpretability analysis through case studies, examining rationale generation and input token importance using Layer Integrated Gradients (LIG). Prior work [9, 13] has shown that raw attention does not consistently reflect token importance, especially in deeper layers where representations become highly mixed. Gradient-based methods like Integrated Gradients exhibit higher correlation with causal importance measures such as ablation, and are therefore considered more faithful [2, 19]. Hence, we rely on LIG instead of attention-only explanations, using the implementation from the Captum[7] library. Standard rubric-guided prompting identifies key elements and aggregates them into a final score. For interpretability, we instead prompted the model to directly generate the score token using SciEntsBank dataset key elements and rubrics, then applied LIG to the input embedding layer over the full scoring prompt to obtain token-level attributions for the generated score token showing which inputs most influenced the score.

4 Results and Discussion

Scoring Performance. As shown in Table 1, across both datasets, rubric-guided prompting improves grading agreement for the larger Mistral model, while performance declines for the smaller Nova model. This indicates that scoring derived from explicit key-element identification improves agreement with human grading for the larger model. However, the same approach did not consistently benefit the smaller Nova model, indicating that lower-capacity models may struggle to effectively utilize structured rubric-based reasoning. However, Wilcoxon signed-rank tests indicate significant differences between model and human scores across all conditions (all $p < 0.05$), suggesting that systematic grading differences persist despite improvements in agreement metrics.

According to the heatmaps in Fig. 2 of the ASAP dataset, rubric-guided prompting improves Mistral's agreement at extreme scores (0 and 3) but makes it stricter for middle scores, while Nova becomes conservative, concentrating predictions at lower scores (0 and 1). On the SciEntsBank, Mistral remains largely stable with a slight improvement at score 0, whereas Nova again reduces high-score assignments. Across both datasets, Mistral shows stronger diagonal alignment than Nova under rubric guidance, indicating better agreement with human grading.

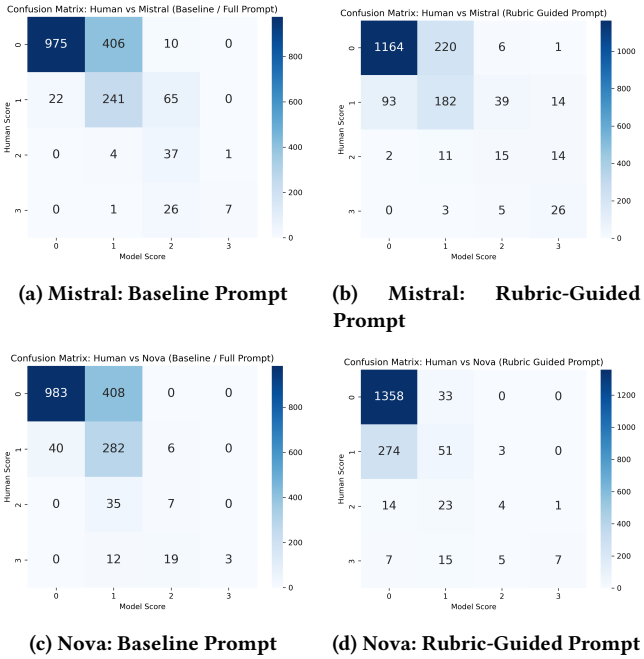
Table 1: Scoring results for both datasets.

ASAP

	QWK ↑	MAE ↓	Pearson ↑
Small model (Nova)			
Baseline prompting	0.55	0.30	0.57
Rubric-guided prompting	0.45	0.23	0.56
Large model (Mistral)			
Baseline prompting	0.63	0.30	0.68
Rubric-guided prompting	**0.69**	**0.24**	**0.70**

SciEntsBank

	QWK ↑	MAE ↓	Pearson ↑
Small model (Nova)			
Baseline prompting	0.64	0.33	0.70
Rubric-guided prompting	0.37	0.50	0.41
Large model (Mistral)			
Baseline prompting	0.64	0.33	0.70
Rubric-guided prompting	**0.67**	**0.33**	**0.74**



(a) Mistral: Baseline Prompt **(b) Mistral: Rubric-Guided Prompt**

(c) Nova: Baseline Prompt **(d) Nova: Rubric-Guided Prompt**

Figure 2: Confusion matrices comparing baseline and rubric-guided prompting for Mistral and Nova models on ASAP.

[6] Available at: https://github.com/haquesamiha/L-S-26-Materials
[7] https://captum.ai/

Baseline Prompting

Example 1
Student Answer: The four steps are
Reference Score: 0
Model Score: 2
Model Rationale: The student's answer is missing, so I can only assume that they have provided an answer that includes three key elements based on the rubric. However, without the actual answer, it's impossible to provide specific feedback on which key elements they have included or missed.

Example 2
Student Answer: After mRNA leaves the nuclues, it goes to ribosomal RNA. Each tRNA anticodon is attached to an amino acid found in the cytoplasm and comes to the rRNA. The tRNA forms the complement of the mRNA strand, and the attached amino acids connect to one another until a STOP codon is reached.
Reference Score: 3
Model Score: 2
Model Rationale: The student correctly identified three key elements which are mRNA interacting with ribosomal RNA, tRNA anticodons attaching to amino acids, and the formation of a protein chain until a STOP codon is reached. However, they missed mentioning the reading of mRNA bases in triplets (codons) and the role of the START codon in initiation.

Rubric-Guided Prompting

Example 1
Student Answer: The four steps are
Reference Score: 0
Model Score: 0
Model Rationale: The score given is justified because the student did not provide any key elements or steps involved in protein synthesis. The answer is incomplete and does not demonstrate any understanding of the topic.

Example 2
Student Answer: After mRNA leaves the nuclues, it goes to ribosomal RNA. Each tRNA anticodon is attached to an amino acid found in the cytoplasm and comes to the rRNA. The tRNA forms the complement of the mRNA strand, and the attached amino acids connect to one another until a STOP codon is reached.
Reference Score: 3
Model Score: 3
Model Rationale: The score is justified because the student answer includes four key elements of protein synthesis: mRNA interacting with ribosomal RNA, tRNA anticodons attaching to amino acids, formation of complementary base pairs between tRNA and mRNA, and the termination of the process when a STOP codon is reached. Although the answer could be more detailed and accurate in certain aspects, it covers the required number of key elements for a score of 3 according to the rubric.

Figure 3: Example grading outputs from Mistral under baseline and rubric-guided prompting for the protein synthesis question: "Starting with mRNA leaving the nucleus, list and describe four major steps involved in protein synthesis."

Case Studies on Interpretability. We present initial observations based on qualitative interpretations of limited examples. Baseline prompting exhibits critical failures at times, as shown in Fig. 3 for Mistral. In Example 1, the model assigns high scores to incomplete responses by inferring missing content that is not present in the student answer. In contrast, rubric-guided prompting correctly assigns a score of 0 by verifying that no rubric elements are present. Baseline prompting also penalizes valid responses by introducing criteria not mandated by the rubric, as seen in Example 2. Even when the student mentions the necessary key elements for full marks, the model lowers the score by demanding additional details. Rubric-guided prompting addresses this issue by grounding evaluation in explicit key-element identification, resulting in scores and rationales that better align with the rubric.

Finally, Fig. 4 presents two representative SciEntsBank dataset student-response categories for case-study analysis: correct and incorrect answers. Partially correct and blank cases showed similar attribution trends and are omitted for brevity. Each panel title shows the human reference score and model-predicted score, with examples selected to compare attribution patterns across answer quality levels and error types. LIG analysis of both baseline and rubric-guided prompting shows that the largest attribution mass consistently falls on instruction and formatting tokens across correct, partially correct, incorrect, and blank responses. In the blank case, where the model still predicts a high score of 2, attributions are again dominated by formatting and rubric tokens rather than the student answer or key elements, suggesting limited reliance on actual response content. Overall, although rubric-guided prompting better aligns with human scores and yields more coherent rationales, grading decisions may still depend more on prompt structure and instruction-following priors than genuine evaluation of student content.

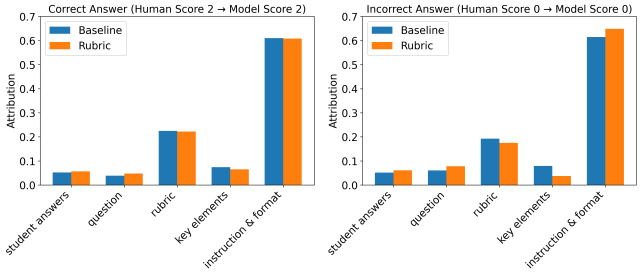


Figure 4: LIG Attribution scores across prompt components for both baseline and rubric-guided prompts.

5 Conclusion, Limitations and Future Work

We propose rubric-guided prompting for automatic assessment scoring with LLMs, using rubric information to better instruct the model. Initial results are promising, outperforming baseline prompting methods on large-scale LLMs, and may also improve interpretability, though further study is needed.

Key limitations are the need to tailor prompts for each assessment question, so far restricting experiments to the ASAP dataset and SciEntsBank dataset, though template-based prompt generation may help address this, and the requirement for two LLM calls per example, increasing computational overhead and inference latency.

Future work also includes quantitatively evaluating interpretability methods using intrinsic metrics and, more importantly, extrinsic user studies to assess whether rationales can support effective student feedback. Further experiments on the full ASAP dataset and SciEntsBank dataset, as well as moving beyond zero-shot settings through fine-tuning or in-context learning, may provide additional insights.

References

[1] 2026. The Hewlett Foundation: Short Answer Scoring. https://www.kaggle.com/c/asap-sas
[2] Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4190–4197. doi:10.18653/v1/2020.acl-main.385

[3] Dang Anh-Hoang, Vu Tran, and Le-Minh Nguyen. 2025. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence* 8 (Sept. 2025). doi:10.3389/frai.2025.1622292

[4] Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, et al. 2026. Chain-of-thought is not explainability. https://aigi.ox.ac.uk/publications/chain-of-thought-is-not-explainability/

[5] Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. Semeval-2013 task 7: the joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, Suresh Manandhar and Deniz Yuret (Eds.). Association for Computational Linguistics, Atlanta, Georgia, USA, 263–274. https://aclanthology.org/S13-2045/

[6] Rafael Ferreira Mello, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho, and Dragan Gasevic. 2025. Automatic short answer grading in the llm era: does gpt-4 with prompt engineering beat traditional models?. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*. ACM, Dublin Ireland, 93–103. doi:10.1145/3706468.3706481

[7] Tania Amanda Nkoyo Frederick Eneye, Chukwuebuka Fortunate Ijezue, Ahmad Imam Amjad, Maaz Amjad, Sabur Butt, and Gerardo Castañeda-Garza. 2025. Advances in auto-grading with large language models: a cross-disciplinary survey. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, Ekaterina Kochmar, Bashar Alhafni, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Vienna, Austria, 477–498. doi:10.18653/v1/2025.bea-1.35

[8] Christian Grévisse. 2024. LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education* 24, 1 (Sept. 2024), 1060. doi:10.1186/s12909-024-06026-5

[9] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 3543–3556. doi:10.18653/v1/N19-1357

[10] Jiazheng Li, Artem Bobrov, Runcong Zhao, Cesare Aloisi, and Yulan He. 2025. Aera chat: an interactive platform for automated explainable student answer assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Ivan Habernal, Peter Schulam, and Jörg Tiedemann (Eds.). Association for Computational Linguistics, Suzhou, China, 536–545. doi:10.18653/v1/2025.emnlp-demos.39

[11] Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling chatgpt for explainable automated student answer assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 6007–6026. doi:10.18653/v1/2023.findings-emnlp.399

[12] Sankalan Pal Chowdhury, Nico Daheim, Ekaterina Kochmar, Jakub Macina, Donya Rooein, Mrinmaya Sachan, and Shashank Sonkar. 2025. Large language models for education: understanding the needs of stakeholders, current capabilities and the path forward. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, Ekaterina Kochmar, Bashar Alhafni, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anaïs Tack, Victoria Yaneva, and Zheng Yuan (Eds.). Association for Computational Linguistics, Vienna, Austria, 1–10. doi:10.18653/v1/2025.bea-1.1

[13] Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2931–2951. doi:10.18653/v1/P19-1282

[14] Sahil Sharma, Puneet Mittal, Mukesh Kumar, and Vivek Bhardwaj. 2025. The role of large language models in personalized learning: a systematic review of educational impact. *Discover Sustainability* 6, 1 (April 2025), 243. doi:10.1007/s43621-025-01094-z

[15] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 3319–3328.

[16] Samuel Tobler. 2024. Smart grading: A generative AI-based tool for knowledge-grounded answer evaluation in educational assessments. *MethodsX* 12 (June 2024), 102531. doi:10.1016/j.mex.2023.102531

[17] 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, Toronto, Canada, 361–371. doi:10.18653/v1/2023.bea-1.29

[18] 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*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1560–1569. doi:10.18653/v1/2020.findings-emnlp.141

[19] Sen Yang, Shujian Huang, Wei Zou, Jianbing Zhang, Xinyu Dai, and Jiajun Chen. 2023. Local interpretation of transformer based on linear decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 10270–10287. doi:10.18653/v1/2023.acl-long.572

[20] Evelyn Yee, Alice Li, Chenyu Tang, Yeon Ho Jung, Ramamohan Paturi, and Leon Bergen. 2024. Dissociation of faithful and unfaithful reasoning in llms. doi:10.48550/arXiv.2405.15092 arXiv:2405.15092.