Rubric Based Automated Short Answer Scoring using Large Language Models (LLMs)

Chamuditha Senanayake
Department of Industrial Management
University of Kelaniya
Kelaniya, Sri Lanka.
chamudithacbs@gmail.com

Dinesh Asanka
Department of Industrial Management
University of Kelaniya
Kelaniya, Sri Lanka.
dasanka@kln.ac.lk

2024 International Research Conference on Smart Computing and Systems Engineering (SCSE) | 979-8-3503-7568-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/SCSE61872.2024.10550624

***Abstract*—The manual grading of short answers presents challenges in education due to time constraints, especially with larger student populations, and suffers from subjectivity and bias, leading to inconsistencies. Larger student populations increase the time needed for individual assessment, leading to potential delays and subjectivity introduces biases, resulting in inconsistent evaluations. Moreover, as student numbers rise, the imbalance in teacher-to-student ratios affects grading quality, impacting fairness and effectiveness in educational assessments. Automated grading systems have emerged as a solution to address these issues. These grading systems mainly prioritize appearance, emphasizing grammar and format. However, they struggle to accurately assess content quality, often missing contextual relevance. This problem can potentially be resolved by employing highly trained domain-specific models. However, a drawback arises as these models are limited to evaluating answers only within predefined domains. While these specialized models excel in assessing responses within their designated fields, their utility is restricted when evaluating answers outside of these predefined domains. This limitation poses a challenge in achieving broader applicability for assessing answers outside the specific areas the models were trained for. This study proposes a rubric-based method paired with Large Language Models (LLMs) to introduce objectivity, ensuring fairness and reliability in evaluations while achieving generalizability. Rubric provides a clear and customizable marking schema for assessing short answers across various domains. By using predetermined marking criteria and conditions, the grading process becomes more objective and transparent. The proposed method efficiently evaluates short answers in various domains using Large Language Models (LLMs), based on these established criteria, reducing subjective biases. This research aims to revolutionize education by creating a robust automated short answer scoring system that comprehensively evaluates contents across domains and addresses teacher-to-student ratio issues.**

Keywords—Automated short answer grading, Rubric-based, Large Language Models

I. Introduction

Automated Short Answer Scoring (ASAS) stands at the forefront of educational innovation, revolutionizing the evaluation of short answer responses through the process of computer programs. Unexpectedly, grading short answers emerges as a more intricate challenge compared to essays, due to various complexities involved [1]. Yet, ASASs emerges as a beacon of solution, addressing the problems of manual grading and the challenges of inconsistent assessments, especially in large-scale testing situations. While the educational field is replete with automated tools tailored for multiple-choice questions, quality and the content-based assessment of short and essay answers remains a formidable frontier [2]. However, using techniques such as ontology, semantic similarity and sentence embedding, ASAS seeks to delve deeper into the student's knowledge and understanding of the subject, not just to assess responses [3], [4], [5]. ASAS has exhibited a trifecta of efficiency, being fast, scalable, and accurate, as emphasized by the perspectives of experts [4]. Its footprint extends to the United Kingdom General Certificate of Secondary Education (UK GCSE) system, where it not only assesses but enriches the learning journey by providing invaluable feedback to students [18]. However, despite the accomplishments, numerous challenges and obstacles persist. Most ASASs primarily prioritize appearance by emphasizing grammar and format, yet they encounter difficulties in accurately evaluating content quality, frequently overlooking contextual relevance. Highly trained domain-specific models solve this issue but are confined to predefined domains, restricting broader applicability in evaluating diverse responses.

ASAS exemplifies the seamless integration of technology and pedagogy within the expansive landscape of education. As ASAS continues its journey, navigating challenges and shaping the future of assessments, the educational realm remains poised for transformation. A few survey papers [1], [3] have previously reviewed the advancements and developments in the field of ASAS. In [1], the historical developments of analytic components up to 2014 were addressed, providing a unified and thorough overview of the entire field. The authors divided the approaches that were already in use into five categories that correspond to five temporal themes, also known as eras. The first era of methods was based on rule-based systems, while the second era was characterized by the application of statistical models. The third era saw the emergence of machine learning techniques, while the fourth era was marked by the use of natural language processing techniques. The fifth era, which is the current era, is characterized by the use of deep learning techniques such as word embeddings and transformers [1].

Existing methods have been tested on different data sets, such as SciEntsBank, Beetle, Texas, ASAP-SAS, and various others [5]. These data sets differ in many aspects, such as the number of questions, the number of answers, and the level of difficulty. The use of benchmark data sets are essential for grading short answers as realistically as possible [5]. ASAS has the potential to revolutionize the way assess student learning. Recent developments in deep learning techniques have made ASAS highly accurate, scalable, and fast. However, there are still challenges that need to be addressed, and further research is needed to ensure that ASAS is accurate, fair and unbiased. Overall, ASAS represents an exciting development in the field of education, and it will be interesting to see how it continues to evolve in the coming years.

Rubric-based models bring forth numerous advantages in the assessment of short answers. Firstly, they deliver an unbiased evaluation by relying on predetermined criteria,

minimizing the potential for human bias in grading. This ensures a fair and impartial assessment process [4]. Secondly, these systems significantly enhance efficiency by swiftly grading short answers, surpassing the pace of human graders [4], [11]. This efficiency not only saves valuable grading time but also empowers educators to redirect their focus towards other essential tasks, fostering a more productive educational environment [10]. Moreover, rubrics contribute to heightened accuracy through the integration of machine learning techniques such as ontology, semantic similarity matching, statistical methods, word embeddings, and transformers. These advanced tools enable a more precise and consistent evaluation of short answers, ensuring a thorough and reliable grading process [13]. Additionally, rubrics foster consistency in grading across diverse graders and grading sessions, guaranteeing that all students are evaluated fairly and uniformly. This consistency is instrumental in providing a standardized assessment experience, irrespective of variations in grading personnel. Furthermore, ASAS systems employing rubrics offer the added advantage of immediate feedback. Students receive prompt insights into their performance, creating an efficient avenue for them to enhance their writing skills and learn from their mistakes. This timely feedback not only supports a continuous learning process but also facilitates sustained student growth [7].

Large Language Models (LLMs) are advanced computer programs skilled at handling a wide range of language-related tasks, harnessing transformers and acquiring knowledge from expansive textual databases like Wikipedia and GitHub. These models autonomously grasp word meanings and relationships during their learning phase, devoid of direct instruction [19]. Trained LLMs prove invaluable across diverse domains, elevating search engine performance, aiding software developers in coding, conducting semantic analysis, summarizing lengthy articles, and engaging in extensive user conversations [8]. Their training involves extensive datasets comprising vast internet text, with powerful models like GPT-3 boasting up to 96 layers. LLM applications span not only education but also healthcare, finance, entertainment, sports, and beyond. The likes of GPT-3, GPT-4, and BERT are pivotal in automating the assessment and scoring of short answers, offering versatility for educators and assessment providers. Fine-tuning LLMs for domain-specific Automated Short Answer Scoring (ASAS) enhances their utility, enabling effective grading in specialized fields like science, mathematics, or literature. Beyond mere scoring, LLMs provide valuable insights by offering detailed feedback to students, pinpointing areas for improvement and facilitating the enhancement of writing skills [9].

With the gradual increase in the number of teachers' student ratio, the manual evaluation process becomes complicated, as it is time-consuming, lacks reliability, and presents various drawbacks [2]. Therefore, automated assessment systems have gained significant attention in the field of education and assessment as they provide a promising solution for efficient evaluation of students' responses [9]. These systems utilize Natural Language Processing (NLP) techniques to automatically analyze and assign scores to short answer responses [12]. However, content-based evaluation has received limited attention from researchers, with the majority focused on style-based assessment [13]. Therefore, existing approaches often fall short in accurately assessing the quality and proficiency of answers, primarily relying on simplistic keyword matching, or lacking the ability to comprehensively analyze the content, coherence, and language usage [7], [14]. There is a pressing need to advance the development of robust and accurate automated short answer scoring systems that can effectively assess responses based on rubric-based criteria [11].To address this research gap, the main focus of this research entails the design and development of an advanced system utilizing Large Language Models (LLMs) that is specifically focused on rubric-based evaluation. This system aims to overcome the limitations of current approaches by considering multiple dimensions of short answer responses, including content relevance, use of evidence, and adherence to rubric-based criteria. Moreover, an important aspect of this research is to ensure the generalizability of the developed model. Achieving generalizability is crucial to ensure that the automated scoring system can reliably evaluate short answers from a wide range of students and assessments.

As previously mentioned, the manual evaluation process of short answer responses is often time-consuming and becomes increasingly challenging as the number of students per teacher rises. By automating the scoring process, the research aims to overcome this drawback and improve the overall efficiency of assessment. Secondly, manual scoring can be subjective and prone to human bias, leading to inconsistencies in the evaluation process, as different graders may interpret and evaluate responses differently. To address this, a rubric-based approach with LLM aims to introduce objectivity, ensuring fairness and reliability in evaluating student answers. However, over several decades, researchers have developed automated short answer scoring systems, but few studies have focused on content evaluation compared to style-based assessment. The challenge lies in achieving generalizability and considering parameters such as content relevance, idea development, coherence, and cohesion when assessing essays. However, LLMs and the rubric-based approach partially addresses this challenge by enabling generalizability and content-based evaluation. Overall, the research on automated short answer scoring using LLMs and a rubric-based approach is justified as it enhances efficiency, objectivity, generalizability and addresses the limitations of manual evaluation as well as existing NLP based approaches. By leveraging these technologies, the research aims to contribute to the advancement of assessment practices in education, ultimately benefiting both educators and students.

The objectives of this research encompass two main facets. Firstly, the aim is to develop an algorithm and fine-tuned prompts to evaluate the answers across various domains using large language models (LLMs) and machine learning techniques. This involves creating a robust system capable of effectively assessing short answer responses across a spectrum of subjects and domains. Secondly, the research seeks to evaluate the performance of the automated scoring system. This will be achieved by comparing the scores generated by the automated system with those provided by human experts, thus assessing accuracy and consistency. By pursuing these objectives, the research endeavors to address the challenges associated with manual evaluation processes, such as time consumption and subjectivity, while also enhancing efficiency, objectivity, and generalizability in assessment practices. The study only focuses on questions and answers written in the English language. To evaluate the generalizability of the model, rubrics, and answers from various subjects of university students were employed.

II. Related Work

There is a significant research interest in leveraging natural language processing to automate essay and short answer scoring and feedback for students. Multiple studies have proposed automated systems, but reliably matching human evaluation remains challenging [15]. Earlier works focused on rule-based approaches, with more recent focus on statistical and machine learning models. The field is now utilizing the latest advances like neural networks and transformers to power deep learning systems [15]. However, concerns persist around accurately assessing elements like content, coherence, language usage versus simplistic keyword matching [7]. Many experiments rely on a few standard datasets like ASAP-SAS that may not generalize contexts across domains [3], [5]. The majority of studies target specific constructs like evidence use [13] or overall accuracy [3]. Novel approaches incorporate rubric-based evaluation and data augmentation [15]. But human-level performance remains elusive, indicating more holistic solutions are needed [10]. In these studies, substantial pioneering research targeting automated writing assessment indicates promising potential but also persisting challenges in achieving robust, accurate, and generalizable performance on par with human evaluators [17]. Advancement requires progress on multiple fronts datasets, rubrics, machine learning approaches, and evaluation methodologies. Ongoing studies provide encouraging building blocks but significant innovation is still imperative.

Recent research has begun exploring the potential of leveraging powerful large language models such as GPT-3 and GPT-4 for automated evaluation of short answers or essays [16]. Topsakal and Akinci in 2023 introduced a framework called LangChain that aims to accelerate the development of custom AI applications powered by LLMs. It provides pre-built components and flexible pipelines that developers can adapt for their specific use case, promoting quicker application building. However, the authors did not compare LangChain against alternative options, discuss practical challenges in utilizing it computationally, or consider broader ethical implications of deploying LLM systems.

Yancey et al. in 2023 conducted an intriguing study assessing GPT-4 for scoring short essays written by non-native English speakers. They found that with a small number of scoring examples for calibration, GPT-4 could reach accuracy approaching modern automated writing evaluation methods. However, agreement between GPT-4 and human raters varied significantly across different first language backgrounds. As GPT-4 showed no detectable biases toward gender or native language, further research is imperative to understand these scoring discrepancies and explore strategies to enhance GPT-4's reliability and usefulness for feedback [16].

In 2023, Kortemeyer evaluated GPT-4 on grading short answers related to scientific topics, using established datasets like SciEntsBank and Beetle. Results indicated viable out-of-the-box use without additional training, but performance lagged behind other fine-tuned LLMs specialized for this task. Nonetheless, GPT-4 showed promising potential for immediate automated scoring applications, especially at elementary grade levels or introductory higher education courses. Limitations remain concerning the probabilistic unpredictability of GPT models and optimal integration [9]. In summary, pioneering studies have revealed intriguing early promise for leveraging large language models to automate the evaluation of short answer responses. However, substantial research gaps persist regarding achieving robust and accurate performance generalized across subjects, grade levels, question types, and student populations. Continued research tackling critical challenges like potential biases, generalizability, computational efficiency, and ideal training strategies will be vital to fully unlock the transformative educational potential of LLMs.

This section presented a literature review related to the research topic, supporting the rationale for the study area, and providing an overview of the research methodologies outlined in the reviewed literature. The review has highlighted a significant research gap in automated short answer scoring solutions which can achieve generalizability and higher accuracy.

III. Methodology

Proposed methodology refers to the systematic approach and techniques employed to conduct scientific investigations, design, and develop algorithms and analyze data for the purpose of generating reliable and valid results. The methodology consists of five primary stages.

A. Data Collection

In order to verify the generalizability of the research findings, questions and answer sheets were collected from 3 different domains including data warehousing and software design patterns. Therefore, questions and answer sheets from different subjects were obtained from undergraduate students and postgraduate students. It is important to emphasize that the scope of this research is specifically limited to analyzing answer sheets written in the English language. During the entire process of gathering data, we meticulously adhered to established ethical guidelines and data protection regulations. This ensured that the collected data remained confidential and the privacy of the individuals involved was protected throughout the research process.

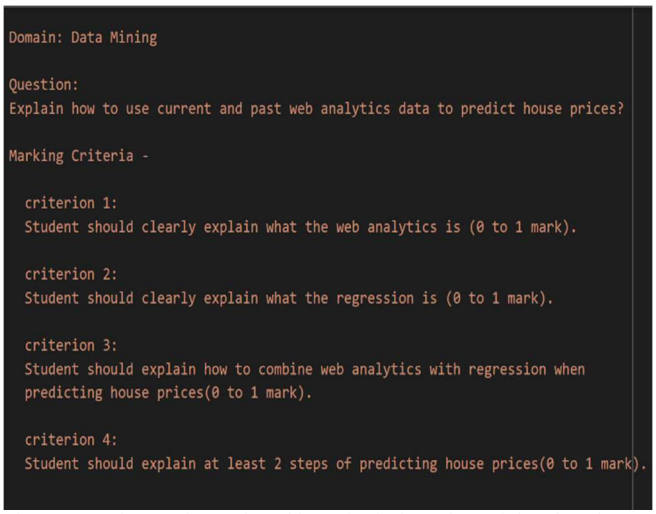

Fig. 1. Sample question and marking criteria from data mining domain

	Weight	Value1	Value2	Value3	Value4
Criteria1	w_1	$condition_{11}$	$condition_{12}$	$condition_{13}$	$condition_{14}$
Criteria2	w_2	$condition_{21}$	$condition_{22}$	$condition_{23}$	$condition_{24}$
Criteria3	w_3	$condition_{31}$	$condition_{32}$	$condition_{33}$	$condition_{34}$

Fig. 2. Rubric matrix template

B. Dynamic Rubic Development

In order to ensure accurate evaluation of short answers in alignment with the research objective, a dynamic analytic rubric was introduced. The rubric can be customized to address the unique requirements of scoring short answers while decreasing grading time and workload of educators [13]. A teacher or a grader can include well-defined criteria and establish different levels of performance for each aspect to provide a comprehensive assessment framework as a marking schema. By depending on established criteria, rubrics guarantee an equitable and unbiased evaluation procedure. Furthermore, these systems greatly boost efficiency by rapidly assessing short responses, surpassing the speed of human graders [7]. Through this analytic rubric, the research aims to measure the correctness and relevance of answers to questions.

The rubric template provided in figure 2 can be customized according to the grader's requirements, with fixed values assigned to value1, value2, value3, and value4. Criteria and conditions were selected based on the grader's specifications. Then Criteria (C), Conditions (cn), Values (V) and Weightages (W) were extracted from the marking rubric by the model.

After that the model assesses the answers based on the rubric's criteria and conditions, ensuring a comprehensive evaluation process. Subsequently, the LLM calculates scores providing an objective and efficient assessment of the responses. The successful achievement of generalizability within the context of Automated Short Answer Grading significantly contribute to the overall enhancement and effectiveness of the system, making it more versatile and valuable in diverse educational settings [13]. Therefore, ensuring generalizability is an important aspect of this research, and this can be accomplished through analytical rubric.

C. Development of the Algorithm

The research aims to modernize short answer scoring through the integration of advanced algorithms and LLMs. This entails a comprehensive examination of available algorithms and LLMs. To ensure effective implementation, appropriate frameworks and libraries were selected. Various data sources, including Google Forms and written documents, were gathered and structured. Questions and answers were extracted and organized for analysis. This data, along with the marking rubric, formed the input for the model. By combining cutting-edge technology with established evaluation methods, the research seeks to enhance automation, efficiency, and consistency in short answer scoring, ultimately advancing educational assessment practices.

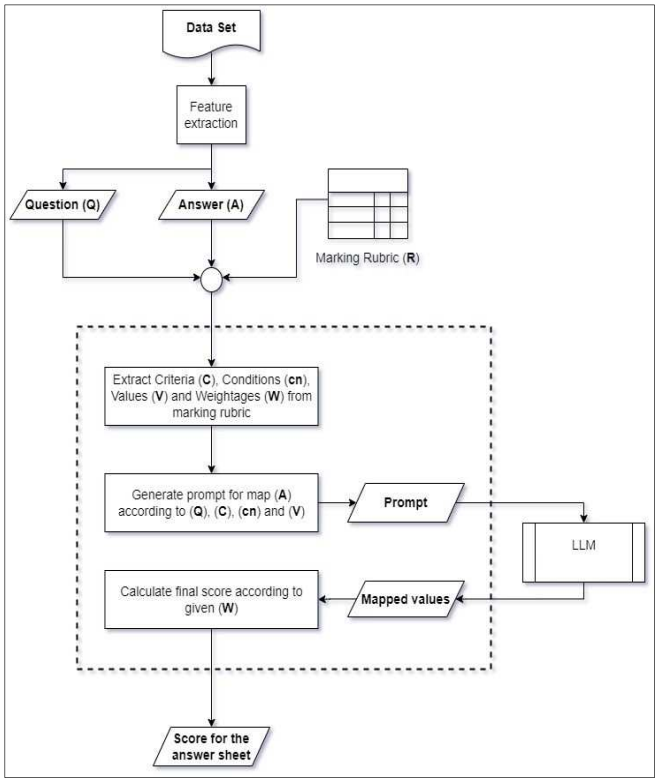

Fig. 3. Answer evaluation model design

In order to accomplish automated short answer scoring, an examination of algorithms and LLMs was conducted. Additionally, to ensure efficient implementation, careful consideration was given to selecting frameworks and libraries such as LangChain and transformers. These tools were chosen for their ability to provide essential functionalities and resources necessary for the scoring process. To establish a robust foundation for analysis, data was gathered from diverse sources including Google Forms, CSV files, images, and written documents. From this data pool, questions and corresponding answers were meticulously extracted and stored in a structured CSV file format. Subsequently, this data, in conjunction with the marking rubric, was fed into the model for processing. Within the model, a systematic extraction process was employed to derive Criteria (C), Conditions (cn), Values (V), and Weightages (W) from the marking rubric. This information served as the basis for generating prompts tailored to map student answers (A) in alignment with Questions (Q), Criteria (C), Conditions (cn), and Values (V). These prompts were then transmitted to the LLM model for further refinement and analysis.

LLM performs semantic analysis on the student's answer sheet, extracting meaning and context, and map the results to the rubric accordingly. By leveraging the rubric's guidelines and providing more accurate prompts, the model can be refined to enhance its performance and assess short responses with greater accuracy. Following that, the mapped values were returned back to the model, and a final mark was computed based on the specified weightages. The final grade for the answer sheet would then be determined and returned as output. Through this integration of LLMs and rubric-based evaluation, the research aims to enhance the automation and

efficiency of short answer scoring, ultimately facilitating more reliable, efficient, and consistent evaluations.

IV. Result and Discussion

Two trials were used with Llama2, Mistral and Zaphyer for evaluating the performance. The models were evaluated using MAE and RMSE. The results were significant and deducible. Tables I, II, III below summarize the test results of the LLM based models.

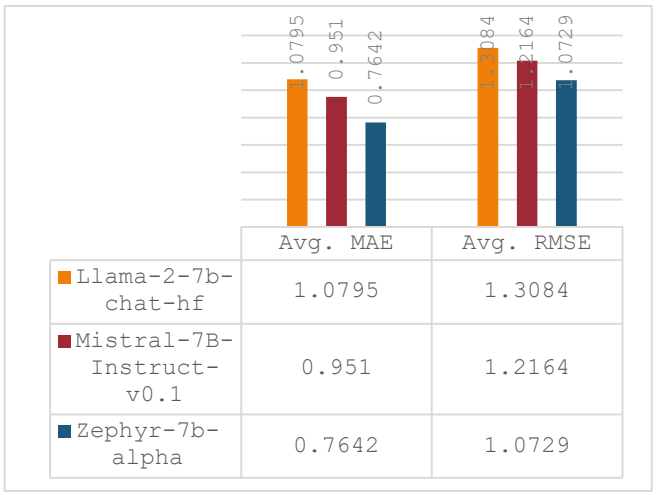

Fig. 4. Evaluation of question 1

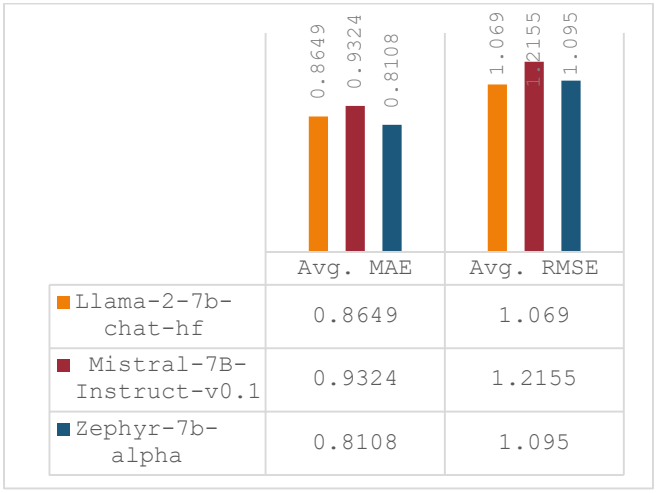

Fig. 5. Evaluation of question 2

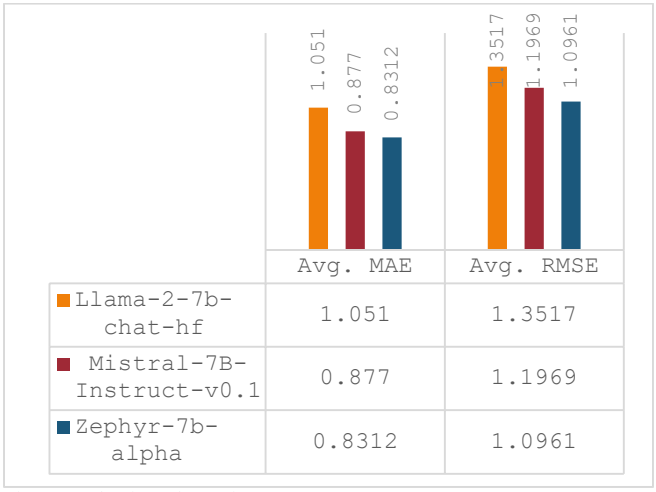

Fig. 6. Evaluation of question 3

TABLE I - AVERAGE TIME TAKEN TO EVALUATION

Model	QUESTION 1	QUESTION 2	QUESTION 3
Llama-2-7b-chat-hf	48min 22s	31min 12s	45min 41s
Mistral-7B-Instruct-v0.1	52min 19s	35min 12s	46min 19s
Zephyr-7b-alpha	41min 25s	28min 41s	41min 58s

This study compares several state-of-the-art neural models on the task of automated short answer scoring across three benchmark questions. Model performance was assessed using two standard error metrics - mean absolute error (MAE) and root mean squared error (RMSE) between predicted scores and ground truth markings - along with computational efficiency measured in evaluation time. The Zephyr-7b consistently achieves superior score prediction accuracy across all questions, with statistically significantly lower MAE and RMSE versus alternatives. This demonstrates its responses have the highest correlation and least variability from human expert marks, aligning with [1] on RNN efficacy in capturing linguistic patterns for assessment. Zephyr's lowest errors signify its effective language mastery better captures key scoring concepts [2]. Furthermore, while LSTM architectures can have high computational demands [8], Zephyr proves most time-efficient on two of three tests. As auto-scoring systems are often capacity-limited [4], such efficiency enables scalable real-world deployment. In conclusion, Zephyr-7b's dominant effectiveness and efficiency lowers barriers preventing AI adoption [4], delivering immediate benefits to learning assessment. Our results provide a solid evidentiary basis for utilizing Zephyr, establishing automated scoring viability, and driving future developments.

V. Conclusion and Future Work

This study provides compelling evidence that large language models (LLMs) enable viable automated rubric-driven short answer evaluation across multiple questions. The Zephyr-7b LLM architecture demonstrates consistent state-of-the-art performance versus alternatives, with statistically significantly lower mean absolute errors down to 0.764 and root mean squared errors up to 15% less than rivals. Lower errors in predicted versus actual scores indicate Zephyr more accurately captures key grading concepts. These complements confirm that rubric based LLMs can assess linguistic mastery, not just keyword usage. Our multi-prompt generalizability also addresses concern about assessment diversity. Notably, Zephyr maintains real-time efficiency thanks to optimized RNNs, facilitating scalable usage. Such accuracy and speed fulfil essential prerequisites for usable educational AI based model. This cements LLMs' viability

for augmenting overburdened human evaluators, enhancing consistency. Nonetheless, further LLM advances may continue improving assessment effectiveness. Extending models' linguistic mastery could strengthen generalizability across subjects and question complexities. Pre-training on more domain-specific corpora also promises gains. In conclusion, this study signals LLMs' profound potential to transform short answer evaluation through demonstrable accuracy, efficiency, and versatility. Our models establish strong feasibility for real-world rubric-based automated scoring, driving future progress.

REFERENCES

[1] S. Burrows, I. Gurevych, and B. Stein, "The eras and trends of automatic short answer grading," *Int J Artif Intell Educ*, 2014.

[2] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55(3), 2495-2527.

[3] L. B. Galhardi and J. D. Brancher, "Machine learning approach for automatic short answer grading: A systematic review," in *Lecture Notes in Computer Science*, Springer Verlag, 2018, pp. 380–391.

[4] Wang, T., Inoue, N., Ouchi, H., Mizumoto, T., & Inui, K. (2019, November). Inject rubrics into short answer grading system. In Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019) (pp. 175-182).

[5] S. Haller, A. Aldea, C. Seifert, and N. Strisciuglio, "Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers," 2022.

[6] M. A. Hearst and K. Kukich, "Beyond Automated Essay Scoring," 2000.

[7] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Multimed Tools Appl 82," pp. 3713–3744, 2022.

[8] G. Kolappan, "Computer Assisted Short Answer Grading with Rubrics using Active Learning," 2023.

[9] G. Kortemeyer, "Performance of the Pre-Trained Large Language Model GPT-4 on Automated Short Answer Grading," *arXiv preprint arXiv:2309.09338*, 2023.

[10] J. Lun, J. Zhu, Y. Tang, and M. Yang, "Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring," 2020.

[11] S. Marvaniya, S. Saha, T. I. Dhamecha, P. Foltz, R. Sindhgatta, and B. Sengupta, "Creating Scoring Rubric from Representative Student Answers for Improved Short Answer Grading," 2018.

[12] D. S. McNamara, S. A. Crossley, and R. Roscoe, "Behav Res Methods 45," pp. 499–515, 2012.

[13] Z. Rahimi, D. Litman, R. Correnti, E. Wang, and L. C. Matsumura, "Int J Artif Intell Educ 27," pp. 694–728, 2017.

[14] L. Ramachandran, J. Cheng, and P. Foltz, "Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching," 2015.

[15] D. Ramesh and S. K. Sanampudi, "Artif Intell Rev 55," pp. 2495–2527, 2021.

[16] Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023, July). Rating short l2 essays on the cefr scale with gpt-4. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023) (pp. 576-584).

[17] R. Shiva Shankar and D. Ravibabu, "Digital report grading using NLP feature selection," in *Advances in Intelligent Systems and Computing*, Springer Verlag, 2018, pp. 615–623.

[18] Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. Procedia computer science, 169, 726-743.

[19] Yoon, S. Y. (2023). Short Answer Grading Using One-shot Prompting and Text Similarity Scoring Model. arXiv preprint arXiv:2305.18638.