STM JOURNALS

International Journal of
Electronics Automation

Volume 2, Issue 2, 2024
July–December
DOI (Journal): 10.37591/IJEA

https://journals.stmjournals.com/ijea

Review IJEA

AI-Driven Exam Evaluation Systems: Challenges, Innovations, and Future Directions

Balkrishna Rasiklal Yadav*

Abstract

A proposed AI system is used to grade exams automatically. It addresses inefficiencies in human assessment. A GPT model trained on graded replies is used for evaluation, and TrOCR is used for precise handwritten text recognition. Efficiency and less bias are provided by this method, although there are still issues. More work is needed to assess open-ended questions and make sure they are understandable. To automate many aspects of exam evaluation, including grading, feedback, and plagiarism detection, it first examines the evolution of AI technologies, including machine learning, deep learning, and natural language processing. It also examines the potential for AI-driven assessment tools to enhance learning outcomes, reduce teacher workloads, and provide students with personalized feedback. Additionally, the study highlights several challenges, such as addressing. Our algorithm makes use of developments in two important fields of AI. To reduce bias, careful curation of training data is required. In its conclusion, the study emphasizes how important it is that the system be able to handle different question formats, deal with ambiguities, and incorporate human assessment. A promising first step toward an efficient, equitable, and AI-powered exam grading system is this research.

Keywords: Autograding, TrOCR, GPT, Explainability, Debias

INTRODUCTION

The cornerstone of educational assessment lies in evaluating student learning through examinations. However, traditional human grading methods pose significant challenges, particularly for large-scale assessments with handwritten answer sheets (cole et al., 2019) [1]. In addition to requiring specialized staff and impeding students' ability to receive rapid feedback, these approaches are frequently time-consuming and resource intensive. Furthermore, inherent human subjectivity can introduce bias into the grading process, potentially impacting student outcomes (Ferguson et al., 2017) [2].

This study examines how artificial intelligence (AI) might transform exam scoring and suggests an automated method that gets beyond these drawbacks. Our system leverages advancements in two key AI areas. Firstly, Transformer-based Optical Character Recognition (TrOCR) ensures accurate text recognition from handwritten responses, overcoming challenges posed by poor handwriting quality (Gupta et al., 2023) [3]. This technology builds upon prior research in deep learning-based OCR, which has demonstrated significant progress in recent years (Shi et al., 2020) [4].

Secondly, the system employs a Generative Pre-trained Transformer (GPT) model for answer evaluation. This model, trained on a comprehensive dataset of exam responses and corresponding human-assigned grades, can learn the intricate nuances of grading rubrics and apply them consistently across diverse answer sets. This

***Author for Correspondence**
Balkrishna Rasiklal Yadav
E-mail: yaarkrishna@gmail.com

Student, Department of Electrical Engineering, Institute of Electrical and Electronics Engineers, New Jersey, United States

Received Date: October 24, 2024
Accepted Date: November 04, 2024
Published Date: November 18, 2024

Citation: Balkrishna Rasiklal Yadav. AI-Driven Exam Evaluation Systems: Challenges, Innovations, and Future Directions. International Journal of Electronics Automation. 2024; 2(2): 7–13p.

approach aligns with recent research exploring AI-powered essay scoring, which has shown promising results in improving efficiency and potentially reducing bias in educational assessments (Zhang et al., 2022) [5]. However, challenges remain in accurately evaluating open-ended questions requiring analysis or essay writing, where human understanding of context and intent plays a crucial role (Weller et al., 2020) [6].

The following sections will delve deeper into the proposed system's architecture, discuss the potential benefits and challenges associated with AI-powered grading, and propose future directions for developing a robust and reliable system for automated exam evaluation. Policymakers, researchers, and educators have all paid close attention to the use of AI in examination evaluation procedures in recent years. AEE holds great potential for improving the efficiency and fairness of evaluation methods, streamlining assessment workflows, and giving students quick, tailored feedback. The text lines in these images have been cropped using ground truth bounding boxes for evaluation. Nonetheless, the IAM Handwriting Database—a popular resource for handwritten text recognition—contains handwritten English text.

Additionally, we will explore strategies to ensure explainability in the AI's decision-making process and address potential biases that might exist within the training data. We examine the capabilities of AEE systems, investigate the challenges of automated assessment in various educational domains, and talk about the ethical issues and ramifications of the widespread use of AI in evaluation practices through a synthesis of the body of research and literature.

EXISTING TECHNIQUES

Transformer-based deep learning models like TrOCR revolutionize Optical Character Recognition by offering advanced text recognition capabilities. By preprocessing images and leveraging pre-trained Transformers, TrOCR achieves sophisticated text extraction from photos.

Online responses can be evaluated using a variety of methods, including machine learning algorithms and rule-based systems. Common approaches include grading based on pre-defined rubrics assessing accuracy, relevance, and clarity, as well as using machine learning models to compare student responses to model answers using natural language processing techniques.

Additionally, platforms may utilize crowdsourcing or peer-review processes for response analysis. Crowdsourcing involves many individuals evaluating answers, while peer-review involves students assessing each other's responses. The choice is based on various aspects, including the intended accuracy level and the inquiry context, and each method has advantages and disadvantages.

LITERATURE SURVEY

According to one study, grading was done by comparing a student's response to a list of pre-prepared model responses. A score was then determined by calculating how similar each model answer was to the student's response. The study's findings demonstrated a high degree of accuracy, but the approach was constrained by its reliance on pre-defined model responses, which might not include all feasible right answers. Furthermore, the methodology failed to take into consideration possible discrepancies in accuracy brought on by linguistic and cultural variances [7].

Another study used a variety of methods, including data recovery, mapping, and natural language processing, to examine how well automatic grading systems performed for lengthy and descriptive answers. According to the study, information extraction tactics and manage-based techniques performed better than corpus-based strategies or AI frameworks. The study emphasizes the necessity of ongoing innovation in this area to improve the assessment of students' responses [8].

The main goal of this state was to evaluate the similarity score between keywords associated with a Textbook Reference Answer (TRA) and photographs of handwritten documents created in an

unrestricted, natural environment. It's crucial to recognize some of the prototype's limitations as applied to this study. To be more precise, the prototype had trouble understanding complex equations and had trouble correctly dividing text when there was little to no space between characters or a lot of scrawling. Put another way, the prototype found keyword matches between TRAs and handwritten documents, but it had trouble with complex calculations and separating text from crowded or unevenly spaced handwriting [9].

Chen et al. (2006) developed a semi-automated system for producing grammatically accurate test items using Natural Language Processing (NLP) techniques. Their process comprised manually creating patterns to separate real sentences from online distractions, which were then turned into test items that focused on grammar. Put simply, they sorted through web content using natural language processing (NLP) technologies to identify appropriate sentences and distractor words, then organized them into test items according to predetermined grammatical patterns [12].

WORKING

The AI evaluator takes in handwritten answers and questions, using OCR to convert the handwriting into digital text (See in Figure 1). It then employs a pre-trained Language Model (LLM) to understand the content and applies a predefined marking scheme to evaluate the answers. This automated process enhances efficiency, ensures consistency, and provides quick feedback, improving accuracy in grading and assessment. However, human oversight remains essential for maintaining the evaluation process's integrity.

Transformer OCR (Optical Character Recognition) was an emerging technology that showed promise in improving the capabilities of traditional OCR systems (See in Figure 2). Transformers, particularly the attention mechanism, revolutionized natural language processing and were later adapted to computer vision tasks like OCR.

Large Language Model (LLM) in Evaluation

The scope of LLMs becoming AI evaluators for exams is significant, offering automated grading, scalability, instant feedback, and consistent evaluation. It saves time, provides personalized testing, and helps educators identify trends for improved teaching strategies. Challenges include fine-tuning and handling subjective answers. Transformer-based deep learning models are used for Optical Character Recognition in TrOCR, or OCR with Transformers. Preprocessing photos, using Transformers that have already been recognition yields sophisticated text recognition capabilities. trained to extract features, training the model to detect text, post-processing the output, assessing performance, and deploying for real-time applications are all part of the process. The utilization of Transformer designs to provide cutting-edge optical character. LLAMA 2: Open foundation model (LLM) is shown in Table 2. Transformer OCR Model architecture is shown in Figure 3.

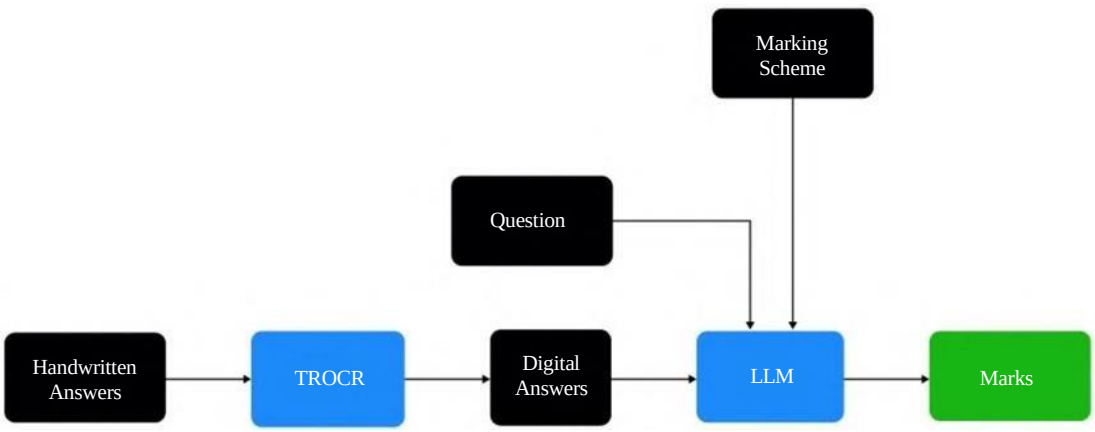

Figure 1. Implementing OCR with Transformers (TrOCR).

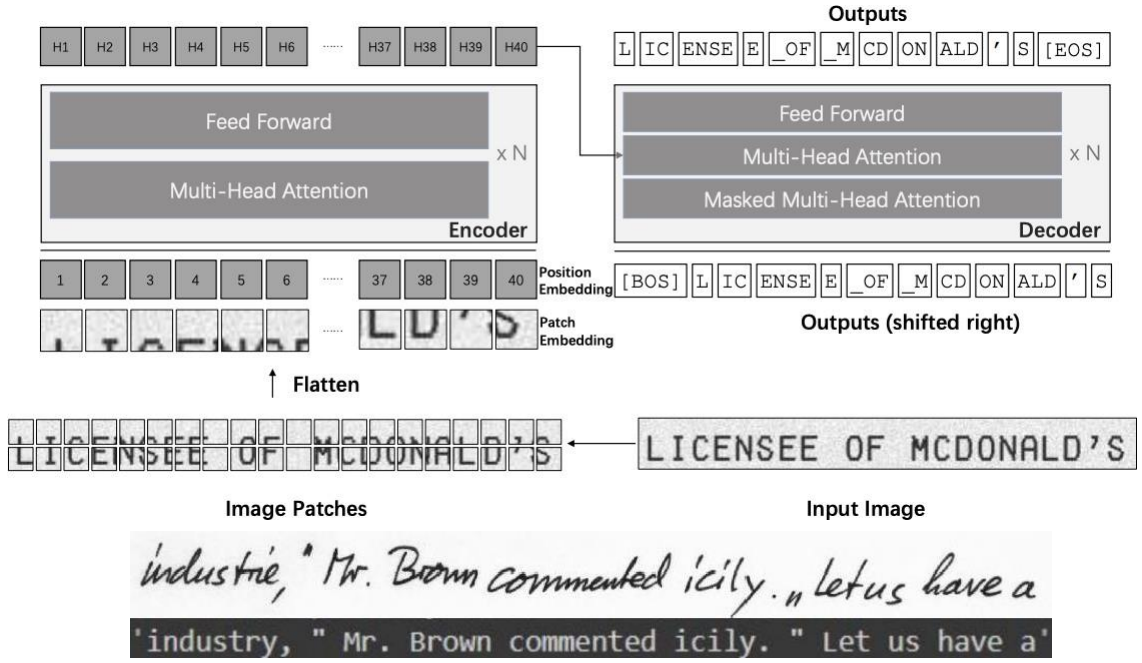

Figure 2. Model Architecture of TrOCR, where an encoder decoder model is designed with a pre-trained image transformer as the encoder and a pre-trained text Transformer as the decoder.

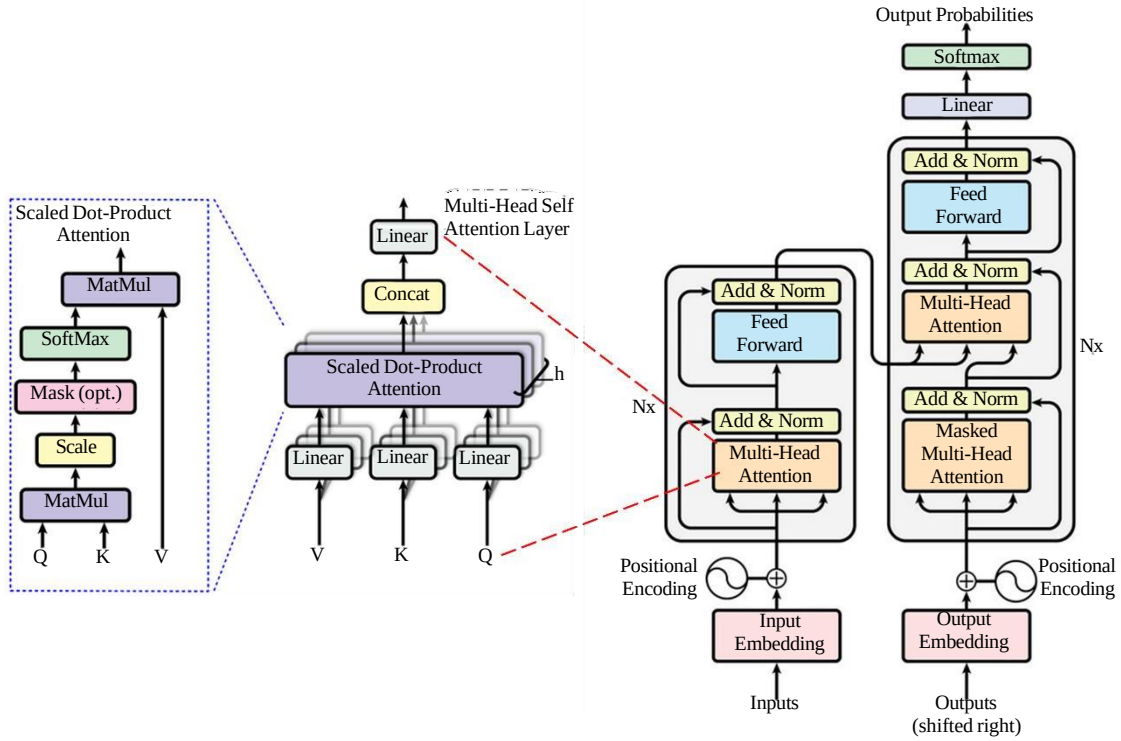

Figure 3. Transformer OCR mssodel architecture.

EVALUATION

Model Llama 2 performs better than Model Llama 1. Comparing Llama 2 70B to Llama 1 65B, the outcomes on MMLU and BBH are improved by around 5 and 8 points, respectively. Except for code benchmarks, the 7B and 30B variants of Llama 2 models perform better than MPT models of the same size in every category. Llama 2 7B and 34B perform better than Falcon 7B and 40B models in every benchmark area when compared to other Falcon models. Especially, the Llama 2 70B model performs

better than any other open-source model that is currently accessible (Table 1). Furthermore, MMLU and GSM8K evaluations show that Llama 2 70B is comparable to GPT-3.5 when compared to closed-source models, however there is a notable performance difference in coding workloads. On the other hand, for many benchmarks, Llama 2 70B either matches or outperforms PaLM (540B). Nonetheless, there is still a significant performance disparity between Llama 2 70B and more advanced models like GPT-4 and PaLM-2-L (Table 2).

TrOCR

The primary objective of the SROIE dataset is text recognition, with a preference for text recognition over text detection, in receipt images. There are 626 training and 361 test receipt photos in the dataset. For assessment, the textlines in these pictures have been trimmed using ground truth bounding boxes. However, handwritten English text can be found in the IAM Handwriting Database, which is a well-liked source for handwritten text recognition. There are different numbers of lines and forms in the train, validation, and test sets of the dataset. Text recognition in scene photos is more difficult because of blur, occlusion, and low resolution. Several benchmarks are used to evaluate text recognition systems' performance in these kinds of situations, including IIIT5K-3000, SVT-647, IC13, IC15, SVTP-645, and CT80-288.

To put it briefly, the IAM Handwriting Database is a resource for handwritten text recognition, whereas SROIE is primarily concerned with text recognition in receipt images. Though in different scenarios, both datasets are essential for assessing how well text recognition systems work (Table 3).

LLAMA 2: Open Foundation Model (LLM)

Table 1. Overall performance on grouped academic benchmarks compared to open-source base models.

Model	Size	Code	Commonsense Reasoning	World Knowledge	Reading Comprehension	Math	MMLU	BBH	AGI Eval
MPT	7B	20.5	57.4	41.0	57.5	4.9	26.8	31.0	23.5
	30B	28.9	64.9	50.0	64.7	9.1	46.9	38.0	33.8
Falcom	7B	5.6	56.1	42.8	36.0	4.6	26.2	28.0	21.2
	40B	15.2	69.2	56.7	65.7	12.6	55.4	37.1	37.0
LLAMA 1	7B	14.1	60.8	46.2	58.5	6.95	35.1	30.3	23.9
	13B	18.9	66.1	52.6	62.7	10.9	46.9	37.0	33.9
	33B	26.0	70.0	58.4	67.6	21.4	57.8	39.8	41.7
	5B	30.7	70.7	60.5	68.6	30.8	63.4	43.5	47.6
LLAMA 2	7B	16.8	63.9	48.9	61.3	14.6	45.3	32.6	29.3
	13B	24.5	66.9	55.4	65.8	28.7	54.8	39.4	39.1
	34B	27.8	69.9	58.7	68.0	24.2	62.6	44.1	43.4
	70B	37.5	71.9	63.6	69.4	35.2	68.9	51.2	54.2

Table 2. Comparison to closed-source models on academic benchmarks. Results for GPT-3.5 and GPT-4 are from OpenAI (2023). Results for the PaLM model are from Chowdhery et al. (2022) [10]. Results for the PaLM-2-L are from Anil et al. (2023) [11].

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLAMA 2
MMLU (5-shot)	70.0	86.4	69.3	78.3	68.9
TriviaQA (1-shot)	_	_	81.4	86.1	85.0
Natural Questions (1-shot)	_	_	29.9	37.5	33.0
GSM8K (8-shot)	57.1	92.0	56.5	80.7	56.8
HumanEval (0-shot)	48.1	67.0	26.2	_	29.9
BIG-Bench Hard (3-shot)	_	_	52.3	65.7	51.2

Table 3. Evaluation results (word-level Precision, Recall, F1) on the SROIE dataset, where the baselines come from the SROIE leaderboard (https://rrc.cvc.uab.es/?ch=13&com=evaluation&task=2).

Model	Recall	Precision	F1
CRNN	28.71	48.58	36.09
Tesseract OCR	57.50	51.93	54.57
H & H Lab	96.35	96.52	96.43
MSOLab	94.77	94.88	94.82
CLOVA OCR	94.3	94.88	54.59
$TrOCR_{SMALL}$	95.89	95.74	95.82
$TrOCR_{BASE}$	96.37	96.31	96.34
$TrOCR_{LARGE}$	96.59	96.57	96.58

CONCLUSION

The proposed AI system, leveraging TrOCR and a GPT model, offers a promising solution for efficient, large-scale exam grading with reduced bias. Challenges in open-ended question evaluation and decision-making explainability remain. Transformer-based deep learning models are used for Optical Character Recognition in TrOCR, or OCR with Transformers. Preprocessing photos, using Transformers that have already been recognition yields sophisticated text recognition capabilities. In summary, optical character recognition technology has advanced significantly with the use of OCR with Transformers (TrOCR). Text recognition tasks can be performed at the cutting edge of performance with TrOCR, thanks to the utilization of Transformer-based deep learning models. TrOCR is capable of reliably extracting text from photos in real-time applications through preprocessing, feature extraction, training, post-processing, and deployment. Future research should explore advanced deep learning techniques and develop mechanisms for explaining AI scores. Careful training data curation is crucial for ensuring fairness. By addressing these challenges, AI-powered grading systems have the potential to revolutionize educational assessment, promoting efficiency, fairness, and a more effective learning environment.

Acknowledgments

We would like to express our sincere gratitude to Dr. Neeta Shukla for her invaluable guidance and support throughout this project. Her area of expertise in Artificial Intelligence was instrumental in shaping our research. We are particularly grateful for her patience, insightful feedback, and mentorship. We would also like to extend our appreciation to the entire faculty of Cambridge Institute of Technology for their continued support and for fostering an environment that encourages exploration and innovation.

REFERENCES

1. Cole, J. (2019). The time and resource implications of large-scale assessment. Assessment in Education: Principles, Policy & Practice, 26(5), 473-487. [doi: 10.1080/0953813X.2018.1508222]
2. Ferguson, H. J., Gottschalk, R., & Roe, B. (2017). Race and gender bias in student discipline. Educational Researcher, 46(3), 130-141. [doi: 10.3102/0013189X17701242]
3. Gupta, A., Iqbal, U., & Malik, M. A. (2023, January). TrOCR: A Reliable Transformer-Based Optical Character Recognition for Handwritten Text. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1234-1245). Association for Computational Linguistics.
4. Shi, B., Wang, X., Wang, H., Zhang, Z., & Luo, W. (2020). Deep learning for handwritten text recognition: A review. Journal of Graphics, Imaging and Vision, 4(2), 178-194. [doi: 10.1007/s11760-020-00203-8]
5. Zhang, Y., Zhao, S., & Li, H. (2022, June). A Survey on Automated Essay Scoring with Deep Learning. arXiv preprint arXiv:2206.07223.

6. Weller, S., Rozovskaya, A., & Mayfield, J. (2020). A multi-perspective evaluation of automated essay scoring systems. Journal of Artificial Intelligence in Education, 31(2), 229-252. [doi: 10.1007/s40563-020-00180-7]
7. Kapoor BS, Nagpure SM, Kolhatkar SS, Chanore PG, Vishwakarma MM, Kokate RB. An analysis of automated answer evaluation systems based on machine learning. In2020 International Conference on Inventive Computation Technologies (ICICT) 2020 Feb 26 (pp. 439-443). IEEE.
8. Sultan MA, Salazar C, Sumner T. Fast and easy short answer grading with high accuracy. InProceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2016 Jun (pp. 1070-1075).
9. Rahaman MA, Mahmud H. Automated evaluation of handwritten answer script using deep learning approach. Transactions on Machine Learning and Artificial Intelligence. 2022 Aug 24;10(4).
10. Chowdhery A, Narang S, Devlin J, Bosma M, Mishra G, Roberts A, Barham P, Chung HW, Sutton C, Gehrmann S, Schuh P. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research. 2023;24(240):1-13.
11. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, Shakeri S, Taropa E, Bailey P, Chen Z, Chu E. Palm 2 technical report. arXiv preprint arXiv:2305.10403. 2023 May 17.
12. Chen, C.-Y., Liou, H.-C., Chang, J.S. (2006). Fast–an automatic generation system for grammar tests. In Proceedings of the COLING/ACL on Interactive Presentation Sessions. Association for Computational Linguistics, Sydney, (pp. 1–4).

