Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #47815

Automated Grading of Engineering Mechanics Assignments Using Large Language Models and Computer Vision: A Work in Progress

Dr. Ahmed Mowafy PEng, University of Alberta

Dr. Ahmed Mowafy Saad is an Assistant Teaching Professor in Civil and Environmental Engineering at the University of Alberta. He teaches large first- and second-year courses such as Engineering Mechanics and Mechanics of Deformable Bodies, reaching over 1,800 students annually. With over 13 years of combined academic and industry experience in the Middle East and Western Canada, he integrates real-world insights into innovative teaching practices. His recent work focuses on AI-assisted grading tools and dynamic, multi-version assessments that improve academic integrity and efficiency at scale. He has also introduced self-marking strategies inspired by Indigenous assessment practices to promote reflection and deeper learning. In addition, he pioneered the Fantasy Mechanics League (FML), a gamified classroom framework that enhances student motivation and engagement in large lecture settings. His work is grounded in the belief that ”teaching empowers you to create countless lives through the brilliant minds you cultivate.”

Mr. Mohammad Talebi-Kalaleh, University of Alberta

Mohammad Talebi-Kalaleh is a self-motivated third-year PhD student in Structural Engineering at the University of Alberta. His research focuses on leveraging Large Language Models (LLMs) in engineering education and exploring their applications in structural engineering. Mohammad’s work primarily revolves around artificial intelligence, smart infrastructure, and bridge monitoring through crowdsensing techniques. He is particularly passionate about integrating advanced AI-driven solutions into structural health monitoring to enhance infrastructure resilience and efficiency.

Mohamed Sabek, University of Alberta

Mohamed Sabek is a Ph.D. student in Construction Management and Engineering at the University of Alberta. He is the lead developer of the AI-driven grading framework presented in this paper, which integrates Optical Character Recognition and Large Language Models (LLMs) to enhance assessment efficiency in large-scale engineering courses. His research lies at the intersection of artificial intelligence, computer vision, and engineering education, with a special focus on fine-tuning vision-language models (VLMs) for real-time construction site monitoring. Mohamed’s broader research goals aim to advance intelligent automation in construction environments to improve safety, productivity, and resource allocation. He holds PMP and RMP certifications and has received multiple awards recognizing his contributions to AI applications in civil engineering and education technology.

Harry Peng, University of Alberta
Mohammad Aqib, University of Alberta
Dr. Samer M. Adeeb P.Eng., University of Alberta
Mohamed Magdy Elgammal, University of Alberta
Dr. Clayton Pettit, University of Alberta

ASEE AMERICAN SOCIETY FOR ENGINEERING EDUCATION

Automated Grading of Engineering Statics Assignments Using Large Language Models and Computer Vision: A Work in Progress

Abstract - This project aims to develop an automated grading framework for the Statics course at a Canadian University, which serves over 1,500 students annually. Grading these complex assignments, often involving handwritten responses and diagrams, currently requires the efforts of over 26 teaching assistants (TAs). The manual system faces challenges including grading inconsistencies, delays in providing feedback, and significant resource strain. To address these limitations, the proposed framework integrates Optical Character Recognition (OCR) and Large Language Models (LLMs) to automate the grading process. The framework begins with anonymization and preprocessing of student submissions, followed by digitization using OCR software such as Mathpix to convert handwritten content into machine-readable formats. The digitized files are evaluated using a grading platform that employs LLMs guided by predefined marking schemes and step-by-step evaluation prompts. The system generates grades, detailed feedback, and consolidated reports for each submission. Preliminary testing demonstrated significant reductions in grading variability and time, while maintaining a strong alignment with human-assigned grades. The findings show that Qwen LLM achieves higher consistency across multiple evaluation metrics, whereas GPT-4 provides superior accuracy in specific scenarios. This approach enhances the scalability and efficiency of grading practices, making it a promising solution in large-scale educational settings.

Keywords: Automated grading, Optical Character Recognition, Large Language Models, Handwritten assignment digitization, Education technology

1. Introduction

The Statics course, has over 1,500 enrolled students annually, relies heavily on detailed grading of complex assignments that include handwritten responses and diagrams. Currently, more than 26 teaching assistants (TAs) manage this process. However, manual grading systems face significant challenges, including grading inconsistencies, limited feedback, and a substantial burden on human resources. Ensuring consistency across a large team of TAs and delivering timely, meaningful feedback to students have become increasingly difficult under these constraints.

The field of Natural Language has seen robust progress over the past few decades, driven by advancements in deep learning, computational resources, and the availability of large data [1]. Progress in the field of NLP dates back to 1950 when researchers at IBM and Georgetown University developed a system and successfully converted the collection of phrases from Russian to English [2]. Pioneering work, such as Shannon's information theory in 1948, laid the groundwork for probabilistic modeling, whereas the introduction of Hidden Markov Models (HMMs) enabled advancements in tasks such as part-of-speech tagging and speech recognition during the mid-20th century [3]. The transition to neural language models (NLMs) in the late 1990s

marked a significant leap forward. Utilizing the increasing computational power and availability of datasets, neural networks have begun to reshape NLP. Early 2000 had witnessed the rise of word embeddings increased with the introduction of Word2vev and GloVe, and they were able to capture semantic and syntactic relationships in dense vector formats [4,5]. Recurrent neural networks (RNNs) further enhance sequential data processing; however, challenges such as vanishing gradients limit their performance [6]. These limitations were addressed by the introduction of networks such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), which facilitated the modeling of long-term dependencies in language sequences [7,8]. A paradigm shift occurred with the advent of the attention mechanism and the transformer architecture in the late 2010s. The transformer model, introduced in 2017, replaced sequential processing with a self-attention mechanism, enabling parallelization of training and efficient context handling across entire text sequences. The self-attention mechanism was introduced in the paper "Attention is all you need" [9]. This innovation laid the foundation for pretrained Large Language Models (LLMs), transfer learning, and revolutionizing NLP. Models such as Bidirectional Encoder Representations from Transformers (BERT) have demonstrated the effectiveness of bidirectional contextual embeddings, excelling in multiple benchmarks. Simultaneously, autoregressive models such as OpenAI's Generative Pre-Trained Transformer (GPT) series can generate coherent and contextually relevant text [10].

LLMs have revolutionized several fields, including programming [11], legal [12], medical [13], and others [14]. One notable advancement brought by LLMs is automatic evaluation, which was shown by a group of researchers that when these LLMs are provided with evaluation criteria and reference samples to be evaluated for a particular domain, they can provide results that are comparable with the experts of that domain [15]. A study presented by a few researchers included automating the scoring of software modules, analyzing them, and checking them for plagiarism [16]. As automatic scoring provides timely feedback on the number of students, it becomes a critical step to include it in today's education system [17]. LLM demonstrates a strong ability to quickly adapt to downstream tasks with limited or no training data [18]. Certain strategies, such as Chain of Thoughts and In-context learning, can help in aligning LLMs evaluators with humans [19,20]. This facilitates the development of systems, while saving extra costs that might incur the development of these systems from scratch.

One of the major issues is automating the scoring process when evaluating students' handwritten assignments. Before asking LLMs to evaluate handwritten assignments, it is necessary to convert them into a machine-readable format. Irregular text, poor assignment format, inconsistency, and bad handwriting pose a major challenge when converting handwritten text into a machine-readable format. In the past, HMMs were used for digitization, but they have the drawback of only considering current observations and not the context in which they occur [21]. Many attempts have been made to correctly identify and digitize handwritten text using a Convolutional Neural Network (CNN) [22]. The task is complex, and there is a need for a language model to assist and provide predictions in the case of visual ambiguity. The current state-of-the-art system utilizes a combined vision and text transformer model based on the TrOCR model [23]. TrOCR is trained on synthetic handwritten data and can be fine-tuned in a specific domain. Currently, multimodal LLMs are also gaining popularity and have shown promising results in the digitization of handwritten text. Multimodality models such as GPT-4v are capable of recognizing text and even tabular structures present in images [10]. Such models can take images of handwritten text and a prompt as input and can provide digitized text as output. Digitizing text

using multi-modal LLMs is not limited to GPT-4v, and nowadays, many models can perform the same task, including Gemini, InternVL, Claude-Sonnet etc. In fact, certain software such as Transcription Pearl are using multi-modal models such as GPT-4o, Claude Sonnet3.5, and Gemini 1.5-Pro to quickly perform digitisation with high degree of accuracy. They reported accuracy levels between 84 and 93%, considering the different methods of evaluation [24].

This study seeks to leverage OCR and LLMs to automate the grading of handwritten engineering assignments, specifically in the Statics course. The proposed framework aims to streamline grading, enhance feedback quality, and alleviate the workload of TAs, allowing them to focus on academic support. The remainder of this paper is structured as follows: Section 1 introduces the project and reviews the relevant literature. Section 2 outlines the methodology, including pre-processing, digitization, and LLM-based grading. Section 3 presents the results and discusses the digitization and grading performances. Section 4 concludes the paper with key findings and potential future directions.

2. Methodology

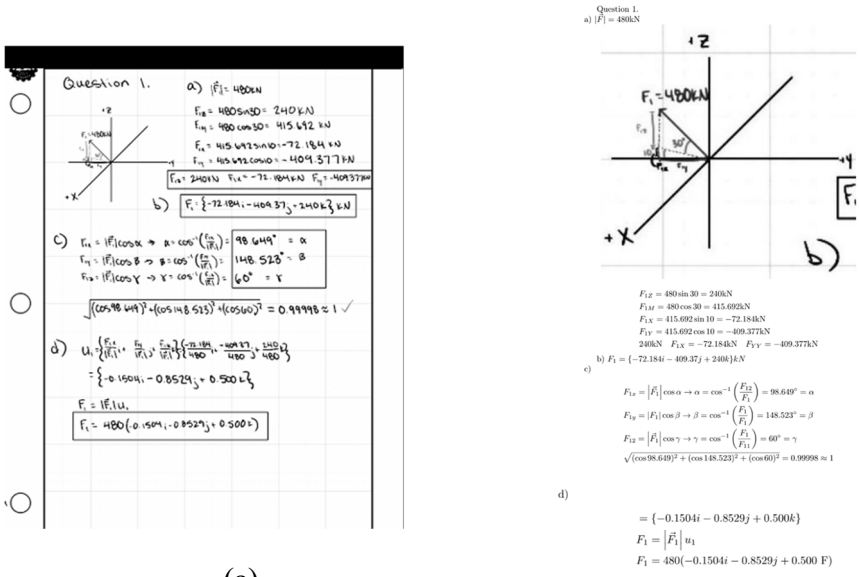
The proposed framework begins by anonymizing student submissions and preprocessing files to ensure privacy and uniformity. This was followed by digitization using OCR software, such as the Mathpix Snipping Tool, to convert handwritten content into machine-readable formats. Finally, the digitized files were fed into the grading platform developed for this study, along with reference solutions and grading scheme prompts. This platform enables users to select an LLM, configure grading parameters, and initiate an automated evaluation process.

2.1. Preprocessing and Digitization

The framework begins with anonymizing student submissions and preprocessing files to ensure privacy and uniformity. The assignments were submitted via the university's learning portal and processed locally using a Python application equipped with OCR libraries for ID detection. The student names and IDs were automatically anonymized (Figure 1a), and images within submissions were preprocessed for digitization. An anonymized PDF was generated for each submission, ensuring that the LLMs evaluating the assignments were not exposed to identifying information. Mathpix Snipping Tool software was used to digitize the handwritten submissions. This OCR software converts handwritten text, equations, and diagrams, including free-body and geometric constructions, into machine-readable formats (Figure 1b). Once digitization is completed, the results are securely saved and linked to anonymized student identifiers.

2.2. LLM-Based Grading Platform

The digitized data were evaluated using LLMs guided by detailed marking schemes tailored to assignment requirements. The grading platform developed for this study allows users to select an LLM, configure grading parameters, and define marking criteria. Figure 2 illustrates the platform's interface, which supports multiple local and online LLMs and generates structured outputs, including individual feedback reports and consolidated grading summaries in Excel format.



(a) (b)

Figure 1. (a) Sample handwritten student submission after pre-processing and anonymization, with student information removed. (b) Submission after digitization using Mathpix Snipping Tool.

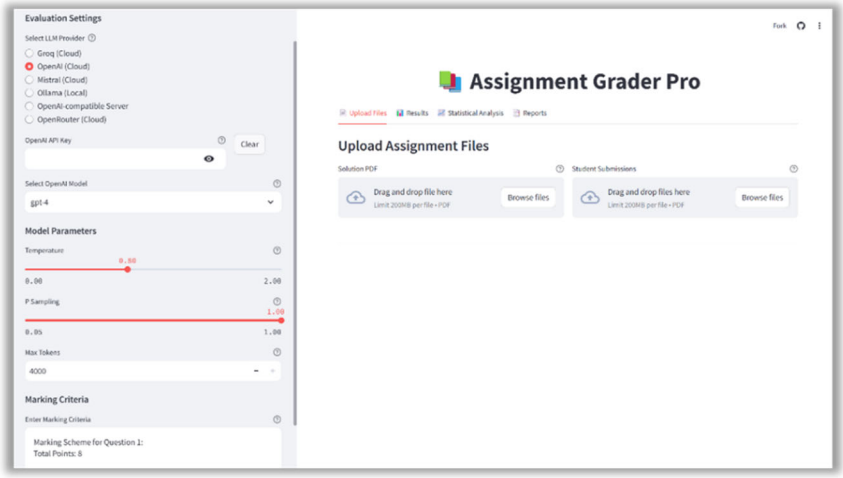


Figure 2. Interactive grading platform for automated evaluation of digitized lab submissions using multiple LLMs.

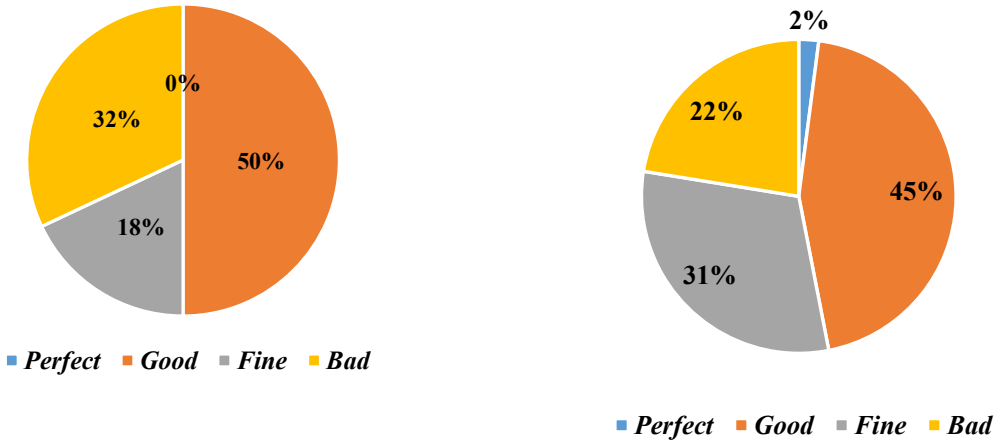
3. Results and Discussion

This section evaluates the reliability of the proposed framework by comparing AI-generated grades with human-graded benchmarks. Discrepancies were flagged for manual review, and iterative refinements were made to the models based on feedback from the TAs.

3.1. Digitization Performance

To interpret the ultimate grading results, a sample of 50 handwritten assignments containing complex equations and diagrams was analyzed in each lab to evaluate the digitization accuracy of Mathpix. Each file was scored from 0 to 10, reflecting the precision of the digitization process. A perfect score of 10 indicated flawless digitization, while lower scores reflected errors such as incomplete or inaccurate content extraction. Scores were categorized as "Perfect" (10), "Good" (8–9), "Fine" (7), and "Bad" (≤6).

Figure 3 shows the percentage distribution of digitization performance for Labs 2 and 5. Approximately 50% of the submissions scored between 8 and 10, primarily due to neat handwriting and consistent text flow. Submissions with lower scores often involve irregular formatting or ambiguous handwriting.



(a) Lab 2 (b) Lab 5

Figure 3. Percentage distribution of digitization performance for Labs 2 (a) and 5 (b).

3.2. Evaluation of the Grading Platform

The grading platform was tested on two sets of 50 assignments graded by the GPT-4 and Qwen. AI-generated grades were compared with the human-graded benchmarks. Figure 4 shows the mean scores and variability (mean ± standard deviation) for Labs 2 and 5, with human scores serving as the reference for comparison. In Lab 2, the human reference mean was 16.89, with Qwen scoring 16.07 and GPT-4 scoring 15.72. Qwen's score was closer to the human reference, indicating better performance than GPT-4 in this laboratory. In Lab 5, the human reference mean was 20.44, and Qwen achieved a mean score of 22.67, whereas GPT-4 scores were 21.17. Although both LLMs score higher than the human reference, GPT-4's score is closer, suggesting that it performs better than Qwen in aligning with the human benchmark. Human evaluations displayed the largest variability in both Lab 2 (Â ±3.48) and Lab 5 (Â ±2.58). By contrast, GPT-4 and Qwen exhibited significantly lower deviations, with Qwen showing the smallest variability (±2.32, Lab 2 and ±1.45 in Lab 5), making it the most consistent performer.

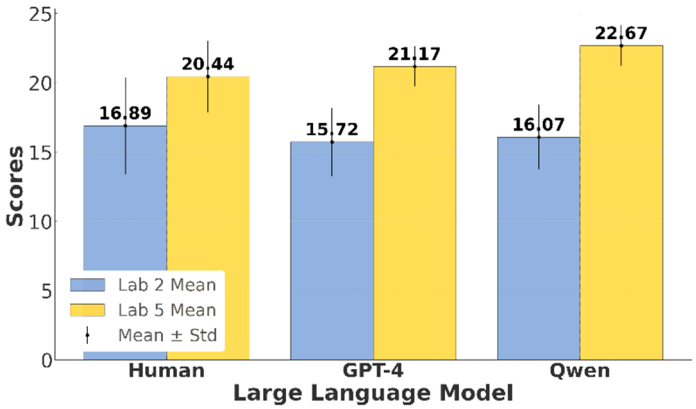


Figure 4. Comparison of mean scores and variability for Lab 2 and Lab 5 across three large language models (Human, GPT-4, and Qwen).

Figure 5 also depicts the correlations between human evaluations and AI-generated scores. In Lab 2, Human-Qwen achieved the highest correlation (0.83), outperforming Human-GPT-4 (0.62), indicating Qwen's closer alignment with human benchmarks. In Lab 5, both Human-Qwen and Human-GPT-4 exhibited identical correlations (0.84), suggesting comparable alignment with the human evaluations in this lab.

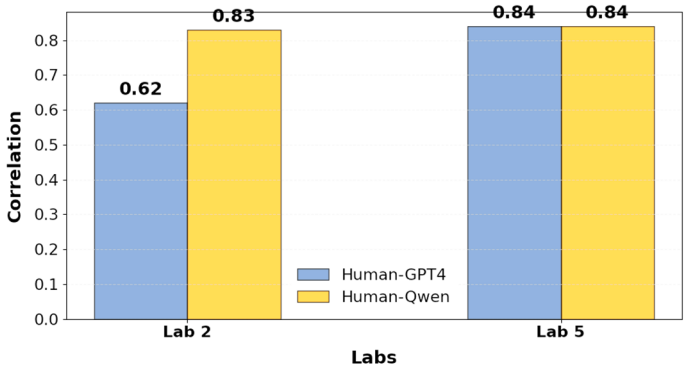
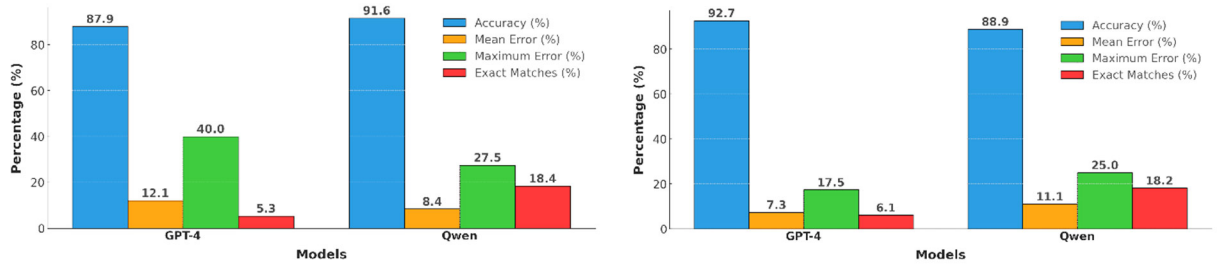


Figure 5. Correlation between human evaluations and scores generated by GPT-4 and Qwen for Labs 2 and 5.

Figure 6 illustrates the key performance metrics for GPT-4 and Qwen in Labs 2 and 5. In Lab 2 (Figure 6a), Qwen outperformed GPT-4, achieving a higher accuracy (91.6% vs. 87.9%), lower mean error (8.4% vs. 12.1%), and smaller maximum error (27.5% vs. 40.0%). Additionally, Qwen demonstrated a higher percentage of exact matches (18.4%) than GPT-4 (5.3%) did. In Lab 5 (Figure 6b), GPT-4 showed a slightly higher accuracy (92.7%) and lower mean error (7.3%) than Qwen (accuracy: 88.9%; mean error: 11.1%). However, Qwen exhibited a smaller maximum error (17.5% vs. 25.0%) and higher percentage of exact matches (18.2% vs. 6.1%). These findings suggest Qwen's consistency and alignment with human grading, whereas GPT-4 demonstrated a stronger accuracy in Lab 5.



(c) Lab 2 (d) Lab 5

Figure 6. Performance metrics for GPT-4 and Qwen in Labs 2 (a) and 5 (b), including accuracy, mean error, maximum error, and exact matches, as percentages.

Overall, the AI grading framework demonstrated strong agreement with the human evaluations. Discrepancies were primarily attributed to handwritten digitization errors, underscoring the importance of accurate pre-processing for optimal AI grading performance.

4. Conclusion and Future Work

This study proposes and evaluates an AI-driven framework for automating the grading of handwritten Statics assignments. By integrating OCR and LLMs, the framework addresses the challenges associated with manual grading, including inconsistencies, delays, and limited feedback. Preliminary results highlight significant reductions in grading time and variability, with Mathpix software achieving high digitization accuracy, and LLMs demonstrating strong alignment with human benchmarks. Despite its potential benefits, the use of AI for grading raises important ethical considerations. One concern is the fear that automated grading could reduce the need for teaching assistants or faculty members, potentially affecting academic employment. However, this system is envisioned as a supplement rather than a replacement. By automating repetitive grading tasks, educators can focus on providing deeper feedback and personalized support to students, thereby enhancing their overall learning experience. Moreover, safeguards to prevent algorithmic bias and to protect student data privacy are critical. Ensuring transparency in how AI-generated scores are derived and maintaining human oversight is essential for responsibly integrating these tools into higher education.

Future work will focus on refining digitization methods, exploring multimodal LLMs, and incorporating additional grading metrics to enhance the reliability and scalability of the framework. These efforts aimed to establish an efficient student-focused assessment system for large engineering courses.

References

[1] Hirschberg, J., and Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), pp.261-266.
[2] Dale, R., 2019. NLP commercialization in the last 25 years. Natural Language Engineering, 25(3), pp.419-426.
[3] Chandrika, V.P., Verma, R., Charan, N., Ditheswar, S., Hansika, S. and Ishwariya, R., 2024, July. POS Tagging Using Hidden Markov Models in Natural Language Processing. In 2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT) (pp. 1-6). IEEE.
[4] Mikolov, T., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 3781.
[5] Pennington, J., Socher, R. and Manning, C.D., 2014, October. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532-1543).
[6] Noh, S.H., 2021. Analysis of gradient vanishing of RNNs and performance comparison. Information, 12(11), p.442.
[7] Hochreiter, S., 1997. Long Short-term Memory. Neural Computation MIT-Press.
[8] Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks in sequence modeling. arXiv preprint arXiv:1412.3555.
[9] Vaswani, A., 2017. Attention is all you need. Advances in Neural Information Processing Systems.
[10] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. and Avila, R., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
[11] Vadim Liventsev, Anastasiia Grishina, Aki Ha¨rma¨, and Leon Moonen. 2023. Fully autonomous programming with large language models. In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '23). Association for Computing Machinery, New York, NY, USA, 1146–1155. https: //doi.org/10.1145/3583131.3590481 [Louis et al. (2023)] Antoine Louis, Gijs van Dijck.
[12] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023. Interpretable Long-Form Legal Question Answering with Retrieval-Augmented Large Language Models. https://doi.org/10.48550/ arXiv.2309.17050 arXiv:2309.17050 [cs].
[13] Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A Survey of Large Language Models for Healthcare: from Data, Technology, and Applications to Accountability and Ethics. https://doi.org/10.48550/arXiv.2310.05694 arXiv:2310.05694 [cs].
[14] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y. and Ye, W., 2024. A Survey on the Evaluation of Large Language Models. ACM Transactions on Intelligent Systems and Technology, 15(3), pp.1-45.
[15] Chiang, C.H., and Lee, H.Y. 2023. Can large-language models be an alternative to human evaluation? arXiv preprint arXiv:2305.01937.
[16] Vimalaraj, H., Thenuwara, T.B.K.P., Wijekoon, V.U., Sathurjan, T., Reyal, S., Kuruppu, T.A. and Tharmaseelan, J., 2022, April. Automated programming assignment marking tool. In 2022, the IEEE 7th International Conference for Convergence in Technology (I2CT) (pp. 1-8). IEEE.
[17] Zhai, X., 2021. Advancing automatic guidance in virtual science inquiry: from ease of use to personalization. Educational Technology Research and Development, 69(1), pp.255-258.
[18] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are based on few-shot learners. Advances in Neural Information Processing Systems, 33, pp.1877-1901.
[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large-language models. arXiv:2201.11903, 2022.
[20] Koike, R., Kaneko, M. and Okazaki, N., 2024, March. Outfox: Llm-generated essay detection through in-context learning using adversarially generated examples. In Proceedings of the AAAI Conference on Artificial Intelligence (vol. 38, No. 19, pp. 21258-21266).
[21] Babu, N. and Soumya, A., 2019, April. Character recognition in historical handwritten documents: A survey. In 2019, the International Conference on Communication and Signal Processing (ICCSP) (pp. 0299-0304). IEEE.
[22] Tran, H.P., Smith, A. and Dimla, E., 2019, November. Offline handwritten text recognition using convolutional recurrent neural networks. In 2019, the International Conference on Advanced Computing and Applications (ACOMP) (pp. 51-56). IEEE.

[23] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, F. Wei, Trocr: Transformer based optical character recognition with pre-trained models, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 13094–13102.
[24] Humphries, M., Leddy, L.C., Downton, Q., Legace, M., McConnell, J., Murray, I., Spence, E. 2024. Unlocking Archives: Using Large Language Models to Transcribe Handwritten Historical Documents. arXiv preprint arXiv:2411.03340.