LLMs in Automated Essay Evaluation: A Case Study

Milan Kostic[1], **Hans Friedrich Witschel**[2], **Knut Hinkelmann**[1, 2], **Maja Spahic-Bogdanovic**[1, 2]
[1]University of Camerino (UNICAM)
[2]FHNW University of Applied Sciences and Arts Northwestern Switzerland
milan.kostic@unicam.it, {hansfriedrich.witschel, knut.hinkelmann, maja.spahic}@fhnw.ch

Abstract

This study delves into the application of Large Language Models (LLMs), such as ChatGPT-4, for the automated evaluation of student essays, focusing on a case study conducted at the Swiss Institute of Business Administration (SIB). It explores the effectiveness of LLMs in assessing German-language student transfer assignments, contrasting their performance with traditional human lecturer evaluations. The primary findings highlight LLMs' challenges in accurately grading complex texts according to predefined criteria and providing detailed feedback. This research illuminates the gap between LLM capabilities and the nuanced requirements of student essay evaluation. The conclusion emphasizes the necessity for ongoing research and development in LLM technology to improve automated essay assessments' accuracy, reliability, and consistency in educational contexts.

Introduction

The ability of students to demonstrate their intellectual development in a specific field by writing represents a critical component of the academic process, as highlighted by Hayland (2013). Writing essays provides a platform for assessment by lecturers and motivates students to engage more deeply with the subject matter, thereby promoting cognitive skills such as analysis and synthesis (Zupanc and Bosnić, 2018). Considering essay assessment's time-consuming and costly nature, Page (1966) presented the "Project Essay Grade" concept for automatically grading students' essays on the English language or literature. Such systems aim to increase assessment efficiency, consistency, and objectivity while minimizing time and resources.

Various systems are available to extract and evaluate various text attributes using algorithms. These systems use NLP algorithms and regression models to assess writing style or content quality (Ifenthaler, 2023). Most of these automatic evaluation systems require a corpus of essays already graded by humans. Based on these human evaluations, the system learns and can assess new essays without human intervention. However, various systems' test quality, accuracy, or correctness are often not publicly accessible or examined in comprehensive empirical studies (Wilson and Rodrigues, 2020).

Ramesh and Sanampudi (2021) conducted a detailed literature analysis. They found that the challenge of automatically evaluating longer texts, such as essays, has yet to be fully mastered. The significant problem is that existing systems have difficulty processing long texts and providing understandable evaluations to both students and lecturers.

The use of Natural Language Processing with Large Language Models (LLMs) is a rapidly growing research field in Artificial Intelligence (Strasser, 2023). It covers a wide range of applications, including automated essay evaluation. LLMs like Google Bard or ChatGPT can process complex sentences and establish relationships between text elements, including the user's intent. Such pre-trained LLMs are an advantage because users do not need extensive programming knowledge and require minimal or no training data to achieve satisfactory results. Users can also provide textual instructions and examples to the model to initiate a desired interaction.

In this paper, we first examine the development of automatic essay evaluation systems before discussing potential ideas for using LLMs to evaluate students' transfer assignments written in German automatically. Transfer assignments are a pivotal evaluation tool at the Swiss Institute of Business Administration (SIB) to assess students' ongoing learning paths. These assignments, which resemble essays in their format, frequently require students to engage with real-world professional scenarios encountered within their employment. We investigated OpenAI's ChatGPT-4 (GPT-4) ability to evaluate student transfer assignments according to predefined assessment criteria and aspects and generate corresponding feedback. Initially, we analyze the consistency of the lecturers when evaluating transfer assignments before tasking GPT-4 with the same. This experiment

used four transfer assignments written in German and of varying quality from the SIB.

This paper reviews state-of-the-art literature in Section 2, presents a practical use case in Section 3, proposes innovative strategies to improve assessment consistency in Section 4, and concludes with key findings and implications in Section 5.

Related Work

Page (1966) introduced the Project Essay Grading concept, catalyzing the development of computer systems capable of evaluating essays without human intervention. This advancement necessitated a development in the terminology of these systems, leading to the interchangeable use of various terms. Initially, Page (1966, 1968) described the process as "Analyzing Students' Essays by Computers" and "Computer Grading of Essays." Subsequently, the field broadened to include terms like "Automated Essay Scoring" (AES) (Attali and Burstein, 2006; Ke and Ng, 2019), "Automated Essay Grading" (AEG) (Valenti, Neri, and Cucchiarelli, 2003), and "Automated Writing Evaluation" (AWE) (Beigman Klebanov and Madnani, 2020). These terminological changes mirror the advancements in automated assessment systems. Contemporary systems aim to design their functionality to provide detailed and transparent assessments. This approach encompasses delivering holistic evaluation results, such as grades or scores, complemented with feedback for improvement. Zupanc and Bosnić (2017) particularly underscore this aspect in their discussion of 'Automated Essay Evaluation' (AEE), illuminating the shift in these systems towards a more nuanced approach.

In their analysis, Zupanc and Bosnić (2015) compared 21 AEE systems based on various criteria, including methodology, main focus, feedback application, required essays for training, prediction model, rank, and average accuracy. Their findings revealed that while these systems broadly tackle similar tasks, they often employ distinct methodologies for extracting attributes and constructing their models. Natural Language Processing (NLP) is the predominant method in this comparison. However, the AEE systems are capable of not only identifying syntactic errors but also providing holistic feedback. Moreover, they currently lack to offer meaningful and content-specific feedback.

Further analysis by Zupan and Bosnić (2018) indicated a growing adoption of these systems linked to their enhanced reliability. High-stakes assessments such as the Graduate Record Examination (GRE), the Test of English as a Foreign Language (TOEFL), and the Graduate Management Admission Test (GMAT) increasingly employ these systems. These assessments require precise and reliable essay grading, a challenge met through the cooperation of automated systems and human graders.

Ramesh and Sanampudi (2021) performed a systematic literature review, including publications from 2010 to 2020, examining AEE systems that process data in English. They identified 62 publications, indicating that most automated grading systems leverage NLP methods for grade prediction. These predictions rely on training with corpora of human-rated essays, enabling the systems to grade autonomously and process large quantities of essays efficiently. They identified several limitations of AEE systems, including the absence of assessments based on content relevance, cohesion, and coherence, a lack of domain knowledge-based evaluation through machine learning models, the inability of NLP libraries to process words with multiple meanings, and a lack of focus on consistency and completeness in the evaluation process.

The utilization of this technology, however, is not limited to English. It encompasses a range of languages, including Arabic, Bahasa Malay, Basque, Chinese, Finnish, French, German, Hebrew, Japanese, Malaysian, Spanish, and Swedish (Hussein, Hassan, and Nassef, 2019). The advancement of automatic essay grading tools in non-English languages faces challenges, primarily due to the complexity of developing NLP tools for each language and the global predominance of English in research.

Evaluating the efficacy of Automated Essay Grading Systems remains a complex task, especially since agreement among human evaluators can be relatively low when grading a given essay or assignment. Williamson, Xi, and Breyer (2012) advocate for a reliable evaluation method that compares the grading outcomes of human graders with those of automated systems and then calculates the interrater agreement. When human and automated grading results align closely, especially within a predefined threshold, the system can be considered adequate and within acceptable levels of prediction accuracy. Ramesh and Sanampudi (2021) have identified that in this domain, researchers predominantly utilize three metrics for evaluation: the quadratic weighted Kappa coefficient, the Pearson Correlation Coefficient, and the Mean Absolute Error, which are necessary for accurately measuring the performance of these systems.

Ramesh and Sanampudi (2021) and Ifenthaler (2023) have highlighted the limitations of AEE systems, including challenges in providing adequate feedback and capturing long-distance dependencies. In contrast, LLMs are gaining popularity due to their capability to handle diverse tasks without specific training on certain datasets. Masikisiki, Marivate, and Hlophe (2023) reported that this increasing prevalence has sparked interest in their potential educational applications, particularly for assessment purposes.

Use Case

The SIB uses so-called "transfer assignments" to test students' practical application of theoretical business knowledge. These assignments necessitate that students demonstrate their comprehension and application skills in real-life contexts. The assignments, typically between 9 and 15 pages, are methodically graded using an analytical rubric. This rubric is designed to capture a multifaceted view of student performance and consists of six criteria. These criteria, defined by specific guiding questions, include Company Description, Stakeholder Analysis, Environmental Analysis, Proposed Measures, Organization and Structure, and Formal Requirements. Each criterion consists of various evaluation aspects, each allowing students to earn different point totals. There are 16 evaluation aspects, with a maximum possible score of 60 points. Lecturers also provide constructive written feedback on each criterion, highlighting areas for student improvement. Each lecturer is responsible for grading their students' assignments to ensure consistency and fairness in assessment. If neither the SIB quality control department nor the students raised any concerns regarding the provided evaluation results, this suggests their overall acceptability.

Working Alone, Together Method

"Working Alone, Together" is a collaborative method for creative thinking or problem-solving (Bruns, 2013). All participants tackle the same challenge in this method by writing down or illustrating their ideas. In the initial solo phase, participants individually develop their solutions, which they share in the subsequent group session. The strength of this method lies in preserving and valuing everyone's creativity, opinion, and viewpoint, ensuring they are not overshadowed or left unexpressed in the group setting, and minimizing mutual influence among participants.

In the workshop, we utilized the "Working Alone, Together" method to assess the consistency of lecturers' evaluation of transfer assignments in General Business Administration. The evaluation included the grading of each aspect and the writing of feedback. Four assignments, previously evaluated by the author of this paper and varying in quality, were selected for evaluation. Neither the SIB quality control department nor the students raised any concerns about these assessments, implying their general acceptability. While aware that these papers had been previously evaluated, the three participating lecturers were not informed of the original evaluation outcomes nor who had initially evaluated them.

Consistency Evaluation

During the workshop, we analyzed the consistency of evaluations of three lecturers by comparing their grades on four transfer assignments (TA1 to TA4) with the original grades.

	TA 1	TA 2	TA 3	TA 4	MAD per Lecturer	PCC
Original Grad	58	26	50	37	-	-
Lecturer 1 Grade	57	29	43	51	6.25	0.7832
Lecturer 2 Grade	57	35	40	51	8.50	0.6571
Lecturer 3 Grade	54	30	46	60	8.75	0.5617
Lecturer 1 AD	1	3	7	14	-	-
Lecturer 2 AD	1	9	10	14	-	-
Lecturer 3 AD	4	4	4	23	-	-
MAD per TA	2.00	5.33	7.00	17.00	-	-

Table 1: Consistency Assessment Results.

The lecturers re-assessed the transfer assignments. Table 1 presents individual evaluations, including grades from both lecturers and original grades. To quantify each lecturer's deviation, we calculated the Mean Absolute Deviation (MAD) from the original grades for each evaluation. The column labeled 'MAD per Lecturer' displays the deviations, while 'MAD per TA' shows the average deviation across all lecturers for each assignment. Although the sample size is small, we also calculated the Pearson Correlation Coefficient (PCC) to measure the strength of the linear relationship between the lecturers' evaluations and the original grades.

The analysis indicates an increasing MAD value for each subsequent assignment, suggesting growing evaluation variance. The PCC values show a positive correlation between lecturer evaluations and original grades, with lecturer 1 having the strongest correlation (PCC = 0.7832) and lecturer 3 having the weakest (PCC = 0.5617).

The sequence effect, where evaluating one assignment influences the perception of subsequent ones, may also be a factor, mainly if an early assignment is exceptionally strong or weak (Attali, 2011). Despite conducting reflections after each evaluation in the workshop to identify differences and minimize deviations, we observed no apparent learning effect among the lecturers. Ironically, although these reflection phases aim to improve evaluation consistency, we noted a paradoxical decrease in consistency from TA1 to TA4. The lecturers cited increasing fatigue and loss of concentration as the reason for the large discrepancy in scores.

The consistency of PCC values shows that lecturers' evaluations correlate positively with the original grades. The subjective evaluation methodology and the lecturers' subtle, individual evaluation styles, which are not immediately apparent in the analysis, may be causing the increase in MAD. Although the sample size was limited, the workshop provided insights into grading consistency across different lecturers. A higher level of grading consistency was observed specifically for TA1 to TA3. However, to determine whether TA4 is an outlier, a larger sample size would be necessary. The workshop highlighted that, despite applying

analytical rubrics, the challenge remains to maintain standardized assessment practices. This challenge is particularly true in an environment characterized by diverse teaching methods and the individual perspectives of lecturers. Confronting the complexities of evaluation, we sought state-of-the-art solutions to enhance evaluation consistency.

LLMs for Consistency Improvement

We chose GPT-4 for its multimodal capabilities to process text and image input without requiring programming skills to evaluate its effectiveness in evaluating transfer assignments. Three experiments were conducted in the German language using TA2.

First, we uploaded three documents to GPT-4, including the requirements for the Transfer Assignments in General Business Administration, the SIB Guide for Written Assignments, the TA2 in PDF format, and the analytical evaluation rubric. On this basis, GPT-4 was instructed to evaluate TA2, which was graded low by all lecturers and had a low MAD deviation of 5.33. On the first attempt, GPT-4 provided minimal feedback but did not fully meet the evaluation rubric, scoring the TA2 with 52 points, approximately 87% of the possible points.

In the second test, we replaced the Excel rubric with a PDF document detailing how to evaluate. GPT-4 graded TA2 with 50 out of 60 points. Again, the feedback for improvement indicated that the evaluation did not fully adhere to the provided assessment rubric.

On the third attempt, all the information necessary for evaluation was included in the prompt, followed by the text of the TA2 itself. GPT-4 graded the paper with the maximum possible points this time, giving feedback that TA2 fully met all evaluation aspects.

Based on our experiment's results, ChatGPT-4 may not be the most suitable tool for evaluating transfer assignments. The feedback and grading provided by GPT-4 were incorrect, suggesting that further investigation is needed to determine its effectiveness. The results differed significantly from the lecturers' evaluation, indicating that there may be some limitations to the current approach.

Future Ideas

While initial testing of GPT-4 has revealed its susceptibility to hallucinations, it is essential to investigate the potential applications of GPT-4 and other LLMs such as Google Bard, Google Gemini, and Meta's Llama 2. The focus should also be identifying prompts that can effectively assess the diverse aspects of transfer assignments, including interpreting and processing visual elements such as figures and tables.

More research is needed to establish the efficiency of different NLP techniques in assessing particular aspects and investigate alternative ways to generate feedback. This can involve analyzing the content of transfer assignments and incorporating previously evaluated transfer assignments into independent LLMs.

Another inquiry is the feasibility of developing a pipeline comprising multiple evaluation methods, with each transfer assignment undergoing these methods to address all relevant evaluation aspects. Investigating a range of evaluation methods for the six criteria and the 16 specific evaluation aspects can enhance the precision and granularity of the results. Transfer assignments would be subjected to various evaluation stages designed for distinct criteria or aspects to achieve a comprehensive assessment, much like traditional lecturer evaluations.

Moreover, it is essential to incorporate the practical experiences of lecturers who evaluate these transfer assignments into the system. Their expertise and insights should be integral to the evaluation process, implying that the expertise and insights of lecturers should be incorporated into training of LLMs. It is essential to distinguish which evaluation aspects are suitable for LLM assessment and which require alternative methods.

Conclusion

In conclusion, exploring automated essay evaluation systems, mainly using LLMs like ChatGPT-4, presents a transformative opportunity in academic assessment. This position paper has delved into the historical and current landscapes of AEE systems, highlighting their evolution from simple grading systems to more complex and nuanced approaches like NLP and LLMs. Through our case study at the SIB, we have demonstrated the potential and challenges of employing LLMs, like GPT-4, in assessing student transfer assignments written in German.

As evidenced by our "Working Alone, Together" method, there is a need for more standardized and objective assessment tools. While LLMs offer promising text processing and understanding capabilities, our experiments revealed their limitations in accurately evaluating complex academic texts according to predefined criteria. These findings highlight the gap between the current abilities of LLMs and the nuanced requirements of academic essay evaluation.

It is essential to continue research in this field, exploring various LLMs and their potential for more sophisticated and reliable essay evaluation. Developing an autonomous, tailored LLM, trained on a corpus of pre-assessed assignments, could pave the way for more accurate and consistent grading systems, streamline the assessment process, and enhance the educational experience by providing students with detailed, constructive feedback.

References

Attali, Y.; and Burstein, J. 2006. Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning, and Assessment* 4(3).

Attali, Y. 2011. Sequential Effects in Essay Ratings. *Educational and Psychological Measurement* 71(1): 68-79. doi.org/10.1177/0013164410387344.

Beigman Klebanov, B.; and Madnani, N. 2020. Automated Evaluation of Writing – 50 Years and Counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics. 10.18653/v1/2020.acl-main.697.

Bruns, H. C. 2013. Working Alone Together: Coordination in Collaboration Across Domains of Expertise. *Academy of Management Journal* 56(1): 62-83. doi.org/10.5465/amj.2010.0756.

Hayland, K. 2013. Writing in the university: education, knowledge and reputation. *Language Teaching* 46(1): 53-70. 10.1017/S0261444811000036

Hussein, M. A.; Hassan, H.; and Nassef, M. 2019. Automated language essay scoring systems: a literature review. *PeerJ Computer Science.* 5:e208. doi.org/ 10.7717/peerj-cs.208.

Ifenthaler, D. 2023. Automated essay grading systems. *In Handbook of open, distance and digital education,* edited by O. Zawacki-Richter and I. Jung, 1-15. Singapore: Springer.

Ke, Z.; and Ng, V. 2019. Automated Essay Scoring: A Survey of the State of the Art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.* International Joint Conferences on Artificial Intelligence Organization. doi.org/10.24963/ijcai.2019/879.

Masikisiki, B.; Marivate, V.; and Hlophe, Y. 2023. Investigating the Efficacy of Large Language Models in Reflective Assessment Methods through Chain of Thoughts Prompting. arXiv preprint. arXiv:2310.00272 [cs.CL]. Ithaca, NY: Cornell University Library.

Page, E. B. 1966. The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan* 47(5): 238-234.

Page, E. B. 1968. The Use of the Computer in Analyzing Student Essay. *International Review of Education* 14(2): 210-225.

Ramesh, D.; and Sanampudi, S. K. 2021. An automatic essay scoring system: a systematic literature review. *Artificial Intelligence Review* 55: 2495-2527. doi.org/10.1007/s10462-021-10068-2.

Strasser, A. 2023. On pitfalls (and advantages) of sophisticated large language models. arXiv preprint. arXiv: 2303.17511 [cs.CY]. Ithaca, NY: Cornell University Library.

Valenti, S.; Neri, F.; and Cucchiarelli, A. 2003. An Overview of Current Research on Automated Essay Grading. *Journal of International Technology Education* 2. doi.org/10.28945/331.

Williamson, D.; Xi, X.; and Breyer, F. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurements: Issues and Practice* 31(1): 2-13. 10.1111/j.1745-3992.2011.00223.x.

Wilson, J.; and Rodrigues, J. 2020. Classification accuracy and efficiency of writing screening using automated essay scoring. *Journal of School Psychology* 82: 123-140. doi.org/10.1016/j.jsp.2020.08.008.

Zupanc, K; and Bosnić, Z. 2015. Advances in the Field of Automated Essay Evaluation. *Informatica* 39(4): 383-395.

Zupanc, K.; and Bosnić, Z. 2017. Automated essay evaluation with semantic analysis. *Knowledge-Based Systems* 120: 118-132. doi.org/10.1016/j.knosys.2017.01.006.

Zupanc, K.; and Bosnić, Z. 2018. Increasing accuracy of automated essay grading by grouping similar graders. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics - WIMS 18*. New York: Association for Computing Machinery. doi.org/10.1145/3227609.3227645.