






EDUCATIONAL DATA MINING AND LEARNING ANALYTICS: TEXT GENERATORS USAGE EFFECT ON STUDENTS' GRADES

Birutė PLIUSKUVIENĖ ¹✉, Urtė RADVILAITĖ ¹, Rasa JUODAGALVYTĖ ¹,
Simona RAMANAUSKAITĖ ², Pavel STEFANOVIČ ¹

¹Department of Information Systems, Vilnius Gediminas Technical University, Vilnius, Lithuania

²Department of Information Technology, Vilnius Gediminas Technical University, Vilnius, Lithuania

Article History:

- received 12 April 2024
- accepted 27 May 2024

Abstract. Today, various types of data are constantly growing, so they can be used for different purposes. In this investigation, educational data has been analyzed to determine the influence of assessment on student knowledge. The newly collected dataset has been prepared and statistically analyzed. The dataset consists of open-question answers collected on one study subject during the midterm exam at Vilnius Gediminas Technical University. The results of the statistical analysis have shown that by using the text generators, students obtained higher grades by paraphrasing the answers to the questions in good quality. Furthermore, research has shown which types of questions are more difficult for students to answer without additional material and using text generation tools. It can be useful for lecturers planning course assessment tasks.

Keywords: educational data mining, learning analytics, statistical analysis, Lithuanian texts, open-questions dataset.

✉Corresponding author. E-mail: birute.pliuskuviene@vilniustech.lt

1. Introduction

Data mining is an essential task in numerous application areas, as it is used to extract interesting patterns or discover anomalies from various datasets. Its capabilities to uncover unknown but useful knowledge make it suitable to be used in such areas as banking, retail, medical, insurance, bioinformatics, etc. Different techniques such as statistics, database systems, machine learning, pattern recognition, visualization, information retrieval, etc., come handy when trying to expose patterns in data (Gupta & Chandra, 2020).

Currently, Machine Learning (ML) is one of the most widely explored fields, largely due to its versatile applicability to address a myriad of daily tasks and significant scientific challenges. In recent years, significant advances in technology and the accessibility of large datasets have contributed to notable progress in the field of ML. Depending on the type of data, the datasets can be further classified into text, image, and audio datasets. Text datasets are essential in Natural Language Processing (NLP) (Fanni et al., 2023) tasks such as sentiment analysis (Kapočiūtė-Dzikiėnė & Salimbajevs, 2022; Shaik et al., 2023; Mercha & Benbrahim, 2023), text classification (Štrimaitis et al., 2022; Palanivinayagam et al., 2023) and semantic analysis (Maulud et al., 2021). These main types of datasets are discussed by Gong et al. (2023). Provides a summary of these datasets, indicating the main types of application tasks, the amount of data, and the data content of the datasets. However, the main aim of this

paper is to provide a comprehensive review of research on dataset quality. Because the effectiveness of ML models greatly hinges on the quality of the dataset used to train and evaluate the models (Gong et al., 2023).

Despite of different types of data mentioned previously, there is a huge amount of information available in natural language in diverse domains. Thus, to process this information by a computer or machine in various applications, it needs to be in a structured format. Knowledge extraction refers to such a process when relevant information from the unstructured data is extracted and represented in a structured form (Nismi Mol & Santosh Kumar, 2023). Data mining and knowledge extraction together allow one to gain more insight for the fields in which they are applied. In recent years, data mining has become more and more applied in the analysis of educational data.

Therefore, it has become a developing research area where researchers focus on using data mining techniques on data collected from an educational environment (Rao & Chen, 2024). Therefore, data mining in education or simpler educational data mining (EDM) usually comes together with another research field, learning analytics (LA). Both fields focus on exploring data to improve learning processes (Baek & Doleck, 2023). Educational Data Mining (EDM) and Learning Analytics (LA) are interdisciplinary fields that draw on information retrieval, recommendation systems, visual data analytics, and more. They represent the intersection of computer science, education, and statistics. This overlap also gives rise to related subfields like computer-based education (CBE), data mining and machine learning, and educational statistics (Romero & Ventura, 2020).

The latest area in EDM is the analysis of students' unfair behavior, where answers are generated by chatbots rather than written by students (Bouaine et al., 2023; Stefanovič et al., 2024). In the field of LA, little attention is paid to estimating how chatbot use affects student evaluation. Therefore, this research aims to investigate the effect of ChatGPT usage on student grades for open question answers written in the Lithuanian language. To achieve this objective, a statistical analysis of the newly collected educational data has been performed. The dataset consists of the students' open-question answers written in the Lithuanian language and the grades for each answer provided by the lecturer. The dataset was prepared in such a way that original and text generator-based answers were collected. The analysis of this research has shown how text generators affect the student's results. Furthermore, research has shown which types of questions are easy to answer for students, no matter whether original or text-generated student answers have been provided. The results are useful for lecturers, as well as to improve the quality of assessment tasks in their courses.

The manuscript is organized as follows. In Section 2, the related works have been reviewed. Section 3 describes the dataset analyzed. The statistical research of the chosen dataset is performed in Section 4. Section 5 concludes the manuscript. The discussion is presented.

2. Related works

2.1. Research in the fields of educational data mining and learning analytics

Educational Data Mining (EDM) and Learning Analytics (LA) are not new topics. The more active article publishing in this field was started in the early 1990s. Although sometimes both terms are presented in the same research, some papers are concentrated just on some

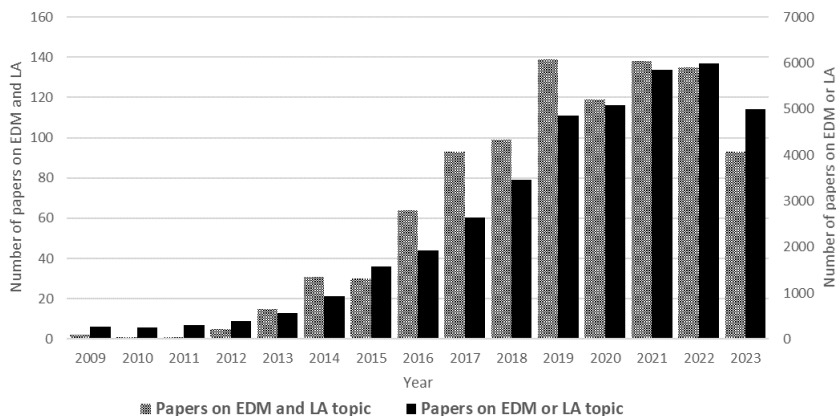


Figure 1. Change in the number of articles published on the Web of Science, search phrase “educational data mining AND learning analytics”

specific aspect of those two. The number of papers published on the Web of Science platform during the last 15 years has grown (see Figure 1). Very similar tendencies are visible for cases where both topics are covered, and at least one of them is covered. The absolute numbers of these two cases are different at different times, whereas the correlation of 0.96 between the data indicates similarities in the trends.

The EDM focuses on creating techniques to analyze unique data from educational environments, essentially applying data mining methods to address significant educational questions (Romero & Ventura, 2020). The recent surge in accessible learning data has increased the importance of EDM in enhancing comprehension and optimizing the learning process and the environments in which it occurs. EDM has emerged as a powerful instrument for uncovering hidden correlations within educational data and forecasting student academic achievements. Modeling the performance of students presents itself as a formidable and widely explored subject within the realm of educational data mining (Khan & Ghosh, 2021).

An overview of the researches carried out between 2010 and 2020 is provided by Namoun and Alshanqiti (2020). The research in this paper describes how a total of 62 relevant papers have been analyzed investigating the prediction of learning outcomes and factors that impact student outcomes (Namoun & Alshanqiti, 2020). Another research by Yağcı (2022) introduces a new model based on machine learning algorithms to anticipate the final exam scores of undergraduate students, utilizing their midterm exam grades as the primary dataset. Based on the results of this research, it can be inferred that the midterm exam scores of the students serve as a significant predictor to forecast their final exam grades (Yağcı, 2022). Khan and Ghosh (2021) presented a review of existing EDM literature on analysis and prediction of student performance. This paper presents a review of 140 relevant studies published between 2000 and 2018. This methodical survey would help EDM researchers progress in the field of predicting grades for the next term (Khan & Ghosh, 2021). Educational data mining is also used in a study by Hasan et al. (2020), which predicts overall performance at the end of the semester using video learning analytics and data mining techniques.

In research conducted by Baek and Doleck (2023), EDM together with LA has been reviewed. The authors systemized 492 LA and 194 EDM articles from Web of Science (WoS) that have been published between 2015 and 2019. Research focused on similarities and differences between these two fields, showing that EDM and LA usually come together (Baek & Doleck, 2023). Rao and Chen (2024) have done a review of the literature retrieved from the Scopus database. However, the author only analyzed data mining in education or educational data mining.

Learning analytics uses data, statistical analysis, and predictive models to understand complex issues and enhance the student learning experience (Hasan et al., 2020). By analyzing student data and activities, higher education institutions can ensure institutional success, retain a diverse student population, and improve resource management based on student success. The field of LA has experienced rapid growth over the past decade. However, the implementation of LA is predominantly limited in scope and often isolated at the instructor level (Tsai et al., 2020). In the research by Tsai et al. (2020), the authors present an exploratory study on institutional approaches to LA in European higher education and discuss the prominent challenges that impede LA from reaching its potential.

Similarly, Márquez et al. (2024) performed the systematic review of LA in higher education institutions. The authors described 14 factors that were identified in their research on the adoption of LA by higher education institutions.

Romero and Ventura (2020) have described 16 of the most popular EDM/LA techniques. Most of these techniques are recognized as universally applicable across various data mining domains, including the following.

- visualization (creating visual representations of data to effectively convey the findings of EDM/LA research to educators);
- clustering (grouping of comparable materials or students according to their learning and interaction behaviors);
- outlier detection (identifying students experiencing challenges or deviations in their learning processes);
- causal mining (discovering which aspects of students' behavior contribute to learning, academic failure, dropout rates, and similar outcomes);
- statistics (analyzing, interpreting, and drawing conclusions from educational data);
- text mining (analyzing the content of documents, chats, and web pages) (Romero & Ventura, 2020).

Gupta and Chandra (2020) performed a similar overview of data mining techniques, applications, and tasks as Gupta and Chandra (2020). The authors systemized the literature on data mining and structured the information that provides the relationship between data mining tasks and data mining techniques, as well as real-life data mining applications (Gupta & Chandra, 2020).

Despite the application areas, data extraction techniques help to extract a huge amount of data. Improvement in technology and the rise of artificial intelligence impact the efficiency of these techniques and enable more accurate knowledge extraction (Nismi Mol & Santosh Kumar, 2023). Most NLP applications in education focus on the automated assessment of essays and open-ended questions. Various methods to improve such evaluations through text mining are suggested in the scholarly literature (Ferreira-Mello et al., 2019).

2.2. Problematics of generative artificial intelligence in student evaluation

Generative Artificial Intelligence (AI), such as ChatGPT, has opened up new possibilities and challenges in traditional education, impacting learning outcomes, teaching methods, and evaluations. The research by Chiu (2024) employs a preliminary conceptual model, based on a comprehensive review of the literature, to examine the potential benefits and difficulties of implementing AI in education. ChatGPT, one of the newest publicly available machine learning solutions, has attracted attention over the past year for its advantages and disadvantages. In particular, in education, a major concern has emerged, as many students have begun to use ChatGPT to address various academic tasks, such as writing essays, answering questions, or completing exams. Often, students do not disclose their use of ChatGPT, leading to concerns about plagiarism. Various studies are being conducted on the use of ChatGPT in the learning process. For example, Baidoo-anu and Owusu Ansah (2023) conducted a study to present the potential advantages and disadvantages of ChatGPT in facilitating teaching and learning. In this context, the question of plagiarism is often investigated. For example, this includes research related to plagiarism detection in academic writing (Jarrah et al., 2023), students answer to open-ended questions (Stefanovič et al., 2024), or student program coding (Hoq et al., 2024). ChatGPT complicates traditional plagiarism detection as user queries affect text generation. Novel solutions are needed to determine whether students wrote the texts themselves or used generators. Although there are some studies performed for plagiarism detection for the English language (Khaled & Al-Tamimi, 2021) or for cross-languages (Bouaine et al., 2023), few exist for the Lithuanian language (Stefanovič et al., 2024). Although more studies are performed for text mining in most spoken languages such as English, less popular languages are not analyzed as often. Further analysis is needed to establish modern solutions based on machine learning to determine similarity. At the same time, it is important to understand how ChatGPT usage affects the students' evaluation. Alneyadi and Wardat (2023) analyzed how ChatGPT affects student achievement in the electronic magnetism unit for 11th grade students in Emirates school. Research on different study areas, national, language, and other study-specific characteristics is not yet investigated. Therefore, research on different conditions, including the usage of the Lithuanian language, would help the Lithuanian education system.

3. Educational context data

In this research, a statistical analysis of educational context data has been performed (Kaggle, n.d.). The published dataset contained only the type of answer (original, generated, rephrased version of the generated answer) and the response text. In this investigation, the grade for each question was estimated. During the first stage, 118 students participated and received scores for the 5 questions received, resulting in 590 scores for the answers in total. In the second stage, 53 students participated. They provided fully generated answers and rephrased their version. They provided 263 answers for each type. Completely generated answers were eliminated from this research, while 263 rephrased answers scores were used to execute the research and compare with the original answers. The dataset has been collected at Vilnius Gediminas Technical University (VILNIUS TECH) during the Fundamentals of Data Mining

course in two stages. In this course, one of the assessment tasks is a midterm exam. In the first stage, students without additional material had to answer five open questions and solve a few practical tasks. During the midterm exam, it was ensured that the students did not cheat: did not use any computer tools, provided the answers in writing on the distributed settlement sheets. The settlement was monitored by multiple teachers. In this research, only open questions have been analyzed. The lecturer prepared a total of 15 questions for this course, where 6 of the questions are worth 3 points, and the rest 9 questions – 4 points. Points were assigned based on the complexity of the questions – standard questions were worth 3 points and more difficult ones – 4 points. Each student was randomly given two questions worth 3 points and three questions worth 4 points. In the further analysis, a relative score was measured for each question to eliminate the impact of the maximum answer mark. The midterm exam questions are presented in Figure 2a and Figure 2b.

At this stage, 118 students attended the midterm exam. The responses of the students were collected and evaluated manually by the lecturer of the course, and the label *Original answers* (standing for original students' answers to the midterm exam) in the dataset was assigned to each answer. In this case, a total of 590 original responses written by the students have been obtained.

In the second stage, the same students had the possibility to participate in additional assessments, where the main idea was to answer the same questions using the ChatGPT text generators provided as a base. The students have been asked to save the original text given by ChatGPT 3.5 because it is free. There were no instructions given to students on how to

ID	Questions
Q1	Savais žodžiais paaiškinkite kas yra dirbtinis intelektas. Pateikite du konkrečius dirbtinio intelekto taikymo pavyzdžius.
Q2	Kokie klasifikavimo modelio vertinimo matai taikomi įvertinti modelio kokybę? Pateikite bent tris matavimus ir ką jie parodo.
Q3	Kokiomis sąlygomis pasižymi didieji duomenys. Išvardinkite bent tris (angl. <i>big data</i>)?
Q4	Kas yra mokymo, testavimo ir validavimo duomenų aibės ir kokia jų paskirtis?
Q5	Savais žodžiais paaiškinkite kas yra populiacija, imtis. Pateikdami pavyzdį paaiškinkite koks esminis skirtumas tarp jų.
Q6	Kuo pasižymi subalansuoti ir nesubalansuoti duomenys? Kaip tai įtakoja apmokamą algoritmą?
Q7	Kuo skiriasi terminai mokymas be mokytojo ir su mokytoju? Apibūdinkite kiekvieną iš jų.
Q8	Kokia yra angl. <i>t-testo</i> paskirtis ir ką jo rezultatas mums parodo? Ką parodo angl. <i>p-value</i> reikšmė? Pateikite gyvenimišką pavyzdį.
Q9	Iš ko susideda dirbtinis neuroninis tinklas bei pakomentuokite jo veikimo principą?
Q10	Kuo skiriasi klasifikavimo nuo klasterizavimo uždaviniai? Pateikite kiekvieno uždavinio pavyzdį.
Q11	Ką statistikoje parodo duomenų asimetrijos ir eksceso koeficientai? Ką reiškia neigiamos ir teigiamos jų reikšmės?
Q12	Kam skirta kryžminė patikra (angl. <i>cross-validation</i>)? Kokie trys kryžminės patikros būdai yra dažniausiai naudojami ir koks jų veikimas?
Q13	Kokius duomenų kintamuosius vadiname kokybiniais ir kiekybiniais? Pateikite realių tokių kintamųjų pavyzdžius.
Q14	Savais žodžiais paaiškinkite kas yra duomenys ir informacija. Esminį skirtumą paaiškinkite pateikdami po pavyzdį.
Q15	Kas yra klasifikavimo matrica? Pateikite jos pavyzdį ir apskaičiuokite pateiktos matricos svorinį tikslumą.

ID	Questions
Q1	In your own words, explain what artificial intelligence is. Give two specific examples of the application of artificial intelligence.
Q2	What evaluation metrics of classification models are used to measure the quality of the models? Give at least three evaluation metrics and explain what they show.
Q3	What is the difference between supervised learning and unsupervised learning? Describe each of them.
Q4	What are the characteristics of big data? Name at least three.
Q5	What is training, testing, and validation datasets? What is their purpose of usage?
Q6	What is the difference between classification and clustering tasks? Give an example of each task.
Q7	In your own words, explain what is a population and a sample in a statistic. Explain the essential difference between them by giving an example.
Q8	What are the characteristics of balanced and unbalanced data? How does this affect the results of trained algorithms?
Q9	What data variables do we call qualitative and quantitative? Give examples of such kinds of variables.
Q10	What is the purpose of the t-test and what does its result tell us? What does the p-value indicate? Give a real-life example.
Q11	What does an artificial neural network consist of and give the comment on its principle of operation?
Q12	What do skewness and kurtosis coefficients show in descriptive statistics? What do their negative and positive values show?
Q13	What is cross-validation used for? What are the three most common types of cross-validation used in practice and how do they work?
Q14	In your own words, explain what data and information are. Explain the essential difference with an example.
Q15	What is a classification matrix? Give an example of it and calculate the weighted accuracy of the given classification matrix.

Figure 2. The midterm exam questions: a – in Lithuanian; b – in English

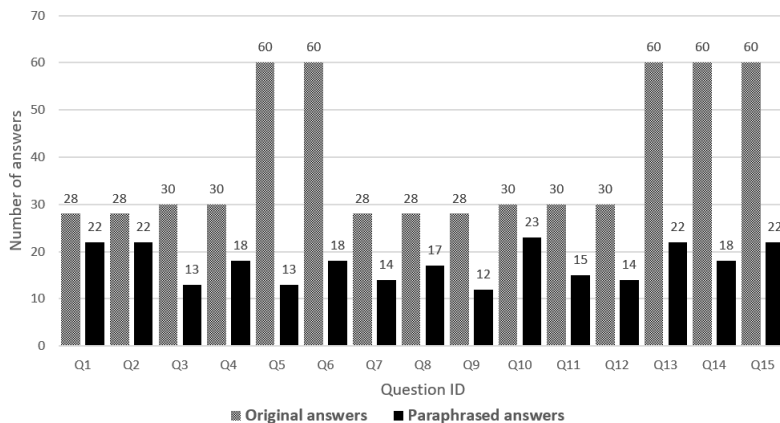


Figure 3. Distribution of the answers to given questions

write prompts. Students could choose the language of prompts, but they had to submit the received answers in Lithuanian and also paraphrase them on their own. All the paraphrased answers have been collected and evaluated by the same lecturer, and the answers have been labeled as paraphrased *answers* (standing for answers from the midterm exam that were generated by a text generator and paraphrased by the students) in the dataset. A total of 263 responses have been obtained in stage 2.

The distribution of the answers to the questions given by the type of answer provided (original and paraphrased) is presented in Figure 3. There were three different cases of the test, where each student was given five questions from the list of 15 possible questions. We can see that in the case of the original answered questions, the highest number (60 answers) of answers has been collected from five questions (ID: 5, 6, 13, 14, 15). The rest of the questions were presented to approximately 30 students.

In the case of the Paraphrased answers, the highest number of answers to the provided questions is equal to 22. This smaller number indicates that not all students participated in the second stage.

4. Comparison of grades for original and paraphrased answers

To highlight the effect of text generator usage and the effect on open question answers grades, a comparative analysis of original and paraphrased answer grades was performed. The grades were expressed as a percentage to reflect the different marking scales (3 points and 4 points) for each question. Taking into account the different number of answers in each type of answer, the relative histogram was analyzed to see the mark distributions (see Figure 4). It illustrates that the grades do not follow a normal distribution (it is leaning toward exponential distribution), and the highest percentage of answers received maximum grade for a question (32% of answers for original answers and 76% for paraphrased answers). This distribution does not reflect the overall midterm complexity, as the second part (55% of possible points) was dedicated to more practical tasks. The open questions reflect the easier part of the midterm test. Therefore, the grade distribution is skewed towards higher grades.

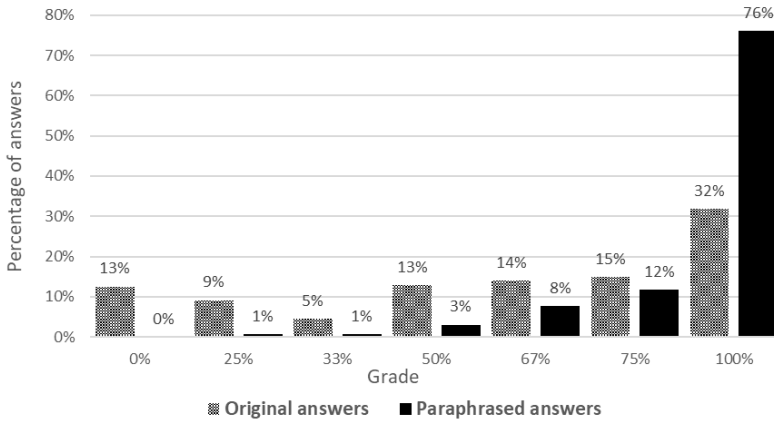


Figure 4. The distribution of the relative grades according to the type of answer – is it a human answer or a rephrased ChatGPT answer

The analysis of original and paraphrased answers indicates very uneven grades for these two types of answers. The average score for the original answer is 63% (standard deviation equal to 0.34), while for the paraphrased answers, 92% (standard deviation – 0.16). The box plot in Figure 5 illustrates that all the grades in the paraphrased answers get the highest grade. Taking into account the grade distributions, the lower grades are interpreted as outliers. This fact illustrates the positive effect of text-generating solutions on higher student marks for open questions. The two cases of students grade according to *the distribution* are statistically significant (the p -value is less than 0.0001).

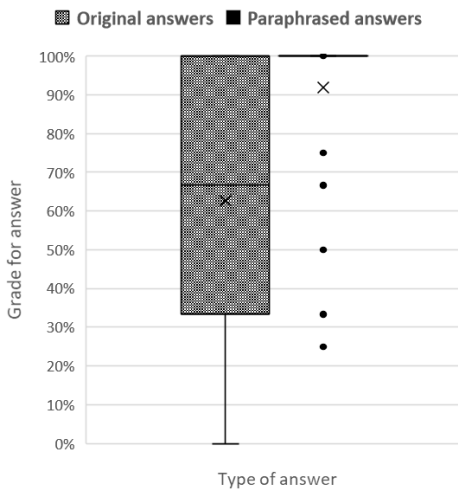


Figure 5. The distribution of the grades according to the type of answer

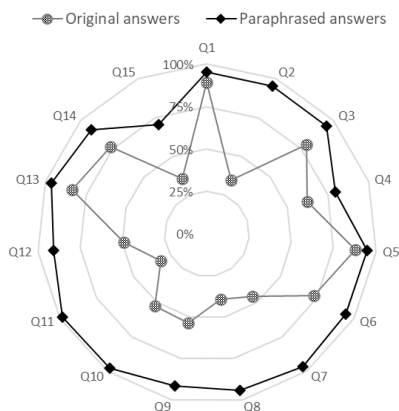


Figure 6. The distribution of the type of grades based on the answer per question

Although the overall grade illustrates the statistically significant difference between the original and paraphrased answer grades, the situation varies for different questions. Question 1 and Question 5 do not have statistically significant differences between grades (p -values are 0.32 and 0.27 respectively). This is affected by the fact that those questions were the easiest for students and the original answer grades average for those to about 90%.

The radar diagram in Figure 6 confirms that the average grade for each question was lower for the original answers. Meanwhile, the difference between individual questions varies and is minimal for Q1 and Q5. The largest difference between the original and paraphrased answers was monitored for Q2, Q8, Q11, and Q15. These four questions were the hardest for the students, and the average grade for each of them did not reach 40%. Those questions are similar in topic as require understanding of some metrics, statistics.

Analyzing the grade distribution for individual questions (see Figure 7) additional insights can be spotted:

- The most balanced questions for the original answers were Q7, Q9, Q10, and Q12. The quartiles of these answer grades are the most even.
- The most difficult question for text generators was Q15. In most cases, the grade for the Q15 paraphrased answer was 75%, while the other cases are assumed to be outliers in this case. The reason behind it is the need to generate a confusion matrix and explain its results specifically, not provide a definition of it.
- Questions Q4, Q12 and Q14 are those, who had the biggest grade distribution for the paraphrased answers. While in other questions the lower grades are statistically assumed as outliers, for those 3 questions, the proportions of lower grades are higher and reflects in a wider grade-quartile distribution.
- The easiest questions for the students were Q1, Q3, Q5, Q6, Q13, and Q14. There are no zero grades for questions that analyze the original answers or those that are extremely rare and are assumed as extremums.

The most stable grades for the original answers were for Q4. From a statistical point of view, all students were able to explain two out of three terms.

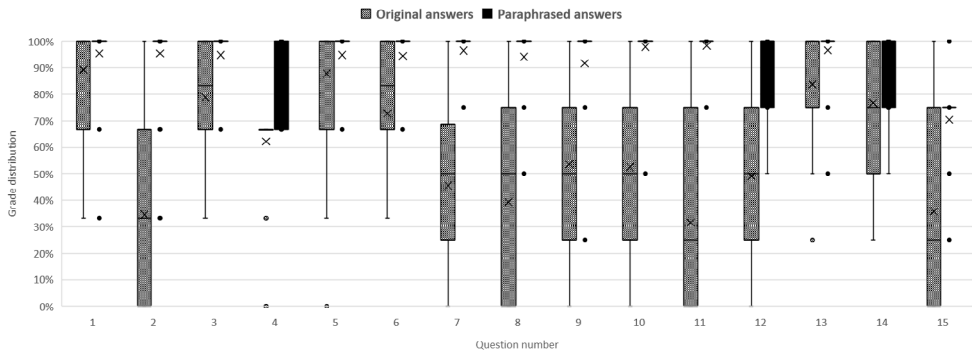


Figure 7. The distribution of the grades according to the type of answer

5. Discussions and conclusions

The analysis of related works indicates that interest in EDM and LA is increasing among researchers and the number of papers on these topics has the potential to grow. Growth may be affected by new possibilities and challenges facing the educational system. The increased availability of generative artificial intelligence solutions, such as ChatGPT, changed the study process. Some research papers have already been published to develop unfair behavior, using ChatGPT for student knowledge evaluation tasks. However, more attention could be paid to estimate the effect these technologies have, in general, on student grades. Therefore, this research contributes in this field, highlighting the possible variations in students' grades in case of fair answers to open questions and when ChatGPT was utilized helping the student answer the questions.

In the research, conducted at VILNIUS TECH, the students answered open questions individually and then used publicly available systems to answer their questions. This reflects the situation of most common cases, when no special resources are needed to get a fast answer to the course related question.

The students did not have dedicated training in prompt generation. They used their personal experience. This usually led to a long answer with multiple details and repetitions. The answer length is one of factors, distinguishing the large language model generated answer from original or rephrased answers. In those cases, generated text paraphrasing was needed to shorten the text or select just the most relevant one. The paraphrased answer grades were compared to original answer grades in this research to reflect a more realistic situation when students adapt and paraphrase the answer (or at least select just part of the generated text) to mimic independent understanding of the question.

The results of the comparative analysis indicate that the use of text generation tools has a significant effect on student grades while answering open questions on the midterm exam. The percentage of maximum grades increases more than twice when students are allowed to use text generation tools to answer questions.

This research was unable to identify the main factors that affect the lower performance of text generation tools, as only one of the questions had a lower than 75% average grade

for the answers. We can guess that it was mostly related to the fact that it required not only to provide an example but also to analyze the data of the provided example. To estimate the limitations of text generation tools, a wider variety of questions should be tested.

In the results, it is evident that the use of text generation tools significantly affects the grades of students. Therefore, to ensure the knowledge of students more accurately on a specific topic, high attention must be paid to the prevention of text generation tools.

References

- Alneyadi, S., & Wardat, Y. (2023). ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools. *Contemporary Educational Technology*, 15(4), Article ep448. <https://doi.org/10.30935/cedtech/13417>
- Baek, C., & Doleck, T. (2023). Educational data mining versus learning analytics: A review of publications from 2015 to 2019. *Interactive Learning Environments*, 31(6), 3828–3850. <https://doi.org/10.1080/10494820.2021.1943689>
- Baidoo-anu, D., & Owusu Ansah, L. (2023). Education in the era of generative Artificial Intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52–62. <https://doi.org/10.61969/jai.1337500>
- Bouaine, C., Benabbou, F., & Sadgali, I. (2023). Word embedding for high performance cross-language plagiarism detection techniques. *International Journal of Interactive Mobile Technologies*, 17(10). <https://doi.org/10.3991/ijim.v17i10.38891>
- Chiu, T. K. (2024). Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence*, 6, Article 100197. <https://doi.org/10.1016/j.caeai.2023.100197>
- Fanni, S. C., Febi, M., Aghakhanyan, G., & Neri, E. (2023). Natural language processing. In *Introduction to Artificial Intelligence* (pp. 87–99). Springer International Publishing. https://doi.org/10.1007/978-3-031-25928-9_5
- Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., & Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6), Article e1332. <https://doi.org/10.1002/widm.1332>
- Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162, Article 107268. <https://doi.org/10.1016/j.infsof.2023.107268>
- Gupta, M. K., & Chandra, P. (2020). A comprehensive survey of data mining. *International Journal of Information Technology*, 12(4), 1243–1257. <https://doi.org/10.1007/s41870-020-00427-7>
- Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences*, 10(11), Article 3894. <https://doi.org/10.3390/app10113894>
- Hoq, M., Shi, Y., Leinonen, J., Babalola, D., Lynch, C., Price, T., & Akram, B. (2024). Detecting ChatGPT-generated code submissions in a CS1 course using machine learning models. In *SIGCSE 2024: Proceedings of the 55th ACM Technical Symposium on Computer Science Education* (pp. 526–532). <https://doi.org/10.1145/3626252.3630826>
- Jarrah, A. M., Wardat, Y., & Fidalgo, P. (2023). Using ChatGPT in academic writing is (not) a form of plagiarism: What does the literature say. *Online Journal of Communication and Media Technologies*, 13(4), Article e202346. <https://doi.org/10.30935/ojcm/13572>
- Kaggle. (n.d.). *Students and chatGPT answers in Lithuanian*. Retrieved April 6, 2024, from <https://www.kaggle.com/datasets/pavelstefanovi/students-and-ChatGPT-answers-in-lithuanian/>
- Kapočiūtė-Dzikiėnė, J., & Salimbajevs, A. (2022). Comparison of deep learning approaches for Lithuanian sentiment analysis. *Baltic Journal of Modern Computing*, 10(3), 283–294. <https://doi.org/10.22364/bjmc.2022.10.3.02>

- Khaled, F., & Al-Tamimi, M. S. H. (2021). Plagiarism detection methods and tools: An overview. *Iraqi Journal of Science*, 62(8), 2771–2783. <https://doi.org/10.24996/ij.s.2021.62.8.30>
- Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and Information Technologies*, 26(1), 205–240. <https://doi.org/10.1007/s10639-020-10230-3>
- Márquez, L., Henríquez, V., Chevreux, H., Scheihing, E., & Guerra, J. (2024). Adoption of learning analytics in higher education institutions: A systematic literature review. *British Journal of Educational Technology*, 55(2), 439–459. <https://doi.org/10.1111/bjet.13385>
- Maulud, D. H., Zeebaree, S. R., Jacksi, K., Sadeeq, M. A. M., & Sharif, K. H. (2021). State of art for semantic analysis of natural language processing. *Qubahan Academic Journal*, 1(2), 21–28. <https://doi.org/10.48161/qaj.v1n2a44>
- Mercha, E. M., & Benbrahim, H. (2023). Machine learning and deep learning for sentiment analysis across languages: A survey. *Neurocomputing*, 531, 195–216. <https://doi.org/10.1016/j.neucom.2023.02.015>
- Namoun, A., & Alshantqi, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), Article 237. <https://doi.org/10.3390/app11010237>
- Nismi Mol, E. A., & Santosh Kumar, M. B. (2023). Review on knowledge extraction from text and scope in agriculture domain. *Artificial Intelligence Review*, 56(5), 4403–4445. <https://doi.org/10.1007/s10462-022-10239-9>
- Palanivinaiyagam, A., El-Bayeh, C. Z., & Damaševičius, R. (2023). Twenty years of machine-learning-based text classification: A systematic review. *Algorithms*, 16(5), Article 236. <https://doi.org/10.3390/a16050236>
- Rao, Y. S. N., & Chen, C. J. (2024). Bibliometric insights into data mining in education research: A decade in review. *Contemporary Educational Technology*, 16(2), Article ep502. <https://doi.org/10.30935/cedtech/14333>
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), Article e1355. <https://doi.org/10.1002/widm.1355>
- Shaik, T., Tao, X., Dann, C., Xie, H., Li, Y., & Galligan, L. (2023). Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, 2, Article 100003. <https://doi.org/10.1016/j.nlp.2022.100003>
- Stefanovič, P., Pliuskvienė, B., Radvilaitė, U., & Ramanauskaitė, S. (2024). Machine learning model for ChatGPT usage detection in students' answers to open-ended questions: Case of Lithuanian language. *Education and Information Technologies*. <https://doi.org/10.1007/s10639-024-12589-z>
- Štrimaitis, R., Stefanovič, P., Ramanauskaitė, S., & Slotkienė, A. (2022). A combined approach for multi-label text data classification. *Computational Intelligence and Neuroscience*, 2022, Article 3369703. <https://doi.org/10.1155/2022/3369703>
- Tsai, Y. S., Rates, D., Moreno-Marcos, P. M., Muñoz-Merino, P. J., Jivet, I., Scheffel, M., Drachsler, H., Delgado Kloos, C., & Gašević, D. (2020). Learning analytics in European higher education – Trends and barriers. *Computers & Education*, 155, Article 103933. <https://doi.org/10.1016/j.compedu.2020.103933>
- Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(1), Article 11. <https://doi.org/10.1186/s40561-022-00192-z>