**Student Assessment Application Development Processes Using LLM Models**

Cezary Orłowski, Marek Ożarowski, Tomasz Wawer

**Abstract**

The article describes the development processes of an application for evaluating student work. The application is designed to support the teaching process in determining students' independent preparation of answers to a question using ChatGPT. The application's support of the didactic process consists in examining, with the help of appropriately selected metrics, the degree of students' independence in their completed works. This means that the lecturer will be able to check whether a given work done by a student using ChatGPT was prepared independently or was prepared only by ChatGPT.

In order to achieve the purpose of the work, it has been divided into the following parts. The introduction presents the state of research in the field of building applications to support teaching processes. Then the research gap in this area was shown, so that in the second chapter a model of application construction processes for supporting student evaluations was presented. This part of the work presents the assumptions for building the model, the research environment and the assessment metrics used.

The next part of the article presents the results of the developed application, in which the evaluation of student work was determined on the basis of selected metrics. The article concludes with a discussion of the significance of the proposed solution and conclusions. The authors' directions for further research are also presented.

***Keywords***: Education, Generative Artificial Intelligence, ChatGPT, Teaching and Learning.

### Introduction

In 2022, 35% of all global businesses were using AI (Artificial Intelligence) based systems. As the trends in the statistics show, in 2025 already more than 80% plan to automate their work and combine AI and automate their tasks. ChatGPT is currently one of the most popular tools for finding information. On its launch day, November 30, 2022, ChatGPT had 153,000 visitors (chat.openai.com). At the end of the first week, the number reached 15.5 million, and in the second week it rose to 58 million. From the beginning, ChatGPT continued on a hyper growth path until May 2023, when it reached its highest level to date - 1.8 billion visits per month. In the process, parent company OpenAI's openai.com website became one of the most visited domains in the world [20].

If the degree of use of ChatGPT continues at this rate, education will have to make big changes in teaching processes. These changes are necessary for education to meet the expectations of employers. Employers would like to be able to hire graduates equipped with the skills to use tools such as ChatGPT.

Education, teaching processes must also meet the expectations of the economy. According to research conducted by the article's authors, currently every student admits to using ChatGPT. In addition, business and the economy are looking for “prompt engineering” specialists and Academic Centers must adjust their course curricula to meet these expectations.

In order for such expectations to be realized practically, the teaching process must enable the use of tools such as ChatGPT. The use of the opportunities offered by ChatGPT in education must therefore be a defined and controlled process. At the same time, the approach in the teaching process itself must be changed. If students are to learn to use ChatGPT in an informed way, the teaching process itself must be rebuilt. This means that lecturers should also use tools such as ChatGPT. If students use such solutions to write evaluation papers, then a problem arises. How to check whether a written evaluation paper is a paper independently written by the student. In other words, it should be resolved whether the student has used ChatGPT-generated texts during the writing of the paper in a way that is considered plagiarism.

The authors of the article found that currently every student uses ChatGPT (self-reported surveys in practical classes with students). The way ChatGPT is used varies. The most common activities with ChatGPT are done in teaching classes and used for writing papers in classes, exercises and labs held at WSB University. A similar degree of use of this tool by students is noted in other academic centers. It seems that the number of daily sessions with ChatGPT is increasing.

Another example of the use of this tool is in evaluation papers. So, can a student use ChatGPT and under what conditions in the process of writing an evaluation paper. Employers and business expect this type of skill from future university graduates. Therefore, evaluation papers written using ChatGPT should also be allowed by academic centers. The way and ability to use ChatGPT should be allowed in a controlled manner. To ensure that the degree of ChatGPT use in evaluation papers written by students is controlled, attempts have been made to build appropriate tools.

## Literature Review

Villagrán, I., Hernández, R., Schuit, G., Neyem, A., Fuentes-Cimma, J., Miranda, C., and Varas, J. (2024) presented a case study using a large language model (LLM) in the field of physical therapy education. The described case demonstrates the effective use of generative AI in a practical training for skills training in health professions education. The proposed solution uses LLM to automatically evaluate feedback provided by instructors based on predefined and literature-based quality criteria and generates practical textual explanations for reformulation. Moreover, if the instructor requires it, the tool supports the generation of summaries for large text data sets to achieve better student reception and understanding. The case study discussed in the article describes how these features were integrated into a feedback-oriented platform. How their effectiveness was assessed in a controlled setting with documented feedback. It presents the results of their implementation with real users through cognitive walkthroughs. Initial results indicate that this innovative implementation has great potential. It improves learning and performance in physical therapy education and has the potential to expand to other medical disciplines where the development of procedural skills is crucial. The described case offers a valuable tool for evaluating and improving feedback based on quality standards for effective feedback processes [1].

Meissner, R., Pögelt, A., Ihsberner, K., Grüttmüller, M., Tornack, S., Thor, A., ... & Hardt, W. (2024, October) described a new LLM-driven process and application, called ItemForge. The application was tailored specifically for the automatic generation of e-assessment items in mathematics. The approach is fully adapted to the levels and hierarchies of cognitive learning goals developed by Anderson and Krathwohl. The application takes into account specific mathematical concepts from the considered courses. The quality of the generated free-text items with their corresponding answers (sample solutions) as well as their adequacy to the designated cognitive level and subject were assessed in a small-scale study. The study included three mathematical experiments, a total of 240 generated items were analyzed. The findings show that the tool is adept at producing high-quality items that are aligned with the selected concepts and target cognitive levels, indicating its potential usefulness for educational purposes. It was observed that the provided answers (sample solutions) were sometimes inaccurate or not completely complete, which signals the need for further refinement of the tool processes [2].

Smolić, E., Pavelić, M., Boras, B., Mekterović, I., & Jagušt, T. (2024, May) used a large language model (LLM) to extract information and/or extrapolate from various sources. In computer science education, a potential application of such technology is automated code review. LLM could then take over the burden of debugging non-compiling code. Such support provided by LLM would consist in detecting overlooked optimization problems such as poor memory management in code, where it would otherwise pass automated testing. LLM could also perform other advanced tasks during code writing. However, currently LLMs are not able to evaluate code or mathematical expressions with 100% reliability. LLM can support token pattern recognition and subsequent generation of probabilistic answers. With this in mind, the risk of incorrect code evaluation by LLM, both descriptive and numerical, was investigated. The paper proposes research on mitigating this risk and suggests further directions. [3]

Joy Kulangara, K. (2024) discusses the potential of LLMs in programming education. It is well-known that large language models (LLMs) have the potential to improve programming education by providing feedback and guidance to students. Despite the potential benefits, integrating LLMs into education poses unique challenges, including the risk of overreliance on their feedback and inconsistent feedback quality. This work presents a flexible platform that can integrate multiple LLMs while providing an experimental space for research. It enables innovative approaches to improve programming education using LLMs. The developed platform effectively demonstrates the feasibility of integrating LLMs into programming education. A small-scale study assessing the overall usability of the platform received an average usability score of 4.21 out of 5.00, while the feedback from LLM received an average usability score of 4.28 out of 5.00, highlighting its effectiveness and value in helping students. Although the sample size of the study was small, the results are encouraging. [4]

Jošt, G., Taneski, V., & Karakatič, S. (2024) The article explores the nuances of the impact of informal LLM use on the learning outcomes of undergraduate students in software development education. An experiment was designed with thirty-two participants over ten weeks to investigate the unrestricted but not specifically encouraged use of LLMs and their correlation with student outcomes. The results reveal a significant negative correlation between increased reliance on LLMs for tasks requiring intensive critical thinking, such as code generation and debugging, and lower final grades. Furthermore, the described experiment observes a downward trend in final grades with increased average use of LLMs across all tasks. However, the correlation between the use of LLMs for seeking additional explanations and final grades was not as strong, indicating that LLMs may serve better as a complementary learning tool. These results emphasize the importance of combining LLM studies with the development of independent problem-solving skills in programming education. [5]

Hu, B., Zheng, L., Zhu, J., Ding, L., Wang, Y., & Gu, X. (2024) This paper describes the study and analysis of the results of large language models (LLM) in teaching design. It aims to reveal their potential strengths and weaknesses. The influence of LLM has gradually increased in many fields, but exploratory studies on its application in education remain relatively rare. Using Generative Pretrained Transformer 4, a dataset of secondary school mathematics teaching plan was generated. Finally, the performance of LLM in teaching design was evaluated. The evaluation results showed that the teaching plans generated by LLM are outstanding in setting teaching objectives, identifying teaching priorities, organizing problem chains and teaching activities, articulating subject content, and selecting methods and strategies. Particularly commendable results were noted in the statistics and function modules. However, there is room for improvement in the aspects related to mathematical culture and interdisciplinary assessment, as well as in the geometry and algebra modules. Finally, this study proposes initiatives such as prompt-based training of teachers using LLMs and integration of mathematics-focused LLMs [6]

Telesko, R., & Wilke, G. (2024) describes a Python-based prototype that semi-automatically generates MCQ-based exams and summaries. The very popular Large Language Models (LLM) are used here. The current prototype software is based on the Python libraries langchain for the LLM structure and streamlit for the user interface (UI). ChromaDB is used as a vector database and Google's Vertex AI as the LLM. The software is tested with selected documents of the course "Machine Learning with Python" (which is part of the specialization within the BIT program) to find out whether the quality of the generated exams and summaries meets the expectations of lecturers and students. The results will help to answer important questions regarding the practical use of LLM in education. The contribution of the discussed work aims to solve the problem in the field of examination support in higher technical education. [7]

Agostini, D., & Picasso, F. (2024) in their paper discuss a literature review in the area of ​​the use of LLMs in education. University lecturers are faced with the assessment of extensive classes. With the rapid advancement of large language models (LLMs) and their increasing availability, part of the solution to these problems may be available in the form of support. In fact, LLMs can process large amounts of text, summarize them, and provide feedback on them according to established criteria. The findings of this analysis can be used both to provide feedback to students and to help lecturers evaluate texts. With the right pedagogical and technological framework, LLMs can move instructors away from some of the sustainability issues, i.e. from the single choice of multiple choice tests and the like. [8]

Huovinen, L. (2024) w pracy opisano projekt wykorzystania LLM w edukacji. Celem projektu końcowego było zbadanie potencjału wykorzystania dużych modeli językowych (LLM) w automatyzacji i ulepszaniu procesu opracowywania programów nauczania. Projekt został zrealizowany poprzez opracowanie i wdrożenie aplikacji internetowej opartej na LLM, a następnie przeprowadzenie testów użyteczności przy użyciu podejścia poznawczego i sesji opinii użytkowników. Użytkownikom testowym przedstawiono narzędzie i zlecono ocenę jego skuteczności, przejrzystości i ogólnej użyteczności w kontekście planowania programu nauczania. Opinie użytkowników podkreśliły znaczenie jasnych wskazówek, integracji zadań w scentralizowanym narzędziu oraz potrzebę udziału człowieka w weryfikacji i udoskonalaniu generowanych treści, aby zapewnić dokładne i znaczące cele w projektowaniu programu nauczania [9]

Yuan, B., & Hu, J. (2024) Recent advances in large language models (LLMs) under artificial intelligence (AI) offer promising new avenues for improving course evaluation processes. This paper describes the application of LLMs in automated course evaluation from multiple perspectives and conducts rigorous experiments on 100 courses at a large university in China. The results indicate that: (1) LLMs can be an effective course evaluation tool; (2) their effectiveness depends on proper tuning and rapid engineering; and (3) the evaluation results generated by LLMs show a significant level of rationality and interpretability. [10]

Mao, Chen, and Liu (2024) address the use of AI tools like ChatGPT in their work. They point out the use of AI to increase efficiency in professional work and in the educational process. At the same time, they highlight the fact that the line between plagiarism and one's own work is blurred. They suggest citing the author by including a reference entry when citing AI-generated text (e.g., citing ChatGPT-generated text by referencing OpenAI). They also recommend that researchers describe their use of AI tools in the methods section if they used such tools in the process. It is not unusual for authors to list ChatGPT or other content generators as collaborators or co-authors. This has become the latest challenge facing researchers and publishers. In addition to warnings from the scientific community, we must be wary of the dangers of relying solely on AI tools for research due to the AI hallucination phenomenon. This phenomenon refers to AI's ability to produce irrelevant or false information using persuasive language [11].

Baidoo-anu, Owusu Ansah (2023) share research with recommendations on how ChatGPT can be used to maximize teaching and learning. Policymakers, researchers, educators and technology experts could work together and begin conversations about how these evolving generative artificial intelligence tools could be safely and constructively used to improve education and support student learning [12].

Prajapati, Kumar, Singh (2024) discuss in their paper the advantages of artificial intelligence that can be seen throughout higher education. New challenges are being generated that give students more time for new methods of learning, benchmarking for jobs in competitive companies with greater efficiency through automation. Increasingly, students are receiving individualized educational instruction from AI tutors that can take into account a student's unique learning style. Such tutors also map learning speed using a variety of interactive tools. There are several potential advantages of AI for the new workplace from a research perspective [13].

Alier, Marc; García-Peñalvo, Francisco; Camba, Jorge D. (2024) They touch on the possibilities of using AI Systems in education in their publication. AI systems, for example, can be used to create custom quizzes, generate essay prompts, and even grade essays. Using AI as an assessment tool can reduce teachers' workload and help students receive quick feedback on their work. Integrating AI into the educational environment also raises challenges of academic integrity. With the availability of AI models, students may use them to study or complete homework assignments, which can raise concerns about the authenticity and authorship of the work provided. Therefore, it is important to ensure that academic standards are maintained and the originality of student work is preserved. This issue, underscores the need to implement ethical practices in the use of AI models and ensure that the technology is used to support, not replace, the student's educational experience [14].

Baidoo-anu, Owusu Ansah, (2023) highlight ChatGPT's remarkable ability to perform complex tasks in education in their work. The use of these capabilities by students and pupils has caused mixed feelings among teachers. Advances in artificial intelligence-based solutions appear to be revolutionizing existing educational practice. In their article, the authors synthesize recent literature to offer potential benefits and drawbacks of ChatGPT in promoting teaching and learning. The benefits of ChatGPT include, among others, promoting personalized and interactive learning, generating prompts for formative assessment activities that provide ongoing feedback to inform teaching and learning [15].

The authors of the articles presented above, as well as the authors' own experiences, highlight the challenges and limitations of these systems and make recommendations for future research. In connection with the above, the main utilitarian issues arise.

1. Currently, every student uses a solution such as ChatGPT to search for knowledge, at the same time, there is no assurance that the knowledge provided by ChatGPT has been verified in any way by the student.

2.Even if a student uses a solution such as ChatGPT and generates material for his or her paper using it, the lecturer is unable to determine how much of the student's work is independent work.

3.To verify such work, the lecturer must also perform additional activities with ChatGPT, which contributes to the lengthening of the process of evaluating the student's work by the lecturer.

4. the provision of a correct answer by a student to a query using ChatGPT, is not equivalent to the fact that the student in question has understood the material included in the work.

For the utilitarian problems thus identified, the article poses the following research questions:

1. what language model will be able to optimally assess the self-efficacy of the generated student work?

2. what metrics can be useful for assessing the independence of student work?

3. if the available metrics prove ineffective, what should be the criteria for evaluating the self-efficacy of student papers?

## Model of application development processes for evaluating student work using LLM models

In this part of the article, for the utilitarian problems posed above and the research questions answering these utilitarian problems, a model of the application development process for the evaluation of student work using LLM models is presented. The article presents the assumptions showing: on the one hand, the state of research, and on the other hand, the possibilities of determining the process of application construction and the criteria for evaluating student works. Subsequently, the research environment in which the model will be implemented is presented. To show the importance of the model itself, the metrics for evaluating student papers are pointed out. The metrics analyzed are relevant to both the application design process and the construction of the model for its subsequent use in evaluating student work.

### Assumptions for the model

In order to study the degree of self-evaluation papers written by students using ChatGPT, the processes of building an application for evaluating these papers were presented. The premise of building the research environment was to allow students to use ChatGPT while writing evaluation papers. The use of ChatGPT was carefully defined by the Lecturers. A student could write a paper using ChatGPT support - but had to follow certain conditions. The evaluation paper must be written independently by the student. If the student copied some passages that ChatGPT suggested to him, he had to provide links to the sources in order to verify the generated information.

The query code was then implemented to evaluate the students' self-evaluation work. The query used the ChatGPT tool. This is a chatbot developed by OpenAI, using a GPT (generative pre-trained transformer) model and used to generate responses to user input. The model was developed based on large data sets so that it can carry on a conversation and engage in a variety of topics, from general conversations to specific areas of expertise. The choice of this tool was a certain consequence of our experiment. For the completeness of our research environment, it was also necessary to have criteria for evaluating the paper written by the student. This was implemented using metrics - algorithms for evaluating distance, similarity of data.

### Research environment

The testing environment included three main components:

* The IoT system scope exam process, in which 154 students participated. Students prepared a paper using their own knowledge and various reference sources including language models
* The development by the authors of the article of an application that allowed the automatic evaluation of 154 student papers. The development took into account the necessary criteria for evaluating student papers, which are based on selected metrics.
* Selection of metrics for evaluating student-submitted works and reference texts generated by language models

Students prepared papers, which were then evaluated by an application developed by the article's authors. The process was carried out in June-September 2024.

### Metrics used to evaluate student evaluation papers

Two metrics, BLEU and ROUGE, were initially used in the process of evaluating papers written by students. The choice of metrics was based on its applications in machine translation tasks, in which the goal is to automatically translate text from one language to another. It was proposed as a way to evaluate the quality of machine-generated translations by comparing them with a set of reference translations provided by human translators. The BLEU score ranges from 0 to 1, with higher values, indicating better translation quality. A perfect translation would have a BLEU score of 1, while a completely incorrect translation would have a BLEU score of 0.

In turn, the selection of the ROUGE metric was predicated on its use in text summarization tasks, where the goal is to automatically generate a concise summary of a longer text. ROUGE was designed to assess the quality of machine-generated summaries by comparing them to reference summaries provided by humans. The ROUGE score ranges from 0 to 1, with higher values indicating better summary quality. Like the BLEU score, an ideal summary would have a ROUGE score of 1, while a completely incorrect summary would have a ROUGE score of 0.

The Euclidean metric is also included in the next step. In machine learning, Euclidean distance is often used as a similarity measure to compare feature vectors. A feature vector represents a data point in a multidimensional space, where each dimension corresponds to a specific feature or attribute. By calculating the Euclidean distance between feature vectors, we can determine how similar or dissimilar they are.

The Braycurtis metric was also used. The Bray-Curtis metric, also known as the Bray-Curtis distance measure, is one way to measure differences between two samples in a multidimensional feature space. It is particularly used in ecology to compare the similarity or diversity of ecological samples, but also has applications in other fields such as data analysis, molecular biology, chemometrics, and others.

Chebyshev metric is also included - Chebyshev metric, also known as Chebyshev distance, is a distance measure used in mathematics and data analysis to determine the maximum difference between two vectors in a multidimensional space. It can also be found in the literature under the name L∞ metric or supremum (supremum) distance.

### The results of using the application development processes

The processes of building the model that forms the basis for building the application included the following steps shown in Figure 1.

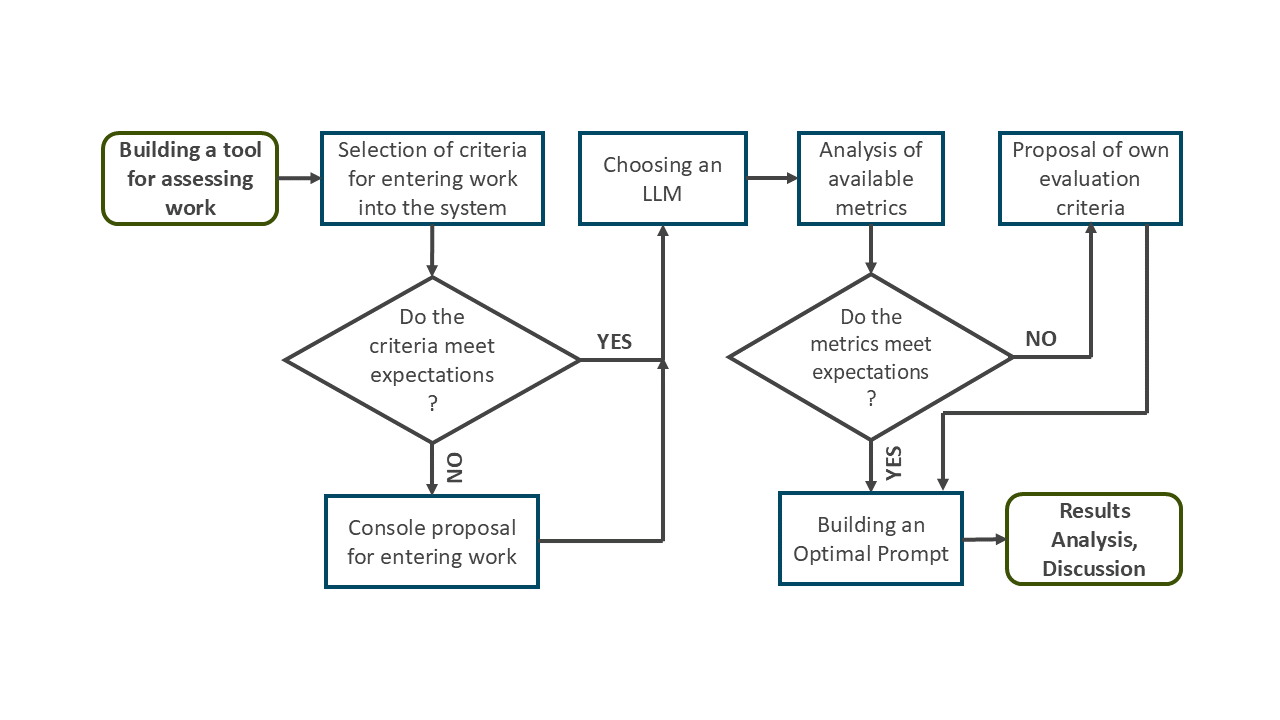


Figure 1 Application development process.

The research goal was to develop criteria in the form of a “prompt for ChatGPT” that could determine the degree of student self-efficacy when writing an evaluation paper using ChatGPT. To this end, an appropriate ChatGPT prompt was designed and developed that takes into account the relevant criteria. Initially, it was assumed that the development of such a “ChatGPT query” should take into account the above-mentioned metrics, which could calculate the distances between the ChatGPT-generated answer and the student-generated answer - the smaller the distance, the more similar the texts are. Such knowledge could automatically support a person evaluating student-generated evaluation papers. The metrics could calculate the distances between the texts - the reference generated by ChatGPT and the work written by the student. At the current state, several metrics were selected to test their usefulness. The entire solution was prepared based on a collected sample of 154 evaluation papers.

The following metrics were used to compare text distances: euclidean, braycurtis, chebyshev, rouge\_1, rouge\_L. A corresponding “prompt” was then offered to ChatGPT, which generated an answer to a question to ChatGPT and compared that answer with an evaluation paper written by the student. The metrics were to calculate the convergence or degree of divergence of the two texts. The evaluation and verification of student papers consisted of providing a score, which is an interpretation of the student's “self-efficacy” in generating his evaluation paper. The system of evaluating independence was prepared so that in the final step of calculating the student's independence, a number was given, which would correspond to the score - 1 non-self-contained work, 5 independent work. If the work was written independently by the student, the evaluation system returned a number corresponding to 1, 2, 3, 4 or 5.

## Results of model application

The developed model was used to evaluate 154 student evaluation papers. Some works were not suitable for analysis because they did not meet the criteria set for students during the generation of evaluation works. These works were omitted. 128 evaluation papers were qualified for analysis. Based on this, the corresponding scripts were prepared for the “prompt” for ChatGPT. The evaluation data for the sample evaluation papers are presented below in Table 1. The personal data of the students were hidden to ensure the protection of personal information. From the tests, one table was generated for the sample results of the experiment. In the experiment, several approaches were performed, but a table with selected results of the experiment was included in the article. Other intermediate results were not included. Several iterations were used for generating the corresponding “prompt”.

The results of the experiment, obtained through successive iterations, made it possible to execute a target application based on the model. The content of one of the “prompts” for ChatGPT is included below:

odp\_kolokwium\_prompt = """

as an IoT engineer, answer the following question briefly:

"""

check\_prompt = """ as an IoT engineer tell me briefly, without details, whether ChatGPT wrote this text? : """

compare\_prompt = """ compare with this text written by ChatGPT:"""

ocen = " Rate it on a scale of 1-5, no description, just numbers, "

wstep = " As a strict academic teacher, evaluate the colloquium given below, on the topic: "

ocen\_strukture = f"{ocen}, is the structure of the colloquium correct? "

ocen\_punkty = f"{ocen} whether the points cover the topic in the answer "

ocen\_zawartosc\_merytoryczna = f"{ocen} the substantive content of the points, whether they cover the topic of the question "

ocen\_odnosniki = f"{ocen} are there references and links to external sources of information such as articles, websites, github repositories "

ocen\_samodzielnosc = f"{ocen}, whether the colloquium was written independently, 1 if it is a copy of ChatGPT, 5 if completely independently. "

parametry\_oceny = f" Please provide your ratings in the following format: str:[ocena struktury], pun[ocena punktow], zaw[ocena zawartości], odn[ocena odnosników], sam[ocena samodzielności]. Don't use line breaks, only commas "

def ocena\_kolokwium(pyt, odp):

return f"""

{wstep}{pyt}

{ocen\_strukture}

{ocen\_punkty}

{ocen\_zawartosc\_merytoryczna}

{ocen\_odnosniki}

{ocen\_samodzielnosc}

{parametry\_oceny}

Rate the colloquium below: {odp}

"""

Table No. 1 Results of evaluation of student papers based on the application developed using the model (sample analysis results for each metric)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **euclidean** | **braycurtis** | **chebyshev** | **rouge\_1** | **rouge\_L** | **struktura** | **punkty** | **zawartość** | **odnośniki** | **zrozumienie** | **samodzielność** |
| **1.** | 14,3037 | 0,5075 | 3,031800032 | 0,205 | 0,114 | 4 | 4 | 4 | 2 | 4 | 3 |
| **2.** | 11,2487 | 0,4304 | 1,973399997 | 0,312 | 0,134 | 3 | 3 | 3 | 4 | 3 | 2 |
| **3.** | 10,0884 | 0,3412 | 1,7227 | 0,054 | 0,037 | 1 | 1 | 3 | 1 | 3 | 2 |
| **4.** | 6,4677 | 0,2329 | 0,917100012 | 0,161 | 0,09 | 3 | 3 | 4 | 1 | 4 | 3 |
| **5.** | 7,1435 | 0,2331 | 1,116000056 | 0,107 | 0,081 | 1 | 1 | 2 | 1 | 3 | 3 |
| **6.** | 17,9935 | 0,6843 | 3,492599964 | 0,229 | 0,102 | 1 | 3 | 3 | 1 | 3 | 3 |
| **7.** | 13,6355 | 0,5182 | 2,859400034 | 0,323 | 0,151 | 2 | 3 | 3 | 1 | 3 | 3 |
| **8.** | 19,0653 | 0,6677 | 4,021299839 | 0,203 | 0,105 | 5 | 5 | 5 | 1 | 5 | 4 |
| **9.** | 6,2819 | 0,1917 | 1,092100024 | 0,1 | 0,074 | 4 | 5 | 4 | 3 | 4 | 3 |
| **10.** | 13,5221 | 0,5087 | 2,707000017 | 0,269 | 0,113 | 3 | 4 | 4 | 1 | 4 | 2 |
| **11.** | 13,0358 | 0,4782 | 2,637000084 | 0,253 | 0,129 | 1 | 2 | 3 | 1 | 3 | 4 |
| **12.** | 15,3231 | 0,5432 | 3,3677001 | 0,221 | 0,111 | 3 | 4 | 4 | 1 | 4 | 4 |

The results of the experiment were compared with the evaluations of the students' evaluation papers done “manually” by the lecturers. The two results were then compared for each evaluation paper to analyze each case. Since the results based on the selected metrics did not give the expected values - meaning that they differed from the expected values, evaluation criteria for evaluation papers were applied. The evaluation criteria were proposed by the article's authors. The proposed sub-assessment criteria by the authors were:

1. the “structure” of the document made by the student (the structure of the evaluation paper, whether the structure of the colloquium is correct)

2. the “points” - whether the points in the structure of the document exhaust the topic in the answer

3. the “substantive content” of the document made by the student (the substantive content of the points, whether they exhaust the topic from the question)

4. “links” contained in the documents made by students (whether the evaluation work made by the student has links and references to external sources of information, such as articles, websites, GitHub repositories)

5. “comprehension” - whether the student answered in the evaluation work in an understandable way

6. “independence” - whether the student generating the evaluation paper who used ChatGPT copied the content or gave the answer independently.

The analysis of the data shows that the use of individual metrics does not give clear results in terms of the usefulness of the metrics used. This means that the authors of the article are not able to clearly indicate which metric gives the expected results. A better solution turned out to be the use of proprietary criteria for evaluating evaluation work done by students. The selected results were aggregated and sample results are placed in Table 1.

## Discussion

To conduct the experiment, the authors of the article decided to use the ChatGPT tool from OpenAI. The choice was motivated primarily by the popularity of this tool and the number of parameters that can make ChatGPT have an advantage over other tools of this type. ChatGPT -3 (and 3.5), which can still be used for free, uses the GPT-3 model. These versions of the chatbot have 175 billion parameters. In turn, the latest version of ChatGPT - available only with a paid subscription - works with the number of parameters included in GPT-4. It is estimated that the number of parameters for the ChatGPT-4 version can reach 100 trillion. However, the specific number of parameters for GPT-4 has not been officially disclosed by OpenAI. This means that the choice of this solution for research purposes is accurate.

The metrics used do not meet the expectations of the authors of the article. The data analysis presented in the table with the results of the experiment does not provide certainty for drawing conclusions. The data concerning the selected metrics seem to generate random results, difficult to interpret. Of course, the topic of metrics still remains open. The authors do not rule out returning to this type of solution in the future, if appropriate software appears. It seems that at the moment there is no appropriate metric that could meet the expectations set in the research problems.

The authors of the article proposed their own criteria for assessing the independence of evaluation papers generated by students. After iterative verification, these criteria can constitute material for development and further research. It seems that the automation of the process of checking the correctness of the generation of an evaluation paper by a student has been achieved. The difficulty concerns the examination of the independence of the generated evaluation papers by students. This topic will be examined and developed in the next steps of the experiment.

## Conclusions

The developed model, its verification processes and the discussion conducted indicate the following conclusions:

1. The aim of the article was to develop a process model for building an application for assessing student papers

2. The model was built, its verification processes were carried out for 154 student papers and its usefulness was assessed using metrics

3. On the one hand, such an approach enabled a preliminary assessment of student papers, which were verified by the instructor and then a face-to-face exam to verify the developed solution.

4. The implementation of this type of experiments will also open the way to changes in the area of ​​didactic classes, which consciously use tools such as ChatGPT by students. In this area, there are no defined rules and criteria for research experiments. There is also no knowledge and experience appropriate for creating new processes in the area of ​​didactics using ChatGPT defined in utilitarian problems

5. Directions for further research include: taking into account diverse metrics for assessing student papers and preliminary questions regarding the state of knowledge of the student before taking the exam and, in the event of a negative assessment, not allowing him to take the exam.

6. Such a situation becomes necessary after conducting an oral exam, where it turned out that students who received insufficient grades do not understand the processes of a given field.

7. The model presented in the work requires a detailed study of its processes, which was not done in this article. This resulted from treating this article as an introduction to the analysis of the possibilities of using language models to evaluate student works.

List of publications

[1] Villagrán, I., Hernández, R., Schuit, G., Neyem, A., Fuentes-Cimma, J., Miranda, C., ... & Varas, J. (2024). Implementing artificial intelligence in physiotherapy education: A case study on the use of large language models (LLM) to enhance feedback. IEEE Transactions on Learning Technologies.

[2] Meissner, R., Pögelt, A., Ihsberner, K., Grüttmüller, M., Tornack, S., Thor, A., ... & Hardt, W. (2024, October). LLM-generated competence-based e-assessment items for higher education mathematics: methodology and evaluation. In Frontiers in Education (Vol. 9, p. 1427502). Frontiers Media SA.

[3] Smolić, E., Pavelić, M., Boras, B., Mekterović, I., & Jagušt, T. (2024, May). LLM Generative AI and Students’ Exam Code Evaluation: Qualitative and Quantitative Analysis. In 2024 47th MIPRO ICT and Electronics Convention (MIPRO) (pp. 1261-1266). IEEE.

[4] Joy Kulangara, K. (2024). Designing and building a platform for teaching introductory programming supported by large language models.

[5] Jošt, G., Taneski, V., & Karakatič, S. (2024). The Impact of Large Language Models on Programming Education and Student Learning Outcomes. Applied Sciences, 14(10), 4115.

[6] Hu, B., Zheng, L., Zhu, J., Ding, L., Wang, Y., & Gu, X. (2024). Teaching Plan Generation and Evaluation With GPT-4: Unleashing the Potential of LLM in Instructional Design. IEEE Transactions on Learning Technologies.

[7] Telesko, R., & Wilke, G. (2024). TOWARDS A SUSTAINABLE OPERATING CONCEPT OF LLM GENERATED EXAMINATION SUPPORT IN HIGHER TECHNICAL EDUCATION. In ICERI2024 Proceedings (pp. 1064-1069). IATED.

[8] Agostini, D., & Picasso, F. (2024). Large language models for sustainable assessment and feedback in higher education. Intelligenza Artificiale, 18(1), 121-138.

[9] Huovinen, L. (2024). Assessing Usability of Large Language Models in Education.

[10] Yuan, B., & Hu, J. (2024). An Exploration of Higher Education Course Evaluation by Large Language Models. arXiv preprint arXiv:2411.02455.

[11] Mao, J., Chen, B. & Liu, J.C. Generative Artificial Intelligence in Education and Its Implications for Assessment. TechTrends 68, 58–66 (2024). <https://doi.org/10.1007/s11528-023-00911-4>

[12] 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>

[13] Prajapati, J.B., Kumar, A., Singh, S. et al. Artificial intelligence-assisted generative pretrained transformers for applications of ChatGPT in higher education among graduates. SN Soc Sci 4, 19 (2024). <https://doi.org/10.1007/s43545-023-00818-0>

[14] Alier, Marc; García-Peñalvo, Francisco; Camba, Jorge D. (2024). "Generative Artificial Intelligence in Education: From Deceptive to Disruptive", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 8, issue Special issue on Generative Artificial Intelligence in Education, no. 5, pp. 5-14. <https://doi.org/10.9781/ijimai.2024.02.011>

[15] 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>

[16] Kisielewicz, A., (2017). *Sztuczna Inteligencja i Logika*. Warszawa, Wydawnictwo Naukowe PWN SA.

[17] Wodecki, A., (2021), *Sztuczna Inteligencja we współczesnych organizacjach*. Warszawa, Wydawnictwo Naukowe PWN SA.

[18] Flasiński, M., (2011), *Wstęp do Sztucznej Inteligencji*. Warszawa, Wydawnictwo Naukowe PWN SA.

[19] Rutkowski, L., (2012), *Metody i techniki sztucznej inteligencji*. Warszawa, Wydawnictwo Naukowe PWN SA.

[20] Carr, David F. (2023). *ChatGPT’s First Birthday is November 30: A Year in Review*, <https://www.similarweb.com/blog/insights/ai-news/chatgpt-birthday/> (dostęp: 15.11.2023).

[21] Vincent, James. (2016). *Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day*, <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist> (dostęp: 15.11.2016).

[22] Strony Parlament UE. (2023). *Akt ws. sztucznej inteligencji: pierwsze przepisy regulujące sztuczną inteligencję*, <https://www.europarl.europa.eu/topics/pl/article/20230601STO93804/akt-ws-sztucznej-inteligencji-pierwsze-przepisy-regulujace-ai> (dostęp: 13.06.2023).

[23] Vincent, James. (2016). *Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day*, <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist> (dostęp: 15.11.2016).

[24] Gates, B. (2023). *The Age of AI has begun*, <https://www.gatesnotes.com/The-Age-of-AI-Has-Begun?WT.mc_id=20230321100000_Artificial-Intelligence_BG-EM_&WT.tsrc=BGEM> (dostęp: 21.03.2023).

[25] Vincent, J. (2022). *AI-generated answers temporarily banned on coding Q&A site Stack Overflow*, <https://www.theverge.com/2022/12/5/23493932/chatgpt-ai-generated-answers-temporarily-banned-stack-overflow-llms-dangers> (dostęp: 05.12.2022)

[26] OpenAI. (2022). *Aligning language models to follow instructions*, <https://openai.com/research/instruction-following> (dostęp: 27.01.2022)