Aritonović, G. (2026). Model of automated test evaluation using generative AI tools, *SCIENCE International journal, 5*(2), 227-232. doi: 10.35120/sciencej0502227a UDK: xxx

# MODEL OF AUTOMATED TEST EVALUATION USING GENERATIVE AI TOOLS

*Goran Aritonović*[1*]

[1]Belgrade Business and Arts Academy of Applied Studies, Belgrade, Serbia, e-mail: goran.aritonovic@bpa.edu.rs

**Abstract:** This paper presents a model for automated test evaluation based on the application of generative AI tools. The proposed approach is structured as a sequential evaluation pipeline that uses predefined prompts to process student responses through clearly defined phases. The evaluation process includes defining the answer key, applying scoring rules, analyzing responses at the question level, generating individual reports, and aggregating results. A key contribution of the paper is the definition of a controlled prompt protocol that decomposes a complex evaluation task into smaller, verifiable steps, improving transparency and reliability. The model also supports human-in-the-loop interaction, allowing instructors to intervene at each stage of the evaluation process and maintain pedagogical control. Unlike traditional approaches where AI is used as a general-purpose assistant, the proposed model defines a structured workflow that ensures consistent and repeatable evaluation. This design reduces the risk of uncontrolled AI behavior and enables precise identification of potential errors within individual phases of the pipeline. The model was applied in a real educational setting, where test evaluation was performed using the proposed pipeline. The results indicate a significant improvement in efficiency, with an approximate 95% reduction in evaluation time compared to manual grading. Additionally, the system demonstrates a high level of consistency in applying grading criteria, reducing subjectivity in assessment. It was observed that system reliability depends on the quality of input data, particularly in cases involving manually entered student identification data. The findings suggest that limitations are not solely related to the AI model itself, but also to data acquisition and preprocessing. The proposed model shows strong potential for application in various educational contexts where efficient and transparent test evaluation is required.

***Keywords:*** *generative AI, automated evaluation, test assessment, prompt engineering, educational technology.*
**Field:** Education

## 1. INTRODUCTION

The development of large language models (LLMs) (Kasneci et al., 2023), such as ChatGPT and Gemini, has significantly influenced the way knowledge is produced, distributed, and used in educational systems. In educational contexts, these models enable the analysis of textual responses, the generation of explanations, and support across various phases of the teaching process (Becker et al., 2023).

Previous research has predominantly focused on students and the ways in which they use AI tools to solve tasks. However, the application of these tools in the work of instructors, particularly in processes related to knowledge evaluation and the analysis of test results, remains insufficiently explored. These processes represent critical points within the teaching cycle, as they directly affect the quality of knowledge assessment and the design of instructional content.

Test evaluation involves a complex set of activities, including the verification of answer accuracy, assessment of conceptual understanding, and consistent application of grading criteria.

In this context, generative AI tools can significantly improve the efficiency of these processes; however, their use raises important questions regarding reliability, transparency, and pedagogical responsibility. The aim of this paper is to present a concrete model for applying AI in the process of test evaluation in education, based on practical experience, and to analyze its advantages and limitations.

## 2. LITERATURE REVIEW

Contemporary literature indicates a complex and often contradictory impact of generative AI systems in education (Holmes et al., 2019; Zhai, 2023). Vaithilingam et al. (2022) show that code generation tools significantly increase user productivity, but do not guarantee a better understanding of problems. This finding is particularly important in education, where efficiency is not the primary goal, but rather the development of knowledge and skills.

Prather et al. (2023) further deepen this perspective by highlighting reduced cognitive engagement among students who use AI tools. Their research suggests that students often accept generated solutions without critical analysis, thereby undermining the learning process and the development of mental models.

*Corresponding author: goran.aritonovic@bpa.edu.rs

Finnie-Ansley et al. (2022) analyze the performance of AI systems on programming tasks and demonstrate that these systems can successfully solve standardized problems, but face difficulties in more complex scenarios. This finding points to limitations of AI in the context of knowledge evaluation (OpenAI, 2023).

Becker et al. (2023) examine the broader context of generative model applications in education and emphasize the need to redefine the role of instructors. Instead of the traditional role of knowledge transmitters, instructors become moderators and evaluators of the learning process.

Kasneci et al. (2023) introduce the concept of "human-in-the-loop," which implies active involvement of instructors in all phases of AI tool application. This approach is shown to be essential for maintaining the quality and reliability of the educational process.

Despite the growing body of research, there is a noticeable lack of integrated models that cover the entire process of test evaluation, including data collection, input processing, and result analysis. This paper contributes to addressing this gap by presenting a unified approach based on practical implementation.

## 3. METHODOLOGY

### 3.1 Hybrid Test Design Model

The test was constructed based on existing questions, with predefined criteria ensuring an appropriate level of difficulty, coverage of the subject matter, and variability of tasks. This approach preserves continuity and evaluation standards while reducing the possibility of answer memorization.

A key element of this model is a clearly defined test preparation process, in which the instructor actively controls the content and structure of the questions, ensuring a balance between standardization and evaluation quality.

### 3.2 Evaluation Pipeline

The evaluation process is organized as a sequential pipeline consisting of several clearly defined phases. In the initial phase, a set of correct answers is defined, which serves as a reference framework for analysis. This is followed by score calculation based on predefined rules, where correct answers are positively scored, incorrect answers are penalized, and unanswered questions are treated as neutral.

In the subsequent phases, a detailed analysis of responses is performed at the question level, followed by the generation of individual reports, aggregation of results, and their ranking. The final phase includes filtering the results and exporting them into a standardized format.

This structure ensures transparency and verifiability at each stage of the process, which is particularly important in the context of automated evaluation.

### 3.3 Structure of the Prompt Protocol for Evaluation

The evaluation of student responses is carried out using a sequential prompt protocol, in which each prompt has a clearly defined function within the data processing process. The process includes defining the answer key, applying scoring rules, analyzing responses at the level of individual questions, generating individual reports, aggregating results, and preparing the output table.

This approach enables the decomposition of a complex evaluation process into smaller, verifiable steps, thereby increasing transparency and reducing the risk of uncontrolled interpretation by the AI system. At the same time, it allows instructor intervention at every stage of the process, ensuring additional control and reliability of the evaluation. The structure of the evaluation pipeline is presented through clearly defined phases, which are systematized in Table 1.

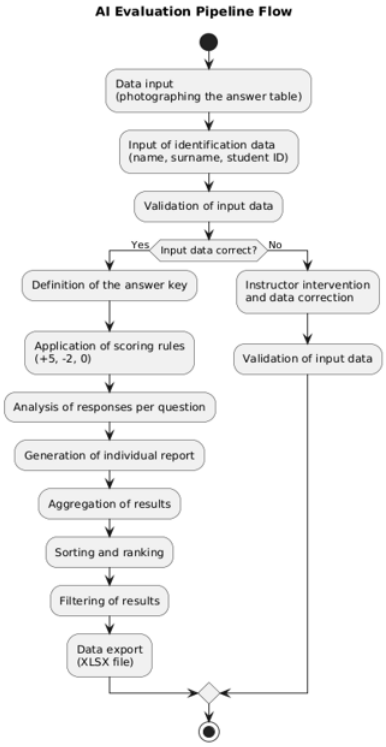
Table 1: Phases of the AI Evaluation Pipeline

| Phase | Function Description | Role in the Process |
|---|---|---|
| F1 | Data input (photographing the answer table) | Collection of input data |
| F2 | Input of identification data (name, surname, student ID) | Linking results to the student |
| F3 | Definition of the answer key | Formation of the reference set |
| F4 | Application of scoring rules | Score calculation (+5, −2, 0) |
| F5 | Analysis of responses per question | Comparison of responses |
| F6 | Generation of individual report | Overview of results per student |
| F7 | Aggregation of results | Summary table |
| F8 | Sorting and ranking | Organization of results |
| F9 | Filtering of results | Selection of passing students |
| F10 | Data export | Generation of XLSX file |

Source: author's own research

The flow of the evaluation process, including the relationships between the phases, is presented in Figure 1.

Figure 1: Flow of the AI Evaluation Pipeline




Source: author's own research

The diagram illustrates the flow of the evaluation process, starting from data input through a photographed answer table and the input of student identification data, through processing within the defined prompt protocol, to the generation of individual reports and the export of results. Particular emphasis is placed on the input data validation phase, in which instructor intervention is possible.

Unlike approaches in which evaluation is performed through a single complex query to an AI system, the applied sequential model enables the decomposition of the process into clearly defined phases. Such decomposition provides several significant advantages. First, each phase can be independently

verified, which increases the transparency and reliability of the system. In the event of an error, it is possible to precisely identify the phase in which the deviation occurs, which significantly facilitates correction. This is consistent with findings in the literature indicating that large language models demonstrate greater reliability when complex tasks are broken down into smaller, structured steps (Kasneci et al., 2023). Finally, the sequential pipeline enables the integration of human control at every stage of the process. The instructor can intervene after any step, which is particularly important in the context of identified limitations of AI systems, such as errors in numerical calculations and insufficiently precise interpretation of responses. In this way, a balance is achieved between automation and pedagogical responsibility.

## 4. RESULTS

In order to analyze the performance of the proposed model, the evaluation results are presented on a representative sample of students who took the theoretical part of the exam. Table 2 shows the relationship between the number of correct, incorrect, and unanswered questions, as well as the corresponding scores and final grades, in accordance with the defined evaluation rules.

Table 2: Student Evaluation Results (20 Questions)

| Student | Correct | Incorrect | Unanswered | Test Score | Practical | Total | Grade |
|---|---|---|---|---|---|---|---|
| S1 | 14 | 4 | 2 | 62 | 70 | 66.0 | 7 |
| S2 | 12 | 6 | 2 | 48 | 75 | 61.5 | 7 |
| S3 | 16 | 3 | 1 | 74 | 80 | 77.0 | 8 |
| S4 | 12 | 5 | 3 | 50 | 65 | 57.5 | 6 |
| S5 | 13 | 4 | 3 | 57 | 72 | 64.5 | 7 |

Source: author's own research

The application of the proposed evaluation model was tested in a real educational environment on a group of students who took the theoretical part of an exam in the field of programming. The evaluation was carried out using the defined sequential pipeline, while the results were compared with the standard grading method.

In order to analyze the system performance, the following aspects were observed: accuracy of score calculation, consistency of evaluation, and the time required for processing results. Particular focus was placed on identifying potential errors in the operation of the AI system. The application of the proposed model in a real educational environment demonstrated significant advantages in terms of efficiency and consistency. One of the key results of the system application relates to a significant reduction in the time required for test evaluation. Based on practical application, it was estimated that processing time is reduced by approximately 95% compared to the manual approach. This result indicates a high potential for applying the system in working with larger groups of students. The estimate is based on a comparison of the time required for manual evaluation and the time required to process the same set of tests using the AI pipeline.

In addition, it was observed that the AI system applies grading criteria consistently to all students, thereby reducing the subjectivity that may be present in manual evaluation. During the application of the system, no inconsistencies in the application of scoring rules or errors in result calculation were observed, provided that the input data were correctly defined. This indicates that the sequential evaluation model ensures stable and deterministic system behavior in the context of clearly defined rules.

However, problems were identified in the data input phase, related to incorrect recognition of student names and identification numbers during the processing of photographed tests, where the issues arise due to manual data entry by students and variability in their writing. These problems are not a consequence of the evaluation logic, but rather limitations in input data processing.

This challenge was resolved by switching to voice input via the mobile version of the system, which significantly improved input accuracy and eliminated the source of errors. This finding indicates that the reliability of the overall system does not depend solely on the evaluation model, but also on the quality of

the input data. In addition to aggregated results, the system generates a detailed report for each student, which includes an overview of responses for individual questions and the corresponding evaluation. This report provides a transparent insight into the grading process and facilitates the verification of results.

Table 3: Example of Response Evaluation for a Single Student

| Question | Correct Answer | Student Answer | Evaluation | Points |
|---|---|---|---|---|
| 1 | A | A | correct | +5 |
| 2 | B | C | incorrect | −2 |
| 3 | C | — | unanswered | 0 |
| 4 | A | A | correct | +5 |
| 5 | B | B | correct | +5 |
| ... | ... | ... | ... | ... |
| 20 | C | B | incorrect | −2 |

Source: author's own research

The presented report enables a detailed analysis of student performance at the level of individual questions. Based on these data, the system automatically calculates the total number of points and generates the corresponding grade, in accordance with the defined evaluation rules.

## 5. DISCUSSION

The results presented in this paper indicate that generative AI tools can have a significantly different and more practical role in education than is most commonly described in the literature. While most research focuses on the use of AI in the learning process or code generation, this paper emphasizes the automation of knowledge evaluation through a concrete workflow. The key difference compared to existing approaches lies in the method of system application. Instead of using AI tools as general-purpose assistants, a clear workflow is defined, which includes data input through photographed tables, their processing through a sequential prompt protocol, and the generation of structured reports. This approach enables the transformation of a complex evaluation process into a series of controlled steps.

The most significant result relates to the reduction in time required for test evaluation, which in practice amounts to approximately 95% compared to the manual approach. This finding has direct implications for the work of instructors, especially in the context of larger groups of students, where manual evaluation represents a significant burden.

At the same time, the results indicate that system reliability largely depends on the quality of input data. The identified issues were not related to the evaluation logic, but to incorrect recognition of student names and identification numbers during the processing of photographs. This finding suggests that system limitations do not necessarily stem from the AI model itself, but from the way data are entered into the system (Susnjak, 2022).

In comparison with findings from the literature, which indicate limitations of generative models in understanding and interpretation (Finnie-Ansley et al., 2022; Kasneci et al., 2023), the results of this study show that these limitations can be significantly mitigated through a structured approach and a clearly defined workflow. In other words, system reliability depends not only on the model, but also on the way it is applied. These findings suggest that future research in this area should focus less on the models themselves and more on system design and the integration of AI tools into specific educational processes.

A limitation of this study is that the proposed model was tested within a single educational context and on a limited dataset. Therefore, the results should not be interpreted as universally applicable without further validation in different educational environments. Nevertheless, the obtained results indicate significant potential for applying the sequential AI pipeline in structured evaluation processes.

## 6. CONCLUSIONS

This paper presents a practical model for the application of generative AI tools in the process of test evaluation in education. Unlike most existing approaches, the central focus of the paper is on the automation of evaluation through a clearly defined workflow. The results show that it is possible to significantly reduce evaluation time while maintaining consistency and control over the process. It is particularly important that system reliability is not based solely on the capabilities of the AI model, but on the way it is integrated into the process. The limitations of the system were identified in the data input phase, which indicates the need for further improvement of input data processing methods. Future work may be directed toward the integration of more reliable data recognition methods and the extension of the system to other forms of evaluation.

## REFERENCES

Becker, B. A., Denny, P., Finnie-Ansley, J., et al. (2023). Generative AI in Computing Education: Opportunities and Challenges. ACM Transactions on Computing Education. https://doi.org/10.1145/3615706.

Cotton, D., Cotton, P., & Shipway, J. (2023). Chatting and Cheating: Ensuring Academic Integrity in the Era of ChatGPT. Innovations in Education and Teaching International. https://doi.org/10.1080/14703297.2023.2190148

Finnie-Ansley, J., Becker, B. A., Denny, P., & Luxton-Reilly, A. (2022). My AI Wants to Know if This Will Be on the Exam: Testing OpenAI's Codex on CS2 Programming Exercises. Proceedings of the 53rd ACM Technical Symposium on Computer Science Education (SIGCSE). https://doi.org/10.1145/3478431.3499348

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign. https://curriculumredesign.org/wp-content/uploads/AIED-Report-2019.pdf

Kasneci, E., et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

OpenAI. (2023). GPT-4 Technical Report. https://arxiv.org/abs/2303.08774

Prather, J., Reeves, B., Denny, P., Becker, B. A., et al. (2023). The Impact of AI Code Generators on Programming Education. Proceedings of the ACM Conference on International Computing Education Research (ICER). https://doi.org/10.1145/3568813.3600137

Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? arXiv preprint arXiv:2212.09292. https://arxiv.org/abs/2212.09292

Vaithilingam, P., Zhang, T., & Glassman, E. (2022). Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3491102.3517506

Zhai, X. (2023). ChatGPT for Teaching and Learning: A Systematic Review. Educational Technology Research and Development. https://doi.org/10.1007/s11423-023-10231-6