

Automated Assessment of Cognitive Levels in Bloom's Taxonomy Through RAG-Based Architecture with LLMs [†]

Anastasia Vangelova  and Veska Gancheva * 

Department of Programming and Computer Technologies, Technical University of Sofia, 1000 Sofia, Bulgaria; avangelova@yahoo.com

* Correspondence: vgan@tu-sofia.bg

[†] Presented at the 15th International Scientific Conference TechSys 2026—Engineering, Technologies and Systems, Plovdiv, Bulgaria, 14–16 May 2026.

Abstract

This paper presents a RAG-based architecture with a large language model for automated assessment of student responses, in which Bloom's Taxonomy is used to structure an analytical rubric and to form a cognitive profile of the performance. The approach integrates Moodle, n8n, a vector database for semantic retrieval of relevant context, and a GPT model for criterion-based assessment. The model's output is formed based on the student response, the RAG-extracted learning context, and a predefined rubric. The system is implemented in a real exam environment in Bulgarian and English. The results of data analysis show differentiated behavior across Bloom's cognitive levels, with higher sensitivity in tasks related to remembering, understanding, and applying, and more conservative behavior in tasks related to evaluating and critical thinking. Expert evaluation of the automatically generated feedback confirms its clarity, usefulness, and correctness.

Keywords: automated grading; Bloom's Taxonomy; LLM; RAG; analytical rubric; Moodle; e-learning; open-ended questions

1. Introduction

The historical evolution of automated assessment—from symbolic and ontological models, through statistical and machine learning approaches, to contemporary large language models—shows an increase in semantic sensitivity, but also the emergence of new limitations [1–3]. Symbolic models provide formal control but are not flexible enough to deal with free text, while statistical and machine learning approaches are scalable but often remain dependent on surface features [4,5]. LLMs introduce contextual and argumentative understanding, but they also suffer from hallucinations, instability, and a lack of built-in pedagogical constraints [3,6].

Recent research confirms that automated assessment of open-ended responses remains a highly active field, driven by the need for scalable, consistent, and pedagogically meaningful evaluation in digital learning environments [1,7]. Contemporary work increasingly moves beyond traditional essay scoring toward AI-based approaches that combine semantic analysis, rubric-based assessment, contextual grounding, and automated feedback [6,8].

At the same time, studies on LLM-based assessment report promising levels of agreement with human raters, but also emphasize persistent challenges related to reliability, validity, fairness, and dependence on prompt design and task formulation [3,9,10]. These findings suggest that the central problem is no longer only whether AI can assign a score,



Academic Editors: Nikola G. Shakev,
Valyo N. Nikolov, Nikolay R.
Kakanakov and Sevil A.
Ahmed-Shieva

Published: 17 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

but under what architectural and pedagogical constraints such scoring can become transparent, reproducible, and educationally valid.

This paper proposes a methodological framework in which RAG provides context grounding and limits the model to the content of the specific course, embeddings support semantic matching between text passages, and a cognitive rubric, based on Bloom's taxonomy, structures the assessment along different levels of cognitive complexity. The number and type of criteria can be adapted according to the goals and difficulty of the specific question.

The developed system represents a real AI-based architecture for automated grading of open-ended student responses in a Moodle environment (available at <https://moodleproject.eu>, accessed on 13 June 2026). It integrates Moodle as the LMS layer, n8n as the orchestration layer, Supabase/pgvector as the semantic base for RAG retrieval, and the OpenAI GPT-4.1-mini language model as the AI layer for rubric-based assessment in an automated n8n workflow. For each student submission, the workflow extracts the response, the analytical rubric, and the relevant learning context, after which the AI model generates a structured criterion-based assessment, a Bloom's cognitive profile, and automatically generated feedback. Each rubric criterion is mapped to a cognitive level in Bloom's taxonomy, which allows for the formation of both a final grade and a cognitive profile of the performance. The methodology is based on a clearly limited set of input objects: a student response, a relevant learning context extracted via RAG, and a configurable rubric with defined criteria and levels. It does not aim for the language model to "imitate" the human rater, but to implement evidence-constrained evaluation, in which each decision is argued through the provided context and the rules of the rubric.

In this context, the study is aimed at the following research questions:

RQ1: How does the system behave towards the different cognitive levels according to Bloom's taxonomy?

RQ2: What is the pedagogical value of the automatically generated feedback?

2. Background and Related Work

One of the key challenges in developing intelligent e-learning systems is how to represent knowledge in a way that is both structured, machine-interpretable, and flexible enough to accommodate linguistic variability in student responses. This is true for both technical disciplines and the humanities and social sciences, where responses often contain interpretations, arguments, and conceptual connections that are difficult to describe through strictly formal structures. In the context of automated assessment of open-ended questions, this problem is particularly significant, as traditional assessment is time-consuming and susceptible to subjective interpretations, which motivates the development of reliable automated approaches [11]. Historically, one of the most common approaches to formalizing knowledge has been ontologies—models that describe concepts in a given domain and the relationships between them through classes, properties, and hierarchies. The development of modern language technologies is gradually directing the representation of knowledge toward semantic vector representations (embeddings). These models transform words, sentences, or entire texts into multidimensional numerical vectors that capture semantic relationships and position phrases with similar content close together, regardless of the specific formulation. This allows for the recognition of different ways of expressing the same idea—a possibility that formal ontological models have difficulty providing in free text.

In an educational context, embeddings allow for more flexible comparisons between student responses and teaching materials, helping to discover key concepts and semantic similarities. At the same time, research shows that the independent use of semantic vector models is often insufficient when assessing a variety of student formulations, as these

models have difficulty capturing the finer contextual and semantic differences in open-ended responses [7]. Therefore, hybrid architectures that combine semantic retrieval with generative language models are increasingly used.

Within the framework of the developed system, embeddings also have a direct technological role: the vector database is built on them, which allows for semantic retrieval of relevant context through Retrieval-Augmented Generation (RAG). Such architectures are considered an effective means of reducing hallucinations and “grounding” the language model in specific learning materials and reference sources [12]. When assessing open-ended student responses, RAG allows the model to work not in the abstract, but on a retrieved context, which improves the reliability and transparency of the assessment [8]. In this sense, embeddings and RAG become a fundamental component of both the theoretical framework and the practical implementation of automated assessment. Bloom’s Taxonomy is a widely used framework for classifying cognitive processes in learning. Originally proposed by Bloom et al. [13] and revised by Anderson and Krathwohl [14], it describes six levels of cognitive complexity: remembering, understanding, applying, analyzing, evaluating, and creating. These levels reflect increasing complexity of thinking operations and are widely used in formulating learning objectives and assessment.

In this work, Bloom’s Taxonomy is used as a conceptual framework for structuring the assessment rubric. The main reason is that it allows a clear distinction between different types of cognitive activity—from knowledge reproduction to analysis, critical evaluation and creation of new solutions. This makes it a suitable basis for automated assessment of open-ended questions. In these, it is necessary to analyze not only the factual correctness, but also the way of reasoning of the student. Bloom’s cognitive levels can be recognized by characteristic language structures in the answer. For example, reproduction of definitions is typical of remembering, explanations and interpretations—of understanding, use of knowledge in a new context—of application, and argumentation and comparisons—of analysis and evaluation. Formulation of new solutions or proposals corresponds to the creation level.

Large language models can recognize such cognitive operations by analyzing the structure and content of the text, making them a suitable tool for automated assessment of open-ended questions.

In the developed system, cognitive levels are operationalized through an analytical rubric. Each criterion in the rubric is associated with a certain type of cognitive activity. For example, the criterion “Accuracy and Correctness” is associated with the level of remembering, “Comprehension and Connection”—with understanding, “Applicability”—with application, and “Analysis”, “Evaluation Skills” and “Creativity” correspond to the higher levels of analysis, evaluation and creation. This structure allows for multidimensional assessment, in which one answer can simultaneously demonstrate different cognitive qualities. Thus, the system does not only assess the correctness of the response, but also takes into account the type of thinking that the student demonstrates.

3. Materials and Methods

3.1. Linking Bloom’s Taxonomy to Rubric Criteria

Table 1 presents an example configuration of a rubric, in which individual criteria are linked to specific Bloom’s cognitive levels.

Table 1. Linking Bloom’s cognitive levels to rubric criteria for automated assessment.

Cognitive Level (Bloom)	Indicators in the Student Response	Component of the Rubric
Remembering	Accurate reproduction of facts	Accuracy and correctness
Understanding	Explanations, examples, interpretations	Understanding and connecting

Table 1. Cont.

Cognitive Level (Bloom)	Indicators in the Student Response	Component of the Rubric
Applying	Solving problems, using knowledge	Applicability
Analyzing	Comparison, justification, argumentation	Analytical ability
Evaluating	Critical evaluation, argumentation	Evaluation skills
Creating	New solutions, proposals, models	Creativity

3.1.1. Relationship Between Cognitive Levels, RAG Process and Language Model

An architecture has been developed that integrates Bloom's Taxonomy, the rubric and RAG-based analysis (see Figure 1). RAG provides relevant context from the learning materials, extracted through vector search, and the language model analyses the student response against this context and the rubric criteria. Thus, the model does not assess in the abstract, but compares the response to the specific learning content. The process includes two main steps:

1. Extraction of relevant passages from the learning materials through embeddings and vector search;
2. Analysis of the student response by the LLM against the extracted context and the rubric.

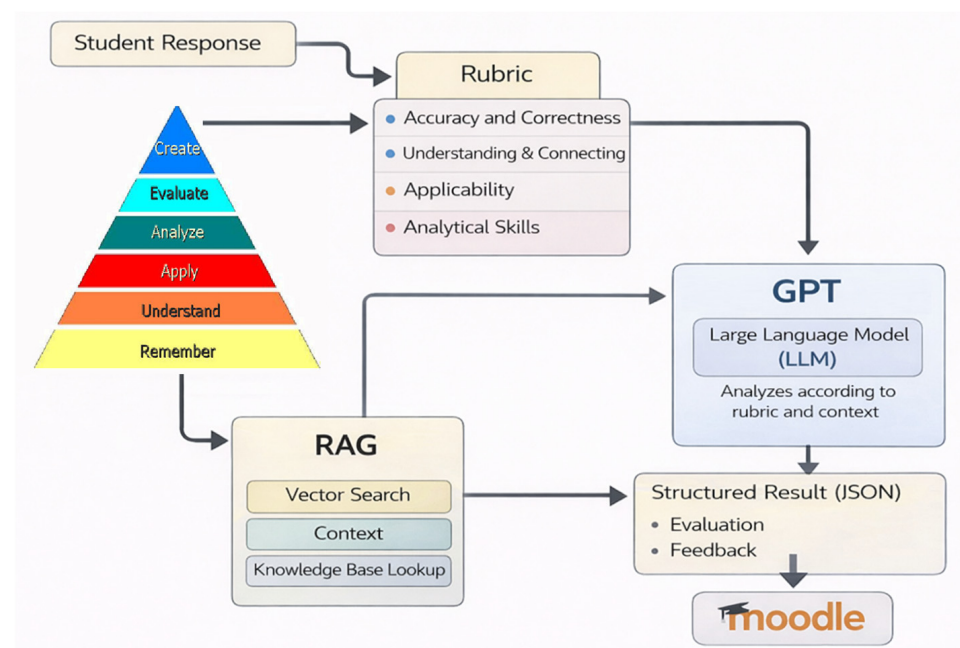


Figure 1. Relationship between Bloom's Taxonomy, the rubric, the RAG process and the language model.

This combination allows the cognitive operations described in Bloom's Taxonomy to be recognized in real student responses. The result is presented in a structured format (JSON) that can be directly saved in the LMS system.

3.1.2. Cognitive Levels as Semantic Patterns

Different cognitive operations leave characteristic linguistic traces in student responses. For example, the reproduction of definitions and key terms is typical of remembering, explanations, interpretations and paraphrases for understanding, the use of knowledge in a new context for application, and comparisons, cause-and-effect relationships and

argumentation for analysis and evaluation. The formulation of new ideas or solutions corresponds to the creation level.

Large language models can recognize such structures by analyzing the content and logical organization of the text. This allows the identification of the predominant cognitive operation in a given student response.

3.1.3. RAG as a Validity Maintenance Mechanism

While a language model can generate and interpret text, it is not reliable if it assesses “from memory”. RAG reduces this problem through semantic retrieval:

1. Embedding models find the most relevant passages from the course materials.
2. These passages are then provided to the LLM as an evidential basis.
3. The model is instructed to score only within this context and the rules of the rubric.

Thus, cognitive levels are assessed not in an abstract sense, but in terms of whether the student accurately understands the material studied in the course.

3.1.4. Linking Cognitive Levels to Rubric Components Through the LLM

In the example configuration used in this study, the rubric includes up to six criteria conceptually related to Bloom’s cognitive levels, each of which is formulated in such a way that it can be recognized and interpreted by the LLM. In other assignments, the rubric may contain only a subset of these components or may aggregate them into broader categories while maintaining the same cognitive logic.

The LLM recognizes these structures but makes its assessment within a context previously retrieved through RAG. This ensures that the assessment is:

- accurate (based on the materials),
- objective (follows a rubric),
- pedagogically valid (measures cognitive levels).

Interactions between cognitive theory and technological implementation

The interaction between the pedagogical framework and the technological architecture is systematized in Table 2.

Table 2. Interaction between Bloom, RAG, and LLM in automated assessment.

Component	Role	How It Contributes to Assessment
Bloom’s Taxonomy	Defines cognitive levels	Provides a pedagogical framework
Embeddings	Discover relevant context	Ensures accuracy and reduces hallucinations
RAG	Connects context + LLM	Ensures assessment is based on the material
LLM	Analyzes student response	Recognizes cognitive operations and assigns a grade

This builds a hybrid assessment model that combines:

- pedagogical theory,
- knowledge structure,
- semantic retrieval,
- generative analysis,
- and rubric-based assessment.

This integration is a key contribution of the methodology and represents an integrated approach to adaptive, cognitively oriented automated assessment.

3.2. Extended Rubric Assessment Model

The proposed model uses an extended but configurable rubric as the basis for automated assessment of open-ended student responses. Rather than relying solely on

superficial keyword matching, the rubric allows for scoring on separate criteria related to different aspects of the quality of the response. Combined with RAG, it provides pedagogical interpretability, transparency, and consistency of scoring.

3.2.1. Formal Definition of a Rubric

In the present system, a rubric is viewed as a structure composed of multiple criteria, multiple levels, and a scoring function. The multiple criteria describe the individual aspects on which a student response is analyzed. The multiple levels determine the level of performance for each criterion, and the scoring function connects the student response, the relevant context, and the active rubric into a structured score. The number of criteria is not fixed and can be adapted according to the type of task and its cognitive goal.

3.2.2. Six-Component Rubric Based on Bloom's Taxonomy

In the example configuration, the rubric includes six components, conceptually related to the cognitive levels in Bloom's Taxonomy. Through this structure, the system assesses not only the correctness of the content, but also the way in which the student understands, applies, analyzes, evaluates and creates. In practical implementation, the criteria are formulated as subject-oriented indicators, so that Bloom's Taxonomy functions as a conceptual framework, and not as a mandatory template for each task.

3.2.3. Criterion-Based Assessment and Level Selection

The assessment is carried out separately for each criterion of the rubric. The language model analyzes the student's response against the context extracted through RAG and determines the level of performance that best corresponds to the requirements of the criterion. In this way, the final result is formed as a set of individual criterion assessments, and not as a general impression. Such logic is consistent with findings that criterion-based and multidimensional scoring allows for a more precise distinction between different aspects of text quality compared to holistic assessments [3].

AI decision-making mechanism

The language model analyzes the student response against the provided context and the rubric criteria. It identifies the cognitive operations manifested in the text and interprets them within the framework of the corresponding criterion. The model compares these manifestations with the descriptions of the levels in the rubric (0, 3 and 5 points). The decision is made according to the logic:

1. Extract the manifested cognitive process in the text.
2. Match this process with the rubric criteria.
3. Select the level that best describes the quality of performance.

The complete AI-based assessment procedure is presented in Algorithm 1. It makes the assessment more transparent and reproducible, since the model works within a predefined pedagogical and algorithmic framework.

Result: (S,C,F)

The inclusion of a mapping between the rubric criteria and the cognitive levels of Bloom's Taxonomy allows the system to adapt the assessment to the type of cognitive activity required of the student. In this way, the language model does not evaluate the answers only at a formal or semantic level, but also takes into account the pedagogical context of the task.

Multidimensionality of assessment

A significant advantage of the rubric model is its multidimensionality. Instead of forming a single assessment based on a general impression, the system takes into account different aspects of cognitive activity and builds a more detailed profile of performance.

Algorithm 1: AI-based assessment using analytical rubrics and Bloom's Taxonomy.

Input:

- Student response: A
- Rubric: R
- Learning materials: M
- Mapping criteria to cognitive levels (Bloom): B

Output:

- Final grade: S
- Criteria scores: C
- Automatically generated feedback: F

Steps:

1. Obtain student response A from the Moodle system.
 2. Extract relevant learning context K from materials M using RAG.
 3. For each criterion k_i from rubric R:
 - 3.1. Determine the corresponding cognitive level using the function $B(k_i)$
 - 3.2. Construct an assessment prompt P_i using A, K, k_i and $B(k_i)$.
 - 3.3. Feed P_i to the language model.
 - 3.4. Obtain a point score s_i and argument j_i .
 4. Formation of criterion-level scores: $C = \{(k_i, s_i) \mid i = 1, \dots, m\}$.
 5. Calculation of raw score: $Score = \sum_{i=1}^m s_i$.
 6. Normalization of $Score$ relative to the maximum possible score.
 7. Aggregation of normalized grades into a final score S.
 8. Generation of criterion-specific feedback F based on the arguments $\{j_i\}$.
 9. Recording of the results (S,C,F) in the Moodle system.
 10. Controlled publication of the results after the exam is completed.
-

Additionally, the multidimensional structure allows for:

- diagnosis of the student's strengths and weaknesses;
- personalized feedback aimed at specific cognitive skills;
- accumulation of data on learning and development over time;
- comparison of cognitive profiles between different tasks, groups or semesters.

3.2.4. Normalization of the Assessment

After determining the points for the individual criteria, a raw score is formed, which is normalized to the maximum possible number of points on the rubric. This allows the results to be comparable between different tasks, rubrics and learning contexts.

3.2.5. Converting the Normalized Score into a Grading Scale

The normalized score can be converted to the grading scale used in the learning environment. In this case, a linear conversion to the Bulgarian six-point system is applied, which allows the result to be presented in a familiar and interpretable form for teachers and students.

This work presents the architecture and technical implementation of the system. The solution integrates the LMS platform Moodle, an n8n orchestration layer, semantic context retrieval via RAG and a large language model GPT, which evaluates the answers in terms of meaning in relation to specific learning materials and an analytical rubric. The system receives student submissions in real time, extracts the rubric definition from Moodle, feeds the model relevant context from the course, and records automatically generated grades

and feedback back to the platform. The architecture implements a fully automated and event-driven assessment process embedded in a predefined pedagogical model.

4. Results and Discussion

The experiment was conducted in the Moodle e-learning environment (accessible via the domain <https://moodleproject.eu>, accessed on 13 June 2026), equivalent to the real university infrastructure. The study was carried out within the framework of a remedial exam session in the discipline “Systems Engineering”, implemented in Bulgarian and English.

The exam was conducted in a controlled presence environment in a computer room. Students were given explicit instructions not to use external AI tools, and during the exam, constant control was exercised by the teaching team so that the collected answers reflected the real knowledge and skills of the students.

Student responses were scored automatically by the proposed AI system and independently by a teacher using the same analytical rubric, for the purpose of scientific comparison. The exam includes structured open-ended tasks requiring analytical thinking, argumentation, application of knowledge and UML modeling. Identical tasks, rubrics and assessment criteria were used in both language versions of the course.

For the purposes of the experimental analysis, data from 32 students were processed—13 from the Bulgarian course and 19 from the English course. Each participant solved an exam with five open-ended tasks assessed using analytical rubrics. The analysis was conducted at two levels: student level ($N = 32$) and task level (160 separate observations). The data were exported from the platform, checked for completeness and correctness, and anonymized using unique identifiers. The prepared database served as the basis for the subsequent analysis of the cognitive profiles.

In this section, two working hypotheses are tested:

H1 (Cognitive Sensitivity). *The system demonstrates differentiated behavior towards the different cognitive levels of Bloom’s Taxonomy.*

H2 (Pedagogical Value). *The generated feedback is clear, criterion-related and useful for learning.*

4.1. Analysis of the Results by Cognitive Levels

The present analysis aims to examine the behavior of the automated system towards the different cognitive levels of Bloom’s Taxonomy. For this purpose, each sub-criterion of the analytical rubric was previously mapped to a dominant cognitive level based on expert-defined pedagogical rules. During the assessment, the system automatically uses this mapping to aggregate the results by cognitive levels and form individual and group cognitive profiles. The analysis was conducted on the automatically generated assessments in order to examine the internal consistency and the assessment profile of the system.

4.2. Bloom Mapping Method

During the assessment, the system automatically identifies the activated cognitive levels for each criterion by analyzing the content of the student’s response and the degree of coverage of the relevant requirements. For each level, a binary or graded indication (✓, ~, ✗) is generated, reflecting full, partial or missing coverage.

The results of this analysis are used to:

- determine the dominant cognitive level of each sub-criterion;
- aggregate scores by level for each task;
- form individual and group cognitive profiles;
- generate extended feedback to students.

The Bloom mapping process is performed completely automatically by the system, without manual intervention by the teacher, which ensures consistency, reproducibility and independence from subjective interpretations.

The obtained data are stored in a structured form and used for subsequent statistical analysis, including calculating averages, standard deviations and coverage of maximum scores by cognitive levels.

4.3. General Bloom's Profile

The indicators Pct_5, Pct_3 and Pct_0 (see Table 3) reflect the relative share of cases in which the system awards a maximum (5 points), average (3 points) and zero (0 points) score for a given cognitive level, respectively. They allow tracking the extent to which the system distinguishes between complete, partial and missing mastery of the relevant cognitive skills.

Table 3. Distribution of scores by cognitive levels (total).

Cognitive Level (Bloom)	N	Mean	Standard Deviation	Pct_5	Pct_3	Pct_0
Remembering	380	3.38	2.33	0.67	0.01	0.32
Understanding	362	3.24	2.31	0.61	0.06	0.33
Applying	323	3.20	2.33	0.60	0.06	0.34
Analyzing	210	3.03	2.22	0.50	0.18	0.32
Evaluating	96	2.68	2.03	0.32	0.35	0.32
Creating	20	3.35	2.13	0.55	0.20	0.25

Figure 2 presents the mean scores of the automated system, aggregated by cognitive levels for all tasks. The results show a relatively even distribution of the mean values in the range of approximately between 2.6 and 3.4 points on a scale from 0 to 5. The highest mean values are observed at the Remembering and Creating levels, and the lowest at Evaluating. This indicates that the system demonstrates higher confidence in assessing tasks related to remembering, understanding and applying, while in tasks requiring critical judgment, it exhibits more conservative behavior.

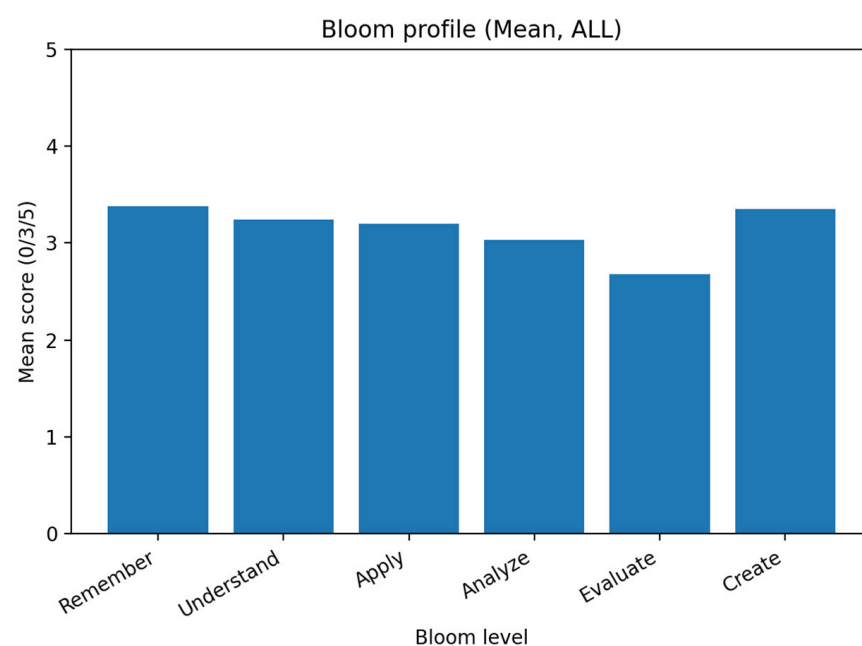


Figure 2. Mean values by cognitive levels (total).

4.4. Bloom's Task Profiles

Figure 3 presents the mean cognitive profiles separately for the five exam tasks. The task analysis shows that the system exhibits different sensitivity depending on the cognitive complexity of the specific task. Higher and more balanced results are observed for tasks oriented towards understanding, application and structural analysis. More conservative behavior is observed for tasks requiring more sophisticated argumentation, critical evaluation or interpretation of logical dependencies, especially for tasks requiring the creation of a diagram (image). These observations are consistent with the general cognitive profile of the system and confirm that the assessment is sensitive to the nature of the task.

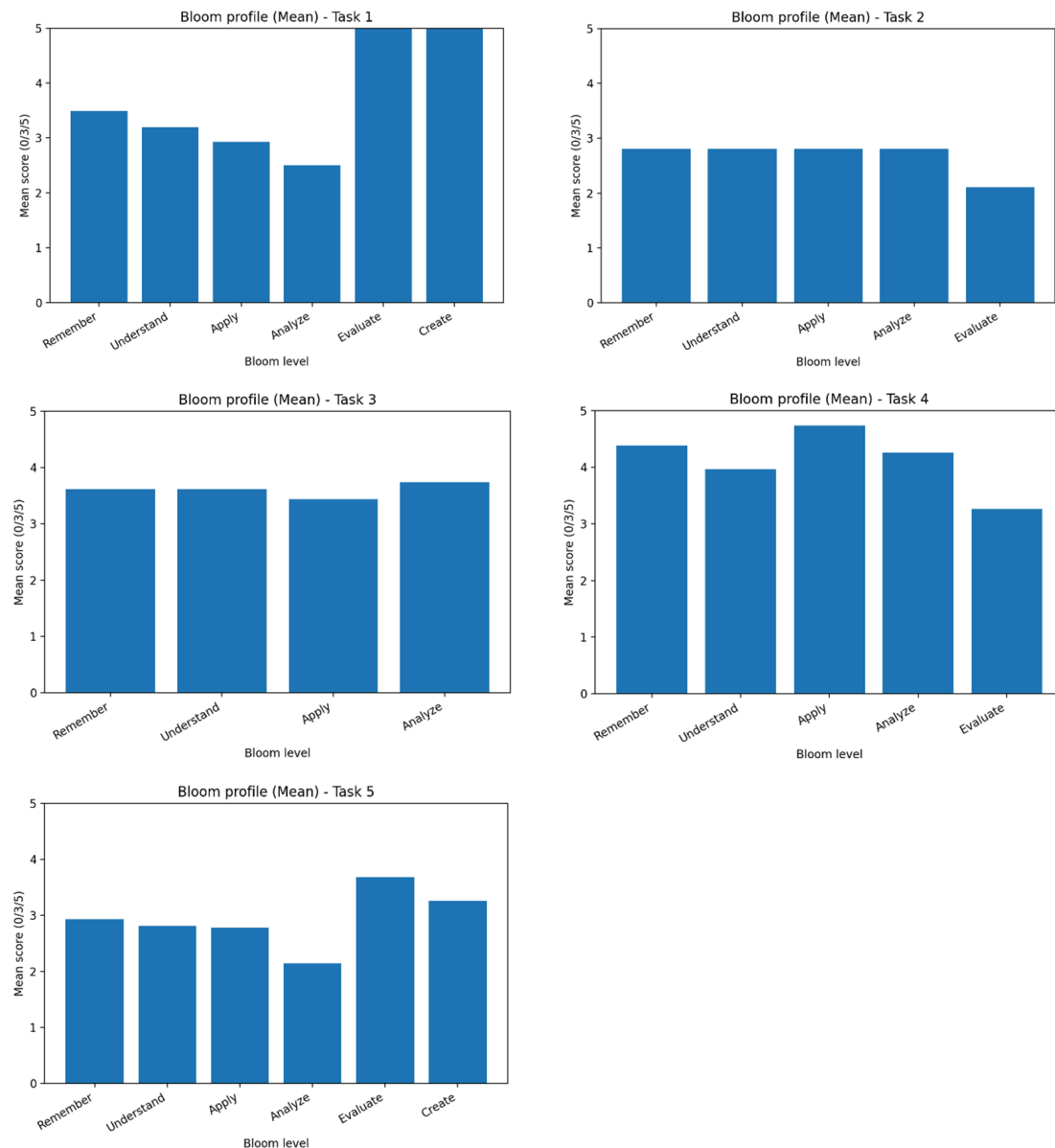


Figure 3. Bloom's task profiles (mean).

4.5. Coverage Analysis

This analysis examines the percentage of cases in which the system awards the maximum score (5 points) for a given cognitive level. As shown in Figure 4, the highest percentage of full ratings is observed in Remembering, Understanding and Applying, while Evaluation demonstrates the lowest coverage. This indicates that the system applies a stricter criterion for tasks requiring critical thinking.

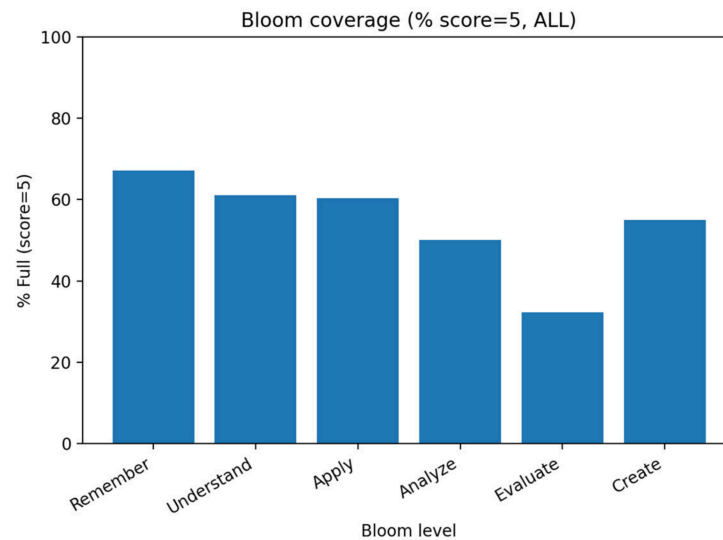


Figure 4. Coverage of maximum scores (overall).

4.6. Assessment of the Pedagogical Effectiveness of Automatically Generated Feedback (RQ2)

The pedagogical value of automatically generated feedback was assessed through expert analysis of a subsample of generated messages on three indicators: clarity, usefulness and accuracy, measured using a five-point Likert scale. For each indicator, mean value and standard deviation were calculated. The resulting descriptive statistics are presented in Table 4.

Table 4. Mean values and standard deviations of feedback quality indicators.

Indicator	Mean $\pm$ Standard Deviations
Clarity	4.70 $\pm$ 0.49
Usefulness	4.67 $\pm$ 0.50
Accuracy	4.88 $\pm$ 0.37

The results obtained show high mean values and low variability for all three indicators, which indicates a stable quality of the feedback generated. The highest value is observed for the correctness indicator, which indicates good coherence between the feedback, the analytical rubric and the learning context. High values for clarity and usefulness indicate that the generated comments are understandable, structured and practically oriented. These results support H2 and show that the system can effectively support learning through pedagogically oriented feedback.

4.7. Conservative Behavior at Higher Cognitive Levels: Interpretation and Future Directions

An additional finding emerging from the analysis concerns the system's more cautious behavior at the higher cognitive levels of Bloom's taxonomy.

The more conservative behavior observed at higher cognitive levels should not be interpreted as an effect of RAG alone. In the proposed architecture, assessment is intentionally implemented as evidence-constrained evaluation, in which the model is restricted to three input sources: the student response, the RAG-retrieved course-specific context, and the predefined analytical rubric. This design increases transparency, reduces hallucinations, and ensures that scoring remains grounded in the instructional material rather than in the model's general world knowledge. Within such a setting, conservative scoring at higher Bloom levels is partly a consequence of methodological control.

At the same time, stronger grounding may reduce the model's freedom in tasks that require more open critical judgment, especially when the rubric demands explicit evidence

for evaluative claims. This effect appears to be more visible for the Evaluating level than for Creating. In the reported results, Evaluating has the lowest mean score, whereas Creating shows a relatively higher mean, although it is represented by a much smaller number of observations. Therefore, the observed pattern is likely influenced not only by contextual grounding, but also by rubric strictness, task design, and the uneven distribution of cases across cognitive levels.

This interpretation suggests an important direction for future work. A controlled comparison between more deterministic grading settings and moderately more flexible decoding settings, such as temperature or top-p adjustment, could clarify whether a limited increase in generative freedom improves sensitivity at the higher cognitive levels. Such an extension should be examined carefully, as greater flexibility may enhance responsiveness to nuanced and original answers, but may also reduce reproducibility, interpretability, and alignment with the retrieved instructional context.

5. Conclusions

A comprehensive methodological framework for automated open-ended question assessment was developed, based on the integration of pedagogical theory, a formalized rubric structure, and modern language models, supported by semantic retrieval through RAG. The methodology is based on Bloom's Taxonomy as a conceptual basis for structuring cognitive levels and transforms these levels into operational assessment criteria. The six-component rubric, based on Bloom's Taxonomy, allows for multidimensional assessment, which takes into account not only factual correctness, but also the quality of cognitive operations—understanding, applying, analysis, evaluation, and creation. In this way, assessment is transformed from content verification to analysis of thinking activity.

The results of the experimental study show that the system exhibits differentiated behavior across cognitive levels and demonstrates higher sensitivity in tasks related to remembering, understanding, and applying, while more conservative behavior is observed particularly at the evaluating level and, more broadly, in tasks requiring complex justification and higher-order critical judgment. The automatically generated feedback shows high values in terms of clarity, usefulness and correctness, which confirms its pedagogical value. The main contribution of the proposed architecture lies in the integration of Bloom, RAG and LLM into a single, controlled and technologically feasible model for automated assessment in a real e-learning environment. Future work could be aimed at expanding the experiment to other disciplines, with larger samples and different types of tasks, in order to further verify the robustness and generalizability of the approach.

Author Contributions: All authors were involved in the full process of producing this paper, including conceptualization, methodology, modeling, implementation, validation, and preparing the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been accomplished with financial support from the European Regional Development Fund within the Operational Program “Bulgarian national recovery and resilience plan”, procedure for direct provision of grants “Establishing of a network of research higher education institutions in Bulgaria”, and under Project BG-RRP-2.004-0005 “Improving the research capacity and quality to achieve international recognition and resilience of TU-Sofia (IDEAS)”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Grévisse, C. LLM-based automatic short answer grading in undergraduate medical education. *BMC Med. Educ.* **2024**, *24*, 1060. [[CrossRef](#)] [[PubMed](#)]
2. Sychev, O.; Anikin, A.; Prokudin, A. Automatic grading and hinting in open-ended text questions. *Cogn. Syst. Res.* **2020**, *59*, 264–272. [[CrossRef](#)]
3. Tang, X.; Chen, H.; Lin, D.; Li, K. Harnessing LLMs for multi-dimensional writing assessment: Reliability and alignment with human judgments. *Heliyon* **2024**, *10*, e34262. [[CrossRef](#)] [[PubMed](#)]
4. Aydın, B.; Kışla, T.; Elmas, N.T.; Bulut, O. Automated scoring in the era of artificial intelligence: An empirical study with Turkish essays. *System* **2025**, *133*, 103784. [[CrossRef](#)]
5. Wang, Y. E-learning application in english writing classroom based on neural machine translation and semantic analysis algorithms. *Entertain. Comput.* **2024**, *51*, 100730. [[CrossRef](#)]
6. Pack, A.; Barrett, A.; Escalante, J. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100234. [[CrossRef](#)]
7. Pecuchova, J.; Benko, L.; Drlik, M. Automated Grading of Open-Ended Questions in Higher Education Using GenAI Models. *Int. J. Artif. Intell. Educ.* **2025**, *35*, 3813–3846. [[CrossRef](#)]
8. Jauhainen, J.S.; Guerra, A.G. Evaluating Students' Open-ended Written Responses with LLMs: Using the RAG Framework for GPT-3.5, GPT-4, Claude-3, and Mistral-Large. *Adv. Artif. Intell. Mach. Learn.* **2024**, *4*, 3097–3113. [[CrossRef](#)]
9. García-Varela, F.; Nussbaum, M.; Mendoza, M.; Martínez-Troncoso, C.; Bekerman, Z. ChatGPT as a Stable and Fair Tool for Automated Essay Scoring. *Educ. Sci.* **2025**, *15*, 946. [[CrossRef](#)]
10. Shermis, M.D. Using ChatGPT to score essays and short-form constructed responses. *Assess. Writ.* **2025**, *66*, 100988. [[CrossRef](#)]
11. Del Gobbo, E.; Guarino, A.; Cafarelli, B.; Grilli, L.; Limone, P. Automatic evaluation of open-ended questions for online learning. A systematic mapping. *Stud. Educ. Eval.* **2023**, *77*, 101258. [[CrossRef](#)]
12. Mendonça, P.C.; Quintal, F.; Mendonça, F. Evaluating LLMs for Automated Scoring in Formative Assessments. *Appl. Sci.* **2025**, *15*, 2787. [[CrossRef](#)]
13. Bloom, B.S.; Engelhart, M.D.; Furst, E.J.; Hill, W.H.; Krathwohl, D.R. Taxonomy of educational objectives: The classification of educational goals. In *Handbook I: Cognitive Domain*; David McKay: New York, NY, USA, 1956.
14. Anderson, L.W.; Krathwohl, D.R. (Eds.) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*; Longman: New York, NY, USA, 2001.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.