A language model-based recommender assessment system for short-answer questions in the intellectual property domain

Recomendador de evaluación para preguntas cortas utilizando modelos de lenguaje en propiedad intelectual

David Bañeres Besora - *Universitat Oberta de Catalunya, UOC (Spain)*
Ana-Elena Guerrero Roldán - *Universitat Autónoma de Barcelona, UAB (Spain)*
M. Elena Rodríguez González - *Universitat Oberta de Catalunya, UOC (Spain)*

ABSTRACT

The use of Artificial Intelligence (AI) in education is growing rapidly, transforming the teaching-learning process as well as the assessment process. This work introduces SLASys, a tool to recommend the assessment of short-answer questions using semantic AI techniques. Unlike other works based on generative AI, SLASys uses the lightweight BERT language model, which better understands specific domain language concepts, improves computational efficiency, and reduces ethical and privacy concerns. SLASys implements semantic comparison and predictive classification models based on BERT. A mixed research methodology was followed, combining action research with a design and creation approach, to develop and refine SLASys over four editions of a master's-level course on patent examination within the intellectual property domain. SLASys has been integrated into Moodle, enabling its use by teachers without technical expertise, and has been tested by 120 students. The results demonstrate its effectiveness even with small datasets and limited participants within the described experience and according to existing literature. Additionally, it has been positively evaluated by both teachers and students. This work shows the feasibility of using AI in higher education, in both hybrid and online environments, offering a practical solution to improve assessment and feedback for short-answer questions in real learning contexts.

Keywords: assessment; feedback; Moodle; test; short-answer; artificial intelligence.

RESUMEN

El uso de la Inteligencia Artificial (IA) en educación está creciendo rápidamente, transformando el proceso de enseñanza-aprendizaje y también el proceso de evaluación. Este trabajo presenta SLASys, una herramienta para recomendar al profesorado la evaluación de preguntas cortas mediante técnicas de IA semántica, difiriendo de otros trabajos basados en IA generativa por el uso del modelo de lenguaje BERT que es más ligero, comprende mejor los conceptos en un contexto específico, mejora la eficiencia computacional y reduce los problemas éticos y de privacidad. SLASys implementa comparación semántica y modelos predictivos de clasificación de respuestas basados en BERT. Se ha seguido una metodología de investigación mixta, combinando investigación de acción con un enfoque de diseño y creación, para desarrollar y perfeccionar SLASys a lo largo de cuatro ediciones de un curso de nivel de máster sobre examen de patentes en el contexto de la propiedad intelectual. SLASys se ha integrado en Moodle, permitiendo su uso por parte de profesorado sin conocimientos técnicos, y ha sido probada por 120 estudiantes. Los resultados muestran su efectividad, tanto en el marco de la experiencia descrita como según la literatura existente, incluso con conjuntos de datos reducidos y un número limitado de participantes, y ha sido valorada positivamente por el profesorado y el estudiantado. Este trabajo contribuye a mostrar la viabilidad del uso de la IA en la educación superior, tanto en entornos híbridos como en línea, ofreciendo una solución para mejorar la evaluación y el *feedback* en preguntas cortas en contextos reales de aprendizaje.

Palabras clave: evaluación; *feedback*; Moodle; test; preguntas cortas; inteligencia artificial.

How to cite: Bañeres Besora, D., Guerrero Roldán, A.-E., & Rodríguez González, M. E. (2026). A language model-based recommender assessment system for short-answer questions in the intellectual property domain. [Recomendador de evaluación para preguntas cortas utilizando modelos de lenguaje en propiedad intelectual]. *RIED-Revista Iberoamericana de Educación a Distancia, 29*(1), 321-352. https://doi.org/10.5944/ried.29.1.45541

INTRODUCTION

Nowadays, educational institutions (including traditional and online) generally adopt information and communication technologies to support their teachers and students. Automating the grading task is one trend followed in many Learning Management Systems (LMS) for efficiently assessing activities and providing timely feedback for formative purposes (György & Vajda, 2007). However, many LMS limit the type of questions to be automatically graded to choice-based or simple text due to the complexity of assessing short-answer questions and essay-type questions.

Although Automatic Short-Answer Grading (ASAG) has been extensively explored in the literature (Burrows et al., 2015), contributions were limited to exploring their accuracy on public datasets without considering their effectiveness in real educational settings. Additionally, the resurgence of Artificial Intelligence (AI) due to the accessibility to computational resources and specific tools for using AI techniques has evolved classic LMS. AI algorithms can power LMS to support the learning process. Recommendations, immediate feedback, or ASAG, are examples of tasks that AI can enhance.

This research aims to provide a starting point for a smart learning system capable of recommending assessment and feedback for short-answer questions using AI techniques. This paper shows how the system is used in a hybrid educational setting and presents the obtained findings. Despite Large Language Models (LLM) and derived generative tools, such as ChatGPT (OpenAI, 2024) currently showing their broad potential in any academic task, they still have limitations for specific tasks or domains like Intellectual Property (IP). Therefore, this work proposes an alternative solution, using a lightweight BERT Language Model (LM) that understands the answers' meaning. Instead of delegating the assessment decision for a short-answer question to a Generative AI tool (GenAI) by defining the assessment criteria as a complex prompt, this work uses BERT, which understands the meaning of any correct and incorrect answer, and can decide whether a student's answer is correct. This latter model is currently more accurate for assessment purposes. It is smaller, contributing to efficient computational and sustainable use; and it can be handled in private instances (i.e., no enterprise solutions are needed), reducing ethical and privacy issues. Thus, this work integrated this AI solution into a Moodle LMS to make it easy and usable for teachers to provide a meaningful assessment to students.

The action research methodology combined with a design and creation approach is used to specify the Smart Learning Automated System (SLASys), explicitly suited for an IP training course, to define a learning system to help students succeed in their learning process using AI techniques for IP training. The study has been carried out in the European Patent Office (EPO) educational area.

This paper is structured as follows. First, it presents the theoretical background of language models and their educational uses, and focusing on ASAG. Then, it details the research design, including methodology, research questions, the SLASys tool, and data analysis. The results section highlights key findings, and the paper concludes with a discussion, limitations, conclusions, and future research.

LITERATURE REVIEW

Large Language Models, BERT, and their application in education

LLMs have emerged as a transformative technology in recent years, demonstrating remarkable capabilities across Natural Language Processing (NLP) tasks. Their rapid progress has led to significant advancements in machine translation, text summarization, and even code generation (Husein et al., 2025; Pang et al., 2025; Zhang et al., 2025). Despite their impressive performance, LLMs face limitations and challenges in real educational settings, such as misinterpretations in specific domain languages (Arefeen et al., 2024) or their interpretability and explainability (Zhao et al., 2024). Recent architectural innovations have improved their performance, reaching the state-of-the-art GenAI tools, such as GPT-4o (OpenAI, 2024). Such tools demonstrate impressive learning capabilities, applicability to limited reasoning tasks, and the ability to generalize across diverse task domains.

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art NLP model developed by Google in 2018 (Devlin et al., 2019). BERT is a transformer-based language model pre-trained on a large collection of textual documents. BERT was considered an LLM but is currently considered an LM compared to its GenAI successors. Its architecture relies on text representation. A sentence is encoded in a special way (i.e., *embeddings*) that stores its semantic meaning, enabling the straightforward application in NLP tasks. Its continued development and application in multiple knowledge domains demonstrate versatility and potential for further advancements in AI and language understanding. BERT has been used for document classification (Adhikari et al., 2019), grammatical category recognition (Souza et al., 2019), logical entailment between pairs of sentences (Wang et al., 2018), or text summarization (Zhang, Cai et al., 2019).

Both BERT and GenAI are being used in education with similar challenges. Scalability and integration with existing infrastructures require significant investment and technical expertise (Xu & Zhu, 2023). Additionally, principles concerning ethics, security, and alignment with core educational objectives must be considered (García-Peñalvo et al., 2024) and developing AI literacy among teachers is essential for fostering critical use (Petridou & Lao, 2024). Moreover, although they seem applicable in any task, GenAI tools are mostly applicable to text generation and analysis, such as concept definition, exercise generation, or improving writing style; while BERT is valid for text classification, sentiment analysis, understanding students' doubts, or providing detailed explanations (Qiu & Jin, 2024).

Assessment is also a relevant task for both models. On the one hand, GenAI tools have started to be explored in assessment as “black boxes” without knowing how they internally work. They have been used to produce feedback, but also for automatic assessment. Automated assessment benefits some domains, such as language learning (Escalante et al., 2023), writing skills (Banihashem et al., 2024), or multimedia designs (Almasre, 2024). On the other hand, BERT can be used on specific assessment tasks. Teachers can generate questions from existing texts, which can be used for tests and exams (Nguyen et al., 2022). BERT models can assess student essays' coherence, relevance, and grammar, helping students improve their writing skills (Wang et al., 2022). BERT-based applications were also designed to understand idiomatic expressions and cultural nuances (Bahdanau et al., 2015), and specific recommenders

can help students practice speaking and comprehending new languages (Zhang, Zhang et al., 2019).

Although automated assessment seems to be one of the future features of GenAI in education, currently, it is not recommended by institutional or even governmental policies (European Commission, 2024; Dai et al., 2025; González Fernández et al., 2025), because GenAI tools are error-prone due to hallucinations (Jia et al., 2024) and their utilization might raise ethical issues. Additionally, GenAI tools may fail on domains with specific language. For instance, some questions related to specific concepts in patent examination in the IP domain, such as Novelty and Clarity, can sometimes result in a GenAI tool interpreting the question as whether the answer is new and clear in natural language. However, such IP-specialized concepts refer to how novel and clear an invention is described in a patent application (or *claim*). GenAI tools could analyze the answer by improving previous context (i.e., *contextualization*), but they sometimes mix patent guidelines from different sources with outdated information. Thus, this requires adding specific documents to the model (i.e., *fine-tuning*) to improve the quality of a proprietary GenAI model, which implies some privacy concerns and discussion about whether it is a cost-effective solution. BERT-based application addresses the assessment task by understanding the answers provided to a question. BERT could be used by performing *semantic comparison* to assess how close the meaning of two sentences are (i.e., the student's answer compared to the teacher's correct answer) or train an *answer classification* AI model to predict whether a sentence meaning is close to a correct or incorrect set of sentences (i.e., decide whether the student's answer is correct or incorrect depending on a set of correct and incorrect answers). Such approaches avoid the variability drawback that GenAI tools currently exhibit.

Automated short-answer questions grading and feedback provision

To assess the acquisition of knowledge, skills, and competences, feedback becomes one of the essential components in the teaching-learning process (Evans, 2013; Hattie & Timperley, 2007). Feedback is a key element to engage students and enhance their learning (Winstone et al., 2017). It aims to provide specific information related to a learning activity to fill the gap between the desired and the actual understanding (Sadler, 1989). Additionally, feedback is crucial in online settings to guarantee a meaningful and timely learning process (Nicol & Macfarlane-Dick, 2006). Elaborating and delivering valuable feedback requires a considerable teacher's effort, which limits its personalization when there are time constraints or a big number of enrolled students (Dhananjaya et al., 2024). However, personalized feedback motivates students to a higher level, better regulates their behavior, contributes to better knowledge acquisition, and enhances engagement with learning materials (Kim, 2023; Wang & Lehman, 2021).

Automating feedback generation has been sought as a tentative solution to support teachers. Although different terminology has been used for denoting such tools in the literature, e.g., intelligent tutoring system, teaching platform, automated feedback system, online feedback tools; all of them focus on enhancing teachers' efficiency while helping students during their learning process (Xie & Li, 2018). Previous works have presented applicable techniques, such as feedback generation for essays (Akçapinar, 2015), for programming exercises (Messer et al., 2024), or collaborative tasks (Zheng et al., 2023).

Automated test grading has also been extensively researched due to its benefits for providing automatic but also formative assessment (Burrows et al., 2015), and the positive impact on learning due to feedback provision (Gaona et al., 2018; Rezaei, 2015). Tests have been mostly simplified using choice-based questions, and short- and essay-based questions were usually evaded since they required higher teacher effort and intervention during the assessment phase, or alternative methods were used, such as peer review (Huisman et al., 2017). However, there is previous work related to ASAG. Answer comparison against one sentence (Siddiqi & Harrison, 2008), against a set of sentences with similar vocabulary (Klein et al., 2011), or against primary keywords (Saha et al., 2018) have been used to detect correct answers. Although these techniques have evidenced good results, they have some deficiencies since semantics is not considered (i.e., one pattern, similar vocabulary, or a set of keywords may not gather all possible correct answers). With GenAI tools, authors investigated their applicability to the ASAG (Aggarwal et al., 2025; Grévisse, 2024). Although its applicability for essay grading seems promising (Senthilnathan et al., 2025), the ASAG results are inconclusive for supporting students' assessment. Hallucinations, decontextualization, outdated information, and limited critical thinking still affect their utilization (De La Cruz Martínez et al., 2024).

Since NLP and semantic LM models are still state-of-the-art techniques for ASAG, this work used BERT because of its capabilities, potential accuracy, and better fit for real educational settings. Two types of models are explored in the literature. Some models seek to predict whether the student's answer is correct (Camus & Filighera, 2020; Liu et al., 2019; Lun et al., 2020; Padó et al., 2024; Schneider et al., 2023; Sung et al., 2019; Wang et al., 2019), while others go further by predicting the grade, demonstrating their potential on public datasets (Baral et al., 2021; del Gobbo et al., 2023; Gaddipati et al., 2020; Metzler et al., 2024; Soulimani et al., 2024). Still, as far as the authors of this research know, no work has tested such tools in real educational settings due to acceptability issues or human intervention needs for their utilization (Hustad & Arntzen, 2013; Xavier et al., 2025). Therefore, this work proposes SLASys as a recommender tool for ASAG, which has been tested in a hybrid educational environment during four editions of a fundamental patent examination course over a year, with 120 students, to gain insights about their utilization. Thus, this research aims to answer the following research questions:

RQ1. How accurate is answer classification compared to semantic comparison for ASAG recommendation?
RQ2. How accurate is SLASys in a real educational setting?
RQ3. Can SLASys be used without previous students' answers?
RQ4. What is the opinion of the students and teachers?

METHODOLOGY

Research design

The SLASys system employs a mixed research methodology that combines action research (Oates, 2006) with a design and creation approach (Kuechler & Vaishnavi, 2012). Action research is appropriate because the system is developed and tested within real educational settings, aiming to address practical challenges in teaching and learning. This methodology emphasizes solving real-world problems through a collaborative, iterative cycle of planning, acting, and reflecting, while collecting data to evaluate outcomes. The design and creation approach complements this by focusing on the development of innovative technological artifacts. It follows a five-step, iterative problem-solving process (Kuechler & Vaishnavi, 2012) to guide the creation and refinement of the system.

This paper focuses on the first two iterations of developing the recommender for ASAG. The first one involved the creation of the first artifact and evaluating the technical solution, while the second evaluated its integration into the teaching-learning process. This last iteration involved testing the artifact in a fundamental course in patent examination for the IP domain. This course, with a duration of 6 weeks, is part of the courses oriented towards the students' training for patent examiners, which comprises six courses with a 2-year duration. There are specific requirements to be eligible for this training. Students should have a master's degree in physics, chemistry, engineering, or natural sciences since knowledge of specific technical fields is required to examine patent application proposals.

The course objective is to prepare the students for the patent examination by introducing the fundamental concepts, the European Patent Convention (EPC) guidelines, systems available for the task, and exercises related to patent examination. The learning methodology combines a flipped classroom strategy (Bergmann & Sams, 2012) with just-in-time teaching (Novak, 2012). Self-learning activities and reading learning resources are performed individually offline. There are also synchronous online sessions with all the students and the teacher, where offline activities are discussed, and practical exercises, questions, and group activities are performed.

Teachers reported issues applying just-in-time teaching since they did not have enough time to assess the students' short-answer activities and provide meaningful feedback before the online sessions. Thus, SLASys answers these needs by providing a solution to efficiently assess and give feedback to each student.

Figure 1
Research design

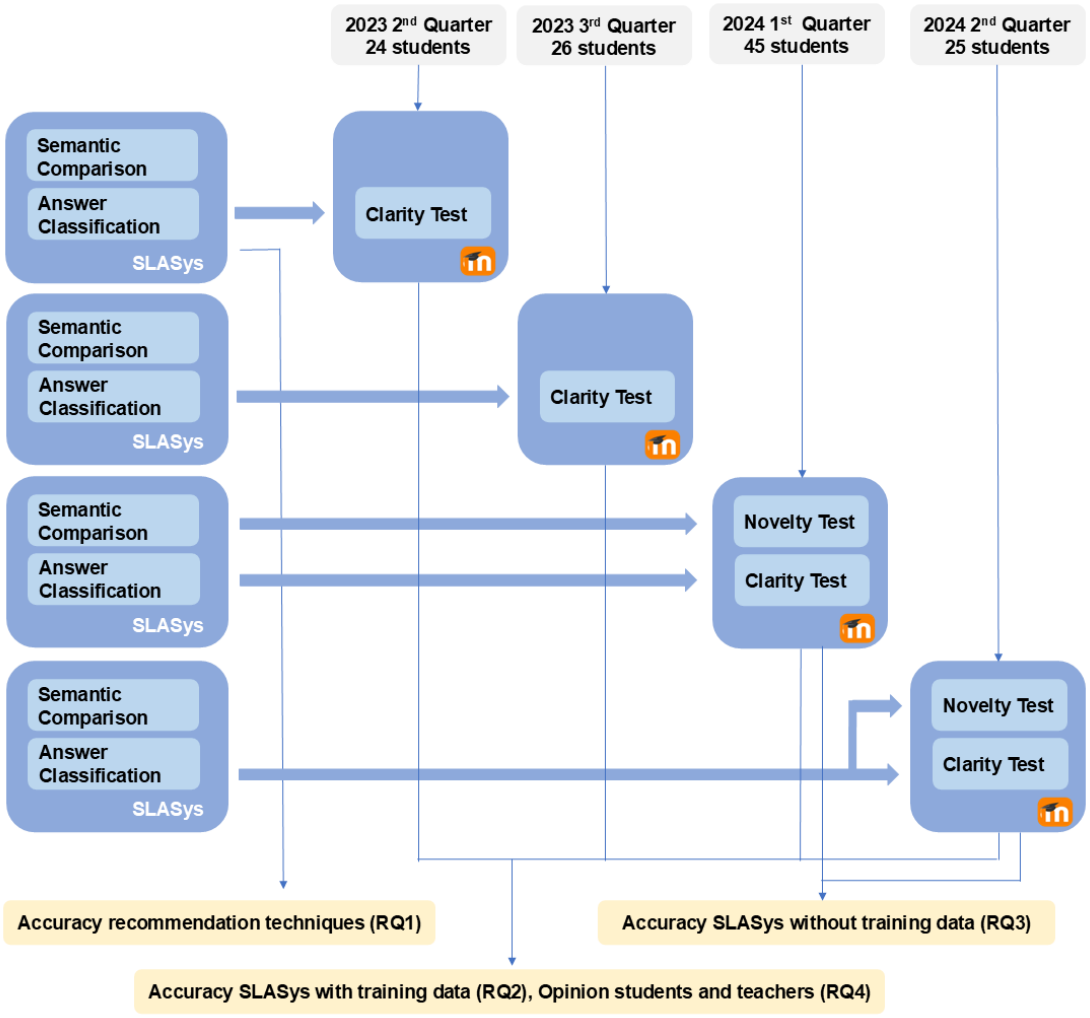


Figure 1 shows the research design. First, the designed approaches are compared to see their accuracy for answering RQ1. Second, during four-course editions, the answer classification approach was tested in a test related to the patent Clarity concept for answering RQ2. In total, 120 students interacted with the tests. Third, a new test concerning the Novelty concept was designed in the last two editions without any training data (i.e., previous students' answers). In this case, the two designed approaches were used to assess the performance of the recommender to answer RQ3. In this case, 70 students answered this second test. Additionally, the students' opinions and the teachers' experiences were collected across the editions to answer RQ4.

The testing with real students was performed online synchronously with a custom Moodle LMS where SLASys was used to recommend the assessment results. The Clarity and Novelty tests comprised 8 and 11 questions, respectively, and students had one hour to answer them. Note that they aim to improve students' knowledge but not students' grades. Students are assessed twice during the complete formation with onsite exams.

SLASys: A Recommender for ASAG

SLASys was designed to find whether a student's short answer is correct, depending on the available information related to the question. Figure 2 shows the two techniques used to detect whether an answer is correct. The question is about the concept of Novelty in the IP domain, asking why the claim is not novel. The comparison is performed by using BERT *embeddings*. Although an embedding is a complex mathematical matrix, it could be seen as an encoding of the sentence's meaning by a set of symbols that define their meaning.

Figure 2
Question example about Novelty

Why is this claim not novel?
A (mobile) phone with dialling means, a microphone and a loudspeaker, characterised in that the dialling means are made from aluminium ...

Golden answer
The claim is unclear because of the bracketed expression (mobile). It is unclear whether the feature in brackets should be limiting or not. See Guidelines F-IV, 4.19.
( ,  ,... , )

	Embedding	Cosine similarity	Probability of being classified correct
Answer 1 (correct) Feature in brackets which is not a reference sign it is not clear whether this is optional or not (F.-IV, 4.19)	( ,  ,... , )	32.2%	99.9%
Answer 2 (incorrect) Expression in brackets are not allowed (I'm not sure if it's falling under the exemption in this technical field).	( ,,... ,)	23.3%	0%

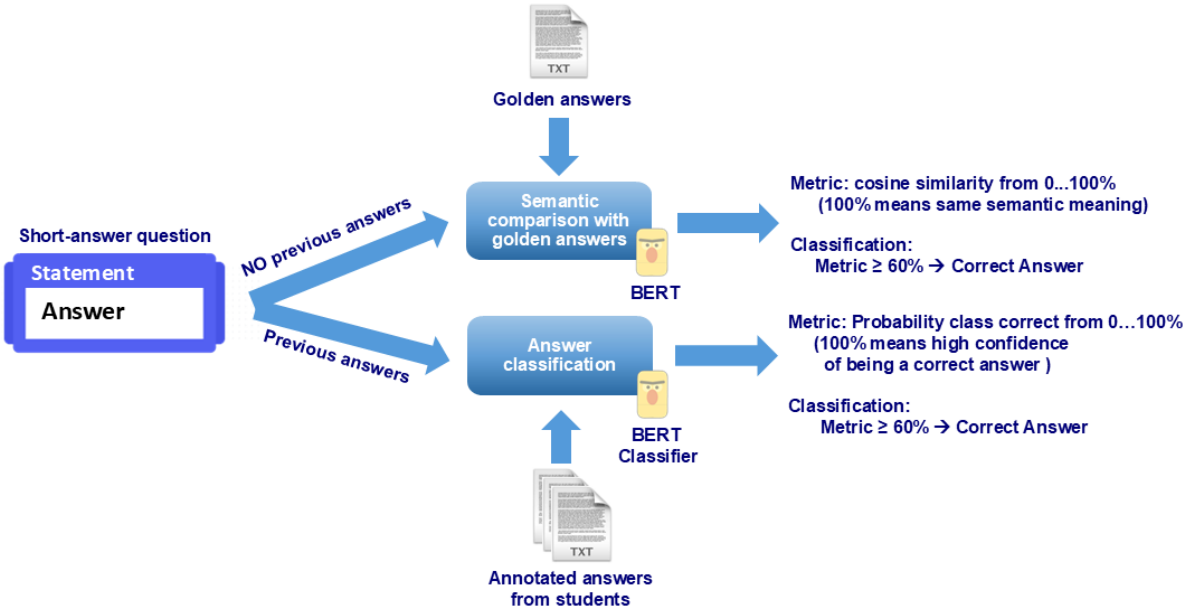
The first technique uses the correct answer embedding provided by teachers, denoted as the *golden answer*, to compare with the answers provided by students. The *cosine similarity* metric is used to evaluate how similar the student's sentence is to the golden answer (i.e., how many symbols are equal in both embeddings). The comparison output (i.e., *metric score*) is a similarity probability, where 100% means the same meaning as the golden answer. Figure 2 shows that this method fails for the correct answer, since there are different correct answers, and the golden answer does not cover all of them.

However, this process could be enhanced by comparing students' answers with a set of correct and incorrect answers (i.e., extracted from real students of previous course editions after being assessed by teachers), denoted as *annotated answers*. Each answer can be annotated, whether it is correct or incorrect. Although the comparison can be done by cosine similarity for each annotated answer, there is a more efficient technique that involves training an AI classifier and fine-tuning based on the embeddings of the annotated information. Fine-tuning specializes the classifier to focus on the specific annotated answers gathered for a question. The model's output is

the probability of the embedding being classified as a correct answer, where 100% means the model is highly confident that the student's answer is correct.

Since a short-answer question could be created from scratch or reused from other learning contexts, the system considers these two possible scenarios (Figure 3). Semantic comparison is used when the question is created and no previous student data are available. The comparison uses cosine similarity with a set of correct answers provided by teachers. The answer classification trains an AI classifier when previous answers are available, providing a prediction of being correct. Since both metrics are percentages, a threshold is defined to decide whether a student's submission is a correct answer. An optimization problem solves the quality threshold identification during performance analysis since the objective is to maximize the threshold while correct classification does not worsen significantly. After some experiments and teachers' validation as domain experts, it was concluded that a threshold greater than 60% mostly identified correct answers.

Figure 3
SLASys technical design

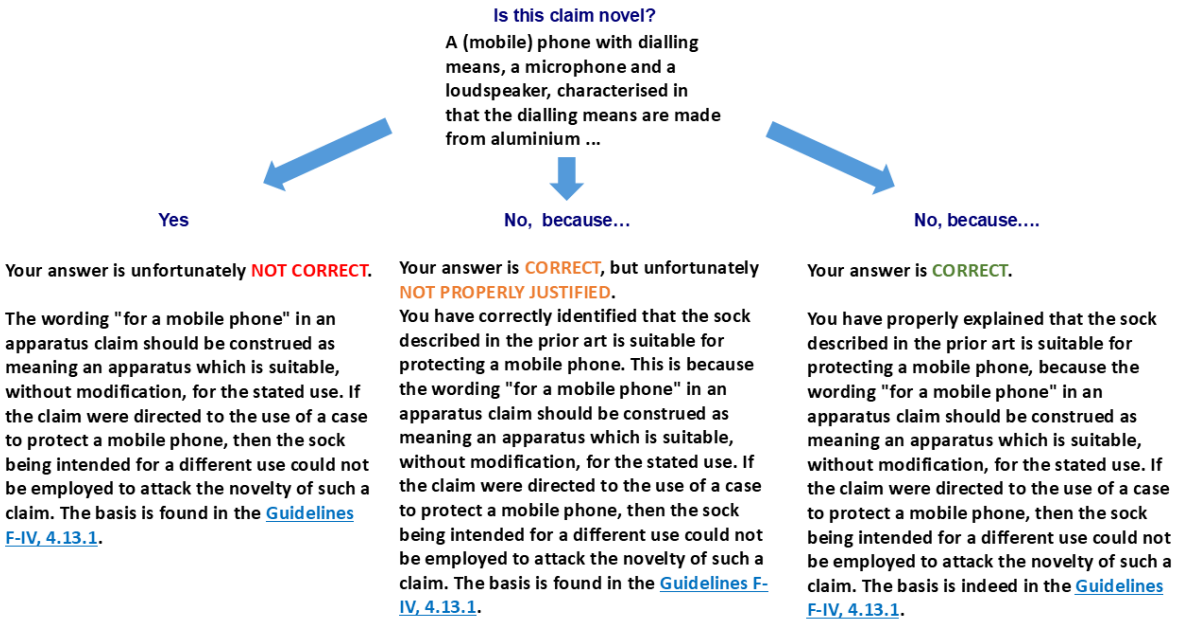


SLASys Moodle Integration for assessment and feedback provision

Since SLASys offers an automatic process to recommend assessment, the assessment process could be enhanced by providing automated feedback. Figure 4 shows an example of the feedback provided on the question associated with the claim of Figure 2. Two types of questions were defined: a simple short answer or a choice question with a justification. This second type is relevant for teachers in patent examination since questions usually state some claim related to a patent definition, and students should examine whether the claim satisfies EPC guidelines, such as novelty or clarity. Figure 4 shows the second type with the feedback provided and the link to the EPC guidelines depending on the assessment result. The feedback is predefined for each option and configured on the question design, and it provides

meaningful information for students to understand the assessment result and improve their learning.

Figure 4
Feedback example about Novelty question



The SLASys system incorporates an API with different operations that can be integrated into any LMS. This work used Moodle integration by developing a custom plugin to access SLASys capabilities. Complexity issues were hidden, providing an easy-to-use interface for teachers' management. A new question type has been integrated into Moodle by adapting the question configuration and the assessment interface for teachers, as well as the interface to submit the answers and review the assessment results for students. The design minimizes teachers' workload with an extremely small learning curve.

A new question can be added to a test by providing mandatory information: type of question, the statement, and the set of golden answers. Optional annotated answers from previous students can be added or retrieved from previous course editions. When this information is not provided, SLASys uses the semantic comparison approach. Otherwise, the corresponding answer AI classifier is automatically trained without technical intervention (Figure 3).

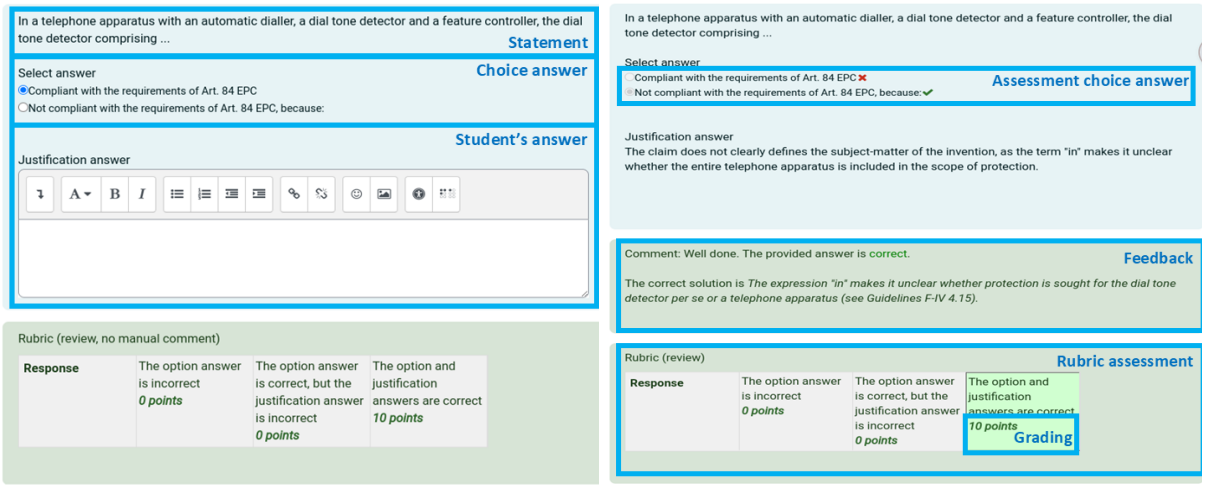
The most powerful part comes from the assessment and feedback configuration. The question has embedded a rubric (i.e., not available for short-answer questions in Moodle), unifying the assessment criteria for teachers and delivering the corresponding feedback depending on assessment results.

Figure 5.a) illustrates the student's interface for submitting an answer to a question with the choice option enabled. The figure also shows each level's rubric and the associated grade. The submission is sent to SLASys, and the recommendation is returned to Moodle when finished. The recommendation is shown in the assessment interface (Figure 6). The teacher sees the golden answers, the SLASys recommendation, the explainable information from the answer classification model,

and the metric score. Such explainable information shows in colors which words impact positive or negative classification decisions. Although this information sometimes includes irrelevant words, teachers can still identify incorrect words that affect how students' answers are classified. The interface also proposes the rubric grading that is automatically selected depending on the recommendation. The grading and the feedback are assigned depending on the rubric selection. The teacher must review this information and select the correct rubric level.

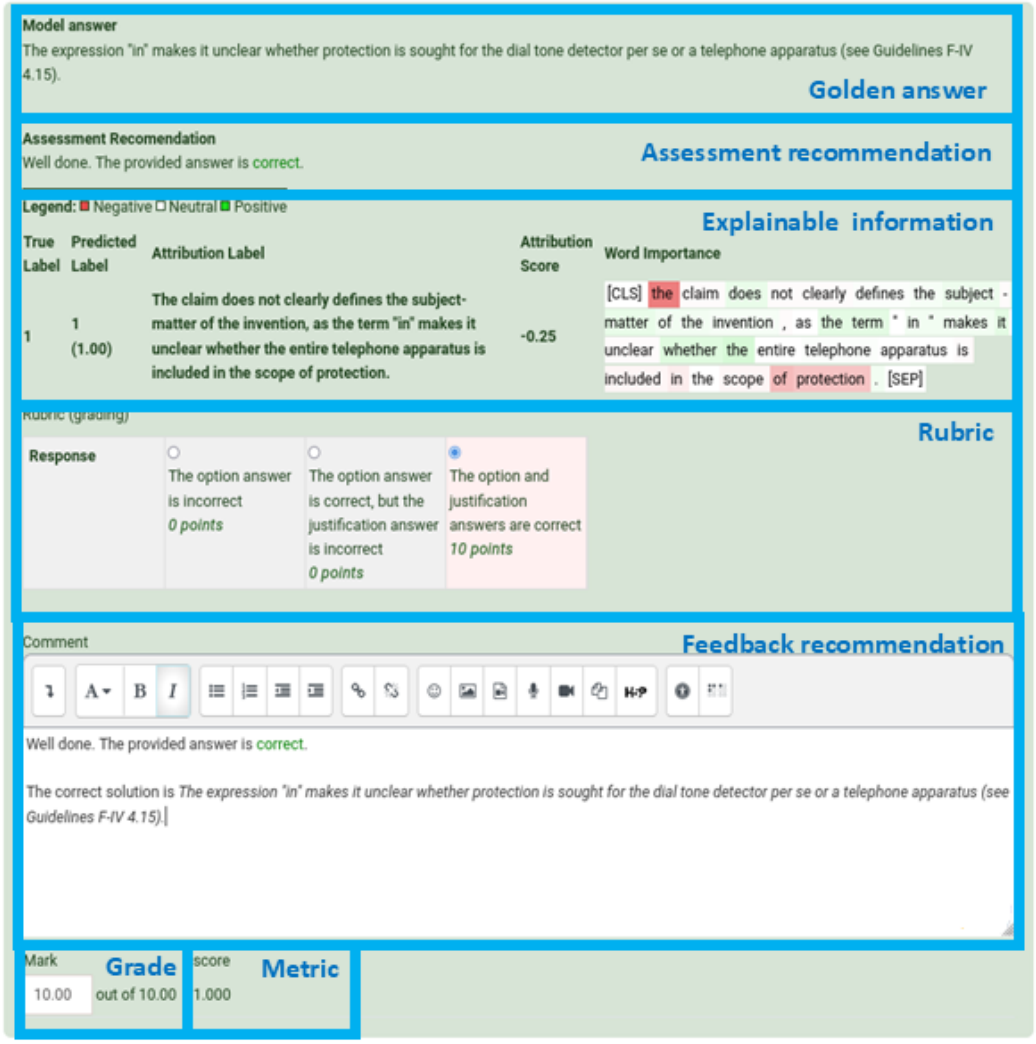
After being assessed, the test results can be reviewed in the students' interface (Figure 5.b). Each student sees the grading and the feedback based on the selected rubric option, validated by the teacher.

Figure 5
Moodle question interface for students



a) Attempt interface b) Feedback interface

Figure 6
Moodle assessment interface for teachers



Data analysis

Data from two sources were used to answer the RQ. First, information from SLASys has been used to analyze the number of correct assessments on answers used for training (i.e., *performance*) (RQ1). Next, using the Moodle plugin, the recommendations provided by SLASys and the assessment performed by the teachers were retrieved to analyze the correct recommendations during the test with students (RQ2 and RQ3). The performance during the training phase and test with students can be computed with the following metrics:

$$TNR = \frac{TN}{TN + FP} \qquad ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

$$TPR = \frac{TP}{TP + FN} \qquad F_{1.5} = \frac{(1 + 1.5^2)TP}{(1 + 1.5^2)TP + 1.5^2 FN + FP}$$

Where TP denotes the number of correct answers correctly identified, TN the number of incorrect answers correctly identified, FP the number of incorrect answers not correctly identified, and FN the number of correct answers not correctly identified. These four metrics allow to compute the accuracy when detecting correct and incorrect answers (ACC), the accuracy when detecting correct answers (True Positive Rate—TPR), the accuracy when distinguishing incorrect answers (True Negative Rate—TNR), and the percentage of correct answers penalizing the incorrect identifications (F score - F1.5). These metrics are used to evaluate the performance between the semantic comparison and answer classification approaches (RQ1) and whether SLASys recommends the expected teacher's assessment in the real educational setting (RQ2 and RQ3).

Finally, the average access time to the feedback interface and students' ratings of the results were gathered. The Moodle plugin added a Likert scale to collect students' appraisals within the feedback interface. The scale ranged from 1 to 5 (5 means the feedback was highly useful). Regarding the teachers' experience, only their feedback on the system usage was collected because only three teachers were involved (RQ4).

RESULTS

RQ1: How accurate is answer classification compared to semantic comparison for ASAG recommendation?

Knowing SLASys' accuracy before the first pilot with students is crucial before analyzing the system's applicability in a real educational setting. The results are detailed in Table 1 where the eight questions from the Clarity test were analyzed.

Table 1
Performance metrics on Clarity training test

	Correct	Incorrect	Semantic comparison				Answer classification			
Exercise	**Train / Test**	**Train / Test**	**ACC**	**TPR**	**TNR**	**$F_{1.5}$**	**ACC**	**TPR**	**TNR**	**$F_{1.5}$**
Question 1	36 / 8	107 / 22	37 %	100 %	14 %	58 %	87 %	88 %	86 %	81 %
Question 2	47 / 10	7 / 2	67 %	70 %	50 %	75 %	100 %	100 %	100 %	100 %
Question 3	3 / 2	46 / 4	42 %	100 %	30 %	48 %	83 %	0 %	100 %	0 %
Question 4	34 / 7	17 / 20	27 %	0 %	75 %	0 %	64 %	86 %	25 %	79 %
Question 5	67 / 14	100 / 20	53 %	100 %	20 %	74 %	82 %	71 %	90 %	75 %
Question 6	82 / 17	36 / 8	84 %	94 %	63 %	91 %	88 %	100 %	63 %	95 %
Question 7	33 / 7	33 / 7	57 %	86 %	29 %	73 %	79 %	71 %	86 %	75 %
Question 8	68 / 14	19 / 4	83 %	93 %	50 %	91 %	83 %	100 %	25 %	94 %
Average			56 %	**80 %**	41 %	64 %	**83 %**	77 %	**72 %**	**75 %**

The table summarizes the number of annotated answers used for fine-tuning the AI classifiers (i.e., *training dataset*), testing their accuracy (i.e., *testing dataset*) and the metrics results. AI classifiers were fine-tuned by taking 80% of the annotated answers (i.e., there is a classifier for each question). The remaining 20% was used for testing and obtaining the results of the table. The split process considers correct and

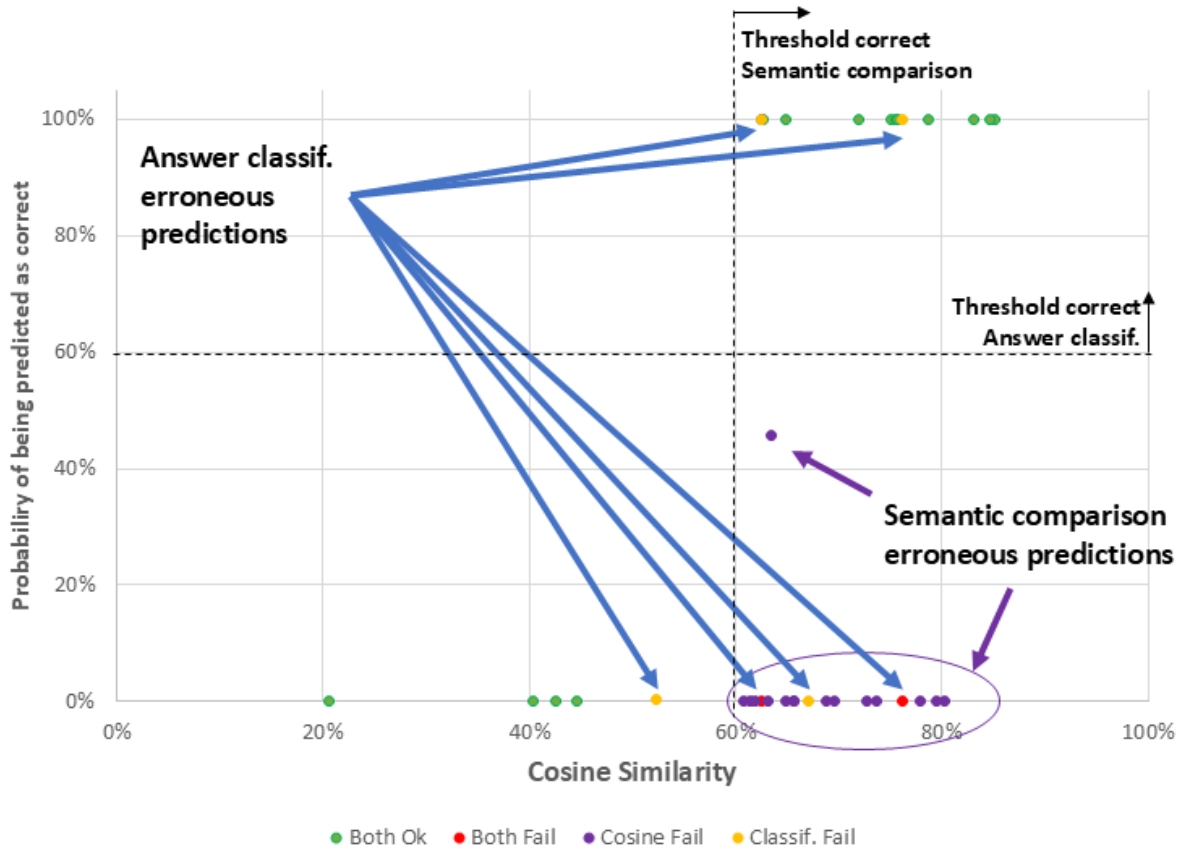
incorrect answers in both datasets. Note that there is no fine-tuning phase on the semantic comparison approach. In such cases, three random correct answers were compared with the testing answers.

Answer classification outperforms semantic comparison because annotated answers provide examples of correct and incorrect answers. Semantic comparison commonly fails when there are multiple ways to answer a question, and they are semantically far from the golden answers. Focusing on metrics to detect correct and incorrect answers (i.e., TPR and TNR), quality results highly depend on the number of answers provided by the teachers for training and the balance between correct and incorrect. Classifiers with few answers have difficulties recommending. It can be observed that semantic comparison can detect more correct answers on average (i.e., 80% compared to 77%). However, this increment is caused by many incorrect answers being incorrectly identified as correct (i.e., 41% compared to 72%).

Individual predictions were analyzed to better understand the techniques' behavior. Figure 7 depicts the analysis on the testing dataset for Question 5. The respective metrics, cosine similarity and probability of being predicted as correct, are plotted for semantic comparison and answer classification, respectively. Different colors were used to identify which predictions are correct using both metrics (green), which ones are incorrectly predicted by semantic comparison (purple), which ones are incorrectly predicted by answer classification (yellow), and which ones cannot be predicted by both methods (red).

Figure 7

Accuracy comparison for Question 5

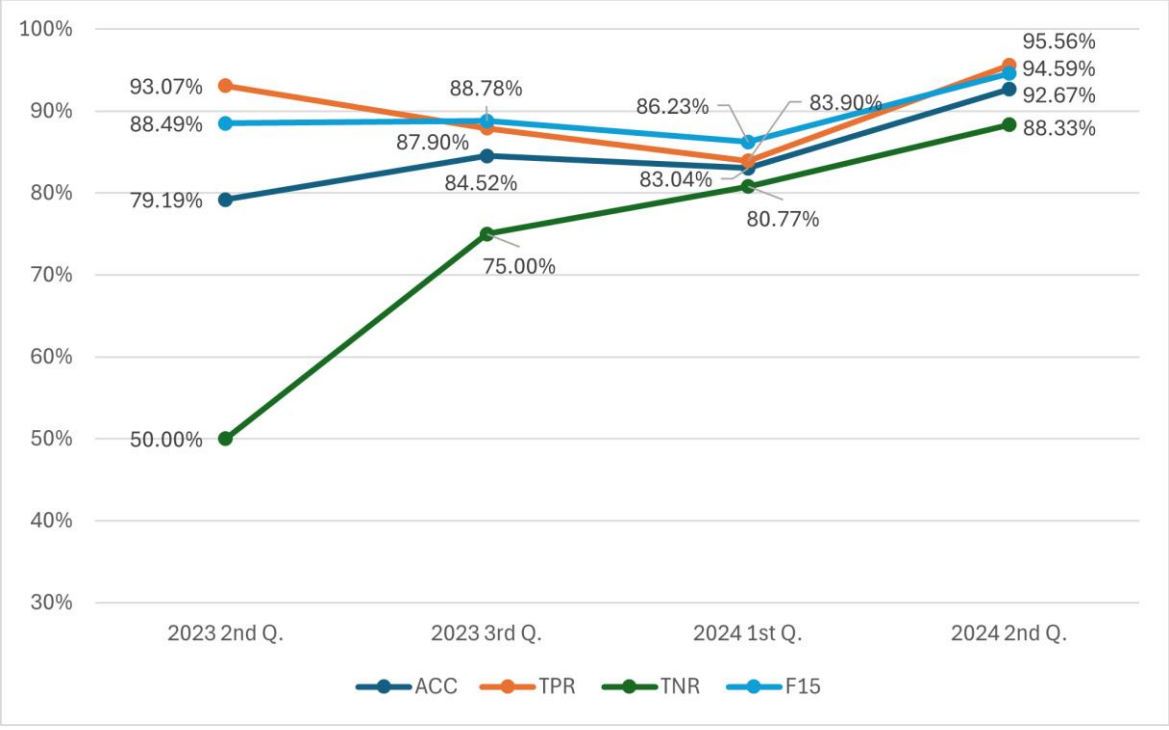


Most errors appear in semantic comparison. Many students' answers are considered correct (i.e., semantically similar to the correct answer provided by the teachers), ranging from 60% to 85%, but are incorrect. Thus, slight changes in the sentence's meaning can generate incorrect predictions. The answer classification also has some erroneous predictions in detecting correct answers, but in fewer cases. Additionally, cases where both methods fail can be observed. Teachers realized that some of these answers were incorrectly predicted because they were not covered in the answers used for the fine-tuning and golden answers.

RQ2: How accurate is SLASys in a real educational setting?

Figure 7 shows some erroneous cases. Adding such answers to the AI classifiers could enhance the identification in new course editions. This is the fundamental idea for enhancing classifiers: classifiers are retrained, including the assessment results performed in the previous editions. Thus, they are refined with new annotated answers that improve identification. RQ2 analyzes whether SLASys answer classification recommends the teachers' assessment during the four-course editions by considering this refinement on each edition. Aggregated results for the different metrics have been summarized in the plot depicted in Figure 8 for the Clarity test. The results for all questions have been aggregated to simplify the evaluation.

Figure 8
Performance metrics on Clarity test



A significant improvement can be observed in the second edition due to new annotated answers from the first edition, which improved the accuracy on questions with an unbalanced number of correct and incorrect answers (i.e., questions 2, 4, and

8 of Table 1). Although there is a slight reduction in the correct detection of correct answers (i.e., TPR), this is due to the refining process.

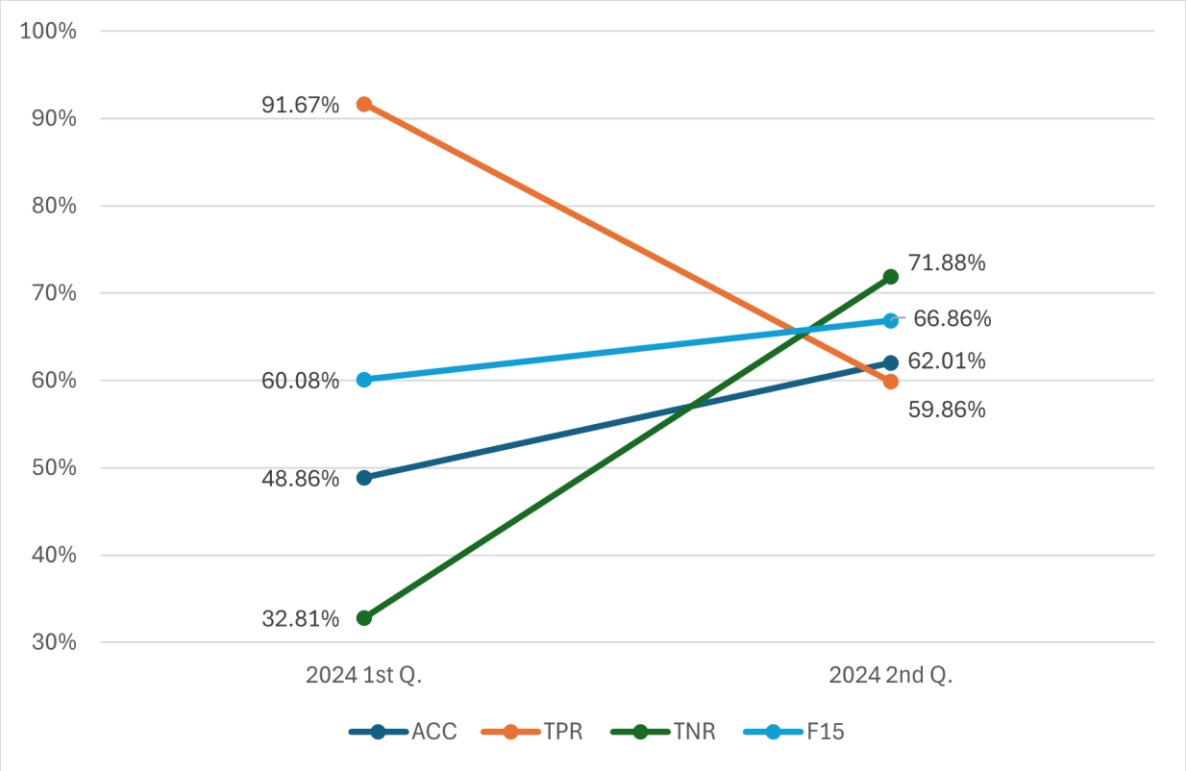
The following course editions contributed to enhancing the overall accuracy, reaching values larger than 88% in the last edition. Thus, including new data can effectively improve the quality of the recommender, which implies that it can be effectively used for assessment purposes. Better accuracy means a recommender with fewer errors that teachers can trust during assessment.

RQ3: Can SLASys be used without previous students' answers?

Good results in the previous section were obtained because teachers performed an initial task to annotate previous students' answers, which contributed to getting initial classifiers.

However, it was relevant to know how SLASys performs on new questions without previous answers. In this case, only semantic comparison can be used with golden answers provided by the teachers in the first edition (i.e., three golden answers were included per question). Nevertheless, in the second edition, answer classification can be applied using the assessment conducted in the first one. The experiment with real students was conducted in the last two editions, i.e., 2024 1st and 2nd quarter, where a new test was designed for the Novelty concept. The experiment depicted in Figure 9 shows how much improvement was reached by the SLASys system when changing from semantic comparison to answer classification in a real educational setting.

Figure 9
Performance metrics on Novelty test.



A significant difference between the two approaches can be observed. Global accuracy (i.e., ACC) improves from 48.86% to 62.01%. The meaningful increment in detecting wrong answers caused this improvement (i.e., TNR increased from 32.81% to 71.88%). However, the detection of correct answers decreases. Semantic comparison tends to recommend that most students' answers are correct since it cannot disaggregate slight semantic differences. This produces the unwanted countereffect that semantic comparison cannot correctly detect most incorrect answers. Thus, changing from semantic comparison to answer classification positively impacts the recommender.

The relevant insight in this experiment is that the answer classification models can produce high-quality recommendations in a few editions, even with a cold start with the semantic comparison approach.

RQ4: What is the opinion of the students and teachers?

Finally, it is of interest whether the new tests and feedback were helpful. Table 2 summarizes the information gathered from students, including the number who accessed the feedback results, the average time spent on the page, and the final rating given to the feedback. Not all students accessed the feedback results, probably because the tests did not impact the course assessment. Additionally, the average time is relatively low. Checking individual access time revealed that the time highly correlates with the number of wrong answers. Thus, students with wrong answers spent more time reading the feedback. Thus, it can be deduced that they are checking what they must learn for improvement. Finally, the average appraisal score is significantly high (i.e., larger than 70%) in all editions and tests.

Table 2
Students' feedback revision time and appraisal

	Clarity			Novelty		
Edition	**2023 2nd Q**	**2023 3rd Q**	**2024 1st Q**	**2024 2nd Q**	**2024 1st Q**	**2024 2nd Q**
Students who Accessed (%)	83.33	84.62	88.89	80.00	77.78	76.00
Avg. Access Time (minutes)	2.33	1.73	3.51	2.91	4.74	4.68
Appraisal (1-5)	3.94	3.84	3.63	3.85	3.55	3.77

Table 3
Comparison with related work

Reference	Technique	Prediction	Feedback	ACC
(Saha et al., 2018)	NLP	Correctness	No	66 %
(Soulimani et al., 2024)	BERT-Classifier	Grade (0-4)*	No	71 %
(Padó et al., 2024)	BERT-Classifier	Correctness	No	72 %
(Wang et al., 2019)	BERT-Classifier	Correctness	No	80 %
(Camus y Filighera, 2020)	BERT-Classifier	Correctness	No	80 %
(Lun et al., 2020)	BERT-Classifier	Correctness	No	82 %
(Sung et al., 2019)	BERT-Classifier	Correctness	No	84 %

Reference	Technique	Prediction	Feedback	ACC
(Schneider et al., 2023)	BERT-Classifier	Correctness	No	86 %
(Liu et al., 2019)	BERT-Classifier	Correctness	No	89 %
(Grévisse, 2024)	GenAI (GPT4)	Grade (0-10)*	Generated	64 %
(Aggarwal et al., 2025)	GenAI (Mistral)	Correctness	Generated	75 %
SLASys	BERT-Classifier	Correctness	Predefined	83 %

(*) Accuracy is computed by assuming correct when the grade is larger than the middle of the maximum points

Furthermore, teachers collected some opinions on the online sessions. Students mostly agreed that the assessment grade, feedback about correct answers, and EPC guidelines to review were highly valuable. Also, obtaining them nearly immediately was positive. However, students complained that the feedback should be more personalized, describing why their answer was wrong.

Teachers also shared their experiences with the recommender. Initially, they were reluctant to use it, arguing that relying on automatic grading could result in wrong assessments. However, the tool was highly accepted as a recommender, where teachers' final revision was mandatory. After assessing the different editions, teachers agreed they gained efficiency since they also acquired experience over time. Related to the experiment with the Novelty test without annotated answers, they complained in the first edition (i.e. 2024 1st Q) since they expressed no gain in workload reduction, since the recommender produced many wrong recommendations. However, they realized the positive benefit in the last edition when the answer classification started to operate.

DISCUSSION

The collected insights allow the research questions to be answered. Related to RQ1, as other work states (Burrows et al., 2015), BERT AI classifiers (i.e., answer classification) show greater precision compared to semantic comparison, and they obtain similar results to other ASAG system evaluations. Table 3 compares with state-of-the-art systems in public datasets, summarizing the technique used, the prediction objective, type of provided feedback, and the obtained accuracy. SLASys' answer classification meets the accuracy performance obtained by related works (i.e., accuracy ranging from 71% to 89%), focusing on predicting the correctness of the answer. However, it cannot be compared with works that focus on predicting the grade since they use other metrics to evaluate how close the predicted grade is to the correct one (i.e., root square mean error). Such techniques have moderate to high error rates ranging from 0.57 (Baral et al., 2021) to values larger than 1 (Gaddipati et al., 2020; Metzler et al., 2024), making their use in a real setting difficult. Note that predicting the grade is a more difficult task since there is more variability in the result (del Gobbo et al., 2023). Therefore, the method proposed in this work simplifies the process by giving the assessment result to the teacher, who decides the grade depending on a rubric. Additionally, the experimental results show insights into the reduced dataset size needed to get a fine-tuned classifier for short-answer questions (Mehrafarin et al., 2022). Typically, it is recommended to have a minimum of 500 answers (i.e., *instances*) to have a proper classifier. However, Table 1 shows that a good classifier can be trained even with fewer answers. Moreover, a comparison was conducted between

both metric computations for a specific question (Figure 7). Teachers can access this explainable information within Moodle to better understand which answers are not well-classified. One should recall that explainable AI is one of the recommendations for AI tools (Zhao et al., 2024). Table 3 also compares with GenAI approaches, which currently perform worse than BERT ones. Although GenAI tools generate great enthusiasm, they still have some drawbacks for assessment. They may have hallucinations (Jia et al., 2024) or conceptual limitations in specific domains (la Cruz Martínez et al., 2024) that can decrease their efficacy for assessment. Some of these GenAI approaches can be used locally. However, enterprise solutions are widely used, which operate remotely, offer a bigger processing capacity, but at higher costs with sustainability issues (van Wynsberghe, 2021). Additionally, they require the analysis of ethical issues and the rethinking of educational and data protection policies. Higher education institutions are encouraged to review García-Peñalvo et al. (2024)'s manifesto for a safe, ethical, and secure system. In this case, SLASys operates locally, reducing processing costs, maintaining its functionality, and not providing data to third parties. Additionally, it is stateless without collecting students' sensitive data, keeping them safe within the LMS.

Concerning RQ2, accurate progression over four editions of the course was shown. As the previous research question claims, the small sets of annotated answers produced highly accurate classifiers, making it available for use in a real educational setting. By generating accurate classifiers, SLASys is an example of an AI tool that can benefit both actors. On the one hand, teachers can benefit from a recommender for ASAG, increasing assessment efficiency and unifying assessment criteria (Xavier et al., 2025). Besides, benefits for teachers are clear as their workload is reduced, allowing them to devote time to other qualitative tasks related to checking feedback, designing new questions, or better-designed learning activities. Due to the integration with Moodle, SLASys provides formative feedback to students compared to other techniques described in Table 3. Related works only provide a recommendation without feedback, which, in the case of deployment, would need additional methods to deliver it. According to Gaddipati et al. (2021), one option could be using GenAI tools that show potential for feedback generation. On the other hand, students gain quality feedback when asked questions that require reasoning (Calimeris & Kosack, 2020). With this feedback, they can learn better with comprehensive and detailed information about their assessment task. Ultimately, such feedback becomes a personalized component that enhances student knowledge (Abu Khurma et al., 2024).

Related to RQ3, results demonstrate that the system can be used even without training data. In the first edition of a new question, the teachers will require more attention due to the low accuracy recommendations. However, the AI classifier will be ready for use in the next edition. It is worth noting the technical simplicity of the system. SLASys follows the good assessment practices described in Petridou & Lao (2024).

Finally, concerning RQ4, opinions and results indicate that recommender mispredictions do not impact students since they obtain the teachers' reviewed assessment. Note that human supervision is mandatory because mistakes in assessment may have negative effects on students (Li et al., 2023). As in Sangapu (2018), opinions are mostly positively related to introducing IA tools in education. Teachers are not overwhelmed with technical issues since the system is embedded in the well-known Moodle LMS. The system will autonomously change from one

technique to another without teachers' intervention. Note that SLASys has been designed as an assessment recommender following the recommendations of the European AI Act (European Commission, 2024), avoiding their utilization as an automatic AI grader.

This study presents limitations related to sample size and non-probability sampling that impact external validity. Due to its application in a real educational setting, a small dataset from a specific domain was used, which restricts the generalizability of the findings and may lead to sample bias. This can be mostly observed on RQ3, where the sample size for the last edition is only 25 students. However, the findings on RQ2 demonstrate that the system improves accuracy and remains applicable and effective over time (i.e., four editions) even when trained on limited data. Regarding non-probability sampling, all students participated in the study, making the knowledge acquisition evaluation with the new tests unfeasible. Note that the first exam is after the second course during the complete training process. A longitudinal analysis until this exam would enrich the understanding of the effect of the new tests on the acquired knowledge. Sample size also affects RQ4. Students' positive opinions could be biased due to the sample size. Additionally, teachers' opinions were collected with simple semi-structured interviews. A qualitative analysis with focus groups would give deeper insights.

However, the obtained experimental results could not lead to generalization, although particular emphasis is placed on the straightforward transferability to any domain with specific language, such as law, history, philosophy, or social science. Transferability would depend on an initial workload to design the questions, add the corresponding golden answers, or gather and review the annotated answers. SLASys allows the creation of questions where students should apply the acquired knowledge instead of answering based on theoretical concepts, leading to an authentic assessment (Villarroel et al., 2018).

CONCLUSIONS

The contribution of this research is threefold. Firstly, a short-answer assessment recommender system guided by three core principles was developed: free BERT techniques suitable for private deployment, running on servers with limited resources, and easy-to-use without requiring technical AI expertise. Secondly, a notable feature is the seamless integration with Moodle, enabling students' tracking, grading, and visualization of learning outcomes through SLASys. Results have shown that, although these results cannot be generalized, answer classification can lead to better recommendations. Nevertheless, SLASys allows the recommender to be refined to produce better results for new cohorts of students. Thirdly, the work underscores the role of AI as a supportive tool in education, empowering teachers to use a tool designed for recommending assessment and feedback that can work in several domains. Teachers can use integrative strategies by highlighting the convergence of technology and pedagogy to enhance the students' learning experience.

Future work will focus on increasing the recommender utilization in other tests in the same fundamental course on patent examination in the IP domain. Such an increment will help to evaluate the system's capability for other concepts. Additionally, how feedback can be personalized based on students' wrong answers and the available learning resources will be explored. BERT can also be used to create semantic search tools based on the course learning materials. Such tools can automatically identify the

resources needed for a specific question, which could leverage the teacher's effort on feedback generation.

Acknowledgments

This work is part of the project PID2023-147592OB-I00 funded by MCIU/AEI/10.13039/501100011033/ FEDER, UE, and the Academic Research Programme of the European Patent Office Grant Agreement No. 2021/8404.

REFERENCES

Abu Khurma, O., Albahti, F., Ali, N., & Bustanji, A. (2024). AI ChatGPT and student engagement: Unraveling dimensions through PRISMA analysis for enhanced learning experiences. *Contemporary Educational Technology, 16*(2). https://doi.org/10.30935/cedtech/14334

Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). *DocBERT: BERT for document classification*. arXiv. https://doi.org/10.48550/arXiv.1904.08398

Aggarwal, D., Sil, P., Raman, B., & Bhattacharyya, P. (2025). "I understand why I got this grade": Automatic short answer grading with feedback. In *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-031-98420-4_22

Akçapınar, G. (2015). How automated feedback through text mining changes plagiaristic behavior in online assignments. *Computers & Education, 87*, 123-130. https://doi.org/10.1016/j.compedu.2015.04.007

Almasre, M. (2024). Development and evaluation of a custom GPT for the assessment of students' designs in a typography course. *Education Sciences, 14*(2), Article 148. https://doi.org/10.3390/educsci14020148

Arefeen, M. A., Debnath, B., & Chakradhar, S. (2024). LeanContext: Cost-efficient domain-specific question answering using LLMs. *Natural Language Processing Journal, 7*, 100065. https://doi.org/10.1016/j.nlp.2024.100065

Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate. *arXiv*. https://doi.org/10.48550/arXiv.1409.0473

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21*(1), 23. https://doi.org/10.1186/s41239-024-00455-4

Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2021). Improving automated scoring of student open responses in mathematics. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*.

Bergmann, J., & Sams, A. (2012). *Flip your classroom: Reach every student in every class every day*. International Society for Technology in Education.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education, 25*(1), 60-117. https://doi.org/10.1007/s40593-014-0026-8

Calimeris, L., & Kosack, E. (2020). Immediate feedback assessment technique (IF-AT) quizzes and student performance in microeconomic principles courses. *Journal of Economic Education, 51*(3–4), 304-319.

https://doi.org/10.1080/00220485.2020.1804501

Camus, L., & Filighera, A. (2020). Investigating transformers for automatic short answer grading. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial intelligence in education* (Lecture Notes in Computer Science, Vol. 12164, pp. 43-48). Springer. https://doi.org/10.1007/978-3-030-52240-7_8

Dai, Y., Lai, S., Lim, C. P., & Liu, A. (2025). University policies on generative AI in Asia: Promising practices, gaps, and future directions. *Journal of Asian Public Policy, 18*(2), 260-281. https://doi.org/10.1080/17516234.2024.2379070

del Gobbo, E., Guarino, A., Cafarelli, B., & Grilli, L. (2023). GradeAid: A framework for automatic short answers grading in educational contexts-Design, implementation and evaluation. *Knowledge and Information Systems, 65*(10), 4479-4507. https://doi.org/10.1007/s10115-023-01892-9

De La Cruz Martínez, G., Eslava-Cervantes, A.-L., & Ramírez, S. (2024, July 1). Analysis of solutions of ChatGPT to logic problems based on critical thinking. *In Proceedings of the 16th International Conference on Education and New Learning Technologies (EDULEARN24)* (pp. 10324-10331). IATED. https://doi.org/10.21125/edulearn.2024.2525

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. *In* J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dhananjaya, G. M., Goudar, R. H., Kulkarni, A. A., Rathod, V. N., & Hukkeri, G. S. (2024). A digital recommendation system for personalized learning to enhance online education: A review. *IEEE Access, 12*, 33591–33615. https://doi.org/10.1109/ACCESS.2024.3369901

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(1), 40. https://doi.org/10.1186/s41239-023-00425-2

European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research, 83*(1), 70–120. https://doi.org/10.3102/0034654312474350

Gaddipati, S. K., Nair, D., & Plöger, P. G. (2020). Comparative evaluation of pretrained transfer learning models on automatic short answer grading. *arXiv*. https://arxiv.org/abs/2009.01303

Gaddipati, S. K., Plöger, P., Hochgeschwender, N., & Metzler, M. (2021, April 5). *Automatic formative assessment for students' short text answers through feature extraction* [Doctoral dissertation, Hochschule Bonn-Rhein-Sieg].

Gaona, J., Reguant, M., Valdivia, I., Vásquez, M., & Sancho-Vinuesa, T. (2018). Feedback by automatic assessment systems used in mathematics homework in the engineering field. *Computer Applications in Engineering Education, 26*(4), 921-934. https://doi.org/10.1002/cae.21950

García-Peñalvo, F. J., Alier, M., Pereira, J., & Casany, M. J. (2024). Safe, transparent, and ethical artificial intelligence: Keys to quality sustainable education (SDG4). *International Journal of Educational Research and Innovation, 22*, 1–21. https://doi.org/10.46661/ijeri.11036

González Fernández, M. O., Romero-López, M. A., Sgreccia, N. F., & Latorre Medina, M. J. (2025). Marcos normativos

para una IA ética y confiable en la educación superior: Estado de la cuestión. *RIED-Revista Iberoamericana de Educación a Distancia, 28*(2), 181-208. https://doi.org/10.5944/ried.28.2.43511

Grévisse, C. (2024). LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education, 24*(1), 1060. https://doi.org/10.1186/s12909-024-06026-5

György, A., & Vajda, I. (2007). Intelligent mathematics assessment in eMax. In *IEEE AFRICON Conference* (pp. 1-7). https://doi.org/10.1109/AFRCON.2007.4401512

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2017). Peer feedback on college students' writing: Exploring the relation between students' ability match, feedback quality and essay performance. *Higher Education Research & Development, 36*(7), 1433-1446. https://doi.org/10.1080/07294360.2017.1325854

Husein, R. A., Aburajouh, H., & Catal, C. (2025). Large language models for code completion: A systematic literature review. *Computer Standards & Interfaces, 92*, 103917. https://doi.org/10.1016/j.csi.2024.103917

Hustad, E., & Arntzen, A. A. B. (2013). Facilitating teaching and learning capabilities in social learning management systems: Challenges, issues, and implications for design. *Journal of Integrated Design and Process Science, 17*(1), 33-46. https://doi.org/10.3233/JID-2013-0003

Jia, Q., Cui, J., Xi, R., Liu, C., Rashid, P., Li, R., & Gehringer, E. (2024). On assessing the faithfulness of LLM-generated feedback on student assignments. In B. Paaßen & C. D. Epp (Eds.), *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 491-499). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.12729868

Kim, T. W. (2023). Application of artificial intelligence chatbots, including ChatGPT, in education, scholarly work, programming, and content generation and its prospects: A narrative review. *Journal of Educational Evaluation for Health Professions, 20*, 38. https://doi.org/10.3352/jeehp.2023.20.38

Klein, R., Kyrilov, A., & Tokman, M. (2011). Automated assessment of short free-text responses in computer science using latent semantic analysis. In *Proceedings of ITiCSE '11: The 16th Annual Conference on Innovation and Technology in Computer Science Education* (pp. 158-162). https://doi.org/10.1145/1999747.1999793

Kuechler, W., & Vaishnavi, V. (2012). A framework for theory development in design science research: Multiple perspectives. *Journal of the Association for Information Systems, 13*(6), 395-423. https://doi.org/10.17705/1JAIS.00300

Li, T. W., Hsu, S., Fowler, M., Zhang, Z., Zilles, C., & Karahalios, K. (2023). Am I wrong, or is the autograder wrong? Effects of AI grading mistakes on learning. In *Proceedings of the 2023 ACM Conference on International Computing Education Research (ICER '23)* (pp. 85-97). https://doi.org/10.1145/3568813.3600124

Liu, T., Ding, W., Wang, Z., Tang, J., Huang, G. Y., & Liu, Z. (2019). Automatic short answer grading via multiway attention networks. In *Lecture Notes in Computer Science* (Vol. 11626, pp. 376–388). https://doi.org/10.1007/978-3-030-23207-8_32

Lun, J., Zhu, J., Tang, Y., & Yang, M. (2020). Multiple data augmentation strategies for improving performance on automatic short answer scoring. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, pp. 13381-13388). https://doi.org/10.1609/aaai.v34i09.7062

Mehrafarin, H., Rajaee, S., & Pilehvar, M. T. (2022). On the importance of data size in probing fine-tuned models. In *Findings of the Association for Computational Linguistics* (pp. 239–248). https://doi.org/10.18653/v1/2022.findings-acl.20

Messer, M., Brown, N. C. C., Kölling, M., & Shi, M. (2024). Automated grading and feedback tools for programming education: A systematic review. *ACM Transactions on Computing Education, 24*(1), Article 1. https://doi.org/10.1145/3636515

Metzler, T., Plöger, P. G., & Hees, J. (2024). Computer-assisted short answer grading using large language models and rubrics. In *INFORMATIK 2024: AI@WORK* (pp. 1383-1393). Gesellschaft für Informatik e.V. https://doi.org/10.18420/inf2024_121

Nguyen, H., Bhat, S., Moore, S., Bier, N., & Stamper, J. (2022). Towards generalized methods for automatic question generation in educational domains. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Intelligent Tutoring Systems. ITS 2022. Lecture Notes in Computer Science* (Vol. 13450, pp. 272-284). Springer. https://doi.org/10.1007/978-3-031-16290-9_20

Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218. https://doi.org/10.1080/03075070600572090

Novak, G. M. (2012). Just-in-time teaching. *New Directions for Teaching and Learning, 2012*(128), 63-73. https://doi.org/10.1002/tl.469

Oates, B. J. (2006). *Researching information systems and computing*. SAGE Publications Ltd.

OpenAI. (2024). *GPT-4o system card*. https://openai.com/index/gpt-4o-system-card/

Padó, U., Eryilmaz, Y., & Kirschner, L. (2024). Short-answer grading for German: Addressing the challenges. *International Journal of Artificial Intelligence in Education, 34*(4), 1488-1510. https://doi.org/10.1007/s40593-023-00383-w

Pang, J., Ye, F., Wong, D. F., Yu, D., Shi, S., Tu, Z., & Wang, L. (2025). Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics, 13*, 73-95. https://doi.org/10.1162/tacl_a_00730

Petridou, E., & Lao, L. (2024). Identifying challenges and best practices for implementing AI additional qualifications in vocational and continuing education: A mixed methods analysis. *International Journal of Lifelong Education, 43*(4), 385-400. https://doi.org/10.1080/02601370.2024.2351076

Qiu, Y., & Jin, Y. (2024). ChatGPT and finetuned BERT: A comparative study for developing intelligent design support systems. *Intelligent Systems with Applications, 21*, 200308. https://doi.org/10.1016/j.iswa.2023.200308

Rezaei, A. R. (2015). Frequent collaborative quiz taking and conceptual learning. *Active Learning in Higher Education, 16*(3), 189-204. https://doi.org/10.1177/1469787415589627

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144. https://doi.org/10.1007/BF00117714

Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R., & Sengupta, B. (2018). Sentence-level or token-level features for automatic short answer grading? Use both. In *Lecture Notes in Computer Science* (Vol. 10947, pp. 475-486). https://doi.org/10.1007/978-3-319-93843-1_37

Sangapu, I. (2018). Artificial intelligence in education: From a teacher and a student perspective. *SSRN*. https://doi.org/10.2139/ssrn.3372914

Schneider, J., Richner, R., & Riser, M. (2023). Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education, 33*(1), 1-29. https://doi.org/10.1007/s40593-022-00289-z

Senthilnathan, V., Sakthi Vaibhav, M., & Alexander, R. (2025). *Semantic refined prompting based automated essay scoring system*. In *Proceedings of the 2025 International Conference on Electronics and Renewable Systems (ICEARS)* (pp. 1344-1348). IEEE. https://doi.org/10.1109/ICEARS64219.2025.10940227

Siddiqi, R., & Harrison, C. (2008). A systematic approach to the automated marking of short-answer questions. In *Proceedings of IEEE INMIC 2008: 12th International Multitopic Conference* (pp. 281-286). https://doi.org/10.1109/INMIC.2008.4777758

Soulimani, Y. A., El Achaak, L., & Bouhorma, M. (2024). Deep learning-based Arabic short answer grading in serious games. *International Journal of Electrical and Computer Engineering, 14*(1), 841-853. https://doi.org/10.11591/ijece.v14i1.pp841-853

Souza, F., Nogueira, R., & Lotufo, R. A. (2019). *Portuguese named entity recognition using BERT-CRF* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1909.10649

Sung, C., Ma, T., Dhamecha, T. I., Reddy, V., Saha, S., & Arora, R. (2019). Pre-training BERT on domain resources for short answer grading. In *Proceedings of EMNLP-IJCNLP 2019* (pp. 6076–6086). https://doi.org/10.18653/v1/D19-1628

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics, 1*(3), 213-218. https://doi.org/10.1007/s43681-021-00043-6

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education, 43*(5), 840-854. https://doi.org/10.1080/02602938.2017.1412396

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 353-355). https://doi.org/10.18653/v1/W18-5446

Wang, H., & Lehman, J. D. (2021). Using achievement goal-based personalized motivational feedback to enhance online learning. *Educational Technology Research and Development, 69*(2), 807-836. https://doi.org/10.1007/s11423-021-09940-3

Wang, Y., Wang, C., Li, R., & Lin, H. (2022). On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of NAACL 2022* (pp. 3432-3444). https://doi.org/10.18653/v1/2022.naacl-main.249

Wang, Z., Lan, A. S., Waters, A. E., Grimaldi, P., & Baraniuk, R. G. (2019). A meta-learning augmented bidirectional transformer model for automatic short answer grading. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*.

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17-37. https://doi.org/10.1080/00461520.2016.1207538

Xavier, C., Rodrigues, L., Costa, N., Neto, R., Alves, G., Falcão, T. P., Gašević, D., & Mello, R. F. (2025). Empowering instructors with AI: Evaluating the impact of an AI-driven feedback tool in learning analytics. *IEEE Transactions on Learning Technologies, 18*, 498-512. https://doi.org/10.1109/TLT.2025.3562379

Xie, X., & Li, X. (2018). Research on personalized exercises and teaching feedback based on big data. In *Proceedings of the ACM International Conference* (pp. 217-221). https://doi.org/10.1145/3232116.3232143

Xu, Z., & Zhu, P. (2023). Using BERT-based textual analysis to design a smarter classroom mode for computer teaching in higher education institutions.

International Journal of Emerging Technologies in Learning, 18(19), 120-133. https://doi.org/10.3991/ijet.v18i19.42483

Zhang, H., Cai, J., Xu, J., & Wang, J. (2019). Pretraining-based natural language generation for text summarization. In *Proceedings of CoNLL 2019* (pp. 789-798). https://doi.org/10.18653/v1/K19-1074

Zhang, H., Yu, P. S., & Zhang, J. (2025). A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*. https://doi.org/10.1145/3731445

Zhang, Z., Zhang, Z., Chen, H., & Zhang, Z. (2019). A joint learning framework with BERT for spoken language understanding. *IEEE Access, 7*, 168849–168858. https://doi.org/10.1109/ACCESS.2019.2954766

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology, 15*(2), Article 26. https://doi.org/10.1145/3639372

Zheng, L., Long, M., Chen, B., & Fan, Y. (2023). Promoting knowledge elaboration, socially shared regulation, and group performance in collaborative learning: An automated assessment and feedback approach based on knowledge graphs. *International Journal of Educational Technology in Higher Education, 20*(1), 12. https://doi.org/10.1186/s41239-023-00415-4

Date of reception: 1 June 2025
Date of acceptance: 6 August 2025
Date of approval for layout: 16 September 2025
Date of publication in OnlineFirst: 14 October 2025
Date of publication: 1 January 2026