applied sciences

MDPI

Article

Application of Artificial Intelligence as an Aid for the Correction of the Objective Structured Clinical Examination (OSCE)

Davide Luordo [1,2], Marta Torres Arrese [3], Cristina Tristán Calvo [1], Kirti Dayal Shani Shani [1], Luis Miguel Rodríguez Cruz [4], Francisco Javier García Sánchez [1,2], Alfonso Lagares Gómez-Abascal [5], Rafael Rubio García [5], Juan Delgado Jiménez [5], Mercedes Pérez Carreras [5], Ramiro Diez Lobato [5], Juan José Granizo Martínez [5], Yale Tung-Chen [6,*] and Mª Victoria Villena Garrido [5]

[1] Infanta Cristina University Hospital, 28981 Madrid, Spain; davidelu@ucm.es (D.L.); cristinatristancalvo@gmail.com (C.T.C.); kirtisshani@gmail.com (K.D.S.S.); franci11@ucm.es (F.J.G.S.)
[2] Department of Medicine, Complutense University of Madrid, 28040 Madrid, Spain
[3] Alcorcón Foundation Hospital, 28922 Alcorcón, Spain; mtorresa@salud.madrid.org
[4] Independent Researcher, 28945 Madrid, Spain; luismi16@gmail.com
[5] 12 de Octubre University Hospital, 28041 Madrid, Spain; algadoc@yahoo.com (A.L.G.-A.); rafaelrubiogarcia@ucm.es (R.R.G.); jfdelgado@med.ucm.es (J.D.J.); mepere03@ucm.es (M.P.C.); ramirodiezlobato@gmail.com (R.D.L.); juanjose.granizo@salud.madrid.org (J.J.G.M.); vvillena@separ.es (M.V.V.G.)
[6] Department of Medicine, Autonomous University of Madrid, 29040 Madrid, Spain
* Correspondence: yale.tung@uam.es; Tel.: +34-917-277-000

Abstract: The assessment of clinical competencies is essential in medical training, and the Objective Structured Clinical Examination (OSCE) is an essential tool in this process. There are multiple studies exploring the usefulness of artificial intelligence (AI) in medical education. This study explored the use of the GPT-4 AI model to grade clinical reports written by students during the OSCE at the Teaching Unit of the 12 de Octubre and Infanta Cristina University Hospitals, part of the Faculty of Medicine at the Complutense University of Madrid, comparing its results with those of human graders. Ninety-six (96) students participated, and their reports were evaluated by two experts, an inexperienced grader, and the AI using a checklist designed during the OSCE planning by the teaching team. The results show a significant correlation between the AI and human graders (ICC = 0.77 for single measures and 0.91 for average measures). AI was more stringent, assigning scores on an average of 3.51 points lower ($t = -15.358$, $p < 0.001$); its correction was considerably faster, completing the analysis in only 24 min compared to the 2–4 h required by human graders. These results suggest that AI could be a promising tool to enhance efficiency and objectivity in OSCE grading.

Keywords: artificial intelligence; objective structured clinical examination (OSCE); medical education; clinical competency assessment; AI in healthcare; AI-assisted grading; human–AI comparison in grading; digital OSCE evaluation; medical report evaluation

Academic Editors: Vladislav Toronov and Gordana Žauhar

Received: 3 December 2024
Revised: 23 December 2024
Accepted: 21 January 2025
Published: 23 January 2025

Citation: Luordo, D.; Torres Arrese, M.; Tristán Calvo, C.; Shani Shani, K.D.; Rodríguez Cruz, L.M.; García Sánchez, F.J.; Lagares Gómez-Abascal, A.; Rubio García, R.; Delgado Jiménez, J.; Pérez Carreras, M.; et al. Application of Artificial Intelligence as an Aid for the Correction of the Objective Structured Clinical Examination (OSCE). *Appl. Sci.* **2025**, *15*, 1153. https://doi.org/10.3390/app15031153

1. Introduction

An essential component of medical education is the assessment of clinical competence. To enhance objectivity in this process, Ronald Harden developed the Objective Structured Clinical Examination (OSCE) in 1975. This approach simulates real clinical

cases using standardized patients who replicate medical conditions, as well as mannequins and simulators for technical procedures. It has proven to be an effective method for evaluating both the theoretical knowledge and practical skills of medical students [1–4].

Universities worldwide, including those in Spain, have adopted the OSCE over the years, although its implementation by faculty may vary slightly.

For years, the OSCE has been used to evaluate the acquisition of clinical competencies at the Teaching Unit of the 12 de Octubre and Infanta Cristina University Hospitals, part of the Faculty of Medicine at the Complutense University of Madrid. Due to the dual qualitative and quantitative nature of this assessment, faculty members must invest significant effort in its development, particularly when providing students with personalized feedback. Exploring ways to optimize this process has been an area of significant research interest to which our teaching unit has made notable contributions [5].

Clinical report writing, a fundamental aspect of daily medical practice, is evaluated in OSCEs.

This process not only assesses specific competencies but also prepares students for the critical skill of transmitting clinical information about patients precisely and systematically. However, grading these clinical reports presents significant challenges. Students must document patient history, physical examination findings, test interpretation, and diagnostic and treatment plans, resulting in unstructured data that must be reviewed using pre-designed checklists [1,2,4,5].

The evaluator's role is crucial. To standardize the grading process, a pre-designed checklist created by the teaching team is typically used. Evaluators read the reports written by students and check off items on this checklist, ensuring that each required element is accounted for during the grading process. Beyond this structured approach, evaluators must also be subject-matter experts capable of accurately assessing the student's writing, especially given the various ways a concept can be articulated and the frequent use of acronyms and abbreviations by students. These issues primarily arise from human factors that affect evaluators' performance, introducing subjectivity and variability into the results [2,6–8].

Manual grading poses significant challenges, including evaluator fatigue. Over time, the repetitive nature of the task may lead to diminished attention to detail, resulting in poor decision-making and inconsistent grading. This variability undermines the exam's fairness, as evaluations can become overly lenient or excessively strict, compromising the validity of the results [6,9,10].

The inherent subjectivity of manual grading is another major concern. Even with efforts to standardize criteria for evaluation, various evaluators may have varying interpretations of the responses from students. The evaluator's emotional state at the moment of grading, their personal interpretation of the responses, or their past experiences can all have an impact on this subjectivity [8,9].

Manual grading also involves prolonged grading times, which not only delay the delivery of results but also increase the operational costs of OSCEs [11].

These challenges highlight the urgent need for alternative methods to improve grading accuracy and efficiency. For decades, artificial intelligence (AI) has transformed various fields, including medical education. As early as the 1950s, researchers began exploring how machines could perform cognitively demanding tasks such as language processing and decision-making. However, it is only in the past ten years that AI, particularly in the area of natural language processing (NLP), has advanced to the point where it can be practically applied in complex fields like medical education [12–16].

The development of OpenAI's Generative Pre-Trained Transformers (GPT), a series of large language models (LLMs) represents one of the most important advances in natural language processing. Trained on vast amounts of textual data, these models are

designed to comprehend and generate text in an organized way. Leveraging transformer architecture, an advanced machine learning technique, GPT models like GPT-3 and GPT-4 excel in evaluating text sequences by identifying patterns and relationships between words.

The integration of AI into medical education has been the subject of numerous recent studies, which have explored its potential to improve education and clinical evaluation. For example, research conducted by Li Sun and associates have shown how AI can improve clinical teaching and diagnosis efficiency and accuracy [13,15–20]. Other studies have evaluated the performance of AI models like ChatGPT in clinical reasoning exams, showing promising results in their ability to achieve competency levels comparable to those of medical students [17,21–29].

These applications not only demonstrate how AI can help with learning and assessment, but they also emphasize how important it is to include AI-related subjects in medical curricula in order to effectively educate future medical professionals.

Currently, there are no studies specifically exploring the use of AI in the grading of OSCE exams. However, this line of research is essential as AI could represent a significant support tool in this stage of medical training. It could help address several issues such as long grading times, increase homogeneity in grading, thereby offering fairer scores, reduce grader fatigue, and lower the costs of OSCEs. Furthermore, it is crucial to clarify the ethical and legal implications that the integration of AI into OSCEs would entail.

This study investigates the potential use of models like ChatGPT for grading clinical reports written by students during the OSCE, taking into consideration the developments of AI in medical education.

2. Materials and Methods

2.1. Study Design

This is a quasi-experimental, analytical, comparative, and explanatory study exploring the ability of AI to grade clinical reports written by students in the OSCE exam.

2.2. Study Population

The study includes all fourth-year medical students from the 2021–2022 cohort of the Teaching Unit of the 12 de Octubre and Infanta Cristina University Hospitals, part of the Faculty of Medicine at the Complutense University of Madrid, who participated in the OSCE exam in May 2022, for a total of 96 participants.

2.3. Planning and Execution of the OSCE

In January 2022, the OSCE was designed by the teaching staff at the 12 de Octubre and Infanta Cristina University Hospitals. The exam consisted of a total of 10 stations: 4 stations with simulated patients, 4 with clinical reports, and 2 with practical skills. Each station lasted 8 min, with a 2 min rest period between each one. Following each station with a simulated patient, there was always a clinical report station, where each student had to write a freestyle report, and all the stations were graded based on checklists. A blueprint was formulated for each station, including all the content, logistics, and evaluation checklists. It is paramount to note that the evaluation items were formulated during the exam design phase with human evaluators in mind and without applying any optimization protocol for AI interpretation.

After the design phase, the blueprints for the clinical cases and skills were integrated into the Digit-ECOE® software database (version 1.0.1.07), which provides different user interfaces to facilitate the execution and evaluation of the exam. In May 2022, medical students took the exam in the Academic Building of the 12 de Octubre University

Hospital. Each station was equipped with a computer running the D-ECOE interface of the Digit-ECOE® program. Specifically, a "student window" (Figure 1) was used in the clinical report stations, allowing students to view the sections to complete for each clinical report. This window was designed so students could only access it using their student number, verifying their identity with their name, surname, and photographs. Once the mandatory sections were completed, the students would sign the report, which would be saved in the program's database. After signing, the fields in the "student window" were reset so that the next Student could perform the task.

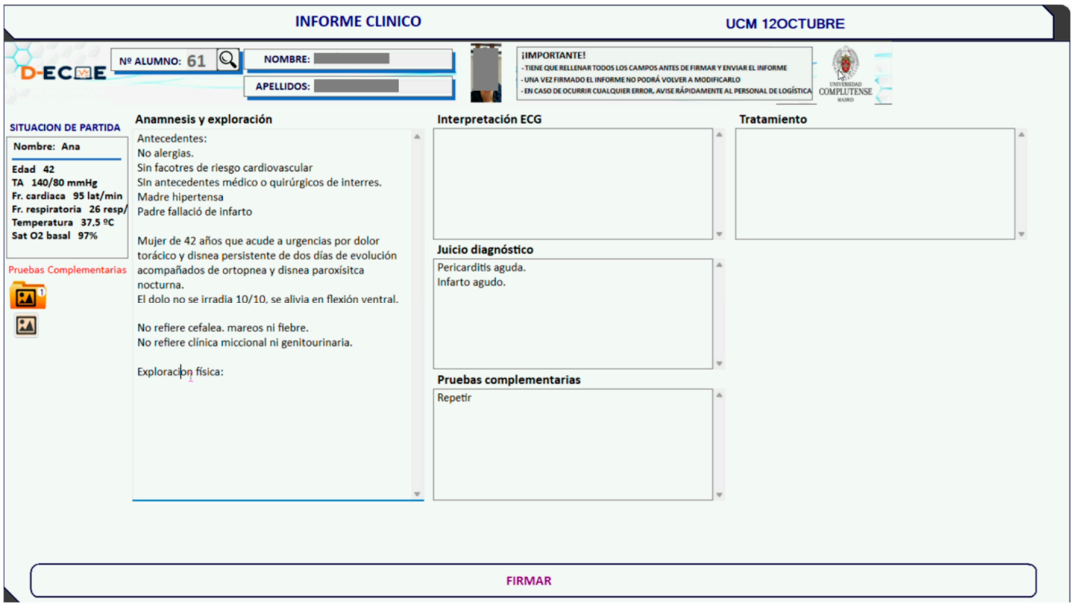

Figure 1. Student interface, D-ECOE program.

2.4. Grading of the OSCE

At the end of the exam, the cardiology station, which involved simulated patients and clinical reports, was selected for this study. In this station, students were required to write a detailed clinical report that included history-taking, a physical examination, the interpretation of complementary tests (including an electrocardiogram), differential diagnosis, and diagnostic and therapeutic plans. Each competency area was assessed using 19 items, adding to a possible score of 19 points (Table 1).

Table 1. Competency areas and evaluation items for the cardiology station.

Subject	Competency Area	Item	Points
Cardiology	Medical History	Patient's age and sex	0.5
		Chief complaint: chest pain	0.5
		Intensity and duration	2
		History of catarrhal symptoms in prior weeks	1
		Single cardiovascular risk factor: smoker	0.5
		Accompanying symptoms of chest pain	1
		Relation between chest pain and posture changes	1
		Relation between chest pain and breathing	1
		Absence of previous heart disease	1
	Physical Examination	Hemodynamic stability	1
		Absence of signs of heart failure	0.5
	Complementary Test Reports	ECG findings: diffuse concave ST elevation	3

Differential Diagnosis	Diagnostic judgment: acute pericarditis	3
	Diagnostic judgment: coronary syndrome/costochondritis	0.5
Diagnostic Plan (to be performed)	Hematology and general biochemistry	0.5
	Chest X-ray	0.5
	Acute phase reactants: ESR, CRP, cardiac enzymes CK-MB, troponins	0.5
	Echocardiogram	0.5
Therapeutic Plan	Treatment: aspirin, ibuprofen, indomethacin, or rest	0.5
Maximum Score		19

The reports were graded by four independent evaluators as follows:

- Two expert human evaluators (Expert 1 and Expert 2):
  - Expert 1: A tenured professor in the Department of Medicine at the Complutense University of Madrid, and a member of the teaching team responsible for designing the OSCE exam.
  - Expert 2: An experienced medical doctor, external to the Complutense University of Madrid, with no involvement in the design of the exam.
- One inexperienced evaluator: An individual without medical training.
- The GPT-4 model from OpenAI®: At the time of the study, this was the most advanced AI model available from OpenAI® and was utilized by the research team for evaluation purposes.

The human evaluators followed the same workflow, performing correction sessions of 2 h with 10 min breaks.

In all four cases, the corrections were made through the C-ECOE interface, which displayed the texts written by the students to the graders, along with the correction checklists for each corresponding section of the report. All human graders followed the same workflow, performing correction sessions of 2 h with 10 min breaks.

For the AI-based correction, the GPT-4 model from OpenAI® was used, with a prompt designed by the Digit-ECOE® team to optimize the accuracy of the results. This prompt defined the AI's role, assigning it the function of a university professor with experience in medical education. The AI's task was to evaluate the clinical reports prepared by the students, following a series of specific items established for each evaluation. These items and the reports were presented in a structured format with labels that facilitated data organization. Without making additional assumptions, the AI was instructed to analyze the reports and provide binary responses (1 or 0), depending on whether each item was present in the students' text. It was also asked to be strict in the correction and not to provide item descriptions.

Furthermore, the correction was carried out using separate queries for each item, ensuring that the AI assessed every clinical report section separately and completed as many "readings" as there were items in that section. This approach stopped AI from looking into pertinent information about an item in sections other than the one to which it belonged. For instance, if the item "relation between chest pain and postural changes" was designated to the anamnesis section, but the student mentioned it only in the physical examination section, the AI would not consider it fulfilled.

To ensure accuracy and avoid potential bias in future interactions, the AI was also instructed to "forget" all processed information at the end of each task.

Lastly, it is important to remember that the items utilized were not AI-optimized. This decision was made deliberately in order to investigate AI behavior in an authentic environment, simulating the exact circumstances and test items utilized in the student-administered exam. The wording of a few of these items was not ideal for automated interpretation. Some of these items had suboptimal wording for automated interpretation.

In particular, certain items were not formulated in strictly binary terms, as they involved multiple conditions to be evaluated within a single statement. Examples of such items include "patient's age and sex" or "Acute phase reactants: ESR, CRP, cardiac enzymes CK-MB, troponins" which could create ambiguity for both the AI and human graders.

2.5. Ethical Considerations and Confidentiality

The study was conducted after the grading and publication of the official OSCE results. The grading performed for the purposes of this study was conducted ex novo and did not influence the final scores of the students who participated in the exam.

All personal data of the students were removed from the database prior to the grading process, ensuring that neither the human graders nor the AI had access to any identifying information about the students.

The clinical reports written by the students were generated within the simulated environment of the OSCE and did not contain data from real patients.

The study protocol was reviewed and approved by the Research Committee of the Infanta Cristina University Hospital in Parla, receiving approval on 8 September 2023.

2.6. Statistical Analysis

All the data obtained were stored in the same database for subsequent statistical analysis.

The data were analyzed using IBM SPSS Statistics software, version 29.0.2.0 (20). Descriptive statistics were calculated for the mean scores. The correlation between the evaluation systems (human graders and AI) was analyzed using the intraclass correlation coefficient (ICC) for single measures and average measures. The differences between means scores were analyzed using the analysis of variance (ANOVA) with a Greenhouse–Geisser correction. Paired-sample t-tests were conducted to compare specific groups. Effect sizes were calculated using Cohen's d and Hedges' correction.

3. Results

A total of 1824 items were analyzed (19 items for each of the 96 students). The grading times are presented in Table 2. Notably, the AI was significantly faster, completing the analysis in just 24 min, which represents less than one-fifth of the time required by an expert evaluator.

Table 2. Total items analyzed and correction times for the AI, an expert, and an inexperienced grader.

Grader	Subject	Total Items	Correction Time	Breaks	Total Time
Expert 2	Cardiology	1824	2 h 15 min	10 min	2 h 25 min
Inexperienced	Cardiology	1824	4 h	10 min	4 h 10 min
AI	Cardiology	1824	2 4 min	N/A	24 min

The average score assigned by the AI was 8.88, with a standard deviation of 2.96. On average, human evaluators are assigned higher scores than AI. The combined average of both experts was 12.39, with a standard deviation of 3.22. The inexperienced grader also had an average score that was higher than the AI's but lower than that of the expert evaluators (Table 3).

Table 3. Comparison of means and standard deviations for the AI, experts, and the inexperienced grader.

	Maximum	Minimum	Mean	Standard Deviation
AI	16.5	3	8.88	2.96
EXP1	18	4.5	12.45	3.42
EXP2	18	5	12.33	3.22
INEXP	17	3	11.05	3.16
EXPERTS' MEAN			12.39	3.22

The ICC was calculated (Table 4) to analyze the correlation between the scores assigned by the different graders. In comparing single measures between the experts and the AI (EXP1-EXP2-AI), an ICC of 0.77 (95% CI: 0.699–0.834) was obtained. The correlation was even higher for average measures, reaching an ICC of 0.91 (95% CI: 0.875–0.938). Notably, the AI showed a higher performance compared to the inexperienced grader, whose correlation with the AI was lower, with an ICC of 0.72 (95% CI: 0.608–0.804) for single measures.

Table 4. Intraclass correlation coefficient displaying the correlation between AI, experts (EXP1, EXP2), and the inexperienced grader (INEXP).

		Single Measures	Average Measures
EXP1-EXP2	IC (95)	0.88 (0.826–0.919)	0.94 (0.905–0.958)
	p	<0.001	<0.001
EXP1-EXP2-IN-EXP-AI	IC (95)	0.79 (0.725–0.842)	0.97 (0.913–0.955)
	p	<0.001	<0.001
EXP1-EXP2-AI	IC (95)	0.77 (0.699–0.834)	0.91 (0.875–0.938)
	p	<0.001	<0.001
INEXP-AI	IC (95)	0.72 (0.608–0.804)	0.84 (0.756–0.892)
	p	<0.001	<0.001

A Greenhouse–Geisser correction was applied to analyze the differences in the mean scores assigned, and a repeated measures analysis of variance (ANOVA) was performed, as the assumption of sphericity was not met according to Mauchly's test. The analysis showed (Table 5) statistically significant differences between graders ($F = 118.117$, $p < 0.001$), indicating that the AI and human evaluators' scores were inconsistent. Specific contrasts between graders revealed that the AI evaluated the reports with significant differences compared to Expert 1 ($F = 206.246$, $p < 0.001$) and Expert 2 ($F = 211.246$, $p < 0.001$) and the combined mean of both experts ($F = 235.862$, $p < 0.001$). Similarly, significant differences were found between the inexperienced grader and the AI, although these differences were minor compared to the experts ($F = 86.486$, $p < 0.001$).

Finally, the scores assigned by the AI were compared to the mean scores of the expert graders, and the mean difference was −3.51, indicating that, on average, the AI-assigned scores were 3.51 points lower than the experts. The correlation coefficient was 0.740, with a *p*-value < 0.001, reflecting a significant positive correlation between the two evaluations. The paired-sample Student's *t*-test yielded a value of $t = -15.358$, with 95 degrees of freedom and a *p*-value < 0.001, confirming that the scores were statistically significant. The 95% confidence interval for the difference was between −3.96420 and −3.05664. The effect

size, calculated through Cohen's d, was 2.23957, while Hedges' correction yielded a value of 2.25745, indicating a large effect size.

Table 5. Repeated measures ANOVA displaying the comparison of scores between the different evaluators and the AI.

	F	*p*
EXP1 vs. AI	206.246	<0.001
EXP2 vs. AI	211.246	<0.001
INEXP vs. AI	86.486	<0.001
MEDIA EXPE vs. AI	235.862	<0.001

4. Discussion

This study assessed the potential of the GPT-4 model as a tool for grading clinical reports in the OSCE, highlighting its strengths and limitations compared to human evaluators. The findings revealed several key points, starting with the significant correlation between AI and human grading, which supports AI's capacity to adhere to predefined evaluation standards. However, the AI's grading was more stringent, assigning lower scores than experts and completing tasks significantly faster.

These results align with previous studies indicating that while AI systems have shown promising levels of accuracy and consistency in educational assessments [15,29], including evaluations of medical reasoning or exam performance [21,23,24], their outcomes are not yet exceptional. For instance, ChatGPT's performance in USMLE-style questions achieved results comparable to medical students [26], which supports its potential in this environment.

The tendency of the AI to assign lower grades than human graders, especially experts, was one of the study's most interesting findings. This discrepancy can be attributed to the inherent rigidity of the AI model, which strictly adheres to the explicit criteria established in the checklists. Although this assures consistent evaluation, it also seems to restrict the AI's capacity to adjust to appropriate variances in medical language or interpret implicit inferences, a skill that human graders are better equipped to manage because of their clinical backgrounds. For instance, in the item "Make a diagnostic judgment: coronary syndrome/osteochondritis", human graders awarded the point if either of the two conditions was mentioned, whereas the AI required explicit mention of both, leading to significant discrepancies.

It is crucial to emphasize that these variations represent opportunities for better AI configuration and design rather than necessarily indicating that its use is invalid. According to published research, AI models like GPT-4 can be made more capable of identifying implicit descriptions or better adjusting to changes in medical language by modifying their prompts and algorithms [20,23,24,30]. According to recent research, AI can perform on clinical reasoning exams at levels comparable to those of medical students, indicating that it might be used in combination with other methods to evaluate clinical competencies [17,20,21,24].

Another relevant finding of this study is that the AI assigned substantially lower scores than human graders, including experts. This result should be viewed as a sign of intrinsic variations in the evaluation process that indicate the need for more research rather than a weakness in the AI. The method by which the items on the checklist are designed may be one important component. Prior research has emphasized the necessity of optimizing item design to ensure clarity, remove ambiguity, and preserve uniqueness [4,9,10], facilitating a review by automated systems as well as humans. In this way, the AI

may classify these ambiguities as errors if it applies the criteria strictly and literally, while human graders are more flexible in their interpretation.

Additionally, there are other elements that might affect how human graders score [6,8]. One such factor is fatigue, which results a decline in accuracy and an increase in leniency as evaluation sessions go on, as demonstrated in in earlier studies [6,7]. This issue is linked to the requirement to maintain effort and concentration over extended periods of time, which could lead to less strict grading. Even though expert human graders are the best at this kind of evaluation, their performance during lengthy and repetitive tasks may suffer from cumulative fatigue.

The OSCE has traditionally been seen as a summative assessment with the main goal of attesting the development of basic competencies, which is another significant historical feature. In this situation, many human graders follow the saying, "when in doubt, favor the student." This approach emphasizes the identification of essential competencies over meticulous analysis, with the goal of not unfairly punishing students for minor errors or variations in their clinical language. Given that the AI lacks this interpretative flexibility, this practice may help to explain why human scores are generally higher.

However, it is important to understand that AI in its current state cannot completely replace human graders, especially for items which require complicated judgments or contextual clinical interpretation.

Additionally, under the European Union's proposed AI Act, it is mandatory to incorporate a "human-in-the-loop" approach to ensure the effective human oversight of AI systems, particularly those categorized as high-risk. This framework, outlined in Article 14, emphasizes the necessity of human supervision to prevent risks and safeguard fundamental rights. While this supervision combines the benefits of algorithmic consistency with the contextual discretion of human graders, it also introduces challenges related to clarifying the responsibilities of the "human-in-the-loop" and ensuring the effectiveness of this oversight [31].

However, integrating this regulatory framework with a hybrid and parallel model, where human oversight remains integral throughout the grading process, offers a practical solution to these challenges by preserving ethical accountability.

Using a hybrid and parallel model, where AI assists in grading initially but a human grader is always present to oversee and finalize scores, represents a robust approach to balancing efficiency with ethical accountability. This model would consistently maintain a "human-in-the-loop", ensuring that contextual discretion and ethical considerations are upheld. Such a framework allows AI to leverage its consistency and speed while human graders provide the critical oversight necessary to interpret complex or nuanced cases.

The advantages of this model include not only the reduction in workload for academic staff but also the assurance that ethical principles in grading are maintained. This is particularly crucial given that exam scores directly impact students' academic trajectories and future opportunities. Entrusting this responsibility solely to an AI could lead to ethical dilemmas, as algorithmic decisions may lack the fairness and empathy required for high-stakes evaluations. By ensuring human involvement throughout the grading process, this hybrid model safeguards the integrity of the evaluation system and builds trust among students and educators alike.

In this regard, a thorough qualitative examination of the differences between AI and human evaluations would be necessary to detect systematic error trends and immediate adjustments to the model's architecture. Additionally, to confirm AI's suitability in various clinical and educational settings, a larger sample of subjects and contexts—including a variety of institutions—remains integral.

Regarding time efficiency, this study confirms that the AI completed item grading significantly faster than human graders, reducing the required time. Notably that this

speed will likely continue to improve in the future, while the speed of human grading will remain constant. The same trend applies to accuracy, where AI is expected to enhance its precision over time.

While acknowledging the significant potential of AI to enhance the evaluation of clinical reports in OSCEs, this study also identifies important areas that require further validation and improvement. An important development in the evaluation of clinical competences may be the integration of AI with human expertise, which would ensure increased efficiency and accuracy in the process.

5. Conclusions

The integration of artificial intelligence in grading clinical reports within the context of OSCEs is a promising tool. It can provide consistent and objective results while alleviating some of the inherent challenges of manual correction, such as the subjectivity and fatigue of human evaluators.

Formulating assessment items plays a crucial role in the effectiveness of automated grading. Ambiguous or low-quality evaluation items can hinder the performance of both AI and human graders.

A hybrid model, where AI and human evaluators collaborate, emerges as the most practical and ethical approach. Such a model maintains human oversight throughout the process, ensuring that ethical considerations are upheld, especially given the high-stakes nature of exam scores on students' futures.

Future studies should focus on refining AI algorithms to better handle nuanced language and complex medical scenarios, as well as exploring ways to optimize hybrid models. Expanding the scope of research to diverse institutional contexts will be crucial to validate the generalizability of these findings. By advancing these efforts, AI can be effectively integrated into the educational landscape, ensuring both efficiency and fairness in student evaluations.

Author Contributions: Conceptualization**:** D.L., M.T.A., C.T.C. and M.V.V.G. Methodology: D.L., M.T.A. and C.T.C. Software: L.M.R.C. Validation**:** K.D.S.S., J.D.J., M.P.C., L.M.R.C., J.J.G.M., F.J.G.S., A.L.G.-A., R.R.G., R.D.L., Y.T.-C. and M.V.V.G. Formal analysis: J.J.G.M., F.J.G.S., A.L.G.-A. and R.R.G. Investigation: D.L., M.T.A., C.T.C., L.M.R.C., J.J.G.M., F.J.G.S., A.L.G.-A., R.R.G., R.D.L., Y.T.-C. and M.V.V.G. Resources: J.J.G.M., F.J.G.S. and A.L.G.-A. Data curation: L.M.R.C. Writing—original draft: D.L., M.T.A. and C.T.C. Writing—review and editing: K.D.S.S., J.D.J., M.P.C., L.M.R.C., J.J.G.M., F.J.G.S., A.L.G.-A., R.R.G., R.D.L., Y.T.-C. and M.V.V.G. Visualization: L.M.R.C. Supervision: D.L., J.J.G.M. and M.V.V.G. All authors have read and agreed to the published version of the manuscript.

Funding: the funding for this study was provided by the Spanish Society for Medical Education (SEDEM) through Grant 3/2024.

Institutional Review Board Statement: The study was conducted in accordance with the guidelines of the Declaration of Helsinki, and the requirement for IRB approval was waived.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Some of the authors of this study (Davide Luordo, Luis Miguel Rodríguez Cruz, Marta Torres Arrese, and Cristina Tristán Calvo) are owners of the company DIGIT-ECOE S.L., which provided the software support used for the development of the exam. However, this article is not intended to promote the applications of DIGIT-ECOE, and the authors do not receive any commercial benefit from its mention in the study. The artificial intelligence technology used for exam correction is ChatGPT-4, owned by OpenAI, which is not related in any way to DIGIT-ECOE. Therefore, the authors declare no conflicts of interest that could influence the results or interpretations of this work.

References

1. Epstein, R.M. Assessment in Medical Education. *N. Engl. J. Med.* **2007**, *356*, 387–396.
2. Patrício, M.F.; Julião, M.; Fareleira, F.; Carneiro, A.V. Is the OSCE a Feasible Tool to Assess Competencies in Undergraduate Medical Education? *Med. Teach.* **2013**, *35*, 503–514. https://doi.org/10.3109/0142159X.2013.774330.
3. Harden, R.M.; Gleeson, F.A. Assessment of Clinical Competence Using an Objective Structured Clinical Examination (OSCE). *Med. Educ.* **1979**, *13*, 39–54. https://doi.org/10.1111/j.1365-2923.1979.tb00918.x.
4. Turner, J.L.; Dankoski, M.E. Objective Structured Clinical Exams: A Critical Review. *Fam. Med.* **2008**, *40*, 574–578.
5. Lobato, R.D.; Lagares, A.; López-Medrano, F.; Villena, V.; Fernández, A.; Martínez-López, J.; Rubio, G.; Munárriz, P.M.; Alen, J.F. Examen clínico objetivo y estructurado formativo tras inmersión clínica precoz empleando estudiantes de sexto curso como observadores y administradores de retroalimentación. *FEM Rev. Fund. Educ. Méd.* **2014**, *17*, 179–186.
6. McLaughlin, K.; Ainslie, M.; Coderre, S.; Wright, B.; Violato, C. The Effect of Differential Rater Function over Time (DRIFT) on Objective Structured Clinical Examination Ratings. *Med. Educ.* **2009**, *43*, 989–992. https://doi.org/10.1111/j.1365-2923.2009.03438.x.
7. Humphris, G.M.; Kaney, S. Examiner Fatigue in Communication Skills Objective Structured Clinical Examinations. *Med. Educ.* **2001**, *35*, 444–449. https://doi.org/10.1046/j.1365-2923.2001.00893.x.
8. Chong, L.; Taylor, S.; Haywood, M.; Adelstein, B.-A.; Shulruf, B. The Sights and Insights of Examiners in Objective Structured Clinical Examinations. *J. Educ. Eval. Health Prof.* **2017**, *14*, 34. https://doi.org/10.3352/jeehp.2017.14.34.
9. García-Puig, J.; Vara-Pinedo, F.; Vargas-Núñez, J.A. Implantación del Examen Clínico Objetivo y Estructurado en la Facultad de Medicina de la Universidad Autónoma de Madrid. *Educ. Méd.* **2018**, *19*, 178–187. https://doi.org/10.1016/j.edumed.2017.01.003.
10. Ramos, J.M.; Martínez-Mayoral, M.A.; Sánchez-Ferrer, F.; Morales, J.; Sempere, T.; Belinchón, I.; Compañ, A.F. Análisis de la prueba de evaluación clínica objetiva estructurada (ECOE) de sexto curso en la Facultad de Medicina de la Universidad Miguel Hernández de Elche. *Educ. Méd.* **2019**, *20*, 29–36. https://doi.org/10.1016/j.edumed.2017.07.020.
11. Cusimano, M.D.; Cohen, R.; Tucker, W.; Murnaghan, J.; Kodama, R.; Reznick, R. A Comparative Analysis of the Costs of Administration of an OSCE (Objective Structured Clinical Examination). *Acad. Med.* **1994**, *69*, 571–576. https://doi.org/10.1097/00001888-199407000-00014.
12. Roberts, J.K.; Sullivan, M.; Atwater, S.; Desai, K.; Prabhu, N.K.; Hertz, J.T.; Buhr, G.T.; Peyser, B.; Weigle, N. Use of Virtual Interactive Patient Encounters to Prepare First-Year Medical Students for Clinical Practice. *Acad. Med.* **2023**, *98*, 1146–1153. https://doi.org/10.1097/ACM.0000000000005286.
13. Zhang, W.; Cai, M.; Lee, H.J.; Evans, R.; Zhu, C.; Ming, C. AI in Medical Education: Global Situation, Effects and Challenges. *Educ. Inf. Technol.* **2024**, *29*, 4611–4633. https://doi.org/10.1007/s10639-023-12009-8.
14. Borakati, A. Evaluation of an International Medical E-Learning Course with Natural Language Processing and Machine Learning. *BMC Med. Educ.* **2021**, *21*, 181. https://doi.org/10.1186/s12909-021-02609-8.
15. Mustafa, M.Y.; Tlili, A.; Lampropoulos, G.; Huang, R.; Jandrić, P.; Zhao, J.; Salha, S.; Xu, L.; Panda, S.; Kinshuk; et al. A Systematic Review of Literature Reviews on Artificial Intelligence in Education (AIED): A Roadmap to a Future Research Agenda. *Smart Learn. Environ.* **2024**, *11*, 59. https://doi.org/10.1186/s40561-024-00350-5.
16. Chan, K.S.; Zary, N. Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. *JMIR Med. Educ.* **2019**, *5*, e13930. https://doi.org/10.2196/13930.
17. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. *PLoS Digit. Health* **2023**, *2*, e0000198. https://doi.org/10.1371/journal.pdig.0000198.

18. Manne, R.; Kantheti, S.C. Application of Artificial Intelligence in Healthcare: Chances and Challenges. *Curr. J. Appl. Sci. Technol.* **2021**, *40*, 78–89. https://doi.org/10.9734/cjast/2021/v40i631320.
19. Ávila-Tomás, J.F.; Mayer-Pujadas, M.A.; Quesada-Varela, V.J. La inteligencia artificial y sus aplicaciones en medicina II: Importancia actual y aplicaciones prácticas. *Aten. Prim.* **2021**, *53*, 81–88. https://doi.org/10.1016/j.aprim.2020.04.014.
20. Sun, L.; Yin, C.; Xu, Q.; Zhao, W. Artificial Intelligence for Healthcare and Medical Education: A Systematic Review. *Am. J. Transl. Res.* **2023**, *15*, 4820.
21. Kufel, J.; Paszkiewicz, I.; Bielówka, M.; Bartnikowska, W.; Janik, M.; Stencel, M.; Czogalik, Ł.; Gruszczyńska, K.; Mielcarska, S. Will ChatGPT Pass the Polish Specialty Exam in Radiology and Diagnostic Imaging? Insights into Strengths and Limitations. *Pol. J. Radiol.* **2023**, *88*, 430–434. https://doi.org/10.5114/pjr.2023.131215.
22. Schaye, V.; Guzman, B.; Burk-Rafel, J.; Marin, M.; Reinstein, I.; Kudlowitz, D.; Miller, L.; Chun, J.; Aphinyanaphongs, Y. Development and Validation of a Machine Learning Model for Automated Assessment of Resident Clinical Reasoning Documentation. *J. Gen. Intern. Med.* **2022**, *37*, 2230–2238. https://doi.org/10.1007/s11606-022-07526-0.
23. Strong, E.; DiGiammarino, A.; Weng, Y.; Basaviah, P.; Hosamani, P.; Kumar, A.; Nevins, A.; Kugler, J.; Hom, J.; Chen, J.H. Performance of ChatGPT on Free-Response, Clinical Reasoning Exams. *Med. Educ. March* **2023**, *29*. https://doi.org/10.1101/2023.03.24.23287731.
24. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv* **2023**, arXiv.2305.09617. https://doi.org/10.48550/arXiv.2305.09617.
25. Mayol, J. Inteligencia artificial generativa y educación médica. *Educ. Méd.* **2023**, *24*, 100851. https://doi.org/10.1016/j.edumed.2023.100851.
26. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med. Educ.* **2023**, *9*, e45312. https://doi.org/10.2196/45312.
27. Cooper, G. Examining Science Education in ChatGPT: An Exploratory Study of Generative Artificial Intelligence. *J. Sci. Educ. Technol.* **2023**, *32*, 444–452. https://doi.org/10.1007/s10956-023-10039-y.
28. Dave, T.; Athaluri, S.A.; Singh, S. ChatGPT in Medicine: An Overview of Its Applications, Advantages, Limitations, Future Prospects, and Ethical Considerations. *Front. Artif. Intell.* **2023**, *6*, 1169595. https://doi.org/10.3389/frai.2023.1169595.
29. Eysenbach, G. The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation with ChatGPT and a Call for Papers. *JMIR Med. Educ.* **2023**, *9*, e46885. https://doi.org/10.2196/46885.
30. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (accessed on 22 December 2024).
31. Key Issue 4: Human Oversight-EU AI Act. Available online: https://www.euaiact.com/key-issue/4 (accessed on 22 December 2024).