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

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

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

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.

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.

INTRODUCCIÓN

Hoy en día, las instituciones educativas (incluidas las tradicionales y las en línea) generalmente adoptan las tecnologías de la información y la comunicación para apoyar a su profesorado y estudiantado. La automatización de la tarea de calificación es una tendencia seguida por muchos Sistemas de Gestión del Aprendizaje (SGA) para evaluar de manera eficiente las actividades y proporcionar un *feedback* oportuno en el momento adecuado con finalidad formativa (György y Vajda, 2007). Sin embargo, muchos SGA limitan el tipo de preguntas que se califican automáticamente a aquellas basadas en la selección de opciones o texto simple, debido a la complejidad de evaluar preguntas cortas o de tipo ensayo.

Aunque la Calificación Automática de Preguntas de Respuesta Corta (CAPRC) ha sido ampliamente explorada en la literatura (Burrows et al., 2015), las contribuciones se han limitado a explorar su precisión en conjuntos de datos públicos sin considerar su efectividad en entornos educativos reales. Además, el renacimiento de la Inteligencia Artificial (IA) debido al acceso a recursos computacionales y a herramientas específicas para aplicar técnicas de IA, han permitido que los SGA clásicos evolucionen. Los algoritmos de IA pueden reforzar los SGA para apoyar el proceso de aprendizaje. Las recomendaciones, el *feedback* inmediato o la CAPRC son ejemplos de tareas que la IA puede mejorar.

Esta investigación tiene como objetivo proporcionar un punto de partida para un sistema de aprendizaje inteligente capaz de recomendar la evaluación y el *feedback* para preguntas cortas utilizando técnicas de IA. Este trabajo también muestra cómo se utiliza el sistema en un entorno educativo híbrido y presenta los resultados obtenidos. A pesar de que los modelos de lenguaje extenso –más conocidos por su denominación en inglés *Large Language Models* (LLM)– y las herramientas generativas derivadas, como ChatGPT (OpenAI, 2024), muestran actualmente su gran potencial en cualquier tarea académica, siguen teniendo limitaciones para tareas o dominios específicos como es el caso de la Propiedad Intelectual.

Por lo tanto, este trabajo propone una solución alternativa, utilizando BERT, un Modelo de Lenguaje (ML) ligero que comprende el significado de las respuestas. En lugar de delegar la decisión de la evaluación de una pregunta de respuesta corta en una herramienta de IA Generativa (IAGen) mediante la definición de criterios expresados como un conjunto complejo de instrucciones (*prompt*), se utiliza BERT. Este modelo comprende el significado de cualquier respuesta correcta e incorrecta, siendo más preciso para fines evaluativos. Al ser menos extenso, contribuye a un uso computacional más eficiente y sostenible, gestionable en instancias privadas (es decir, no se necesitan soluciones empresariales), reduciendo también los problemas éticos y de privacidad. Por lo tanto, este trabajo ha integrado esta solución de IA en un SGA (concretamente, Moodle) para que el profesorado pudiera proporcionar una evaluación significativa al estudiantado.

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), pp. 321-352. https://doi.org/10.5944/ried.45541

Se sigue una metodología de investigación de acción combinada con un enfoque de diseño y creación para crear el sistema automatizado de aprendizaje inteligente (denominado *Smart Learning Automated System*, SLASys), explícitamente adaptado a un curso de formación en Propiedad Intelectual, para definir un sistema de aprendizaje que facilite el éxito del estudiantado mediante el uso de técnicas de IA aplicadas a la enseñanza de la Propiedad Intelectual. El estudio se ha llevado a cabo en el área educativa de la Oficina Europea de Patentes (OEP).

Este trabajo se estructura de la siguiente manera. En primer lugar, se presentan los antecedentes teóricos de los modelos de lenguaje y sus usos educativos, centrándose en la CAPRC. A continuación, se detalla el diseño de la investigación, incluyendo la metodología, las preguntas de investigación, la herramienta SLASys y el análisis de datos. La sección de resultados destaca los principales hallazgos. El trabajo concluye con la discusión, limitaciones, contribuciones e investigación futura.

REVISIÓN DE LA LITERATURA

Modelos de Lenguaje Extenso, BERT y su aplicación en educación

Los LLM han surgido como una tecnología transformadora en los últimos años, demostrando capacidades notables en diversas tareas de Procesamiento del Lenguaje Natural (PLN). Su rápido progreso ha propiciado avances significativos en la traducción automática, la síntesis de textos e incluso la generación de código (Husein et al., 2025; Pang et al., 2025; Zhang et al., 2025). A pesar de su impresionante rendimiento, los LLM se enfrentan a limitaciones y desafíos en entornos educativos reales, como interpretaciones erróneas en lenguajes específicos de dominio (Arefeen et al., 2024) o su interpretabilidad y explicabilidad (Zhao et al., 2024). Las innovaciones recientes en computación han mejorado su rendimiento, permitiendo alcanzar el nivel de las herramientas de IAGen de última generación, como GPT-4o (OpenAI, 2024). Estas herramientas demuestran capacidades impresionantes de aprendizaje, aplicabilidad a tareas de razonamiento limitadas y capacidad de generalización en diversos dominios.

BERT (*Bidirectional Encoder Representations from Transformers*) es un modelo de PLN de última generación desarrollado por Google en 2018 (Devlin et al., 2019). BERT es un ML basado en transformadores previamente entrenado con una gran colección de documentos de texto. BERT se consideró inicialmente un LLM, pero actualmente se considera un ML en comparación con sus sucesores de IAGen. Su arquitectura se fundamenta en la representación del texto. Una frase está codificada de una manera especial (denominada *embeddings*) que almacena su significado (es decir, su semántica), lo que permite la aplicación directa en tareas de PLN. Su desarrollo y aplicación continuados en múltiples dominios de conocimiento demuestran su versatilidad y potencial para impulsar nuevos avances en IA y en la comprensión del lenguaje. BERT se ha utilizado para la clasificación de documentos

(Adhikari et al., 2019), el reconocimiento de categorías gramaticales (Souza et al., 2019), la implicación lógica entre pares de oraciones (Wang et al., 2018) o la síntesis de textos (Zhang, Cai et al., 2019).

Tanto BERT como las herramientas de IAGen se utilizan en educación afrontando desafíos similares. La escalabilidad y la integración con las infraestructuras existentes requieren de una inversión significativa y de conocimiento técnico (Xu y Zhu, 2023). Además, deben considerarse principios relacionados con la ética, la seguridad y la alineación con los objetivos educativos fundamentales (García-Peñalvo et al., 2024), y desarrollar la alfabetización en IA entre el profesorado es esencial para fomentar un uso crítico (Petridou y Lao, 2024). Asimismo, aunque parecen aplicables a cualquier tarea, las herramientas de IAGen son principalmente aplicables a la generación y el análisis de texto, la definición de conceptos, la creación de ejercicios o la mejora del estilo de escritura; mientras que BERT es válido para la clasificación de texto, el análisis de sentimientos, la comprensión de las dudas del estudiantado o la provisión de explicaciones detalladas (Qiu y Jin, 2024).

La evaluación también es una tarea relevante en ambos modelos. Por un lado, las herramientas de IAGen han empezado a explorarse en la evaluación como "cajas negras", es decir, sin saber cómo funcionan internamente. Se han utilizado para producir *feedback*, pero también para la evaluación automática. La evaluación automática beneficia a algunos dominios, como el aprendizaje de idiomas (Escalante et al., 2023), la adquisición de habilidades de escritura (Banihashem et al., 2024) o el diseño multimedia (Almasre, 2024). Por otro lado, BERT se puede utilizar en tareas de evaluación específicas. El profesorado puede generar preguntas a partir de textos existentes, que se pueden utilizar en tests y exámenes (Nguyen et al., 2022). Los modelos basados en BERT también pueden evaluar la coherencia, la relevancia y la gramática de los ensayos del estudiantado, lo que ayuda a que este pueda mejorar sus habilidades de escritura (Wang et al., 2022). También se han diseñado aplicaciones basadas en BERT para comprender expresiones lingüísticas y matices culturales (Bahdanau et al., 2015), y recomendadores específicos que pueden ayudar al estudiantado a practicar la expresión oral y la comprensión de nuevos idiomas (Zhang, Zhang et al., 2019).

Aunque la evaluación automatizada parece ser una de las características futuras de la IAGen en educación, actualmente no está recomendada por las políticas institucionales ni gubernamentales (European Commission, 2024; Dai et al., 2025; González Fernández et al., 2025), porque las herramientas de IAGen son propensas a errores debido a alucinaciones (Jia et al., 2024) y su utilización podría plantear dilemas éticos. Además, las herramientas de IAGen pueden fallar en dominios con un lenguaje específico. Por ejemplo, algunas preguntas relacionadas con conceptos específicos en el examen de patentes dentro del dominio de la Propiedad Intelectual, como Novedad y Claridad, a veces son interpretadas por una herramienta de IAGen como si se tratara de evaluar si la respuesta es nueva y clara en lenguaje natural. Sin embargo, dichos conceptos especializados en Propiedad Intelectual

se refieren a cuán novedosa y clara está descrita una invención en una solicitud de patente (*reivindicación*). Las herramientas de IAGen podrían analizar la respuesta mejorando el contexto anterior (es decir, su *contextualización*), pero a veces mezclan directrices de patentes de diferentes fuentes con información obsoleta. Por lo tanto, se requiere añadir documentos específicos al modelo (es decir, es necesario realizar un *ajuste fino*) para mejorar la calidad de un modelo propietario de IAGen, lo que implica algunas preocupaciones acerca de la privacidad, y un debate sobre si es una solución rentable. En cambio, una aplicación basada en BERT aborda la tarea de evaluación comprendiendo las respuestas proporcionadas a una pregunta. BERT podría utilizarse para realizar una *comparación semántica* para evaluar lo cerca que está el significado de dos frases (la respuesta de un estudiante en comparación con la respuesta correcta del profesorado) o entrenar un modelo de IA de *clasificación de respuestas* para predecir si el significado de una frase se acerca a un conjunto correcto o incorrecto de frases (es decir, decidir si la respuesta de un estudiante es correcta o incorrecta en base a un conjunto de respuestas correctas e incorrectas). Estos enfoques evitan, además, el inconveniente de la variabilidad que las herramientas de IAGen exhiben actualmente.

Calificación automática de preguntas cortas y provisión de feedback

El *feedback* constituye uno de los componentes esenciales tanto del proceso de aprendizaje (Evans, 2013; Hattie y Timperley, 2007) como de la evaluación de la adquisición de conocimientos, habilidades y competencias. El *feedback* también es un elemento clave para fomentar la implicación del estudiantado y mejorar su aprendizaje (Winstone et al., 2017). Su objetivo es proporcionar información específica vinculada a una actividad de aprendizaje, con el fin de reducir la brecha entre la comprensión deseada y la real (Sadler, 1989). Además, el *feedback* es crucial en entornos en línea para garantizar un proceso de aprendizaje significativo y oportuno (Nicol y Macfarlane-Dick, 2006). La elaboración y entrega de *feedback* útil requiere de un esfuerzo considerable por parte del profesorado, lo que restringe su nivel de personalización cuando existen restricciones de tiempo o un gran número de estudiantes matriculados (Dhananjaya et al., 2024). Sin embargo, el *feedback* personalizado motiva al estudiantado en mayor medida, regula mejor su comportamiento, contribuye a una mejor adquisición de conocimientos y mejora el compromiso con los materiales de aprendizaje (Kim, 2023; Wang y Lehman, 2021).

La automatización de la generación de *feedback* se ha planteado como una solución tentativa para apoyar al profesorado. Aunque en la literatura se ha utilizado una terminología diversa para referirse a estas herramientas –por ejemplo, sistemas de tutoría inteligente, plataforma de enseñanza, sistema automatizado de *feedback*, herramientas de *feedback* en línea– todas ellas se centran en mejorar la eficiencia del profesorado mientras ayudan al estudiantado durante su proceso de aprendizaje (Xie y Li, 2018). Trabajos previos han presentado las técnicas aplicables en diferentes

contextos, como sería la generación de *feedback* para ensayos (Akçapinar, 2015), ejercicios de programación (Messer et al., 2024) o tareas colaborativas (Zheng et al., 2023).

La calificación automatizada de tests también ha sido ampliamente investigada debido a sus beneficios, tanto para proporcionar una evaluación automática como para facilitar los procesos de evaluación formativa (Burrows et al., 2015), así como su impacto positivo en el aprendizaje debido a la provisión de *feedback* (Gaona et al., 2018; Rezaei, 2015). Para simplificar su implementación, los tests han recurrido principalmente a preguntas de opción múltiple, mientras que las preguntas cortas y de ensayo han sido usualmente evitadas, ya que requieren de un mayor esfuerzo e intervención del profesorado durante la fase de evaluación, o bien se han utilizado métodos alternativos como la evaluación por pares (Huisman et al., 2017). A pesar de ello, existe trabajo previo relacionado con la CAPRC. Se ha utilizado la comparación de respuestas con una frase (Siddiqi y Harrison, 2008), con un conjunto de frases con un vocabulario similar (Klein et al., 2011) o con palabras clave principales (Saha et al., 2018) para detectar respuestas correctas. Aunque estas técnicas han demostrado buenos resultados, presentan algunas deficiencias, ya que no consideran la semántica (un patrón, un vocabulario similar o un conjunto de palabras clave pueden no abarcar todas las posibles respuestas correctas). Algunos autores han investigado la aplicabilidad de herramientas de IAGen para la CAPRC (Aggarwal et al., 2025; Grévisse, 2024). Aunque su aplicabilidad para la calificación de ensayos parece prometedora (Senthilnathan et al., 2025), los resultados para la CAPRC no son concluyentes para apoyar la evaluación del estudiantado. Las alucinaciones, la descontextualización, la información obsoleta y las limitaciones en el pensamiento crítico siguen afectando a su utilización (De La Cruz et al., 2024).

Dado que el PLN y los ML semánticos siguen siendo técnicas de vanguardia para la CAPRC, se utiliza BERT debido a sus capacidades, a su potencial precisión y a su mejor adaptación a entornos educativos reales. En la literatura se han explorado dos tipos de modelos. Algunos buscan predecir si la respuesta del estudiantado es correcta (Camus y 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), mientras que otros van más allá al predecir la calificación, demostrando su potencial en conjuntos de datos públicos (Baral et al., 2021; del Gobbo et al., 2023; Gaddipati et al., 2020; Metzler et al., 2024; Soulimani et al., 2024). Sin embargo, hasta donde llega el conocimiento de los autores de esta investigación, ningún trabajo previo ha probado estas herramientas en entornos educativos reales debido a los problemas de aceptabilidad o a la necesidad de intervención humana para su uso (Hustad y Arntzen, 2013; Xavier et al., 2025). Por lo tanto, este trabajo propone SLASys como herramienta de recomendación para la CAPRC, la cual ha sido probada en un entorno educativo híbrido durante cuatro ediciones de un curso fundamental de examen de patentes a lo largo de un año, con 120 estudiantes, con el fin de obtener conocimiento sobre su uso. Así, el objetivo es responder a las siguientes preguntas de investigación:

PI1. ¿Cómo de precisa es la clasificación de respuestas con respecto con la comparación semántica para la recomendación en la CAPRC?
PI2. ¿Cómo de preciso es SLASys en un entorno educativo real?
PI3. ¿Puede utilizarse SLASys sin respuestas previas del estudiantado?
PI4. ¿Cuál es la opinión del estudiantado y del profesorado?

METODOLOGÍA

Diseño de la investigación

SLASys sigue una metodología de investigación mixta que combina investigación de acción (Oates, 2006) con un enfoque de diseño y creación (Kuechler y Vaishnavi, 2012). La investigación de acción es adecuada porque el sistema se desarrolla y se prueba en entornos educativos reales, con el objetivo de abordar desafíos prácticos en la enseñanza y el aprendizaje. Esta metodología enfatiza la resolución de problemas del mundo real a través de un ciclo colaborativo e iterativo de planificación, acción y reflexión, al tiempo que recopila datos para evaluar los resultados. El enfoque de diseño y creación complementa dicha metodología, al centrarse en el desarrollo de artefactos tecnológicos innovadores, siguiendo un proceso iterativo de resolución de problemas de cinco pasos (Kuechler y Vaishnavi, 2012) para guiar la creación y el perfeccionamiento del sistema.

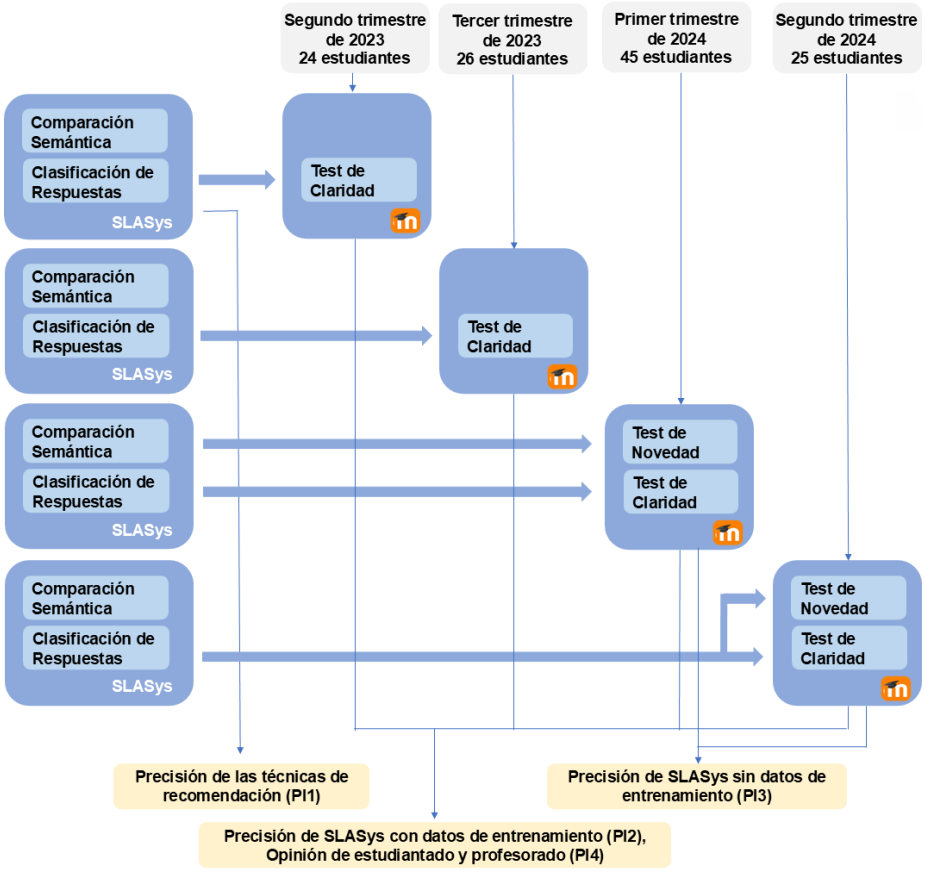
Este trabajo se centra en las dos primeras iteraciones del desarrollo del recomendador para la CAPRC. La primera consistió en la creación del primer artefacto y la evaluación de la solución técnica, mientras que la segunda evaluó su integración en el proceso de enseñanza-aprendizaje. Esta última iteración implicó probar el artefacto en un curso fundamental de examen de patentes en el dominio de la Propiedad Intelectual. Este curso, con una duración de 6 semanas, forma parte de los cursos orientados a la formación del estudiantado como examinador de patentes, que, a su vez, consta de seis cursos con una duración total de 2 años. Existen requisitos específicos para ser elegible para esta capacitación. El estudiantado debe tener un máster en física, química, ingeniería o ciencias naturales, ya que se requiere conocimiento de campos técnicos específicos para examinar las propuestas de solicitud de patentes.

El objetivo del curso es preparar al estudiantado para el examen de patentes, introduciendo los conceptos fundamentales, las directrices de la Convención Europea de Patentes (CEP), los sistemas disponibles para dicha tarea y los ejercicios relacionados con el examen de patentes. La metodología de aprendizaje combina una estrategia de clase invertida (Bergmann y Sams, 2012) con aprendizaje justo a tiempo (Novak, 2012). Las actividades de autoaprendizaje y la lectura de recursos de aprendizaje se realizan de forma asíncrona. También hay sesiones en línea síncronas con la participación del estudiantado y del profesorado, donde se debaten

las actividades individuales realizadas previamente de forma asíncrona y se realizan ejercicios prácticos, preguntas y actividades grupales.

El profesorado informó de dificultades al aplicar la estrategia de aprendizaje de justo a tiempo, ya que no disponían de tiempo suficiente para evaluar las actividades de respuesta corta del estudiantado y proporcionar un *feedback* significativo antes de las sesiones en línea. Por lo tanto, SLASys responde a estas necesidades proporcionando una solución para evaluar y proporcionar *feedback* a cada estudiante de manera eficiente.

Figura 1
Diseño de la investigación



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), pp. 321-352. https://doi.org/10.5944/ried.45541

La Figura 1 muestra el diseño de la investigación. En primer lugar, los enfoques diseñados se comparan para ver su precisión para responder a la PI1. En segundo lugar, durante cuatro ediciones del curso, se probó el enfoque de clasificación de respuestas en un test relacionado con el concepto de Claridad de patentes para responder a la PI2. En total, 120 estudiantes realizaron el test. En tercer lugar, se diseñó un nuevo test sobre el concepto de Novedad en las dos últimas ediciones sin datos de entrenamiento (es decir, sin respuestas del estudiantado de ediciones anteriores). En este caso, se utilizaron los dos enfoques diseñados para evaluar el rendimiento del recomendador y responder a la PI3. Un total de 70 estudiantes respondieron este segundo test. Además, se recopilaron las opiniones del estudiantado y las experiencias del profesorado en todas las ediciones para responder a la PI4.

Las pruebas con estudiantado real se realizaron en línea de forma síncrona usando un Moodle adaptado como SGA que integraba SLASys para recomendar los resultados de la evaluación. Los tests de Claridad y Novedad incorporaron 8 y 11 preguntas, respectivamente, y el estudiantado dispuso de una hora para responderlos. Su objetivo fue mejorar su conocimiento, pero no sus calificaciones. El estudiantado es evaluado en dos ocasiones durante su formación mediante exámenes presenciales.

SLASys: un recomendador para la CAPRC

SLASys se diseñó para determinar si la respuesta de cada estudiante a una pregunta de respuesta corta es correcta, dependiendo de la información disponible relacionada con la pregunta. La Figura 2 muestra las dos técnicas utilizadas para detectar si una respuesta es correcta. La pregunta es sobre el concepto de Novedad en el dominio de la Propiedad Intelectual, preguntando por qué la reivindicación de la patente no es novedosa. La comparación se efectúa utilizando *embeddings* de BERT. Aunque un *embedding* es una matriz matemática compleja, se podría ver como una codificación del significado de la frase mediante un conjunto de símbolos que definen su significado.

Figura 2
Ejemplo de pregunta sobre Novedad

¿Por qué esta reivindicación no es novedosa?
Un teléfono (móvil) con medios de marcación, un micrófono y un altavoz, caracterizado porque los medios de marcación están hechos de aluminio...

Respuesta modelo
La reivindicación no está clara debido a la expresión entre paréntesis (móvil). No está claro si la función entre paréntesis debe ser limitante o no. Consulta las directrices F-IV, 4.19.
(★, ,... ,)

	Embedding	Similitud del coseno	Probabilidad de ser clasificado como correcto
Respuesta 1 (correcta) La característica entre paréntesis que no es un signo de referencia; no está claro si es opcional o no (F.-IV, 4.19)	(★, ,... ,)	32,2 %	99,9 %
Respuesta 2 (incorrecta) No se permite la expresión entre paréntesis (no estoy seguro de si se encuentra dentro de la exención en este campo técnico).	(, ,... ,)	23,3 %	0 %

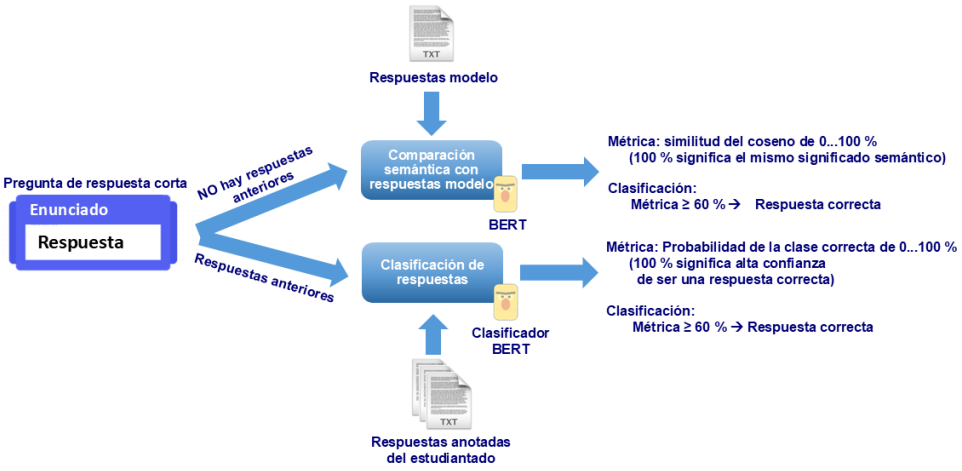
La primera técnica usa el *embedding* de la respuesta correcta proporcionada por el profesorado (denominada *respuesta modelo*), para compararla con las respuestas proporcionadas por el estudiantado. Para evaluar lo similar que es la respuesta de cada estudiante con respecto a la respuesta modelo del profesorado (es decir, cuántos símbolos son iguales en ambos *embeddings*) se utiliza la métrica de *similitud del coseno*. El resultado de la comparación (es decir, la *puntuación de la métrica*) es una probabilidad de la similitud, donde el 100 % indica que el significado es el mismo que el de la respuesta modelo del profesorado. La Figura 2 muestra que este método falla para la respuesta correcta, ya que hay diferentes respuestas correctas y la respuesta modelo del profesorado no cubre todas ellas.

Sin embargo, este proceso podría mejorarse comparando las respuestas del estudiantado con un conjunto de respuestas correctas e incorrectas (extraídas del estudiantado real de ediciones anteriores del curso después de ser evaluadas por el profesorado), denominadas *respuestas anotadas*. Cada respuesta puede ser anotada como correcta o incorrecta. Aunque la comparación puede realizarse usando la similitud del coseno para cada respuesta anotada, existe una técnica más eficiente que implica el entrenamiento de un clasificador de IA mediante un *ajuste fino* basado en los *embeddings* de la información anotada. Dicho ajuste especializa el clasificador para centrarse en las respuestas específicas anotadas y recopiladas para una pregunta. El resultado del modelo es la probabilidad de que el *embedding* se clasifique como una respuesta correcta, donde un valor del 100 % significa que el modelo tiene una alta confianza en que la respuesta sea correcta.

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), pp. 321-352. https://doi.org/10.5944/ried.45541

Dado que una pregunta de respuesta corta puede crearse desde cero (es decir, se trata de una pregunta nueva) o reutilizarse de otros contextos de aprendizaje, el sistema considera estos dos posibles escenarios (Figura 3). La comparación semántica se utiliza cuando se crea la pregunta y no hay datos disponibles del estudiantado anterior. La comparación semántica utiliza la similitud del coseno con un conjunto de respuestas modelo correctas proporcionadas por el profesorado. La clasificación de respuestas entrena un clasificador de IA cuando hay respuestas anteriores disponibles, proporcionando una predicción sobre si la respuesta es correcta. Como ambas métricas se expresan en porcentajes, la identificación del umbral de calidad para decidir si la respuesta enviada por el estudiante es correcta se resolvió como un problema de optimización, ya que el objetivo es maximizar el umbral mientras que la clasificación correcta no empeora significativamente. Después de algunos experimentos y de la validación con profesorado experto en el dominio, se concluyó que un umbral superior al 60 % identificó principalmente respuestas correctas.

Figura 3
Diseño técnico de SLASys

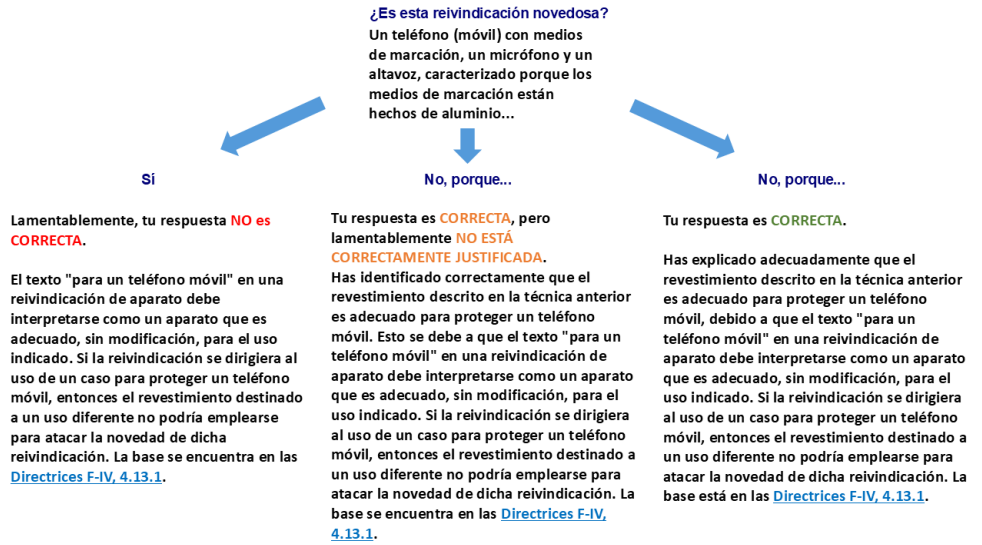


Integración de SLASys con Moodle para la evaluación y provisión de *feedback*

Dado que SLASys ofrece un proceso automático para recomendar la evaluación, el proceso de evaluación podría mejorarse proporcionando *feedback* automatizado. La Figura 4 muestra un ejemplo del *feedback* suministrado sobre la pregunta asociada a la reivindicación de la Figura 2. Es posible definir dos tipos de preguntas:

preguntas de respuesta corta simple o preguntas de opción con una justificación. Este segundo tipo de pregunta es relevante para el profesorado en el examen de patentes, ya que las preguntas suelen plantear alguna reivindicación relacionada con la definición de una patente, y el estudiantado debe examinar si la reivindicación cumple con las directrices de la CEP, como la novedad o la claridad. La Figura 4 muestra un ejemplo del segundo tipo de pregunta con el *feedback* proporcionado y el enlace a las directrices de la CEP según el resultado de la evaluación. El *feedback* está predefinido para cada opción y se ha configurado en el diseño de la pregunta, proporcionando información significativa para que el estudiantado comprenda el resultado de la evaluación y mejore su aprendizaje.

Figura 4
Ejemplo de feedback de una pregunta sobre Novedad



SLASys incorpora una API con diferentes servicios que se pueden integrar en cualquier SGA. En concreto, se utilizó una integración con Moodle mediante el desarrollo de un complemento (*plugin*) adaptado para acceder a las capacidades de SLASys. Esto permitió ocultar los aspectos más complejos, proporcionando una interfaz fácil de usar y de gestionar por el profesorado. Se integraron los nuevos tipos de pregunta en Moodle, adaptando la configuración de preguntas y la interfaz de evaluación para el profesorado, así como la interfaz del estudiantado para enviar las respuestas y revisar los resultados de la evaluación. El diseño minimiza la carga de trabajo del profesorado con una curva de aprendizaje extremadamente baja.

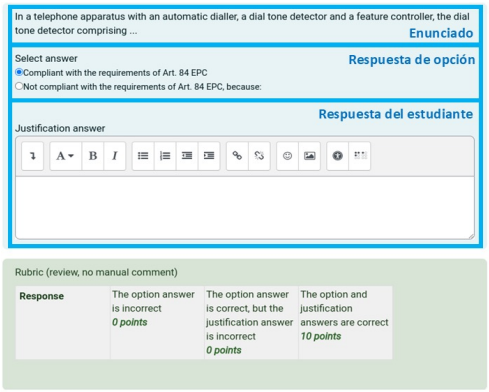
Se puede añadir una nueva pregunta a un test, proporcionando la información obligatoria siguiente: el tipo de pregunta, el enunciado y el conjunto de respuestas modelo del profesorado. Opcionalmente, se pueden agregar o recuperar las respuestas anotadas del estudiantado de ediciones anteriores del curso. Cuando no se proporciona esta última información, SLASys utiliza el enfoque de comparación semántica. En el caso contrario, el clasificador de IA de respuestas correspondiente se entrena automáticamente sin intervención técnica (Figura 3).

La particularidad más destacada está en la configuración de la evaluación y el *feedback*. La pregunta incorpora una rúbrica (un elemento no disponible para preguntas de respuesta corta en Moodle), permitiendo unificar los criterios de evaluación entre el profesorado y ofrecer el *feedback* correspondiente según los resultados de la evaluación.

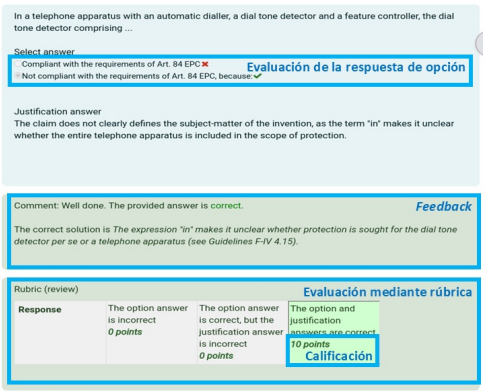
La Figura 5.a) ilustra la interfaz del estudiantado para enviar una respuesta a una pregunta de opción habilitada. La figura también muestra los diferentes niveles de la rúbrica y las calificaciones asociadas. La respuesta entregada se envía a SLASys, y la recomendación de evaluación se devuelve a Moodle cuando se ha procesado. La recomendación se muestra en la interfaz de evaluación (Figura 6). El profesorado puede visualizar las respuestas modelo, la recomendación de SLASys, la información interpretable del modelo de clasificación de respuestas y la puntuación de la métrica. La información interpretable utiliza colores para indicar qué palabras impactan positiva o negativamente en las decisiones del clasificador. Aunque a veces se incluyen palabras irrelevantes, el profesorado aún puede identificar las palabras incorrectas que afectan la forma en que se clasifican las respuestas del estudiantado. La interfaz también sugiere la calificación, así como el *feedback* según la rúbrica que se selecciona automáticamente dependiendo de la recomendación. El profesorado debe revisar toda esta información y seleccionar el nivel de rúbrica correcto.

Después de ser evaluados, los resultados del test se pueden revisar en la interfaz del estudiantado (Figura 5.b), de manera que cada estudiante puede ver su calificación y el *feedback* con base en la opción de rúbrica seleccionada, validada por el profesorado.

Figura 5
Interfaz de la pregunta en Moodle para el estudiantado

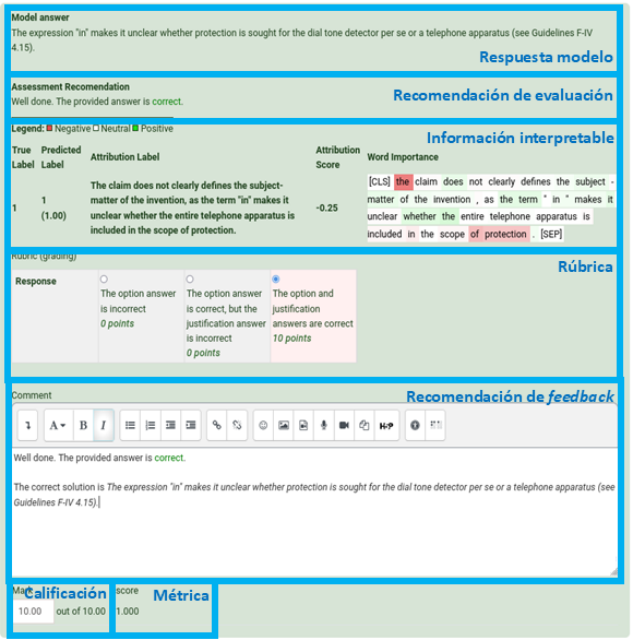


a) Interfaz del intento



b) Interfaz del *Feedback*

Figura 6
Interfaz de evaluación en Moodle para el profesorado



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), pp. 321-352. https://doi.org/10.5944/ried.45541

Análisis de datos

Para responder a las preguntas de investigación se han utilizado dos fuentes de datos. En primer lugar, la información de SLASys se ha empleado para analizar el número de evaluaciones correctas sobre las respuestas utilizadas para el entrenamiento (*rendimiento*) (PI1). A continuación, usando el complemento de Moodle, se recuperaron las recomendaciones proporcionadas por SLASys y la evaluación realizada por el profesorado para analizar las recomendaciones correctas durante la prueba con los estudiantes (PI2 y PI3). El rendimiento durante la fase de entrenamiento y prueba con el estudiantado se puede calcular a partir de las siguientes métricas:

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

$$TPR = \frac{TP}{TP + FN} \qquad F_{1,5} = \frac{(1 + 1{,}5^2)TP}{(1 + 1{,}5^2)TP + 1{,}5^2 FN + FP}$$

Donde TP denota el número de respuestas correctas correctamente identificadas, TN el número de respuestas incorrectas correctamente identificadas, FP el número de respuestas incorrectas no identificadas correctamente y FN el número de respuestas correctas no identificadas correctamente. Estas cuatro métricas permiten calcular la precisión en la detección de respuestas correctas e incorrectas (ACC), la precisión en la detección de respuestas correctas (*True Positive Rate*—TPR), la precisión en la detección de respuestas incorrectas (*True Negative Rate*—TNR) y el porcentaje de respuestas correctas con penalización por las identificaciones incorrectas (*F score* - F1,5). Estas métricas se utilizan para evaluar el rendimiento de los enfoques de comparación semántica y de clasificación de respuestas (PI1) y para determinar si SLASys recomienda la evaluación esperada por el profesorado en un entorno educativo real (PI2 y PI3).

Por último, se recopiló el tiempo medio de acceso del estudiantado a la interfaz de *feedback* y sus valoraciones sobre los resultados proporcionados. Para ello, el complemento de Moodle añadió una escala Likert para recopilar las valoraciones del estudiantado en la interfaz de *feedback*. La escala osciló entre 1 y 5 (donde 5 significa que el *feedback* se considera muy útil). En cuanto a la experiencia del profesorado, solo se recopilaron sus comentarios sobre el uso del sistema porque únicamente estuvieron implicados tres profesores (PI4).

RESULTADOS

PI1: ¿Cómo de precisa es la clasificación de respuestas con respecto a la comparación semántica para la recomendación en la CAPRC?

Conocer la precisión de SLASys antes del primer piloto con el estudiantado es crucial para analizar la aplicabilidad del sistema en un entorno educativo real. Los resultados se detallan en la Tabla 1, donde se analizaron las ocho preguntas del test de Claridad.

Tabla 1
Métricas de rendimiento en el test de Claridad con los datos de entramiento

	Correcto	Incorrecto	Comparación semántica				Clasificación de respuestas			
Ejercicio	**Entren. / Valid.**	**Entren. / Valid.**	**ACC**	**TPR**	**TNR**	**F1,5**	**ACC**	**TPR**	**TNR**	**F1,5**
Pregunta 1	36 / 8	107 / 22	37 %	100 %	14 %	58 %	87 %	88 %	86 %	81 %
Pregunta 2	47 / 10	7 / 2	67 %	70 %	50 %	75 %	100 %	100 %	100 %	100 %
Pregunta 3	3 / 2	46 / 4	42 %	100 %	30 %	48 %	83 %	0 %	100 %	0 %
Pregunta 4	34 / 7	17 / 20	27 %	0 %	75 %	0 %	64 %	86 %	25 %	79 %
Pregunta 5	67 / 14	100 / 20	53 %	100 %	20 %	74 %	82 %	71 %	90 %	75 %
Pregunta 6	82 / 17	36 / 8	84 %	94 %	63 %	91 %	88 %	100 %	63 %	95 %
Pregunta 7	33 / 7	33 / 7	57 %	86 %	29 %	73 %	79 %	71 %	86 %	75 %
Pregunta 8	68 / 14	19 / 4	83 %	93 %	50 %	91 %	83 %	100 %	25 %	94 %
Promedio			56 %	**80 %**	41 %	64 %	**83 %**	77 %	**72 %**	**75 %**

La tabla resume el número de respuestas anotadas utilizadas para el ajuste fino de los clasificadores de IA (es decir, el *conjunto de datos de entrenamiento, entren.* en la tabla), para evaluar su precisión (es decir, el *conjunto de datos de validación, valid.* en la tabla) y los resultados de las métricas. Los clasificadores de IA se ajustaron usando el 80 % de las respuestas anotadas (nótese que existe un clasificador para cada pregunta). El 20 % restante se utilizó para realizar la validación y obtener los resultados de la tabla. El proceso de división de los datos considera respuestas correctas e incorrectas en ambos conjuntos de datos. Cabe destacar que no hay fase de ajuste de la precisión en el enfoque de comparación semántica. En este caso, se compararon tres respuestas correctas seleccionadas de forma aleatoria con las respuestas de validación.

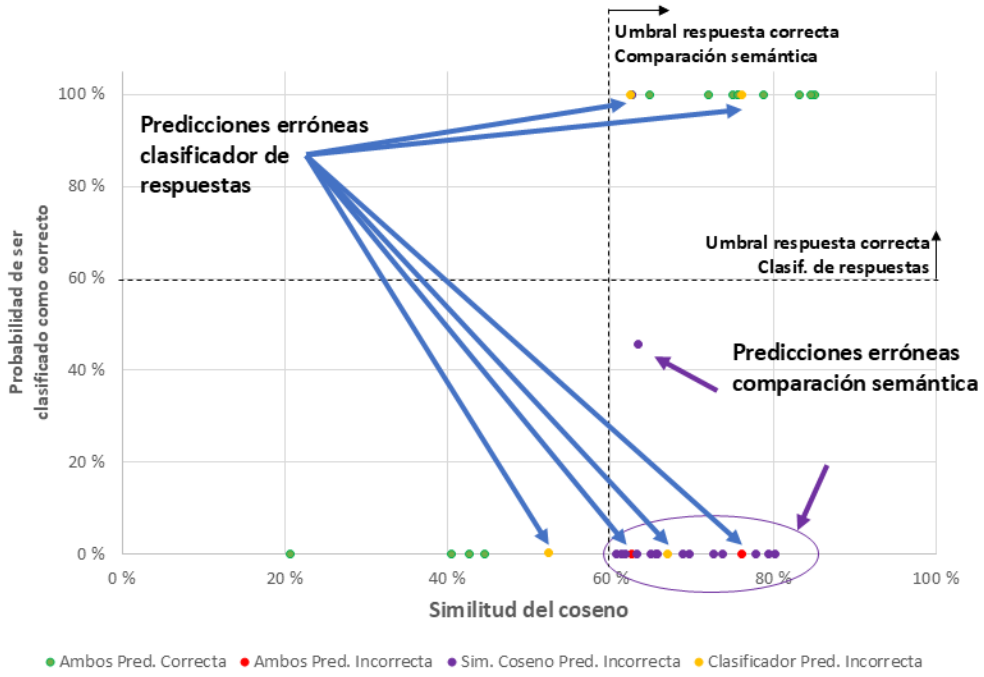
La clasificación de respuestas supera a la comparación semántica porque las respuestas anotadas proporcionan ejemplos de respuestas correctas e incorrectas.

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), pp. 321-352. https://doi.org/10.5944/ried.45541

La comparación semántica suele fallar cuando existen múltiples formas de responder a una pregunta, y las respuestas están semánticamente alejadas de las respuestas modelo del profesorado. Con respecto a las métricas para detectar respuestas correctas e incorrectas (es decir, el TPR y el TNR), la calidad de los resultados depende en gran medida del número de respuestas proporcionadas por el profesorado para el entrenamiento y del equilibrio entre respuestas correctas e incorrectas. Los clasificadores con pocas respuestas tienen dificultades para realizar recomendaciones. Se puede observar que la comparación semántica puede detectar más respuestas correctas en promedio (un 80 % en comparación con el 77 %). Sin embargo, este incremento se debe a que muchas respuestas incorrectas se identificaron incorrectamente como correctas (un 41 % en comparación con el 72 %).

Para comprender mejor el comportamiento de las técnicas, se analizaron las predicciones individuales. La Figura 7 muestra el análisis del conjunto de datos de validación para la Pregunta 5. Las métricas respectivas, la similitud del coseno y la probabilidad de que se predigan como correctas, se utilizan para la comparación semántica y la clasificación de respuestas, respectivamente. Se han utilizado colores diferentes para identificar qué predicciones son correctas utilizando ambas métricas (verde), cuáles se predicen incorrectamente mediante la comparación semántica (morado), cuáles se predicen incorrectamente mediante la clasificación de respuestas (amarillo) y cuáles no se pueden predecir por ninguno de los dos métodos (rojo).

Figura 7
Comparación del rendimiento de las métricas para la Pregunta 5



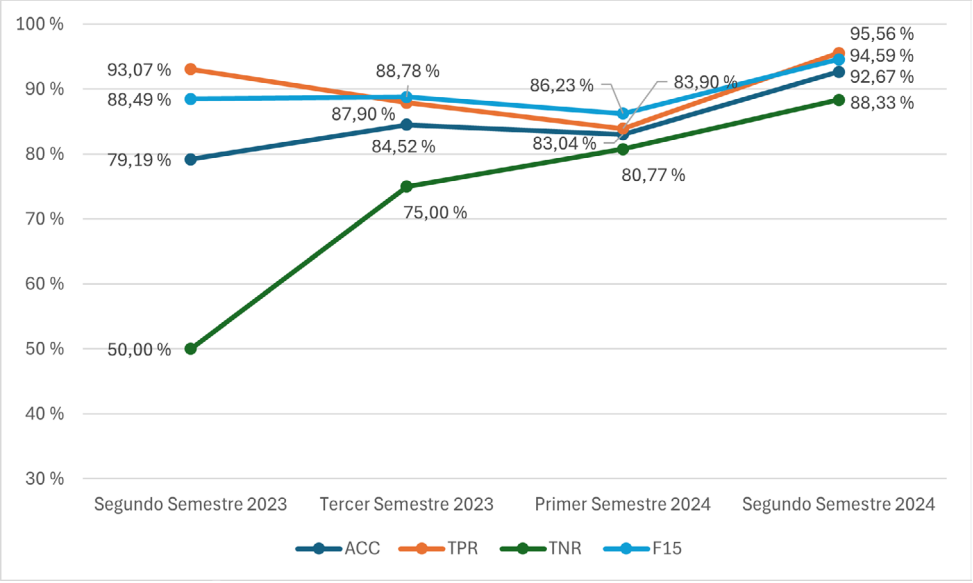
La mayoría de los errores aparecen en la comparación semántica. Las respuestas de buena parte del estudiantado se consideran correctas (es decir, semánticamente similares a la respuesta modelo proporcionada por el profesorado), con un rango del 60 % al 85 %, pero son incorrectas. Por lo tanto, pequeños cambios en el significado de la frase pueden generar predicciones incorrectas. La clasificación de respuestas también proporciona algunas predicciones erróneas en la detección de respuestas correctas, pero en menos casos. Además, existen casos en los que ambos métodos fallan. El profesorado observó que algunas de estas respuestas se predijeron incorrectamente porque no se cubrieron en las respuestas utilizadas para el ajuste fino ni en las respuestas modelo del profesorado.

PI2: ¿Cómo de preciso es SLASys en un entorno educativo real?

La Figura 7 muestra algunos casos erróneos. Añadir dichas respuestas a los clasificadores de IA podría mejorar la identificación en nuevas ediciones del curso. Esta es la idea fundamental para mejorar los clasificadores: estos se reentrenan, incluyendo los resultados de la evaluación realizados en las ediciones anteriores.

De este modo, se perfeccionan con nuevas respuestas anotadas que mejoran la identificación. La PI2 analiza si la clasificación de evaluación de respuestas que SLASys recomienda al profesorado durante las cuatro ediciones del curso mejora, teniendo en cuenta el ajuste fino efectuado en cada edición. Los resultados agregados de las distintas métricas se han resumido en el gráfico de la Figura 8 para el test de Claridad. Los resultados de todas las preguntas se han agregado para simplificar la evaluación.

Figura 8
Rendimiento de las métricas en el test de Claridad



Se puede observar una mejora significativa en la segunda edición debido a las nuevas respuestas anotadas de la primera edición, que mejoraron la precisión en las preguntas con un número desequilibrado de respuestas correctas e incorrectas (es decir, las preguntas 2, 4 y 8 de la Tabla 1). Aunque hay una ligera reducción en la detección correcta de las respuestas correctas (es decir, el TPR), esto se debe al proceso de refinamiento.

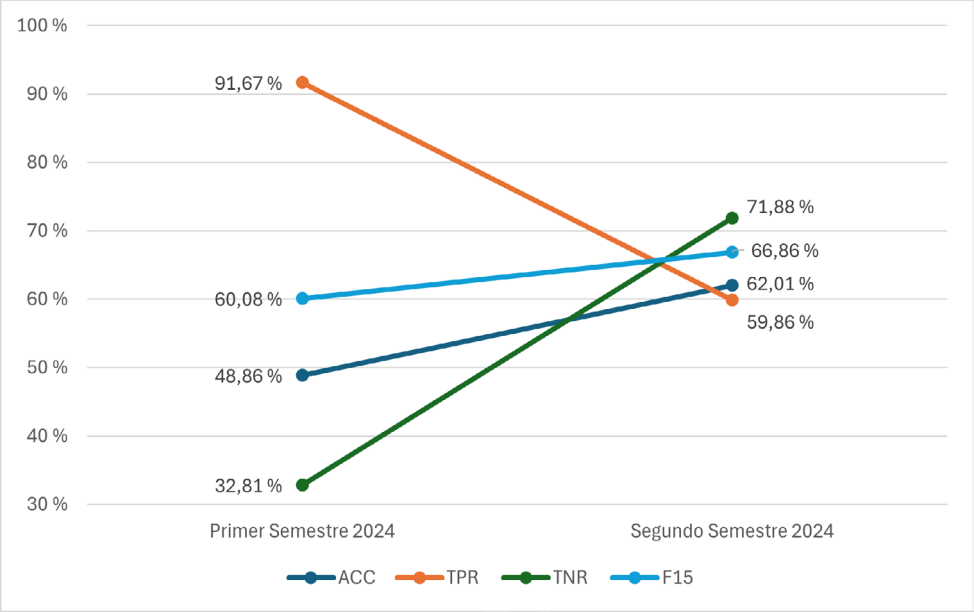
Las siguientes ediciones del curso contribuyeron a mejorar la precisión general, alcanzando valores superiores al 88 % en la última edición. Por lo tanto, incluir nuevos datos puede mejorar de forma eficaz la calidad del recomendador, lo que implica que puede utilizarse de forma efectiva con fines de evaluación. Una mayor precisión significa disponer de un recomendador con menos errores en el que el profesorado puede confiar durante el proceso de evaluación.

PI3: ¿Puede utilizarse SLASys sin respuestas previas del estudiantado?

Los buenos resultados mostrados en la sección anterior se obtuvieron porque el profesorado realizó una tarea inicial de anotación de las respuestas proporcionadas por el estudiantado anterior, lo que contribuyó a generar los clasificadores iniciales.

Sin embargo, se desea saber cómo funciona SLASys ante nuevas preguntas para las que no se dispone de respuestas anteriores. En este caso, solo se puede utilizar la comparación semántica con las respuestas modelo proporcionadas por el profesorado en la primera edición (se incluyeron tres respuestas modelo por pregunta). Sin embargo, en la segunda edición, la clasificación de respuestas puede aplicarse utilizando la evaluación realizada en la primera edición. El experimento con estudiantes reales se llevó a cabo en las dos últimas ediciones, es decir, en el primer y segundo trimestre de 2024, donde se diseñó un nuevo test para el concepto de Novedad. La Figura 9 muestra la mejora de SLASys al cambiar de la comparación semántica a la clasificación de respuestas en un entorno educativo real.

Figura 9
Rendimiento de las métricas en el test de Novedad



Se puede observar una diferencia significativa entre ambos enfoques. La precisión global (es decir, el ACC) mejora del 48,86 % al 62,01 %. Esta mejora se debe al incremento significativo en la detección de respuestas incorrectas (el TNR

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), pp. 321-352. https://doi.org/10.5944/ried.45541

aumentó del 32,81 % al 71,88 %). Sin embargo, la detección de respuestas correctas disminuye. La comparación semántica tiende a recomendar que la mayoría de las respuestas del estudiantado son correctas, ya que no puede distinguir las ligeras diferencias semánticas. Esto produce el efecto no deseado de que la comparación semántica no puede detectar correctamente la mayoría de las respuestas incorrectas. Por lo tanto, el cambio de la comparación semántica a la clasificación de respuestas tiene un impacto positivo en el recomendador.

El hallazgo relevante de este experimento es que los modelos de clasificación de respuestas pueden producir recomendaciones de alta calidad en pocas ediciones, incluso comenzando desde cero con un enfoque basado en comparación semántica.

PI4: ¿Cuál es la opinión del estudiantado y del profesorado?

Finalmente, se desea determinar si los nuevos tests y el *feedback* fueron útiles. La Tabla 2 resume la información recopilada del estudiantado, incluyendo el número de estudiantes que accedió al *feedback*, el tiempo promedio dedicado a la página y la valoración final otorgada al *feedback* recibido. No todo el estudiantado accedió al *feedback*, probablemente porque los tests no afectaron a la superación del curso. Además, el tiempo promedio de dedicación es relativamente bajo. Al comprobar el tiempo de acceso individual, se observa que este se correlaciona en gran medida con el número de respuestas incorrectas. En consecuencia, el estudiantado con respuestas incorrectas dedicó más tiempo a leer el *feedback*. Por lo tanto, se puede deducir que comprobó lo que debe aprender para mejorar. Por último, la valoración promedio es significativamente alta (superior al 70 %) en todas las ediciones y tests.

Tabla 2
Tiempo de acceso al feedback y valoración del estudiantado

Test	Claridad			Novedad		
Edición	**Segundo Semestre 2023**	**Tercer Semestre 2023**	**Primer Semestre 2024**	**Segundo Semestre 2024**	**Primer Semestre 2024**	**Segundo Semestre 2024**
Acceso estudiantado (%)	83,33	84,62	88,89	80,00	77,78	76,00
Tiempo medio de acceso (minutos)	2,33	1,73	3,51	2,91	4,74	4,68
Valoración (1-5)	3,94	3,84	3,63	3,85	3,55	3,77

Tabla 3
Comparación con trabajos relacionados

Referencia	Técnica	Predicción	Feedback	ACC
(Saha et al., 2018)	PLN	Corrección	No	66 %
(Soulimani et al., 2024)	Clasif. BERT	Calificación (0-4)*	No	71 %
(Padó et al., 2024)	Clasif. BERT	Corrección	No	72 %
(Wang et al., 2019)	Clasif. BERT	Corrección	No	80 %
(Camus y Filighera, 2020)	Clasif. BERT	Corrección	No	80 %
(Lun et al., 2020)	Clasif. BERT	Corrección	No	82 %
(Sung et al., 2019)	Clasif. BERT	Corrección	No	84 %
(Schneider et al., 2023)	Clasif. BERT	Corrección	No	86 %
(Liu et al., 2019)	Clasif. BERT	Corrección	No	89 %
(Grévisse, 2024)	IAGen (GPT4)	Calificación (0-10)*	Generado	64 %
(Aggarwal et al., 2025)	IAGen (Mistral)	Corrección	Generado	75 %
SLASys	Clasif. BERT	Corrección	Predefinido	83 %

(*) La precisión se calcula asumiendo como correcto cuando la calificación es superior a la mitad de los puntos máximos.

Además, el profesorado recopiló algunas opiniones durante las sesiones en línea. El estudiantado estuvo de acuerdo en que la calificación de la evaluación, el *feedback* sobre las respuestas y las directrices de la CEP a revisar fueron muy valiosos. Además, obtenerlos casi de forma inmediata fue positivo. Sin embargo, el estudiantado se quejó de que el *feedback* debería ser más personalizado, describiendo por qué su respuesta era incorrecta.

El profesorado también compartió su experiencia con el recomendador. Inicialmente, era reacio a usarlo, argumentando que confiar en la calificación automática podría dar lugar a evaluaciones erróneas. Sin embargo, la herramienta fue altamente aceptada como recomendador, puesto que la revisión final del profesorado era obligatoria. Después de evaluar las diferentes ediciones, el profesorado coincidió en que ganó eficiencia, ya que también adquirió experiencia con el tiempo. Con relación al experimento con el test de Novedad sin respuestas anotadas, se quejó en la primera edición (es decir, en el primer trimestre de 2024) puesto que no observó ninguna mejora en la reducción de su carga de trabajo, ya que el recomendador produjo muchas recomendaciones incorrectas. Sin embargo, constató el beneficio en la última edición cuando se pudo utilizar la clasificación de respuestas.

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), pp. 321-352. https://doi.org/10.5944/ried.45541

DISCUSIÓN

Los resultados obtenidos permiten responder a las preguntas de investigación. En relación con la PI1, como señalan otros trabajos (Burrows et al., 2015), los clasificadores de IA de BERT (es decir, la clasificación de respuestas) muestran una mayor precisión con respecto a la comparación semántica, y obtienen resultados similares a otros sistemas de CAPRC. La Tabla 3 compara SLASys con otros sistemas que usan conjuntos de datos públicos y que han sido reportados en la literatura. En concreto, se resume la técnica utilizada, el objetivo de predicción, el tipo de *feedback* proporcionado y la precisión obtenida. La clasificación de respuestas de SLASys alcanza una precisión similar a la obtenida por los trabajos relacionados que se centran en predecir la correctitud de respuestas, con una precisión que oscila entre el 71 % y el 89 %. Sin embargo, no es posible comparar SLASys con trabajos centrados en la predicción de calificaciones, ya que emplean métricas distintas para evaluar lo cerca que está la calificación predecida de la correcta (concretamente, se usa la *raíz cuadrada del error cuadrático medio*). Tales técnicas tienen tasas de error que oscilan entre moderadas y altas, con valores que van desde el 0,57 (Baral et al., 2021) hasta valores mayores que 1 (Gaddipati et al., 2020; Metzler et al., 2024), lo que dificulta su uso en un entorno educativo real. Cabe señalar que predecir la calificación es una tarea más compleja, ya que existe más variabilidad en el resultado (del Gobbo et al., 2023). Por lo tanto, el método propuesto en este trabajo simplifica el proceso, al proporcionar el resultado de la evaluación al profesorado, que decide la calificación en función de una rúbrica. Además, los resultados experimentales ofrecen información sobre el tamaño reducido del conjunto de datos necesario para obtener un clasificador ajustado para preguntas de respuesta corta (Mehrafarin et al., 2022). Por lo general, se recomienda tener un mínimo de 500 respuestas (es decir, *instancias*) para disponer de un clasificador adecuado. Sin embargo, la Tabla 1 muestra que se puede entrenar un buen clasificador incluso con menos respuestas. Además, se han comparado los métodos de clasificación semántica y de clasificación de preguntas, y sus métricas asociadas, para una pregunta específica (Figura 7). El profesorado puede acceder a esta información en Moodle para comprender mejor qué respuestas no están bien clasificadas. Cabe recordar que la interpretabilidad de la IA es una de las recomendaciones para las herramientas de IA (Zhao et al., 2024). La Tabla 3 también compara SLASys con trabajos basados en IAGen, que actualmente muestran un rendimiento inferior que los modelos basados en BERT. Aunque las herramientas IAGen generan un gran entusiasmo, todavía tienen algunos inconvenientes en el ámbito de la evaluación. Pueden incurrir en alucinaciones (Jia et al., 2024) o mostrar limitaciones conceptuales en dominios específicos (De La Cruz et al., 2024), lo que puede disminuir su eficacia para evaluar. Algunos de estos modelos de IAGen se pueden utilizar localmente. Sin embargo, las soluciones empresariales son ampliamente utilizadas. Estas operan de forma remota, ofreciendo una mayor capacidad de procesamiento, pero a mayores costes y con problemas de sostenibilidad

(van Wynsberghe, 2021). Además, requieren de un análisis de cuestiones éticas y el replanteamiento de las políticas educativas y de protección de datos. Es importante que las instituciones de educación superior revisen el manifiesto propuesto por García-Peñalvo et al. (2024) para obtener un sistema seguro y ético. SLASys opera localmente, reduciendo los costes de procesamiento, manteniendo su funcionalidad y no proporcionando datos a terceros. Además, no recopila datos confidenciales del estudiantado, dado que estos se mantienen seguros dentro del SGA.

Con respecto a la PI2, se ha mostrado la progresión de SLASys a lo largo de cuatro ediciones del curso. Como afirma la pregunta de investigación anterior, los conjuntos de respuestas anotadas (aunque sean reducidos) producen clasificadores de alta precisión, lo que hace disponible SLASys para su uso en un entorno educativo real. Al generar clasificadores precisos, SLASys es un ejemplo de una herramienta de IA que puede beneficiar a ambas partes. Por un lado, el profesorado puede beneficiarse de un recomendador para la CAPRC, lo que aumenta la eficiencia de la evaluación y unifica los criterios de evaluación (Xavier et al., 2025). Además, los beneficios para el profesorado son más claros a medida que se reduce su carga de trabajo, lo que le permite dedicar más tiempo a otras tareas cualitativas como la revisión del *feedback*, el diseño de nuevas preguntas o actividades de aprendizaje mejor diseñadas. Debido a la integración con Moodle, SLASys proporciona *feedback* formativo al estudiantado en comparación con los trabajos relacionados descritos en la Tabla 3. Estos trabajos únicamente proporcionan una recomendación sin *feedback*, que, en el caso de su aplicación en un entorno real, necesitaría de métodos adicionales para proporcionarlo. Según Gaddipati et al. (2021), una opción podría ser utilizar herramientas IAGen que muestran potencial para la generación de *feedback*. Por otro lado, con SLASys, el estudiantado obtiene *feedback* de calidad cuando se le hacen preguntas de razonamiento (Calimeris y Kosack, 2020). Con este *feedback*, el estudiantado puede aprender mejor con información más completa y detallada sobre su tarea de evaluación. En última instancia, este *feedback* se convierte en un componente personalizado que mejora el conocimiento del estudiantado (Abu Khurma et al., 2024).

En relación con la PI3, los resultados demuestran que SLASys se puede utilizar incluso sin datos de entrenamiento. En la primera edición que se proponga una nueva pregunta, el profesorado necesitará prestar más atención debido a las recomendaciones de baja precisión. Sin embargo, el clasificador de IA estará listo para su uso en la próxima edición. Es importante destacar también la simplicidad técnica del sistema. SLASys sigue las buenas prácticas de evaluación descritas en Petridou y Lao (2024).

Por último, en relación con la PI4, las opiniones y los resultados indican que las predicciones erróneas del recomendador no afectan al estudiantado, ya que este obtiene la evaluación revisada por el profesorado. En este sentido, la supervisión humana es obligatoria porque los errores en la evaluación pueden tener efectos negativos en el estudiantado (Li et al., 2023). Al igual que en Sangapu (2018), las

opiniones son principalmente positivas con la introducción de herramientas de IA en la educación. El profesorado no se siente abrumado por los problemas técnicos, ya que SLASys está integrado en Moodle. De forma similar, SLASys cambiará automáticamente de una técnica a otra sin la intervención del profesorado. Igualmente relevante, SLASys se ha diseñado como un recomendador de evaluación siguiendo las recomendaciones de la Ley Europea de IA (European Commission, 2024), evitando su uso como calificador automático basado en IA.

Este estudio presenta limitaciones relacionadas con el tamaño de la muestra y el muestreo no probabilístico que afectan a la validez externa. Debido a su aplicación en un entorno educativo real, se utilizó un conjunto de datos reducido de un dominio específico, lo que restringe la generalización de los hallazgos y puede generar sesgo muestral. Esto se puede observar principalmente en la PI3, donde el tamaño de la muestra para la última edición del curso es de solo 25 estudiantes. Sin embargo, los hallazgos que se derivan de la PI2 demuestran que SLASys mejora su precisión y sigue siendo aplicable y eficaz a lo largo del tiempo (es decir, en las cuatro ediciones del curso) incluso cuando los clasificadores de IA se entrenan con datos limitados. En cuanto al muestreo no probabilístico, todo el estudiantado participó en el estudio, lo que hizo inviable el análisis de adquisición de conocimientos con los nuevos tests. Cabe señalar que el primer examen dentro de la formación completa se realiza después del segundo curso. Un análisis longitudinal hasta ese examen enriquecería la comprensión del impacto de los nuevos tests sobre el conocimiento adquirido. El tamaño de la muestra también afecta a la PI4. Las opiniones positivas del estudiantado podrían estar sesgadas debido al tamaño de la muestra. Además, las opiniones del profesorado se recopilaron con entrevistas semiestructuradas simples. Un análisis cualitativo con grupos focales proporcionaría una visión más detallada y profunda.

Aunque los resultados experimentales obtenidos no se pueden generalizar, se destaca la facilidad de transferencia a distintos dominios con un lenguaje especializado, como Derecho, Historia, Filosofía o Ciencias Sociales. La transferibilidad dependería de la carga de trabajo inicial para diseñar las preguntas, añadir las respuestas modelo del profesorado correspondientes, y el esfuerzo de recopilar y revisar las respuestas anotadas. SLASys permite la creación de preguntas en las que el estudiantado debe aplicar el conocimiento adquirido, en lugar de responder basándose en conceptos teóricos, lo que conduce a una evaluación auténtica (Villarroel et al., 2018).

CONCLUSIONES

Esta investigación ofrece tres contribuciones principales. En primer lugar, se desarrolla un sistema de recomendación para la evaluación de preguntas de respuesta corta (SLASys) guiado por tres principios fundamentales: uso de técnicas gratuitas (BERT) y adecuadas para implementaciones privadas, ejecución en servidores con recursos limitados y facilidad de uso sin necesidad de conocimientos técnicos en

IA. En segundo lugar, una característica destacada es la integración con Moodle, que permite el seguimiento del estudiantado, la calificación y la visualización de los resultados de aprendizaje a través de SLASys. Los resultados han demostrado que, aunque no puedan generalizarse, la clasificación de respuestas puede conducir a mejores recomendaciones. Asimismo, SLASys permite refinar el recomendador para producir mejores resultados con nuevas cohortes de estudiantado. En tercer lugar, el trabajo subraya el papel de la IA como herramienta de apoyo en la educación, empoderando al profesorado para utilizar una herramienta diseñada para recomendar las evaluaciones y el *feedback* que puede aplicarse en diversos dominios. El profesorado puede emplear estrategias integradoras, destacando la convergencia de la tecnología y la pedagogía para mejorar la experiencia de aprendizaje del estudiantado.

Como trabajo futuro, se plantea extender la utilización del recomendador en otros tests del mismo curso fundamental de examen de patentes en el dominio de la Propiedad Intelectual. Esto permitirá evaluar la capacidad del sistema para abordar otros conceptos. Asimismo, se propone explorar cómo se puede personalizar el *feedback* en función de las respuestas incorrectas del estudiantado y de los recursos de aprendizaje disponibles. BERT se puede emplear para crear herramientas de búsqueda semántica basadas en los materiales del curso, capaces de identificar automáticamente los recursos necesarios para una pregunta específica, lo que podría contribuir a optimizar el esfuerzo del profesorado en la generación del *feedback*.

Agradecimientos

Este trabajo forma parte del proyecto PID2023-147592OB-I00 financiado por MCIU/AEI/10.13039/501100011033/ FEDER, UE, y del Programa de Investigación Académica de la Oficina Europea de Patentes, Acuerdo de Subvención Nº 2021/8404.

REFERENCIAS

Abu Khurma, O., Albahti, F., Ali, N. y 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. y Lin, J. (2019). *DocBERT: BERT for document classification***.** arXiv. https://doi.org/10.48550/arXiv.1904.08398

Aggarwal, D., Sil, P., Raman, B. y Bhattacharyya, P. (2025). "I understand why I got this grade": Automatic short answer grading with feedback. En *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. y 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. y 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. y 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. y Heffernan, N. T. (2021). Improving automated scoring of student open responses in mathematics. En *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*.

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

Burrows, S., Gurevych, I. y 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. y 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. y Filighera, A. (2020). Investigating transformers for automatic short answer grading. En I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin y 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. y 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. y 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. y Ramírez, S. (2024, July 1). Analysis of solutions of ChatGPT to logic problems based on critical thinking. En *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. y Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. En J. Burstein, C. Doran y 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. y 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. y 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. y 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. y 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. y 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. y 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. y 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. y Vajda, I. (2007). Intelligent mathematics assessment in eMax. En *IEEE AFRICON Conference* (pp. 1-7). https://doi.org/10.1109/AFRCON.2007.4401512

Hattie, J. y 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. y 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. y 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. y 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. y Gehringer, E. (2024). On assessing the faithfulness of LLM-generated feedback on student assignments. En B. Paaßen y

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. y Tokman, M. (2011). Automated assessment of short free-text responses in computer science using latent semantic analysis. En *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. y 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. y Karahalios, K. (2023). Am I wrong, or is the autograder wrong? Effects of AI grading mistakes on learning. En *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. y Liu, Z. (2019). Automatic short answer grading via multiway attention networks. En *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. y Yang, M. (2020). Multiple data augmentation strategies for improving performance on automatic short answer scoring. En *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. y Pilehvar, M. T. (2022). On the importance of data size in probing fine-tuned models. En *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. y 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. y Hees, J. (2024). Computer-assisted short answer grading using large language models and rubrics. En *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. y Stamper, J. (2022). Towards generalized methods for automatic question generation in educational domains. En S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren y 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. y 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. 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. y 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. y 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. 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. y Sengupta, B. (2018). Sentence-level or token-level features for automatic short answer grading? Use both. En *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. y 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. y Alexander, R. (2025). *Semantic refined prompting based automated essay scoring system*. En *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. y Harrison, C. (2008). A systematic approach to the automated marking of short-answer questions. En *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. y 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. y 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. y Arora, R. (2019). Pre-training BERT on domain resources for short answer grading. En *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. y 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. y Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. En *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. y 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. y Lin, H. (2022). On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. En *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. y Baraniuk, R. G. (2019). A meta-learning augmented bidirectional transformer model for automatic short answer grading. En *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*.

Winstone, N. E., Nash, R. A., Parker, M. y 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. y 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. y Li, X. (2018). Research on personalized exercises and teaching feedback based on big data. En *Proceedings of the ACM International Conference* (pp. 217-221). https://doi.org/10.1145/3232116.3232143

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

Zhang, H., Yu, P. S. y 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. y 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. y 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. y 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

Fecha de recepción del artículo: 1 de junio de 2025
Fecha de aceptación del artículo: 6 de agosto de 2025
Fecha de aprobación para maquetación: 16 de septiembre de 2025
Fecha de publicación en OnlineFirst: 14 de octubre de 2025
Fecha de publicación: 1 de enero de 2026