

# ¿Puede la inteligencia artificial proporcionar un feedback más sostenible?

Eloi Puertas Prats<sup>1,\*</sup>, Maria Elena Cano García<sup>2</sup>

<sup>1</sup> Facultat de Matemàtiques i Informàtica. Universitat de Barcelona, Spain, epuertas@ub.edu, <https://orcid.org/0000-0001-6292-6448>

<sup>2</sup> Facultat d'Educació. Universitat de Barcelona, Spain, ecano@ub.edu, <https://orcid.org/0000-0003-2866-5058>

## RESUMEN

La evaluación entre pares es una estrategia mediante la cual los estudiantes evalúan el nivel, valor o calidad del trabajo de sus compañeros dentro del mismo entorno educativo. La investigación ha demostrado que los procesos de evaluación entre pares impactan positivamente en el desarrollo de habilidades y el rendimiento académico. Al aplicar criterios de evaluación al trabajo de sus compañeros y ofrecer comentarios, correcciones y sugerencias para mejorar, los estudiantes no solo mejoran su propio trabajo, sino que también cultivan habilidades de pensamiento crítico. Para nutrir eficazmente el papel de los estudiantes como evaluadores, son esenciales oportunidades deliberadas y estructuradas para practicar, junto con capacitación y orientación. La inteligencia artificial puede ofrecer un medio para evaluar automáticamente las evaluaciones entre pares, asegurando su calidad y ayudando a los estudiantes a realizar evaluaciones con precisión. Este enfoque permite a los educadores centrarse en evaluar las producciones de los estudiantes sin necesidad de una formación especializada en evaluación de retroalimentación. Este documento presenta el proceso desarrollado para automatizar la evaluación de la calidad de la retroalimentación. A través de la utilización de fragmentos de retroalimentación evaluados por investigadores basados en criterios preestablecidos, se entrenó un Modelo de Lenguaje Masivo (LLM) de Inteligencia Artificial para lograr la evaluación automatizada. Se discuten los desafíos y perspectivas de este proceso, junto con recomendaciones para optimizar los resultados.

**PALABRAS CLAVE:** Inteligencia artificial, evaluación, feedback, educación superior

## 1 INTRODUCCIÓN

Los procesos de evaluación entre iguales poseen beneficios para los procesos de aprendizaje, pero deben de darse bajo ciertas condiciones, como una mínima alfabetización del estudiantado o la necesaria revisión de los comentarios que se proporcionan los pares. Esta revisión suele hacerse por parte del profesorado tanto para asegurar su pertinencia como para contribuir a formar al estudiantado en la provisión de comentarios de calidad. Esta tarea requiere tiempo y esfuerzo por parte del profesorado. El trabajo que se presenta intenta responder a si la Inteligencia artificial (IA) puede proporcionar información sobre la calidad del feedback., permitiendo así que el profesorado dedique su tiempo a tareas insustituibles de seguimiento y acompañamiento. Por ello el objetivo de este estudio es diseñar y validar el prompt más adecuado para evaluar la calidad del feedback., contribuyendo así a automatizar este proceso. Se estudia la similitud humano-máquina a través del desarrollo y refinamiento de un prompt que permita emular la valoración realizada por las evaluadoras humanas previamente.

## 2 MARCO TEÓRICO

### 2.1 Los procesos de evaluación entre iguales en el marco de la evaluación para el aprendizaje

La evaluación puede entenderse como el proceso de acreditar o certificar el alcance de unos resultados o puede entenderse como una oportunidad para promover el aprendizaje. Se trata de recoger y valorar información sobre el aprendizaje para guiar los siguientes pasos. Es la evaluación para el aprendizaje (AfL del inglés Assessment for Learning) (Hawe & Dixon, 2017), en la que no se

penaliza el error y se supera el sentido de vigilancia o surveillance para comprender el proceso evaluativo como un proceso de aprendizaje en sí mismo.

La AfL tiene que darse en un contexto de evaluación continua (Fernández Ferrer & Cano, 2019), implica un rol activo por parte del estudiantado, quien va desarrollando su juicio evaluativo mediante estrategias de evaluación entre iguales (Panadero & Alqassab, 2019) como a través de estrategias de autoevaluación (Panadero et al., 2022), que van fortaleciendo la competencia en feedback.

La competencia en feedback pasa por saber buscar información, apreciarla e interpretarla y ser capaz de aplicarla a futuras tareas y procesos. En este sentido, el feedback no se concibe ya como la información o comentarios que se proporcionan al estudiantado sino como la acción que el estudiantado realiza con dichos comentarios (Carless & Boud, 2018).

Así pues, la evaluación formativa es aquella que promueve el aprendizaje. En ella el estudiante debe de tener un papel activo. En el caso de procesos de evaluación participados entre iguales, sus posibilidades dependen no solo de factores personales (Panadero et al., 2023) sino también de que se proporcione una formación al respecto (Anderson et al., 2020) y de que se ejercite o practique (Carless & Winstone, 2023). Es decir, se requiere tanto capacitación sobre el feedback como oportunidades, diseñadas intencionalmente, para: (a) comprender y comprometerse con unos criterios de calidad acerca de qué significa valorar la tarea de los compañeros/as; (b) aplicarlos y (c) reflexionar sobre el proceso.

### 2.2 La relevancia de los criterios de evaluación.

Estos criterios pueden ofrecerse de manera pautada (Bearman & Ajjwi, 2018; Gielen y De Weber, 2015; Panadero et al., 2022) y progresivamente ir retirando este andamiaje y deben de observar

ciertos principios, como el foco (Hattie y Timperley, 2007); su naturaleza (Chi, 1996) o su categoría (formato/contenido). Chi (1996) diferencia *feedback*:

- **Correctivo.** Es aquel que suele darse cuando las respuestas del alumno son incorrectas e, indirectamente, refuerza lo que sí se ha hecho bien. Puede ser muy simple/corto y darse de forma oral rápida para evitar progresar en procesos mal enfocados.
- **Didáctico.** Es aquel al que se asocian explicaciones extensas para complementar la identificación de los errores cometidos.
- **Sugestivo.** Es aquel que, en primer lugar, alerta de la existencia de un problema y, en segundo lugar, realiza sugerencias para orientar el pensamiento o actuación del alumnado. Es más probable que se dé si la respuesta preliminar de un estudiante es incompleta y no manifiestamente incorrecta y suele tener más sentido en contextos de *feedback* dialógico, con intercambios sucesivos.

Hattie y Timperley (2007) establecieron cuatro niveles de *feedback* sobre la tarea, sobre el proceso, autorregulador y sobre la persona:

- El *feedback* sobre la tarea se realiza sobre el trabajo o producto que el estudiantado entrega. Es muy usual que sea finalista y que posea una naturaleza correctiva, identificando información errónea o ausente.
- El *feedback* sobre el proceso se refiere al procesamiento o desarrollo de la tarea, refiriéndose a las estrategias empleadas para progresar en su elaboración.
- El *feedback* autorregulador proporciona información para que el estudiantado pueda planificar, monitorear y reflexionar sobre su trabajo. Por ello estimula los procesos de fijación de metas, apropiación de criterios, autoevaluación y atribución de éxito o fracaso por parte del propio estudiantado.
- El *feedback* sobre la persona es aquel que elogia al estudiantado. Esta felicitación puede entenderse como algo vinculado a las características personales y no como algo que se puede entrenar o aprender, generando con ello una falta de compromiso.

Por su parte, Sadler (2010), indica que:

- Algunos "criterios" (longitud del texto, referencias, aspectos formales) son fáciles de aplicar y tienen límites definidos. Deberían considerarse reglas, condiciones o requisitos. Son fáciles de enunciar, reconocer y valorar.
- Lo que son verdaderamente criterios son más abstractos, con límites difusos, interpretables en contextos sociales o profesionales. El estudiantado debe de ir reconociéndolos, dotándolos de significado, construyéndolos e incorporándolos en sus producciones en base a un análisis de cada contexto o demanda particular. Una comunicación unidireccional de los criterios no es suficiente para lograrlo.

Estos criterios se pueden aplicar sobre tareas o sobre procesos. Cuando se incorporan procesos de evaluación entre iguales habrá que considerar tanto los criterios de una buena ejecución de la tarea que los iguales tienen que aplicar sobre el trabajo de sus compañeros como los criterios de calidad del proceso de evaluación en sí mismo.

### 2.3 Los criterios de calidad del feedback

Teniendo en cuenta esta triple clasificación, Cano et al. (2024) emplearon esta clasificación para la evaluación de los criterios del proceso de evaluación entre iguales:

| Categoría                                                                                               | Definición: comentarios...                                                            | Los |
|---------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-----|
| 1) La tarea se evalúa considerando el propósito de la tarea y los criterios de evaluación establecidos. | 1) Se realizan por separado en función de criterios preestablecidos.                  |     |
| 2) Se identifican fortalezas y debilidades de la tarea evaluada.                                        | 2) Destacan los puntos de mejora y los fuertes de la tarea.                           |     |
| 3) Justifica con razones por que algo es correcto o necesita cambiarse.                                 | 3) No se dan sin justificación, sino que argumentan explícitamente su (in)corrección. |     |
| 4) Se proporcionan sugerencias para mejorar la tarea.                                                   | 4) Se ofrecen pistas y recursos para mejorar.                                         |     |
| 5) Se proporcionan comentarios constructivos y objetivos.                                               | 5) Son sustanciales y se ofrecen en un tono amigable.                                 |     |

Tabla 1. Criterios sobre la calidad del feedback entre iguales  
Fuente: Cano et al. (2024)

Contribuir a acordar estos criterios y/o aplicarlos fortalece el juicio evaluativo (Tai et al., 2018). Pero tomar decisiones sobre la calidad del propio trabajo y del de los demás implica disponer de ciertas habilidades o competencia en feedback, que hay que planificar explícitamente. De ahí la importancia de que el profesorado diseñe experiencias de evaluación participada por el estudiantado, como la evaluación entre iguales, que se entiende como "un mecanismo diseñado para que los aprendices analicen y precisen el nivel, valor o calidad de un producto o desempeño de otros aprendices de igual estatus" (O'Donnell, 1998, 256). Esta evaluación entre iguales debe de ser supervisada valorando su calidad con los criterios de la tabla 1.

Es decir, el juicio evaluativo se va desarrollando mediante actividades de evaluación formativa explícitamente diseñadas en las que el estudiantado construye y/o se apropia de los criterios de evaluación y los aplica a su propio trabajo o al de los demás. En el caso de tratarse de prácticas de evaluación entre iguales toman importancia tanto los criterios que se aplican sobre las tareas como los criterios de calidad bajo los que deben proporcionarse los comentarios entre iguales y que, de algún modo, también deben de ser objeto de evaluación.

En síntesis, a pesar de los beneficios de los procesos de evaluación entre iguales (Panadero et al., 2023), estos procesos suelen hallar cierta resistencia, en parte por la falta de cultura evaluativa y, en parte, porque su diseño, seguimiento y evaluación supone una carga de trabajo considerable. Concretamente, los procesos de evaluación entre iguales deben ser monitoreados. Por ejemplo, los LMS, como Moodle, requieren tanto de la evaluación entre compañeros sobre la tarea (que, por defecto, es un 80% de la calificación) como de la revisión de la calidad de dichas evaluaciones. En este sentido, la IA puede contribuir a reducir esa carga.

### 2.4 El uso de los modelos de lenguaje masivo en la evaluación de feedback

Desde su aparición, hace escasamente dos años, los modelos de lenguajes masivos (LLM, por sus siglas en inglés -Large Language Model-) usados en el campo de la Inteligencia Artificial han significado una revolución en el procesamiento y generación de lenguaje natural (Zhao et al., 2023).

Herramientas como Chat-GPT han llevado al gran público una nueva manera de interactuar con estos sistemas informáticos inteligentes mediante prompts. La facilidad de crear estos prompts sin tener que ser un especialista en Inteligencia Artificial o en informática ha permitido que estas herramientas se conviertan en

asistentes artificiales para numerosas tareas (Imani et al., 2023; Kamaloo et al., 2023; Li et al., 2023).

Así pues, la facilidad con las que se puede interactuar con estos asistentes mediante prompts más la capacidad de entender y crear lenguaje humano que tienen estos de forma inherente, hace que puedan ser usados como evaluadores de la calidad del lenguaje escrito por humanos (Mendonça et al., 2023; Wang et al., 2023)

Sin embargo, hay que tener en cuenta las limitaciones que estos sistemas tienen (Bender et al., 2021). A parte del sesgo que puedan tener los datos o la imposibilidad que tienen todavía de razonar como lo haríamos los humanos hay que considerar que diseñar un buen prompt que sea capaz de realizar exactamente la tarea que se desea no es sencillo. Pequeños cambios en los prompts pueden dar resultados muy diferentes. Por eso es importante entender el funcionamiento de estos y seguir los consejos que los desarrolladores de los modelos hayan sugerido (Bsharat et al., 2024; OpenAI Platform, s.d.).

Así pues, teniendo en cuenta tanto las posibilidades como las limitaciones de estos modelos de lenguaje masivo, nuestra pregunta de investigación es: ¿Puede la IA proporcionar información sobre la calidad del feedback?

### 3 METODOLOGÍA

El objetivo de este estudio es diseñar y validar el prompt más adecuado para evaluar la calidad del feedback., contribuyendo así a automatizar este proceso.

La metodología empleada para responder a este objetivo ha sido la siguiente. En primer lugar, se procedió a recoger datos reales de procesos de evaluación entre iguales. A continuación, se empleó el modelo de lenguaje masivo (LLM) basado en lenguaje natural GPT-4 de OpenAI (OpenAI et al., 2024) para automatizar la evaluación sobre la calidad del feedback siguiendo los criterios mencionados anteriormente en la Tabla 1. Para ello se diseñaron los distintos prompts necesarios para que el modelo llevase a cabo dicha tarea.

#### 3.1 Datos

Los datos proceden de una intervención organizada en una secuencia didáctica (Imagen 1) en la que una de las fases era la evaluación entre iguales. Los estudiantes debían de proporcionarse comentarios cualitativos relativos a una tarea compleja e iterativa en una o dos ocasiones a lo largo del curso. El lenguaje en que se expresaban los alumnos podía ser en catalán o castellano indistintamente.

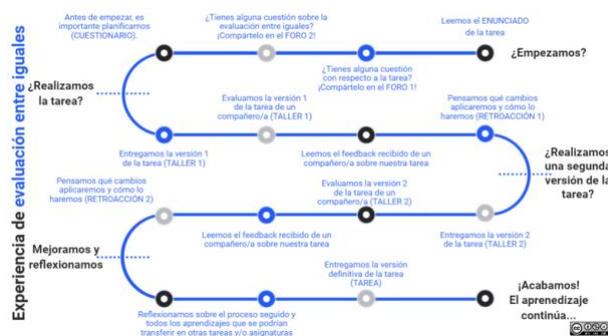


Imagen 1. Secuencia didáctica de la intervención en la que se realizó la evaluación entre iguales por Lluch, L. (2023), <https://www.ub.edu/digital-feedback/es/secuencia-didactica>, CC-BY-NC-ND.CC-BY-NC-ND.

Para este proceso de evaluación entre iguales se empleó la herramienta Taller de Moodle (*Using Workshop - MoodleDocs*, s.d.). Para cada uno de los criterios de ejecución de la tarea, los iguales (aparejados aleatoriamente) debían de revisar la tarea de otro compañero/a y ofrecer comentarios cualitativos.

Se recogieron los comentarios realizados y se analizaron los fragmentos correspondientes a diversas asignaturas de grados de educación de la Universitat de Barcelona (Tabla 2).

| Asignatura (grado - curso)                                                | Curso | Grado                  | Tarea                                                             |
|---------------------------------------------------------------------------|-------|------------------------|-------------------------------------------------------------------|
| Educación de personas adultas                                             | 1º    | Educación Social       | Biografía educativa de una persona adulta a partir de entrevistas |
| Teoría y Práctica de la Escuela Inclusiva                                 | 2º    | Educación Primaria     | Talleres a partir de casos                                        |
| Planificación, diseño y evaluación del aprendizaje y la actividad docente | 2º    | Educación Primaria     | Propuesta Educativa Curricular Integrada (Programación)           |
| Sistemas Distribuidos                                                     |       | Ingeniería Informática | Revisión de Código entre iguales                                  |

Tabla 2. Asignaturas implicadas

Algunos de los 1173 fragmentos recopilados se categorizaron deductivamente por parte de diversos miembros del equipo. Una vez asegurada la concordancia (coeficiente Kappa K = 0,82), dos investigadores clasificaron el resto de los fragmentos considerando la tabla 1.

#### 3.2 Evaluación de la calidad del feedback

Como se ha indicado, los modelos de lenguaje masivo (LLM) están entrenados para entender y generar lenguaje natural. En nuestro caso se usó el modelo de lenguaje masivo gpt4-turbo de OpenAI (OpenAI et al., 2024). para que actuase como nuestro asistente en la evaluación de la calidad del feedback partiendo de los criterios establecidos en la Tabla 1.

Para ello se emplearon dos estrategias de prompting distintas, zero-shot y few-shot, con el fin de evaluar cuál de las dos funciona mejor para la tarea de evaluación del feedback.

##### 3.2.1 Diseño del prompt

Se pueden diseñar *prompts* que sean efectivos siguiendo las recomendaciones de los desarrolladores del modelo de lenguaje masivo (Bsharat et al., 2024; OpenAI Platform, s.d.). Si bien es cierto que no todos los componentes son necesarios para un *prompt*, aquellas tareas que sean más complejas necesitarán de *prompts* más detallados. En términos generales, Se sugiere que el *prompt* se estructure en los siguientes componentes:

- Contexto: Introducción de un preámbulo que añade contexto adicional que puede ayudar al modelo a dar mejores respuestas.

- Instrucciones: Descripción de la tarea que se desea realizar.
- Ejemplos o datos de entrada: Inclusión de ejemplos de cómo debería realizar la tarea.
- Indicador de salida: Indicaciones relativas al tipo o formato de salida.

Aunque los LLM han demostrado ser una potente herramienta en tareas multilingües (OpenAI et al., 2024), debido a que los comentarios de los alumnos pueden estar escritos tanto en catalán como español, se ha optado por escribir los prompts en español ya que este lenguaje cuenta con una buena representación de textos en ChatGPT (Lai et al., 2023).

Para la tarea de evaluación del feedback se usaron dos prompts complejos diferentes, uno para cada estrategia de prompting usada, zero-shot y few-shots. El contexto, las instrucciones y el indicador de salida son los mismos para cada estrategia. Solo cambian los ejemplos de entrada, que en zero-shot no se emplearon.

El contexto especificado en ambos prompts fue el siguiente:

Buenos días. Estoy realizando una investigación en el campo de las ciencias de la educación sobre feedback en educación superior y te pido por favor que actúes como mi ayudante en esta investigación, que será de vital importancia en el futuro de la educación. Se ha realizado una intervención con diversos grupos de estudiantes en la que se ha propuesto un proceso de evaluación entre iguales durante la realización de una tarea iterativa, de modo que los comentarios que se proporcionan entre pares han de ser útiles para la realización de la siguiente fase de la tarea. Como es muy importante que los estudiantes den sentido a la información que reciben de sus iguales, la calidad del feedback es algo muy relevante en nuestra investigación y me gustaría poder clasificarla. Los comentarios o feedbacks estarán escritos íntegramente en catalán o en castellano.

### 3.2.2 Estrategia Zero-Shot

En la ingeniería de prompts la estrategia zero-shot (Liu et al., 2021) se refiere a que no se proporciona al modelo ningún ejemplo de la tarea que queremos que realice, simplemente se le anuncia que debería hacer.

En nuestro caso las instrucciones especificadas en la estrategia zero-shot fueron las siguientes:

Basándonos en estudios precedentes, los investigadores hemos establecido 5 criterios para valorar lo que es un buen feedback. Estos 5 criterios son:

Criterio 1. El evaluador considera los objetivos generales de la tarea y los criterios específicos de evaluación al proporcionar su retroalimentación.

Criterio 2. El evaluador proporciona comentarios que incluyen lo que se hizo bien (fortalezas o puntos fuertes) y lo que podría mejorarse (debilidades o puntos débiles) en la tarea evaluada.

Criterio 3. El evaluador explica de manera razonada y basada en evidencia por qué considera que ciertos aspectos de la tarea son correctos o necesitan modificarse, motivando sus valoraciones.

Criterio 4. El evaluador ofrece recomendaciones específicas sobre cómo el autor de la tarea podría mejorar su trabajo actual y futuras asignaciones/tareas.

Criterio 5. El evaluador expresa sus puntos de vista de manera positiva y objetiva, ofreciendo críticas útiles y centrándose en el trabajo o en el proceso en lugar de en la persona.

A continuación, realiza los siguientes pasos:

Paso 1: Actúa como 1 investigador que actúa de manera responsable y debe realizar su tarea íntegramente.

Paso 2: Como investigador aplica los cinco criterios, tal y como se han descrito, puntuando con 1 o 0 según si cumplen o no el criterio en todos los [TEXTOS] siguientes.

Paso 3: Cada investigador justifica su puntuación en cada uno de los criterios para todos los textos

Paso 4: Solo devuelve todos los resultados de la investigación y nada más en el siguiente formato JSON:

```
{
  "Investigador 1": {
    "Texto 0": {
      "Criterio 1": Nota entre 0 y 1,
      "Criterio 2": Nota entre 0 y 1,
      "Criterio 3": Nota entre 0 y 1,
      "Criterio 4": Nota entre 0 y 1,
      "Criterio 5": Nota entre 0 y 1,
      "Justificaciones": {
        "Criterio 1": "Justificacion 1",
        "Criterio 2": "Justificacion 2",
        "Criterio 3": "Justificacion 3",
        "Criterio 4": "Justificacion 4",
        "Criterio 5": "Justificacion 5"
      }
    },
  },
  [TEXTOS]
  ...
}
```

A continuación, se tienen que añadir los textos que se vayan a evaluar.

### 3.2.3 Estrategia Few-Shots

En la estrategia de prompting Few-Shots (Schick & Schütze, 2021) se dan algunos ejemplos de la tarea que se quiere realizar. De esta forma el modelo de lenguaje masivo puede aprender de estos ejemplos para realizar mejor la tarea a ejecutar. En nuestro caso, se añadieron algunos ejemplos de evaluaciones realizadas a mano por los investigadores en el prompt. El diseño del prompt es igual que en el caso de zero-shot pero añadiendo los ejemplos antes de declarar los pasos a seguir. El texto añadido en el prompt es el siguiente:

... Hemos clasificado manualmente algunos feedbacks otorgando 1 punto si cumple un criterio concreto y 0 puntos si no lo cumple, de manera que un feedback que cumple los 5 criterios de calidad tendría 5 puntos y un feedback que no cumple ninguno, 0 puntos

Se te mostraran algunos [EJEMPLOS] de feedback al que hemos asignado manualmente la puntuación correspondiente a cada criterio y como para ayudarte a llevar a cabo la [TAREA]

```
[EJEMPLOS]
[Texto 7: "Descriuen conseqüències en l'aprenentatge, ja que ens diuen que per culpa dels plans individualitzats i reforços fora del grup aula, es donen menys expectatives i, per tant, no obtenim tot el potencial de l'infant i li donem a entendre que ell no és capaç. També comenten que la participació de les famílies ha de millorar i que han de mostrar-se més oberts, tot i que no descriuen les conseqüències d'aquest fet."
Criterio 1: 1
Criterio 2: 1
Criterio 3: 0
Criterio 4: 0
Criterio 5: 1]
...
```

## 4 RESULTADOS

En la tabla 3 se aprecian los comentarios de feedback que se han introducido en el modelo de ChatGPT para evaluar. Se han usado

los mismos textos tanto para el experimento Zero-shot como Few-Shots.

| Comentarios |                                                                                                                                                                                                                                                                                                                                                                                 |
|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1           | Cada barrera té el seu origen darrere tot i no ser explicat amb detall.                                                                                                                                                                                                                                                                                                         |
| 2           | Pel que fa a l'aspecte 1, fa una introducció molt correcta on exposa clarament el supòsit presentat, ho justifica amb la citació de diversos articles i fa una breu exposició justificada de les que després s'identificaran com a barreres.                                                                                                                                    |
| 3           | s'ha analitzat correctament el que cal fer. La proposta i situació plantejada és correcte. Tot i això, penso que potser s'hauria d'haver marcat de manera més clara els criteris per avaluar l'assoliment de l'alumnat. D'altra banda, crec que els objectius marcats per la professora d'aprofundir en la higiene bucal i la higiene personal, són indispensables per la vida. |
| 4           | Els aspectes formals estan correctes. El format i la presentació és adequada i es veu neta. Potser en trobo a faltar alguna referència o cita bibliogràfica.                                                                                                                                                                                                                    |
| 5           | Han analitzat correctament les característiques de la tasca i han omplert totes les fases.                                                                                                                                                                                                                                                                                      |
| 6           | Penso que aquest aspecte s'ha assolit de forma suficient, podria haver-hi més reflexió sobre les conseqüències de les barreres detectades.                                                                                                                                                                                                                                      |
| 7           | Si                                                                                                                                                                                                                                                                                                                                                                              |
| 8           | Nuestra conexión queda pendiente una vez que se inicia el proceso manual.                                                                                                                                                                                                                                                                                                       |
| 9           | S'han trobat diferents tipus de barreres i les han argumentat a més de fer-ho més visual amb exemples i una graella al final de la redacció de les diferents barreres esmentades.                                                                                                                                                                                               |

Tabla 3. Comentarios introducidos

En las tablas 4 y 5 se muestran los resultados obtenidos en los comentarios evaluados para el experimento Zero-Shot. En la tabla 4 aparece la justificación que ha dado ChatGPT, mientras que en la tabla 5 se ve cómo han evaluado cada texto los evaluadores humanos y el modelo ChatGPT. Se puede observar que la diferencia en la puntuación global entre los dos no es superior a dos en estos ejemplos. En cuanto a la coincidencia de criterios vemos que en la mayoría de los casos coinciden entre 3 y 5 criterios en cada comentario. Solo en un caso vemos una coincidencia de solo 2 criterios

| Justificaciones Zero-shot ChatGPT |                                                                                                       |                                                                         |                                                                     |                                                     |                                                         |
|-----------------------------------|-------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|---------------------------------------------------------------------|-----------------------------------------------------|---------------------------------------------------------|
| ID                                | Criterio 1                                                                                            | Criterio 2                                                              | Criterio 3                                                          | Criterio 4                                          | Criterio 5                                              |
| 1                                 | El texto es demasiado vago y no especifica cómo se relaciona con los objetivos generales de la tarea. | No se mencionan fortalezas ni debilidades específicas.                  | No hay explicación razonada ni basada en evidencia.                 | No ofrece recomendaciones específicas para mejorar. | El texto es neutral y no ofrece críticas constructivas. |
| 2                                 | Se mencionan los objetivos y se justifican con citas de artículos.                                    | Identifica barreras, lo que implica reconocer fortalezas y debilidades. | Justifica las barreras con citas, lo que muestra una base razonada. | La mención de barreras sugiere áreas de mejora.     | El tono es objetivo y se centra en el trabajo.          |
| 3                                 | Menciona los objetivos de la tarea y sugiere la necesidad de criterios claros de evaluación.          | Identifica lo que está bien y sugiere mejoras.                          | Proporciona una razón para la necesidad de criterios claros.        | Ofrece recomendaciones específicas para mejorar.    | El tono es constructivo y se centra en el trabajo.      |
| 4                                 | No se menciona                                                                                        | Menciona una                                                            | No explica por qué las                                              | No ofrece recomendaciones                           | El tono es objetivo y                                   |

|   |                                                                                                       |                                                                          |                                                             |                                                                  |                                                          |
|---|-------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|-------------------------------------------------------------|------------------------------------------------------------------|----------------------------------------------------------|
|   | n los objetivos generales de la tarea.                                                                | debilidad (falta de referencias)                                         | referencias son importantes.                                | ones específicas, solo señala una falta.                         | se centra en el trabajo.                                 |
| 5 | Afirma que se han analizado correctamente las características de la tarea.                            | No menciona especificamente las fortalezas ni debilidades.               | No proporciona una explicación razonada de la evaluación.   | No ofrece recomendaciones para mejorar.                          | El tono es positivo y se centra en el trabajo.           |
| 6 | No se menciona cómo se relaciona con los objetivos generales de la tarea.                             | Identifica una debilidad y sugiere la necesidad de más reflexión.        | Proporciona una razón para la necesidad de más reflexión.   | Sugiere indirectamente cómo mejorar (más reflexión).             | El tono es constructivo y se centra en el trabajo.       |
| 7 | El texto no proporciona información relevante.                                                        | No se mencionan fortalezas ni debilidades.                               | No hay explicación alguna.                                  | No ofrece recomendaciones.                                       | No es aplicable debido a la falta de contenido.          |
| 8 | No se relaciona con los objetivos de una tarea educativa.                                             | No menciona fortalezas ni debilidades en un contexto educativo.          | No proporciona una base razonada en un contexto educativo.  | No ofrece recomendaciones para mejorar en un contexto educativo. | No es aplicable debido a la falta de contexto educativo. |
| 9 | Identifica y argumenta diferentes tipos de barreras, lo que implica una comprensión de los objetivos. | Menciona fortalezas (argumentación y ejemplos) y debilidades (barreras). | La argumentación y ejemplos proporcionan una base razonada. | La identificación de barreras sugiere áreas para mejorar.        | El tono es constructivo y se centra en el trabajo.       |

Tabla 4. Justificaciones experimento zero-shot

| ID | Criterio 1 |        | Criterio 2 |        | Criterio 3 |        | Criterio 4 |        | Criterio 5 |        | Puntuación Total |        |            |              |
|----|------------|--------|------------|--------|------------|--------|------------|--------|------------|--------|------------------|--------|------------|--------------|
|    | ChatGPT    | Humano | ChatGPT          | Humano | Diferencia | Coincidencia |
| 1  | 0          | 0      | 0          | 1      | 0          | 0      | 0          | 0      | 0          | 0      | 0                | 1      | -1         | 4            |
| 2  | 1          | 1      | 1          | 1      | 1          | 0      | 1          | 0      | 1          | 1      | 5                | 3      | 2          | 3            |
| 3  | 1          | 1      | 1          | 1      | 1          | 0      | 1          | 1      | 1          | 1      | 5                | 4      | 1          | 4            |
| 4  | 0          | 1      | 1          | 1      | 0          | 0      | 0          | 0      | 1          | 1      | 2                | 3      | -1         | 4            |
| 5  | 1          | 1      | 0          | 1      | 0          | 0      | 0          | 0      | 1          | 0      | 2                | 2      | 0          | 3            |
| 6  | 0          | 1      | 1          | 1      | 1          | 0      | 1          | 0      | 1          | 1      | 4                | 3      | 1          | 2            |
| 7  | 0          | 0      | 0          | 0      | 0          | 0      | 0          | 0      | 0          | 0      | 0                | 0      | 0          | 5            |
| 8  | 0          | 0      | 0          | 0      | 0          | 0      | 0          | 0      | 0          | 0      | 0                | 0      | 0          | 5            |
| 9  | 1          | 1      | 1          | 1      | 1          | 0      | 1          | 0      | 1          | 1      | 5                | 3      | 2          | 3            |

Tabla 5. Resultados experimento zero-shot

En las tablas 6 y 7 se muestran los resultados obtenidos en los comentarios evaluados para el experimento Few-Shots. En la tabla 6 vemos la justificación que ha dado ChatGPT, mientras en la tabla 7 vemos como han evaluado cada texto los anotadores humanos y el modelo ChatGPT. Se puede observar que la diferencia en la

puntuación global entre los dos no es superior a dos en estos ejemplos. En cuanto a la coincidencia de criterios vemos que en todos los casos coinciden entre 3 y 5 criterios en cada comentario.

| Justificaciones ChatGPT |                                                                                                             |  |                                                                     |  |                                                                             |  |                                                                                |  |                                                                                            |  |  |  |  |  |
|-------------------------|-------------------------------------------------------------------------------------------------------------|--|---------------------------------------------------------------------|--|-----------------------------------------------------------------------------|--|--------------------------------------------------------------------------------|--|--------------------------------------------------------------------------------------------|--|--|--|--|--|
| ID                      | Criterio 1                                                                                                  |  | Criterio 2                                                          |  | Criterio 3                                                                  |  | Criterio 4                                                                     |  | Criterio 5                                                                                 |  |  |  |  |  |
| 1                       | No menciona la relación con los objetivos generales de la tarea ni los criterios específicos de evaluación. |  | No se proporciona un balance entre fortalezas y debilidades.        |  | No hay una explicación razonada y basada en evidencia.                      |  | No se ofrecen recomendaciones específicas para mejorar.                        |  | No hay suficiente información para determinar si se expresa de manera positiva y objetiva. |  |  |  |  |  |
| 2                       | Se consideran los objetivos generales al justificar con artículos.                                          |  | Se identifican aspectos positivos y se anticipan barreras.          |  | Se justifica la introducción con la citación de artículos.                  |  | No se ofrecen recomendaciones específicas para mejorar.                        |  | El tono es positivo y se centra en el trabajo.                                             |  |  |  |  |  |
| 3                       | Se mencionan los objetivos y la situación planteada.                                                        |  | Se señalan aspectos positivos y se sugieren mejoras.                |  | Se razona sobre la necesidad de marcar criterios de evaluación.             |  | Se sugieren objetivos específicos para mejorar.                                |  | El feedback es constructivo y se centra en el trabajo.                                     |  |  |  |  |  |
| 4                       | No se menciona cómo se relaciona el formato con los objetivos de la tarea.                                  |  | Se señala una fortaleza y se sugiere una mejora.                    |  | No se proporciona una justificación basada en evidencia para la sugerencia. |  | Se sugiere añadir referencias o citas bibliográficas.                          |  | El tono es positivo y constructivo.                                                        |  |  |  |  |  |
| 5                       | Se menciona que se han analizado correctamente las características de la tarea.                             |  | No se mencionan específicamente fortalezas y debilidades.           |  | No se proporciona una justificación razonada y basada en evidencia.         |  | No se ofrecen recomendaciones específicas para mejorar.                        |  | El tono es generalmente positivo.                                                          |  |  |  |  |  |
| 6                       | No se menciona cómo se relaciona con los objetivos generales de la tarea.                                   |  | Se sugiere que podría haber más reflexión, indicando una debilidad. |  | Se razona sobre la necesidad de más reflexión sobre las consecuencias.      |  | No se ofrecen recomendaciones específicas para mejorar.                        |  | El tono es constructivo.                                                                   |  |  |  |  |  |
| 7                       | No proporciona información relevante para evaluar este criterio.                                            |  | No proporciona información relevante para evaluar este criterio.    |  | No proporciona información relevante para evaluar este criterio.            |  | No proporciona información relevante para evaluar este criterio.               |  | No proporciona información relevante para evaluar este criterio.                           |  |  |  |  |  |
| 8                       | No se menciona la relación con los objetivos de la tarea.                                                   |  | No se mencionan fortalezas ni debilidades.                          |  | No se proporciona una justificación razonada y basada en evidencia.         |  | No se ofrecen recomendaciones específicas para mejorar.                        |  | No hay suficiente información para determinar si se expresa de manera positiva y objetiva. |  |  |  |  |  |
| 9                       | Se relaciona con los objetivos al identificar y argumentar diferentes tipos de barreras.                    |  | Se identifican aspectos positivos y se sugieren mejoras.            |  | Se razona sobre las barreras identificadas con ejemplos.                    |  | Se hace más visual con ejemplos y una tabla, lo que sugiere formas de mejorar. |  | El tono es constructivo y se centra en el trabajo.                                         |  |  |  |  |  |

Tabla 6. Justificaciones experimento few-shots

| ID | Criterio 1 |        | Criterio 2 |        | Criterio 3 |        | Criterio 4 |        | Criterio 5 |        | Total Scoring |        |            |              |
|----|------------|--------|------------|--------|------------|--------|------------|--------|------------|--------|---------------|--------|------------|--------------|
|    | ChatGPT    | Humano | ChatGPT       | Humano | Diferencia | Coincidencia |
| 1  | 0          | 0      | 0          | 1      | 0          | 0      | 0          | 0      | 0          | 0      | 0             | 1      | -1         | 4            |
| 2  | 1          | 1      | 1          | 1      | 1          | 0      | 0          | 0      | 1          | 1      | 4             | 3      | 1          | 4            |
| 3  | 1          | 1      | 1          | 1      | 1          | 0      | 1          | 1      | 1          | 1      | 5             | 4      | 1          | 4            |
| 4  | 0          | 1      | 1          | 1      | 0          | 0      | 1          | 0      | 1          | 1      | 3             | 3      | 0          | 3            |
| 5  | 1          | 1      | 0          | 1      | 0          | 0      | 0          | 0      | 1          | 0      | 2             | 2      | 0          | 3            |
| 6  | 0          | 1      | 1          | 1      | 1          | 0      | 0          | 0      | 1          | 1      | 3             | 3      | 0          | 3            |
| 7  | 0          | 0      | 0          | 0      | 0          | 0      | 0          | 0      | 0          | 0      | 0             | 0      | 0          | 5            |
| 8  | 0          | 0      | 0          | 0      | 0          | 0      | 0          | 0      | 0          | 0      | 0             | 0      | 0          | 5            |
| 9  | 1          | 1      | 1          | 1      | 1          | 0      | 1          | 0      | 1          | 1      | 5             | 3      | 2          | 3            |

Tabla 7. Resultados experimento few-shots

La Tabla 8 compara los resultados de los dos experimentos. Se puede ver como la estrategia Few-Shots se acerca más a la valoración hecha por los expertos humanos en dos de los

comentarios. En el comentario 3 da la misma puntuación total que el humano, pero no acierta en el criterio.

| I<br>D | ZeroShot   |              | FewShot    |              | Comparativa |              |
|--------|------------|--------------|------------|--------------|-------------|--------------|
|        | Diferencia | Coincidencia | Diferencia | Coincidencia | Diferencia  | Coincidencia |
| 1      | -1         | 4            | -1         | 4            | 0           | 0            |
| 2      | 2          | 3            | 1          | 4            | 1           | 1            |
| 3      | 1          | 4            | 1          | 4            | 0           | 0            |
| 4      | -1         | 4            | 0          | 3            | 1           | -1           |
| 5      | 0          | 3            | 0          | 3            | 0           | 0            |
| 6      | 1          | 2            | 0          | 3            | 1           | 1            |
| 7      | 0          | 5            | 0          | 5            | 0           | 0            |
| 8      | 0          | 5            | 0          | 5            | 0           | 0            |
| 9      | 2          | 3            | 2          | 3            | 0           | 0            |

Tabla 8. Comparación de resultados

## 5 DISCUSIÓN Y CONCLUSIONES

La evaluación requiere de recogida, análisis y valoración de información para tomar decisiones. En ese sentido, una de las ventajas de la IA es su capacidad para integrar más datos en su proceso de toma de decisiones que los humanos jamás podrían gestionar, lo que conduce a calificaciones más ajustadas y precisas (Ruge y Mayweg, 2023). Sin embargo, según estos autores la evaluación tiene una dimensión interactiva y comunicativa que peligra con el empleo de la IA, puesto que su empleo puede tener efectos menos beneficiosos en la motivación, la autoeficacia y las emociones, dado que esto depende, en gran parte, de las características del proveedor (Ruge y Mayweg, 2023). Por otra parte, si bien es cierto que se evitan prejuicios, los matices y ajustes que se producen en los procesos de evaluación como consecuencia del grado de conocimiento que el profesorado tiene sobre cada alumno se pierden con los procesos automatizados (Swiecki et al., 2022).

Concretamente, respecto a los prompts, la necesaria formación del profesorado y el tiempo que hay que invertir en crear y refinar prompts válidos son limitaciones ya consignadas por estudios previos (Eager & Brunton, 2023). A pesar de la limitación del análisis, disponer de esta herramienta puede contribuir a valorar la capacidad de los estudiantes de emitir juicios ajustados a los objetivos, de forma objetiva, con un tono amable, indicando lo que se hace bien y lo que se tendría que mejorar, aportando sugerencias (Cano et al., 2024). Esta capacidad de emitir juicios permite obtener información sobre el juicio evaluativo, que es una competencia transversal vinculada al aprendizaje a lo largo de la vida (Tai et al, 2018) y que puede desarrollarse mediante procesos de evaluación entre iguales o de autoevaluación (Panadero & Alqassab, 2019; Panadero et al., 2022). Esta funcionalidad, integrable en el LMS que cada institución posea, permitiría operativizar los procesos de participación del estudiantado en la evaluación.

Aún y el del valor de la propuesta, se hace necesario consignar algunas limitaciones. En primer lugar, los datos usados en el diseño

del prompt del Chat GPT-4 se ha realizado en base a una clasificación humana de alcance limitado (con asignaturas del campo de conocimiento de ciencias de la educación, en un contexto particular y relativas a tres tipos de tareas puntuales). Probablemente diversificar los fragmentos de feedback con otras áreas, tareas y contextos culturales podría llevar a resultados más generalizables. En segundo lugar, algunas dificultades técnicas del propio Chat GPT han sugerido depurar el análisis. En un principio, se intentó usar la estrategia de prompting tree of thoughts (Yao et al., 2023) donde se le pide al modelo que actúe como diferentes agentes y así poder comparar distintas salidas del modelo de forma simultánea. En nuestro caso, la idea era la de solicitar una valoración por parte de 3 evaluadores independientes. El inconveniente que hubo es que usando el modelo predefinido de Chat GPT no alcanzaba a responder a más de un par de evaluaciones por experimento, volviéndose esta estrategia muy ineficiente. Aparte, las diferentes salidas de los evaluadores no diferían en las valoraciones, sino que tan solo daban justificaciones diferentes.

Por último, hay que destacar como limitación, que se ha empleado una herramienta de pago que, con grandes cantidades de datos, puede llegar a ser muy costosa.

Así pues, se ha realizado un trabajo de naturaleza exploratoria sobre una selección de fragmentos analizados con Zero-shot y Few-shots y se aportan los resultados de un primer análisis, que sería necesario ampliar con el fin de poder generalizar los hallazgos.

El siguiente paso sería desarrollar un fine-tuning (Mecklenburg et al., 2024) del modelo con un mayor volumen de datos. El proceso de fine-tuning consiste en entrenar un modelo pre-entrenado como GPT-4, pero usando datos propios para mejorar la tarea que se desea solucionar. En nuestro caso se podría entrenar con comentarios de feedback anotados de evaluadores o de los propios profesores.

En síntesis, ulteriores análisis se hacen necesarios para disponer de una propuesta generalizable pero el análisis iniciado parece sugerir prometedores resultados respecto a la posibilidad de automatizar la evaluación de los feedbacks proporcionados por el estudiantado.

## Agradecimientos

Esta publicación es parte del proyecto de I+D+i PID2019-104285GB-I00, financiado por MCIN/ AEI/10.13039/501100011033

## REFERENCIAS

- Anderson, O. S., El Habbal, N., & Bridges, X. D. (2020). A peer evaluation training results in high-quality feedback, as measured over time in nutritional sciences graduate students. *Advances in Physiology Education*, 44(2), 203-209. <https://doi.org/10.1152/ADVAN.00114.2019>
- Bearman, M., & Ajjawi, R. (2019). Can a rubric do more than be transparent? Invitation as a new metaphor for assessment criteria. *Studies in Higher Education*, 46(7), 1-10. <https://doi.org/10.1080/03075079.2019.1637842>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623. <https://doi.org/10.1145/3442188.3445922>
- Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2024). Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4 (arXiv:2312.16171). *arXiv*. <http://arxiv.org/abs/2312.16171>
- Cano, E.; Jardí, A.; Lluch, L.; Martins, L. (2024). Improvement in the Quality of Feedback as an Indication of the Development of Evaluative Judgement.

- Assessment and Evaluation in Higher Education [10.1080/02602938.2024.2329170](https://doi.org/10.1080/02602938.2024.2329170)
- Carless, D. and Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment and Evaluation in Higher Education*, 43(8), 1315-1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Eager, B., & Brunton, R. (2023). Prompting Higher Education Towards AI-Augmented Teaching and Learning Practice. *Journal of University Teaching & Learning Practice*, 20(5). <https://doi.org/10.5376/j.1.20.5.02>
- Fernández-Ferrer, M., & Cano, E. (2019). Feedback experiences to improve the continuous assessment: The use of Twitter as an emerging technology. *Educator*, 55(2). <https://doi.org/10.5565/rev/educar.872>
- Hawe, E., & Dixon, H. (2017). Assessment for learning: a catalyst for student self-regulation. *Assessment and Evaluation in Higher Education*, 42(8), 1181–1192. <https://doi.org/10.1080/02602938.2016.1236360>
- Gaynor, J. (2020). Peer review in the classroom: student perceptions, peer feedback quality and the role of assessment. *Assessment & Evaluation in Higher Education*, 45(5), 758-775. <https://doi.org/10.1080/02602938.2019.1697424>
- Gielen, M., & De Wever, B. (2015). Structuring the peer assessment process: A multilevel approach for the impact on product improvement and peer feedback quality. *Journal of Computer Assisted Learning*, 31(5), 435-449. <https://doi.org/10.1111/jcal.12096>
- Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1), 81-112. <https://doi.org/10.3102/003465430298487>
- Imani, S., Du, L., & Shrivastava, H. (2023). MathPrompter: Mathematical Reasoning using Large Language Models (arXiv:2303.05398). arXiv. <http://arxiv.org/abs/2303.05398>
- Kamalloo, E., Dziri, N., Clarke, C. L. A., & Rafiei, D. (2023). Evaluating Open-Domain Question Answering in the Era of Large Language Models (arXiv:2305.06984). arXiv. <http://arxiv.org/abs/2305.06984>
- Lai, V. D., Ngo, N. T., Veysah, A. P. B., Man, H., Deroncourt, F., Bui, T., & Nguyen, T. H. (2023). ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning (arXiv:2304.05613). arXiv. <http://arxiv.org/abs/2304.05613>
- Li, R., Allal, L. B., Zi, Y., Muennighoff, N., Kocetkov, D., Mou, C., Marone, M., Aikiki, C., Li, J., Chim, J., Liu, Q., Zheltonozhskii, E., Zhuo, T. Y., Wang, T., Dehaene, O., Davaadorj, M., Lamy-Poirier, J., Monteiro, J., Shliazhko, O., ... de Vries, H. (2023). StarCoder: May the source be with you! (arXiv:2305.06161). arXiv. <http://arxiv.org/abs/2305.06161>
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing (arXiv:2107.13586). arXiv. <http://arxiv.org/abs/2107.13586>
- Meckenburg, N., Lin, Y., Li, X., Holstein, D., Nunes, L., Malvar, S., Silva, B., Chandra, R., Aski, V., Yannam, P. K. R., Aktas, T., & Hendry, T. (2024). Injecting New Knowledge into Large Language Models via Supervised Fine-Tuning (arXiv:2404.00213). arXiv. <http://arxiv.org/abs/2404.00213>
- Mendonça, J., Pereira, P., Moniz, H., Carvalho, J. P., Lavie, A., & Trancoso, I. (2023). Simple LLM Prompting is State-of-the-Art for Robust and Multilingual Dialogue Evaluation (arXiv:2308.16797). arXiv. <http://arxiv.org/abs/2308.16797>
- O'Donnell, A.M.; Topping, K.J. (1998). Peers assessing. peers: Possibilities and problems. In K. Topping, & S. Ehly, S., Eds., *Peer-Assisted Learning*. Lawrence Erlbaum.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2024). GPT-4 Technical Report (arXiv:2303.08774). arXiv. <http://arxiv.org/abs/2303.08774>
- OpenAI Platform. (s.d.). Recuperat 14 abril 2024, de <https://platform.openai.com>
- Panadero, E., & Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment and Evaluation in Higher Education*, 44(8), 1253-1278. <https://doi.org/10.1080/02602938.2019.1600186>
- Panadero, E., Alqassab, M., Fernández Ruiz, J., & Ocampo, J. C. (2023). A systematic review on peer assessment: Intrapersonal and interpersonal factors. *Assessment & Evaluation in Higher Education*, 1-23. <https://doi.org/10.1080/02602938.2023.2164884>
- Panadero, E., Pérez, D. G., Ruiz, J. F., Fraile, J., Sánchez-Iglesias, I., & Brown, G. T. L. (2022). University students' strategies and criteria during self-assessment: Instructor's feedback, rubrics, and year level effects. *European Journal of Psychology of Education*. <https://doi.org/10.1007/s10212-022-00639-4>
- Ruwe, T., & Mayweg-Paus, E. (2023). "Your argumentation is good", says the AI vs humans – The role of feedback providers and personalised language for feedback effectiveness. *Computers and Education: Artificial Intelligence*, 5, 100189. <https://doi.org/10.1016/j.caeai.2023.100189>
- Sadler, D.R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5), 535-550. <https://doi.org/10.1080/02602930903541015>
- Schick, T., & Schütze, H. (2021). Few-Shot Text Generation with Pattern-Exploiting Training (arXiv:2012.11926). arXiv. <http://arxiv.org/abs/2012.11926>
- Swiecki, Z., Khosravi, H., Chen, G., Martínez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3, 100075. <https://doi.org/10.1016/j.caeai.2022.100075>
- Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. *Higher Education*, 76(3), 467-481. <https://doi.org/10.1007/s10734-017-0220-3>
- Using Workshop—MoodleDocs. (s.d.). Recuperat 14 abril 2024, de [https://docs.moodle.org/403/en/Using\\_Workshop](https://docs.moodle.org/403/en/Using_Workshop)
- Wang, J., Liang, Y., Meng, F., Sun, Z., Shi, H., Li, Z., Xu, J., Qu, J., & Zhou, J. (2023). Is ChatGPT a Good NLG Evaluator? A Preliminary Study (arXiv:2303.04048). arXiv. <http://arxiv.org/abs/2303.04048>
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models (arXiv:2305.10601). arXiv. <http://arxiv.org/abs/2305.10601>
- Zawacki-Richter, O., Marin, V.I., Bond, M. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators?. *Int J Educ Technol High Edu*, 16, 39. <https://doi.org/sire.ub.edu/10.1186/s41239-019-0171-0>
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023). A Survey of Large Language Models (arXiv:2303.18223). arXiv. <http://arxiv.org/abs/2303.18223>
- Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. En B. J. Zimmerman y D. H. Schunk (Eds), *Self-regulated Learning and Academic Achievement: Theoretical Perspectives* (pp. 1-37). Lawrence Erlbaum.

## POT LA INTEL·LIGÈNCIA ARTIFICIAL PROPORCIONAR UN FEEDBACK MÉS SOSTENIBLE?

L'avaluació entre iguals és una estratègia mitjançant la qual els estudiants avaluen el nivell, valor o qualitat del treball dels seus companys dins del mateix entorn educatiu. La recerca ha demostrat que els processos d'avaluació entre iguals impacten positivament en el desenvolupament de habilitats i el rendiment acadèmic. En aplicar criteris d'avaluació al treball dels seus companys i oferir comentaris, correccions i suggeriments per millorar, els estudiants no només milloren el seu propi treball, sinó que també cultiven habilitats de pensament crític. Per a alimentar eficaçment el paper dels estudiants com a avaluadors, són essencials oportunitats deliberades i estructurades per practicar, juntament amb formació i orientació. La intel·ligència artificial pot oferir un mitjà per avaluar automàticament les avaluacions entre iguals, assegurant-ne la qualitat i ajudant els estudiants a realitzar avaluacions amb precisió. Aquest enfocament permet als educadors centrar-se en avaluar les produccions dels estudiants sense necessitat d'una formació especialitzada en avaluació de retroalimentació. Aquest document presenta el procés desenvolupat per automatitzar l'avaluació de la qualitat de la retroalimentació. A través de l'utilització de fragments de retroalimentació avaluats per investigadors basats en criteris preestablerts, es va entrenar un Model de Llenguatge Massiu (LLM) d'Intel·ligència Artificial per aconseguir l'avaluació automatitzada. Es discuteixen els reptes i perspectives d'aquest procés, juntament amb recomanacions per optimitzar els resultats.

**PARAULES CLAU:** Intel·ligència artificial, avaluació, feedback, educació superior

## CAN ARTIFICIAL INTELLIGENCE HELP PROVIDE MORE SUSTAINABLE FEEDBACK?

Peer assessment is a strategy wherein students evaluate the level, value, or quality of their peers' work within the same educational setting. Research has demonstrated that peer evaluation processes positively impact skill development and academic performance. By applying evaluation criteria to their peers' work and offering comments, corrections, and suggestions for improvement, students not only enhance their own work but also cultivate critical thinking skills. To effectively nurture students' role as evaluators, deliberate and structured opportunities for practice, along with training and guidance, are essential. Artificial Intelligence (AI) can offer a means to assess peer evaluations automatically, ensuring their quality and assisting students in executing assessments with precision. This approach allows educators to focus on evaluating student productions without necessitating specialized training in feedback evaluation. This paper presents the process developed to automate the assessment of feedback quality. Through the utilization of feedback fragments evaluated by researchers based on pre-established criteria, an Artificial Intelligence (AI) Large Language Model (LM) was trained to achieve automated evaluation. The findings show the similarity between human evaluation and automated evaluation, which allows expectations to be generated regarding the possibilities of AI for this purpose. The challenges and prospects of this process are discussed, along with recommendations for optimizing results.

**KEYWORDS:** Artificial Intelligence, Assessment, Feedback, Higher Education

The authors retain copyright and grant the journal the right of first publication. The texts will be published under a Creative Commons Attribution-Non-Commercial-NoDerivatives License.

