

Generación de rúbricas con herramientas de inteligencia artificial para la evaluación de aprendizajes en educación superior

Rubric Generation with Artificial Intelligence Tools for Learning Assessment in Higher Education

Beatriz Ortega Estrada

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO (UNAM), MÉXICO

betyortegae@gmail.com

ORCID: <https://orcid.org/0009-0009-7264-2241>

Abraham Daniel Hernández Fabián

UNIVERSIDAD IBEROAMERICANA CIUDAD DE MÉXICO, MÉXICO

abraham.hernandez@ibero.mx

ORCID: <https://orcid.org/0009-0002-7196-0108>

https://doi.org/10.48102/didac.2024..84_JUL-DIC.211



RESUMEN

Este artículo recupera la implementación de herramientas de inteligencia artificial generativa, específicamente ChatGPT y Bard, en el proceso de evaluación de un proyecto colaborativo final en un curso de pedagogía en la Universidad Nacional Autónoma de México (UNAM). El proyecto consistió en el diseño de un organismo internacional especializado en educación. Se utilizaron programas de inteligencia artificial generativa para elaborar rúbricas de evaluación en el trabajo del estudiantado. Se reflexiona sobre las experiencias con ChatGPT y Bard, resaltando sus fortalezas y limitaciones, a partir de lo cual se enfatiza la importancia de instrucciones claras y específicas al utilizar herramientas de IA generativa con fines educativos. Se concluye que estas herramientas pueden ser útiles para generar instrumentos con mayor eficiencia, aunque siempre es necesaria la supervisión y modificación del docente para obtener resultados confiables.

Palabras clave: Evaluación del aprendizaje; Rúbricas; Inteligencia artificial; Educación y tecnología; Docencia universitaria.

ABSTRACT

This article examines the use of Generative Artificial Intelligence (GAI) tools, specifically ChatGPT and Bard, in assessing a collaborative final project in a Pedagogy course at the Universidad Nacio-

nal Autónoma de México (UNAM). The project involved students designing and establishing an international education organization. ChatGPT and Bard were employed to develop assessment rubrics for students' work. The article reflects on the experiences with these tools, highlighting their strengths and limitations in this context. It emphasizes the importance of clear and specific instructions when using GAI for educational purposes. The conclusion posits that while these tools can help generate instruments with greater efficiency, the supervision and modification of the professor are always necessary to obtain reliable results.

Keywords: Learning Assessment; Rubrics; Artificial Intelligence; Education and Technology; University Teaching.

Fecha de recepción: 10/01/2024

Fecha de aceptación: 06/03/2024

Introducción

A lo largo del desarrollo de la asignatura obligatoria de quinto semestre en el plan de estudios de Pedagogía de la Facultad de Filosofía y Letras de la Universidad Nacional Autónoma de México (UNAM), Organismos y Sistemas Internacionales de Educación, uno de los principales retos es construir junto con el estudiantado los lineamientos y características del trabajo final colaborativo que permitiera observar y valorar su aprendizaje. El trabajo final consistió en generar una propuesta de diseño de un organismo internacional especializado en educación definiendo su nombre, eslogan, clasificación, concepto de la educación, líneas de acción, marco teórico, acciones principales, la manera en la que colabora con los países y la presentación de un proyecto impulsado desde dicho organismo.

El diseño de un organismo internacional, como producto de aprendizaje y evaluación sumativa, tuvo la finalidad de poner en práctica los conocimientos adquiridos durante las clases y verificar el logro del objetivo fijado en la planeación didáctica de la asignatura. La evaluación en el área educativa representa el fundamento de este trabajo y es definida por Miller como: “un término genérico que incluye un rango de procedimientos para adquirir información sobre el aprendizaje del estudiante y la formación de juicios de valor” (Miller, 2012, como se citó en Sánchez, 2022, p. 20). Por este motivo, era necesario basarse en los principios generales de este autor, que indican que se requiere especificar claramente lo que se va a evaluar, considerar que la evaluación es un medio para un fin, definir sus métodos, procedimientos e instrumentos para que ésta sea útil y efectiva.

Por esta razón se definió que la evaluación final sería la construcción del diseño de un organismo internacional por equipo, la presentación en plenaria de las características del mismo, la realización de una coevaluación entre pares y la heteroevaluación del equipo docente, con base en una rúbrica (compartida con el grupo, previo a la presentación en plenaria) elaborada con ayuda de inteligencia artificial (IA), tomando en cuenta que este instrumento “define tareas, actividades o comportamientos específicos que se desean valorar, así como los niveles de desempeño asociados a cada uno de éstos. Es una guía articulada y precisa que ilustra los objetivos de cada tarea y su relevancia en el proceso de evaluación” (Sierra, Sosa & González en Sánchez, 2022, p. 234).

El proceso de coevaluación tuvo la intención de realizar la valoración como aprendizaje; es decir, orientar al estudiantado “a aprender a aprender, fomentar la meta-cognición [*sic*] y el aprendizaje autorregulado” (Sánchez, 2022, p. 28) al realizar una evaluación del trabajo de sus colegas al mismo tiempo que realizaban una autoevaluación tanto de sus procesos como de sus resultados de aprendizaje. El presente artículo muestra un análisis de la experiencia como equipo docente, integrado

por el profesor titular y la ayudante docente en educación superior, al utilizar dos herramientas de inteligencia artificial generativa en la construcción de un instrumento que sirvió para evaluar el trabajo final colaborativo del estudiantado. Se inició destacando y definiendo algunos conceptos clave para, posteriormente, desarrollar la experiencia en el uso de dichos programas y así definir las razones por las que fueron elegidas para realizar un primer experimento de integrar las IA generativas en la evaluación educativa. Finalmente, se mencionan algunas de las limitaciones de estos programas y se brindan unas recomendaciones para superarlas, específicamente por medio de dos modelos de redacción de instrucciones (*prompts*).

Una experiencia del uso de ChatGPT y Bard en la generación de rúbricas

La inteligencia artificial es el resultado del trabajo de una serie de disciplinas y ciencias, lo que la hace un campo interdisciplinario de la informática que desarrolla sistemas y algoritmos para emular algunas capacidades humanas tales como el aprendizaje, el reconocimiento perceptual, el razonamiento y el análisis de datos, la toma de decisiones y la inteligencia a través del uso de técnicas avanzadas para el procesamiento de datos y modelado de lenguaje natural (Chao, Caudillo, Caletti & Mayorga, 2023).

Existe una diversidad de clasificaciones de IA; sin embargo, Chao, Caudillo, Caletti y Mayorga (2023) la catalogan en tres tipos: inteligencia artificial especializada, inteligencia artificial generativa e inteligencia artificial súper poderosa. La primera, realiza tareas específicas por medio de información en distintos formatos, ya sea gráfico, auditivo o por texto; por este hecho, su ejecución es limitada. La IA generativa se caracteriza por procesar y generar información multimodal con ayuda de modelos de lenguajes sofisticados (*Large Language Models*, LLM) lo que origina que sea capaz de entablar conversaciones complejas, un ejemplo de este tipo es ChatGPT, de OpenAI. La inteligencia artificial súper poderosa hace referencia a que es posible que la IA supere a la inteligencia humana. Las dos primeras inteligencias actualmente son predominantes y éstas sólo responden a las necesidades y usos de las personas y pueden observarse comúnmente en los teléfonos celulares, correos electrónicos, buscadores de información en la web, entre otros.

Las herramientas de IA generativa se popularizaron al final de 2022 en el ámbito académico y profesional. Fue durante el siguiente año que se empezaron, poco a poco, a incorporar en el quehacer docente, hasta que finalmente se tomó la decisión de utilizarlas para construir un instrumento de evaluación y sistematizar la experiencia descrita en este texto.

Esta decisión puso como reto el familiarizarse aún más con estas herramientas y empezar a diversificar los programas que se utilizan, puesto que únicamente se había utilizado ChatGPT de OpenAI para generar textos, ideas de propuestas de actividades de aprendizaje y resúmenes, así como reelaborar escritos, detectar errores ortográficos y de redacción, entre otras tareas simples.

El trabajo académico de analizar y proponer los usos de IA generativa en la educación se ha materializado en diversos documentos de trabajo, artículos y talleres. En particular, el Programa Universitario de Estudios sobre Democracia y Justicia Social (PUEDJS-UNAM) generó un taller con el título *Interactuando con la Inteligencia Artificial: Taller Práctico sobre ChatGPT y Bard*, el cual fungió como un espacio para aprender a redactar instrucciones para obtener el resultado deseado.

Para el desarrollo de la rúbrica, se decidió emplear ChatGPT de OpenAI, versión 3.5 y Bard de Google, versión 2. ChatGPT es una herramienta que forma parte de la familia de modelos de lenguaje GPT (Generative Pre-trained Transformer) y su objetivo principal es “comprender y generar texto de manera coherente y contextual” (OpenAI, 2023a). Hay que tomar en cuenta que esta he-

herramienta no tiene acceso a internet en tiempo real. Sus datos de entrenamiento, obtenidos de diversas fuentes de internet, fueron actualizados por última vez en enero de 2022. Por su parte, Bard es un modelo de lenguaje factual de Google AI, con acceso a internet en tiempo real.

En particular, ambos requieren que el usuario escriba una serie de instrucciones para que la inteligencia genere lo que se le está solicitando. A esas instrucciones se les denomina *prompts*, y se refieren a

frases o preguntas que se utilizan para dar dirección a un modelo de lenguaje para que produzca una respuesta ... Los *prompts* bien definidos y precisos pueden ayudar a guiar la conversación de manera efectiva, asegurándose de que los temas de interés del usuario sean abordados. (Morales-Chan, 2023, p. 2)

El taller del PUEJDS y la práctica en sí misma, enseñaron estrategias de diseño de instrucciones o *prompts*. Los más utilizados son RTF (*role, task, format*) que en español significa “rol”, “tarea” y “formato”; y RISEN (*role, instructions, steps, end goal, narrowing*) que en español refieren a “rol”, “instrucciones”, “pasos a seguir”, “objetivo” y “restricciones” (Kremb, 2023).

a) Experimentando con ChatGPT

El primer *prompt* ingresado en el chat de OpenAI fue el siguiente: *soy una profesora de la licenciatura en Pedagogía, de la asignatura Organismos y Sistemas Internacionales de Educación, quiero que me ayudes a generar una rúbrica para evaluar el trabajo final de mis estudiantes, que consiste en realizar un organismo internacional especializado en educación (Prompt 1)*; se le asignó la lista de lineamientos a evaluar, ya mencionados en la introducción. El resultado fue una tabla con 15 criterios, uno para cada uno de los lineamientos, como se muestra a continuación.

Tabla 1
Resultado del Prompt 1 en ChatGPT

Fuente: OpenAI.

| CRITERIO | EXCEPCIONAL (5) | NOTABLE (4) | ACEPTABLE (3) | BÁSICA (2) | INSATISFACTORIA (1) |
|---|--|----------------------------|---------------------------|--|--------------------------------|
| 1. Nombre del organismo internacional | Claro, relevante y original | Claro y relevante | Adecuado | Vago o poco relevante | Ausente o inapropiado |
| 2. Slogan | Creativo y representativo | Creativo | Aceptable | Poco original | Ausente o poco relevante |
| 3. Clasificación (gubernamental o no gubernamental) | Claramente justificado | Justificación adecuada | Justificación básica | Falta justificación | No clasificado |
| 4. Concepción de la educación | Profundidad y claridad en la explicación | Explicación clara | Adecuada explicación | Superficial | No definido o confuso |
| 5. Líneas de acción (3) | Detalladas y coherentes con la concepción | Coherentes | Adecuadas | Poco claras o vagas | Ausentes o irrelevantes |
| 6. Marco teórico | Relacionado con las líneas de acción, cita de autores relevantes | Cita de autores relevantes | Cita de autores adecuados | Poca relación con las líneas de acción | Ausente o sin citas relevantes |

| | | | | | |
|---|--|--|--|--|-------------------------|
| 7. Acciones de los organismos internacionales | Detalladas y relacionadas con las líneas de acción | Relacionadas con líneas de acción | Adecuadas | Poco relacionadas | Ausentes o irrelevantes |
| 8. Colaboración con México | Detallada y sustentada con ejemplos | Sustentada con ejemplos | Mencionada con ejemplos | Poco clara o sin ejemplos | No mencionada |
| 9. Nombre del proyecto | Claro, relevante y refleja la esencia del proyecto | Claro y relevante, conectándose con los objetivos del proyecto | Adecuado, pero puede faltar originalidad o conexión directa con los objetivos del proyecto | Adecuado, pero puede faltar originalidad o conexión directa con los objetivos del proyecto | Ausente o incoherente |
| 10. Objetivo general y específicos | Detallados y completamente desarrollados | Detallados y desarrollados, aunque puede haber áreas que necesiten más información | Detallados, pero puede haber áreas que necesiten más desarrollo | Pueden estar ausentes o carecer de detalle y desarrollo | Ausentes o incoherentes |
| 11. Indicadores para evaluación | Claros, específicos y medibles | Claros y específicos, pero pueden necesitar más detalle | Proporcionados, pero poco claros o necesitan más especificidad | Ausentes o poco claros y específicos | Ausentes |
| 12. Población objetivo | Definida y precisa | Definida de manera clara, pero puede haber áreas que necesiten más precisión | Definida, pero vaga o necesita más precisión | Ausente o poco clara | Ausente |
| 13. Periodo | Claramente establecido | Establecido, pero puede necesitar más detalle | Establecido, aunque puede necesitar más detalle | Ausente o poco claro | Ausente |
| 14. Financiamiento | Detalles sobre la fuente y cantidad de financiamiento son proporcionados | Detalles sobre la fuente de financiamiento, pero puede necesitar más información | Fuente de financiamiento mencionada, pero puede necesitar más información | Ausente o insuficientemente detallado | Ausente |
| 15. Instituciones gubernamentales y/o asociadas | Mencionados [sic] e integrados [sic] de manera clara en el proyecto | Mencionados [sic], aunque la integración puede necesitar más detalle | Mencionados [sic], pero puede faltar integración clara en el proyecto | Ausentes o poco relevantes para el proyecto | Ausente |

Sin embargo, debido a que la redacción del *prompt* no fue lo suficientemente específica, los criterios resultaron demasiado genéricos, en algunos casos repetidos, y los niveles de desempeño eran demasiados, por lo que se tuvo que solicitar su reelaboración.

El segundo *prompt* fue el siguiente: *soy una profesora de la licenciatura en Pedagogía, de la asignatura Organismos y Sistemas Internacionales de Educación, quiero que me ayudes a generar una rúbrica con 4 niveles de desempeño para evaluar el trabajo final de mis estudiantes, que consiste en realizar un organismo internacional especializado en educación. Los lineamientos del trabajo final son los siguientes, por favor sé muy específico en los elementos a evaluar.*

La respuesta fue más atinada que la anterior; no obstante, los criterios que antes eran demasiado genéricos, ahora se presentaron de manera muy extensa. Por ejemplo, para un desempeño excepcional del nombre del organismo internacional, se establecía lo siguiente: “El nombre es creativo, relevante y refleja una profunda comprensión de los objetivos del organismo, proporcionando una

identidad clara y distintiva” (OpenAI, 2023b). Fue a partir de este resultado que se empezó a examinar criterio por criterio y, en caso de creerlo necesario, solicitar diversas modificaciones tanto de redacción como del uso de elementos que conforman el organismo internacional a evaluar (creatividad, claridad, relevancia, coherencia, justificación, relación y precisión, entre otros).

Se solicitaron entre una y tres reelaboraciones de los elementos a evaluar de algunos criterios y, finalmente, tuvieron que modificarse sin ayuda de la IA, sobre todo aquellos criterios que la inteligencia artificial ignoraba y que estaban estrechamente relacionados con los objetivos y características de la asignatura y del estudiantado. La falta de contexto de las herramientas de IA generativa es un elemento que hay que tomar en cuenta durante su uso.

Utilizando lo aprendido para interactuar con Bard

En el caso del uso de Bard, el primer paso fue proporcionarle la misma lista de lineamientos otorgados a ChatGPT (Prompt 2). La respuesta de Bard fue una tabla (Tabla 2) con cuatro niveles de desempeño para cada criterio: Excelente, Bueno, Suficiente e Insuficiente. Sin embargo, la tabla no estaba organizada de acuerdo con los lineamientos, por lo que se le solicitó reestructurarla.

Tabla 2
Resultado del Prompt 2 en Bard

Fuente: Google, 2023.

| NIVEL DE DESEMPEÑO | PORCENTAJE | DESCRIPCIÓN |
|--------------------|------------|--|
| Excelente | 90-100% | El trabajo cumple con todos los requisitos establecidos en los lineamientos, además de: <ul style="list-style-type: none"> • Presenta una concepción de la educación coherente con sus líneas de acción. • El marco teórico es pertinente y utiliza autores reconocidos en el campo de la educación. • Las acciones que realiza el organismo internacional son relevantes y están alineadas con sus líneas de acción. • La colaboración con México es efectiva y se basa en un diálogo respetuoso. • El proyecto impulsado desde el organismo es novedoso, viable y tiene un impacto positivo en la educación. |
| Bueno | 70-89% | El trabajo cumple con la mayoría de los requisitos establecidos en los lineamientos, pero: <ul style="list-style-type: none"> • La concepción de la educación no es totalmente coherente con sus líneas de acción. • El marco teórico es pertinente, pero no utiliza autores reconocidos en el campo de la educación. • Algunas de las acciones que realiza el organismo internacional no están alineadas con sus líneas de acción. • La colaboración con México es adecuada, pero podría mejorarse. • El proyecto impulsado desde el organismo es viable y tiene un impacto positivo en la educación, pero no es novedoso. |
| Regular | 50-69% | El trabajo cumple con algunos de los requisitos establecidos en los lineamientos, pero: <ul style="list-style-type: none"> • La concepción de la educación no es coherente con sus líneas de acción. • El marco teórico no es pertinente o no utiliza autores reconocidos en el campo de la educación. • Algunas de las acciones que realiza el organismo internacional no están alineadas con sus líneas de acción. • La colaboración con México es deficiente. • El proyecto impulsado desde el organismo no es viable o no tiene un impacto positivo en la educación. |
| Insuficiente | 0-49% | El trabajo no cumple con los requisitos establecidos en los lineamientos. |

En la segunda versión, Bard proporcionó una tabla organizada de acuerdo con los lineamientos. Los elementos por evaluar eran relevantes y compatibles con los fines de aprendizaje de la asigna-

tura. Aun así, se solicitó la reelaboración de algunos criterios y, tal como sucedió con ChatGPT, se editaron otros sin ayuda de la IA. Es rescatable que esta herramienta de Google brinda una serie de recomendaciones en torno al uso de fuentes de información confiables, la correcta citación, la organización del trabajo y el esfuerzo del estudiantado para crear un proyecto realista que tenga un impacto positivo en la educación (Google, 2023).

La tarea final fue comparar ambas rúbricas, criterio por criterio, para elegir los que estuvieran mejor definidos, tomando en cuenta las relaciones entre éstos y así tuviera sentido el instrumento de evaluación en el contexto de la asignatura. La rúbrica final fue, en su mayoría, producto de las interacciones previas con Bard de Google.

Tabla 3
Rúbrica final

Fuente: Elaboración propia con base en OpenAI (2023a) y Google (2023).

| CRITERIO | EXCEPCIONAL (4) | NOTABLE (3) | ACEPTABLE (2) | INSATISFACTORIA (1) |
|---|--|---|---|---|
| 1. Nombre del organismo internacional | Creativo, relevante y con conexión clara con los objetivos del organismo | Claro, relevante y con conexión adecuada con los objetivos del organismo | Adecuado, pero puede faltar originalidad o conexión directa con los objetivos | Vago, poco relevante o inapropiado |
| 2. Slogan | Impactante, claro, relevante y contribuye significativamente a la identidad del organismo | Claro, relevante y contribuye a la identidad del organismo | Aceptable, pero puede faltar originalidad o claridad en su conexión con el organismo | Poco claro, poco relevante o no contribuye significativamente [a la identidad del organismo] |
| 3. Clasificación (gubernamental o no gubernamental) | Clasificación claramente justificada con comprensión profunda de las implicaciones | Justificación adecuada, mostrando comprensión de las diferencias | Justificación básica, puede faltar profundidad en la comprensión | Justificación insuficiente o no presente |
| 4. Concepción de la educación | Refleja excepcionalmente la esencia y valores del nombre y slogan del organismo, estableciendo una conexión clara. | Refleja notablemente la esencia y valores del nombre y slogan del organismo, estableciendo una conexión clara. | Adecuada, aunque puede necesitar más énfasis en reflejar la esencia y valores del nombre y slogan del organismo. | Vaga o poco clara, y la conexión con el nombre y slogan del organismo no está adecuadamente especificada ni es coherente. |
| 5. Líneas de acción (3) | Claramente definidas, son coherentes con la concepción del organismo y son altamente relevantes | Definidas, se alinean con la concepción del organismo y son relevantes | Definidas, aunque pueden necesitar más claridad, se alinean generalmente con la concepción del organismo y son relevantes | Vagas o poco claras, no están adecuadamente alineadas con la concepción del organismo o no son relevantes |
| 6. Marco teórico | Bien desarrollado, relacionado directamente con las líneas de acción y respaldado por citas de autores relevantes | Desarrollado y relacionado con las líneas de acción, aunque puede faltar profundidad o citas de autores relevantes en algunas áreas | Básico, puede carecer de conexión directa con las líneas de acción o [carecer de] citar autores relevantes | Insuficiente, confuso o no relacionado con las líneas de acción |
| 7. Acciones de los organismos internacionales | Son relevantes, y coherentes con las líneas de acción del organismo y producen un beneficio a la educación del país. | Son relevantes y están alineadas con las líneas de acción del organismo. | Son poco relevantes o no son coherentes con las líneas de acción. | Están incompletas |
| 8. Colaboración con México | Es coherente con la clasificación del organismo. | | | No es coherente con la clasificación del organismo. |

| CRITERIO | EXCEPCIONAL (4) | NOTABLE (3) | ACEPTABLE (2) | INSATISFACTORIA (1) |
|---|---|---|--|---|
| 9. Nombre del proyecto | Claro, relevante y refleja la esencia del proyecto y del organismo. | Claro y relevante, conectándose con los objetivos del proyecto. | Adecuado, pero puede faltar conexión directa. | Vago, poco relevante o inapropiado |
| 10. Objetivo general y específicos | Objetivo general claramente definido y los objetivos específicos están directamente alineados, contribuyendo de manera excepcional al logro del objetivo general. | Objetivo general claro y objetivos específicos alineados, contribuyendo de manera notable al logro del objetivo general, aunque puede haber áreas que necesiten más precisión o conexión. | Objetivo general y objetivos específicos definidos, aunque pueden necesitar más detalle y la conexión entre ellos puede ser más clara. | Definición vaga o poco clara de objetivos generales y específicos, con una falta de coherencia y conexión clara entre ellos |
| 11. Indicadores para evaluación | Claros, específicos y medibles | Claros y específicos, pero pueden necesitar más detalle. | Proporcionados, pero pueden ser poco claros o necesitar más especificidad | Pueden estar ausentes o ser poco claros y específicos. |
| 12. Población objetivo | Población precisa y coherente con los objetivos e indicadores y en conexión directa con la esencia del proyecto. | Población clara y en conexión directa con los objetivos e indicadores. | Población definida, pero falta precisión. Conexión incompleta con los objetivos e indicadores. | Definición de población vaga, sin conexión con los objetivos y coherencia limitada o inexistente con los indicadores |
| 13. Periodo de implementación | Claramente establecido y coherente con los objetivos. | Está establecido y tiene relación con algunos objetivos. | Está establecido, pero muy poco coherente con los objetivos. | Puede estar ausente o no ser coherente con los objetivos. |
| 14. Financiamiento | Fuente de financiamiento claramente especificada y altamente coherente con la clasificación del organismo, sus acciones y las instituciones asociadas | Fuente de financiamiento especificada y en coherencia con dos de los elementos | Fuente de financiamiento especificada y en coherencia con uno de los elementos. | Puede estar ausente o sin coherencia con los elementos |
| 15. Instituciones gubernamentales o asociadas | Se mencionan más de dos y es clara su pertinencia en el proyecto. | Se mencionan dos o menos y es clara su pertinencia. | Se mencionan dos o menos, pero no es clara la pertinencia de alguna. | Se mencionan dos o menos, pero no es clara la pertinencia de ninguna. |

Cómo usar las IA generativas de manera efectiva: consejos para superar sus limitaciones

A partir de las interacciones con las IA generativa y de la revisión bibliográfica, se han notado algunas limitaciones de estas herramientas que es necesario tener en cuenta. Cabe resaltar que no todos los grandes modelos de lenguaje tienen acceso a internet en tiempo real. Pérez Salazar (2023, p. 4) menciona que, como las compañías no revelan de qué sitios web obtuvieron la gran cantidad de datos para entrenar a las IA, éstas no pueden verificar la fuente de las respuestas que arrojan; tal es el caso de ChatGPT. Por su parte, Bard sí hace referencia a la fuente de consulta en algunos casos; sin embargo, puede mostrar información inexacta.

Por otra parte, estas herramientas son muy sensibles a la redacción de las instrucciones o *prompts* y tienen una limitada comprensión contextual. Hay que ser sumamente claros o interactuar varias veces para obtener alguna respuesta cercana a la deseada. En esa misma línea, cada respuesta que generan es relativamente única, aunque se le dé la misma instrucción en diferentes momentos (Pérez Salazar, 2023, p. 5).

Además, las IA pueden “desvariar” y generar respuestas poco lógicas o coherentes, incluso dar una misma respuesta a diferentes preguntas. Aunque esto sucede poco debido a que están entrenadas para generar texto de manera contextualmente relevante, hay que ser precavidos y siempre revisar las respuestas.

A continuación, se comparten algunos ejemplos como recomendaciones de *prompts* empleando los distintos modelos para generar rúbricas como instrumentos de evaluación, reiterando que se debe adaptar según la finalidad de la valoración que se requiera hacer y las características de la asignatura y el estudiantado. Hay que recordar que, mientras más específica sea la instrucción, mejor será el resultado.

Como recomendaciones generales, para el modelo RFT, es necesario definir e instruir el rol que la IA generativa asumirá y lo que se quiere hacer; describir la tarea o instrumento que se requiere y las características de éste y, por último, mencionar su formato. Un ejemplo específico de este modelo es el siguiente:

Rol: Eres un(a) profesor(a) de licenciatura en Pedagogía y quieres evaluar el trabajo final de tus estudiantes.

Tarea: Realiza una rúbrica con cuatro niveles de desempeño para evaluar un trabajo final que consiste en una investigación socioeducativa sobre una entidad de la República mexicana. Los criterios son los siguientes: a) Nombre del estado, b) Datos demográficos, c) Contexto socioeconómico, d) Contexto cultural, e) Dimensión del sistema educativo, f) Indicadores educativos, g) Políticas y programas educativos públicos. También debe llevar introducción, conclusiones y referencias.

Formato: Exprésalo en una tabla separando cada criterio en una fila distinta.

(Se puede consultar la respuesta en el siguiente enlace: <https://chat.openai.com/share/87fce6b4-3228-4cd7-8240-9d691371486a>)

Para el Modelo RISEN, se tiene que establecer el rol, dar las instrucciones precisas, la serie de pasos que la inteligencia debe realizar, la meta que se pretende alcanzar y las restricciones. Para ejemplificar este modelo se propone lo siguiente:

Rol: Eres un(a) profesor(a) de la licenciatura en Pedagogía. Instrucciones: Genera una rúbrica para evaluar la exposición de un programa educativo impulsado por la Secretaría de Educación Pública.

Serie de pasos: 1. Distingue los criterios que son los siguientes: nombre del proyecto o programa, objetivo general, objetivos específicos, justificación haciendo referencia a organismos internacionales y datos del sistema educativo nacional, marco normativo (documentos legales), marco conceptual, población objetivo, origen de los recursos, convenios de colaboración con la sociedad civil, organismos internacionales e instituciones gubernamentales. 2. Genera elementos o indicadores para evaluar cada uno de los criterios. 3. Asegúrate de que esos elementos o indicadores tengan relación entre sí.

Meta: Una rúbrica en formato de tabla con cuatro niveles de desempeño por cada criterio del programa educativo.

Restricciones: 1. Toma en cuenta que se evaluará el contenido de una presentación frente a grupo, no el trabajo escrito, por lo que se debe[n] evitar indicadores como redac-

ción [o] citación correcta, entre otros que se usen para evaluar textos. 2. No usar más de 4 indicadores por criterio. 3. El marco de actuación del programa está limitado al sistema educativo mexicano. 4. No se evaluará tampoco la dicción o habilidades de habla en público. 5. Evita utilizar indicadores que no sean fácilmente identificables. 6. La estructura de la tabla debe ser la siguiente: los criterios representan las filas y cada columna representa un nivel de desempeño.

(Se puede consultar la respuesta en el siguiente enlace: <https://g.co/bard/share/498a28e46c82>)

También es importante mencionar que es posible compartir el enlace de las interacciones sostenidas con alguno de los modelos de lenguaje revisados. Es recomendable guardar los enlaces por si se requiere retomar dicha conversación otro día o solamente revisarla de nuevo. Bard, al final de cada respuesta, muestra un menú con el ícono para compartir; la interfaz de ChatGPT lo muestra en la esquina superior derecha.

Conclusiones

La emergencia y popularización de los programas de IA generativa han convocado a las comunidades educativas a pensarlas e incorporarlas en el quehacer pedagógico. Este tipo de herramientas es, sin duda alguna, un aliciente y ayuda para el desarrollo de diversas actividades docentes. Su uso debe estar respaldado por las cuestiones teóricas y metodológicas que determinen su uso, ya que estas herramientas están resolviendo y eficientizando problemáticas, no sustituyéndolas. Éstas presentan una oportunidad para los grupos de docentes con interés en experimentar con las nuevas tecnologías.

Específicamente, como equipo docente, se reconoce que pueden ser de gran ayuda en la automatización de algunos procesos e incluso un apoyo importante en la generación de material didáctico, actividades e instrumentos de evaluación; así como una oportunidad para repensar las prácticas docentes y de evaluación. En particular, la incorporación de herramientas de IA en el proceso evaluativo permite generar diferentes instrumentos con mayor eficiencia y seleccionar aquel que demuestre la validez y confiabilidad necesaria para el producto o proceso de aprendizaje que será evaluado.

Generar un instrumento de evaluación con ayuda de la IA generativa pone en juego la validez de éste, ya que puede estar dissociado de lo que realmente se pretende evaluar. Para asegurar la validez del instrumento, se analizó si la evidencia se basaba en el contenido de la prueba; es decir, si el instrumento aludía directamente a los conocimientos que se pretendían evaluar (Carrillo & Sánchez en Sánchez, 2022, p. 39). A su vez, al realizar la coevaluación y la heteroevaluación del producto de aprendizaje, se confirmó que los puntajes obtenidos fueran consistentes para procurar su confiabilidad. La rúbrica generada con IA en este estudio brindó un marco claro y transparente para la evaluación, aumentando la seguridad y confianza del alumnado.

No obstante, hay que señalar que las IA generativas siempre tendrán una comprensión contextual limitada, a menos que sea un programa entrenado específicamente con datos de la asignatura. Por ello es importante ser cuidadosos en la redacción de instrucciones para conseguir resultados pertinentes usando el menor tiempo y esfuerzo. Aunque, como se relató en esta experiencia, es difícil obtener una respuesta satisfactoria de rúbrica a partir de una única interacción. Lo anterior no significa que la calidad de las respuestas impida su aprovechamiento, sino que, dado el estado actual de estas tecnologías, la supervisión humana es absolutamente necesaria.

Si bien los instrumentos no son replicables en su totalidad debido a que el funcionamiento de las IA generativa brinda respuestas únicas de acuerdo con cada interacción, consideramos que son un buen incentivo o base que debe adecuarse a las distintas características de la asignatura, la institución en la que se instaura y el número de estudiantes, así como sus características e intereses y, sobre todo, a la experiencia profesional del profesorado. Por ello, lo presentado en este escrito pretende brindar una metodología para incentivar la incorporación de estas herramientas en la labor docente, haciendo énfasis en la redacción de *prompts* con base en modelos predefinidos (RTF y RISEN) para que las respuestas de la IA se apeguen a lo requerido.

REFERENCIAS

- Carrillo Ávalos, B., & Sánchez Mendiola, M. (2022). Validez, confiabilidad y amenazas a la validez. En M. Sánchez Mendiola, & A. Martínez González (Eds.), *Evaluación del aprendizaje en educación universitaria: estrategias e instrumentos* (pp. 37-52). UNAM-CUAIEED. <https://cuaed.unam.mx/publicaciones/libro-evaluacion/>
- Chao Rebolledo, C., Caudillo Zambrano, L., Caletti González, & Mayorga Ayala, A. (2023). *Orientaciones pedagógicas para el uso de la inteligencia artificial en la Educación Superior*. IBERO-DIE Idea. <https://idea.ibero.mx/recurso/orientaciones-pedagogicas-para-el-uso-de-la-inteligencia-artificial-ia-en-la-educacion-superior/>
- Google. (2023). *Bard* (versión 20 de noviembre) [Large language model]. Recuperado de <https://bard.google.com/chat>
- Kremb, M. (13 de octubre de 2023). 5 prompt frameworks to level up your prompts. The Prompt Warrior. <https://www.thepromptwarrior.com/p/5-prompt-frameworks-level-prompts>
- Morales-Chan, M. A. (2023). Explorando el potencial de ChatGPT: Una clasificación de Prompts efectivos para la enseñanza. GES Department, Galileo University, Guatemala. <https://biblioteca.galileo.edu/tesario/handle/123456789/1348>
- OpenAI. (2023a). *ChatGPT* (versión 20 de diciembre) [Large language model]. <https://chat.openai.com/>
- OpenAI. (2023b). *ChatGPT* (versión 17 de noviembre) [Large language model]. <https://chat.openai.com/>
- Pérez Salazar, G. (2023). *Uso del ChatGPT y otros LLMs en los entornos educativos*. Universidad Autónoma de Coahuila.
- Sánchez Mendiola, M. (2022). Evaluación del, para y como aprendizaje. En M. Sánchez Mendiola, & A. Martínez González (Eds.), *Evaluación del aprendizaje en educación universitaria: estrategias e instrumentos* (pp. 17-35). UNAM-CUAIEED. <https://cuaed.unam.mx/publicaciones/libro-evaluacion/>
- Sierra González, R., Sosa Ramírez, K., & Gonzáles Garibay, V. (2022). Rúbricas. En M. Sánchez Mendiola & A. Martínez González (Eds.) *Evaluación del aprendizaje en educación universitaria: estrategias e instrumentos* (pp. 233-249). UNAM-CUAIEED. <https://cuaed.unam.mx/publicaciones/libro-evaluacion/>

BIBLIOGRAFÍA RECOMENDADA

- Bozkurt, A., & Sharma, R. C. (2023). Generative AI and prompt engineering: The art of whispering to let the genie out of the algorithmic world. *Asian Journal of Distance Education*, 18(2), i-vii. <https://asianjde.com/ojs/index.php/AsianJDE/issue/view/40>
- Nerantzi, C., Abegglen, S., Karatsiori, M., & Martínez-Arboleada, A. (Eds.). (2023). *101 Creative ideas to use AI in education*. Zenodo. <https://zenodo.org/records/8072950>
- UNESCO. (2022). *Recomendaciones sobre la ética de la inteligencia artificial*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137_spa
- Universidad Nacional Autónoma de México (UNAM). (2023). *Recomendaciones para el uso de Inteligencia Artificial Generativa en la docencia de la UNAM*. <https://cuaieed.unam.mx/de-scargas/recomendaciones-uso-ia-gen-docencia-unam-2023.pdf?fbclid=IwAR33uFXDD3aKzxOLdETULjK-zXFSShGtoC2AvRJleve2tHxwUtoAcSTnTaY>

SEMBLANZAS

Beatriz Ortega Estrada es egresada de la licenciatura en Pedagogía de la Facultad de Filosofía y Letras de la UNAM, ayudante de profesor en las asignaturas Sistema Educativo Nacional y Organismos y Sistemas Internacionales de Educación desde agosto de 2022, así como ayudante de investigador SNI 3 desde octubre de 2023. Fue miembro del comité organizador y evaluador del Encuentro Nacional de Estudiantes de Pedagogía y Ciencias de la Educación en sus ediciones 19°, 20° y 21°, así como del Primer Coloquio de Historiografía de la Educación y la Pedagogía en México.

Abraham Daniel Hernández Fabián es pedagogo, egresado de la maestría en Pedagogía de la UNAM. Actualmente es académico de tiempo completo en la Universidad Iberoamericana Ciudad de México. Desde 2015 es profesor en el Colegio de Pedagogía de la UNAM impartiendo asignaturas como Sistema Educativo Nacional y Organismos y Sistemas Internacionales de Educación. Con más de 10 años de experiencia, se ha destacado en la concreción del currículo y la didáctica en instituciones a nivel nacional e internacional para la educación media y superior. Ha participado como profesor invitado en diversas instituciones de educación superior nacionales y colaborado en eventos académicos como organizador, tallerista y ponente.

¹A partir del 18 de enero de 2024, Gemini (también creado por Google AI) sustituyó a Bard. A lo largo del texto se mantendrá el término Bard debido a que esta herramienta se utilizó al momento de redactar el artículo, sobre todo, durante la creación de *prompts* para la generación de rúbricas.