

DOI: <https://doi.org/10.56712/latam.v4i2.1040>

## **El uso de ChatGPT 4.0 para la elaboración de exámenes: crear el prompt adecuado**

The Use of ChatGPT 4.0 for Test Development: Creating the Right  
Prompt

**Karla Karina Ruiz Mendoza**

[ruiz.karla32@uabc.edu.mx](mailto:ruiz.karla32@uabc.edu.mx)

<https://orcid.org/0000-0001-8978-8364>

Instituto de Investigación y Desarrollo Educativo

Ensenada – México

Artículo recibido: 14 de agosto de 2023. Aceptado para publicación: 26 de agosto de 2023.

Conflictos de Interés: Ninguno que declarar.

### **Resumen**

El objetivo de este estudio fue explorar y determinar las mejores prácticas para formular prompts que permitan a ChatGPT 4.0 generar el diseño de un examen, de forma genérica para pruebas de evaluación. Esto permitiría que los encargados de su diseño simplemente intercambien detalles de la plantilla, personalizándola para su uso específico, agilizando así los procedimientos tradicionales de creación de exámenes. Los hallazgos mostraron que el ChatGPT-4 pudo responder satisfactoriamente a las solicitudes, pero exhibió ciertas irregularidades y redundancias en sesiones de interacción más extensas. Aunque los resultados indican que la inteligencia artificial puede desempeñar un papel relevante en la estandarización de la creación de exámenes, también enfatizan la importancia de planificar la duración de las sesiones de interacción con el ChatGPT-4 para evitar incongruencias.

*Palabras clave:* ChatGPT-4, especificidad, sistematización, generación de ítems, examen educativo

### **Abstract**

The objective of this study was to explore and determine the best practices for formulating prompts that allow ChatGPT 4.0 to generate a generic design for evaluation tests. This would enable those in charge of their design to simply exchange details of the template, personalizing it for their specific use, thereby streamlining traditional exam creation procedures. The findings showed that ChatGPT-4 was able to respond satisfactorily to requests, but exhibited certain irregularities and redundancies in longer interaction sessions. Although the results indicate that artificial intelligence can play a significant role in standardizing the creation of exams, they also emphasize the importance of planning the duration of interaction sessions with ChatGPT-4 to avoid inconsistencies.

*Keywords:* ChatGPT-4, specificity, systematization, item generation, assessment

Todo el contenido de LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades, publicados en este sitio está disponibles bajo Licencia Creative Commons .



Como citar: Ruiz Mendoza, K. K. (2023). El uso del ChatGPT 4.0 para la elaboración de exámenes: crear el prompt adecuado. *LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades* 4(2), 6142–6157. <https://doi.org/10.56712/latam.v4i2.1040>

## INTRODUCCIÓN

La inteligencia artificial (IA) ha sido una fuerza disruptiva en muchos ámbitos, incluyendo la educación y la evaluación educativa. La IA, término acuñado por John McCarthy en 1956, alude a la capacidad de los sistemas informáticos para realizar tareas humanas, como aprender y pensar (Sadiku et al., 2021). A partir de los años 70, la Inteligencia Artificial en Educación (IAED) ha comenzado a implementarse en la enseñanza y el aprendizaje con el objetivo de optimizar estos procesos y potenciar los logros de los estudiantes de educación superior (Southgate et al., 2019). La IAED busca desarrollar sistemas basados en IA, tales como agentes pedagógicos virtuales, robots de IA y sistemas inteligentes, que faciliten un aprendizaje personalizado y flexible, y que permitan automatizar tareas educativas cotidianas, como la retroalimentación y evaluación (AlFarsi et al., 2021; Saini & Goel, 2019) con el fin de permitir un aprendizaje más personalizado, flexible, inclusivo y atractivo y automatizar tareas de enseñanza rutinarias a través de la evaluación y retroalimentación automáticas (Southgate et al., 2019).

Por otro lado, Ralph W. Tyler, uno de los padres de la evaluación educativa moderna, destacó la importancia de la evaluación sumativa en la toma de decisiones sobre el progreso de los estudiantes en el cumplimiento de metas educativas (Tyler, 1949). La evaluación sumativa es un tipo de evaluación que se lleva a cabo al final de un período de enseñanza o un curso, con el propósito de medir y valorar el nivel de aprendizaje alcanzado por los estudiantes en relación con los objetivos educativos establecidos (Russell & Airasian, 2012; Theal & Franklin, 2010), por ejemplo, al concluir una actividad instruccional de una unidad, curso, o programa académico (Bennett, 2015) y que es sumamente importante para aproximar al estudiantado a la obtención de mejores resultados.

La capacidad de las IA para generar contenido de forma autónoma ha abierto posibilidades para la generación automática de ítems de prueba, lo que podría tener implicaciones significativas para la eficiencia y la diversidad de las evaluaciones (Tlili et al., 2023). En este contexto, se vuelve imperativo explorar y determinar las mejores prácticas para formular prompts que permitan a ChatGPT 4.0 generar ítems de prueba de manera eficiente y precisa. Dicha exploración podría ayudarnos a comprender mejor las posibilidades y limitaciones de esta innovadora herramienta IA en la esfera educativa, ya que esto nos podría permitir contrarrestar, en verdad, contrarrestar la deserción estudiantil o el agotamiento de los profesores y cerrar las brechas de rendimiento entre los estudiantes debido a diferencias individuales o sociales gracias a la IAED. Por ende, hemos establecido el objetivo de explorar y determinar las mejores prácticas para formular prompts que permitan a ChatGPT 4.0 generar el diseño de un examen, de forma genérica, es decir, donde el usuario solo tenga que sustituir de manera eficiente y precisa; y de esta manera encaminarnos a procesos sistematizados en la elaboración de exámenes.

Este análisis permitirá no solo evaluar las afirmaciones críticas, sino también descubrir formas efectivas de aprovechar ChatGPT en el entorno educativo. Por ello, a continuación, se presentan algunos estudios con relación al uso de la IAED así como del ChatGPT y sus diferencias entre las versiones 3.5 y 4.

### Algunos estudios recientes

Actualmente hay limitaciones sobre la cantidad de investigaciones relacionadas al tema de evaluación educativa con el uso de IA, sin embargo, existen algunos trabajos que nos pueden ayudar a elaborar este tipo de investigaciones. Dimitriadou & Lanitis (2023) elaboraron un análisis FODA (Fortalezas, Oportunidades, Debilidades y Amenazas) para analizar el uso de IA en las denominadas SMART Classroom (Aula Inteligente); el acrónimo representa Mostrando (Showing), Gestionable (Manageable), Accesible (Accessible), Interactiva en tiempo real (Real-

time Interactive) y Pruebas (Testing). Desde su perspectiva, el futuro de las aulas inteligentes se enfrenta a desafíos críticos, especialmente en cuanto a la ética y el manejo de datos en los sistemas de IA. Por ende, es esencial establecer y adoptar regulaciones sobre la recopilación y uso de datos para proteger la privacidad.

Por otro lado, Dimitriadou & Lanitis (2023), especifican que se deben abordar los sesgos en la IA, requiriendo que los desarrolladores eliminen cualquier sesgo potencial derivado de sus creencias personales para evitar la discriminación. Para una toma de decisiones más justa y efectiva, los sistemas de IA deben incorporar técnicas de aprendizaje automático con IA explicable. Aunque la recolección de datos, la creación de perfiles de usuarios y el aprendizaje adaptativo pueden mejorar la personalización y efectividad del aprendizaje, es fundamental ponderar cuidadosamente las implicaciones éticas y garantizar la protección de la privacidad de los estudiantes.

Por otro lado, Tlili et al. (2023), en su análisis sobre el uso de ChatGPT, describen diversos escenarios en los que se ve involucrado su uso. Uno de ellos, el quinto, es el diseño estructural de las evaluaciones de aprendizaje. Un diseño bien estructurado es crucial para que los estudiantes comprendan y resuelvan las evaluaciones de manera efectiva, por ende, al usar ChatGPT para diseñar posibles cuestionarios de evaluación de aprendizaje que podrían ayudar a los educadores a preparar sus materiales de enseñanza, aún se observa una inconsistencia en el diseño de la evaluación de aprendizaje, lo que puede complicar más que facilitar las tareas de los profesores.

Además de ello, Tlili et al. (2023), estos diez escenarios exploran las diversas implicaciones y desafíos que presenta ChatGPT en el ámbito educativo. Los problemas van desde la posibilidad de facilitar la trampa en exámenes y trabajos escritos, hasta cuestiones sobre la precisión, equidad y adecuación del contenido de aprendizaje proporcionado. Otros problemas abordados incluyen la inconsistencia en la estructura de las evaluaciones de aprendizaje, la necesidad de maximizar la utilidad del asistente de aprendizaje y la falta de capacidad de ChatGPT para reflejar las emociones o el compromiso de los estudiantes. La honestidad y veracidad de ChatGPT también son áreas de preocupación, especialmente en relación con respuestas incompletas o justificaciones insuficientes. Además, se mencionan preocupaciones sobre la privacidad y cómo se manejan las conversaciones de los usuarios, así como la posibilidad de que ChatGPT proporcione información incorrecta o excesiva.

En este sentido, Gregorcic y Pendrilla (2023), en el ámbito de la física a un nivel universitario, advierten que la habilidad de ChatGPT para proporcionar respuestas articuladas y razonadas puede ser perjudicial si enseña conceptos incorrectos a los estudiantes desprevenidos, destacando la importancia de la educación en el pensamiento crítico. Sin embargo, también se sugiere que ChatGPT puede ser útil en la formación de docentes, ya que puede proporcionar respuestas incorrectas a preguntas conceptuales de física que ayuden a los profesores a reconocer y comprender la argumentación problemática. El hecho de que estas respuestas incorrectas estén bien formuladas obliga a los profesores a comprometerse con el contenido en lugar de centrarse en los aspectos gramaticales o estilísticos.

Por último, Issayeva (2023) exploró la utilidad del ChatGPT para la creación de preguntas de examen en áreas como matemáticas, inglés y ruso. A través de experiencias prácticas, se resaltaron diversas fortalezas de ChatGPT, tales como su cortesía, su habilidad para generar respuestas de manera rápida, y su capacidad de recordar información previa de las conversaciones, aspectos que pueden contribuir a una notable reducción del tiempo requerido para la elaboración de ítems de exámenes. No obstante, también se identificaron ciertas limitaciones, como la generación ocasional de respuestas inexactas, su conocimiento limitado

del mundo después de 2021, y errores en la creación de ítems de examen en las áreas mencionadas. A pesar de estos desafíos, la conclusión fue que el uso de ChatGPT para la creación de ítems de examen puede ser de gran beneficio, dada su capacidad para reducir el tiempo de desarrollo de nuevos ítems, siempre que se le proporcionen preguntas bien estructuradas y la guía adecuada.

Aunado a todo lo anterior, se puede decir que los autores apuntan a que ChatGPT no se comporta como un estudiante desinformado, sino más bien como un interlocutor ignorante y seguro de sí mismo sin una experiencia confiable en la vida real que pueda ser útil para construir una comprensión conceptual de la física (Gregorcic & Pendrill, 2023; Tlili et al., 2023; Issayeva, 2023). Sin embargo, no todo es negativo, estos escenarios son de gran importancia para comprender los aspectos éticos de su uso, y esto no quiere decir que debamos de dejar de utilizar el ChatGPT como una herramienta, sino de comprender y dominar el nivel de demanda cognitiva. Por ello, recomendamos utilizar esta herramienta siempre y cuando se utilice su propio criterio y conocimiento para hacer cada vez más certeras las instrucciones (prompts) a desarrollar.

### Diferencias entre ChatGPT-3.5 y 4

En marzo de 2023 se lanzó la cuarta iteración de ChatGPT: ChatGPT-4. El cual tiene una mejora sustancial en comparación con su predecesor, ChatGPT-3.5. Aunque ambas versiones comparten similitudes, existen diferencias significativas en cuanto a exactitud, creatividad, seguridad, y capacidad de procesamiento que marcan un avance notable en la tecnología de inteligencia artificial. En la Tabla 1 se pueden observar estas diferencias expresadas de forma sencilla, sin embargo, es solo una aproximación para comprender los cambios tan rápidos de esta herramienta.

**Tabla 1**

*Diferencias entre ChatGPT 3.5 y 4*

Característica	ChatGPT-3.5	ChatGPT-4
Multimodalidad	No	Sí
Exactitud	Menos exacto	Más exacto
Creatividad	Alta	Más alta
Capacidad de Colaborar	Menor	Mayor
Sesgos	Presentes	Reducidos pero aún presentes
"Alucinaciones de IA"	Mayor probabilidad	Menor probabilidad
Procesamiento de Imagen	No	Sí
Seguridad	Basada en la moderación	Integrada en el sistema
Contexto	Limitado	Mejorado
Longitud de Texto	Limitada	Mayor capacidad

**Nota:** Elaboración propia a partir de Rodríguez (2023) y Roose (2023).

### METODOLOGÍA

Por lo anterior y con el fin de optimizar los resultados para obtener los prompts generales para el diseño de un examen de principio a fin a partir del uso de ChatGPT, se ha optado por seleccionar a la versión de paga ChatGPT4. Asimismo, esta investigación seguirá un enfoque exploratorio y descriptivo. Un enfoque exploratorio se utiliza cuando un área de investigación es relativamente nueva y se entiende poco (Teijlingen & Hundley, 2001) el cual permite evidenciar cuestiones metodológicas, descubrir posibles problemas técnicos, éticos y más. En este caso, aunque se ha trabajado con ChatGPT y otras IA para la generación de texto, la tarea

específica de utilizar ChatGPT 4.0 para crear ítems de prueba para un examen es un área de investigación que todavía se está explorando.

Por otro lado, un enfoque descriptivo se utiliza cuando el objetivo de la investigación es proporcionar una descripción detallada de la situación o fenómeno bajo estudio (Babbie, 2010). En este estudio, una vez que se haya explorado el uso de ChatGPT 4.0 para generar ítems de prueba, el siguiente paso será describir en detalle las características de los ítems generados, cómo varían en función del prompt dado, y cómo estas variaciones pueden influir en la eficacia del ítem. De esta manera, el enfoque exploratorio nos permitirá descubrir nuevas formas de interactuar con ChatGPT 4.0 y ver cómo responde a diferentes prompts, mientras que el enfoque descriptivo nos permitirá caracterizar y entender a fondo estos hallazgos. En conjunto, estos dos enfoques proporcionarán una visión completa de la capacidad de ChatGPT 4.0 para generar un examen. A continuación, se expresan la serie de pasos que se siguieron para obtener un examen de forma genérica, siguiendo las etapas de construcción de una prueba del INEE (2019):

Primero se le solicitaron las recomendaciones generales para la elaboración de prompts

Después, se le solicitó que definiera un objetivo general donde se pudieran sustituir los conceptos clave.

Una vez teniendo bien pensado qué grado, asignatura y grado evaluaremos, entonces se le pidió elaborar una instrucción para la elaboración de la tabla de especificaciones. En este caso, el ejemplo será sobre primer grado de primaria, área de matemáticas.

Así, como muestra, se le solicitó crear 20 ítems a partir de ella, siendo preguntas de opción múltiple de tres opciones a elegir con una respuesta correcta.

**Figura 1**

*Etapas de construcción de una prueba del INEE*



**Nota:** Tomado de INEE (2019).

## RESULTADOS

A continuación, se muestran los resultados por cada una de las etapas que se realizaron para obtener un examen con el uso de ChatGPT 4 optimizando su uso con sus mismas respuestas y lógica.



#### Las recomendaciones de ChatGPT 4

Para obtener las recomendaciones de obtención de mejores resultados en la elaboración de instrucciones o prompts, se realizaron las siguientes preguntas: ¿Cuáles son las directrices o recomendaciones que los usuarios deben considerar para formular un prompt efectivo al interactuar con ChatGPT? ¿Existe algún método o estrategia preferible para maximizar la precisión y la eficiencia de las respuestas generadas por el modelo? En la Tabla 2 se pueden observar los resultados obtenidos, lo cual evidencia que la especificidad, la claridad, el contexto, el formato solicitado y el feedback son elementos clave para obtener los mejores resultados (Issayeva, 2023).

**Tabla 2**

*Recomendaciones del ChatGPT-4 para obtener un prompt más preciso*

Nº	Recomendación	Descripción
1	Ser específico	Cuanto más específico sea el prompt, más fácil será para ChatGPT entender lo que se le pide. Los prompts específicos reducen la ambigüedad y ayudan a obtener respuestas más precisas.
2	Proporcionar contexto	Si el prompt se basa en información previa, generará una respuesta más precisa. proporcionar ese contexto puede ayudar a ChatGPT a esto es especialmente útil para preguntas de seguimiento.
3	Uso de lenguaje claro y conciso	El uso de un lenguaje sencillo y directo puede ayudar a evitar malentendidos. Los giros idiomáticos, las metáforas y las expresiones coloquiales pueden llevar a respuestas inesperadas.
4	Solicitar formatos específicos si es necesario	Si se necesita una respuesta en un formato específico (como una lista, un párrafo, un poema, etc.), es útil especificarlo en el prompt.
5	Iteración de los prompts	Si la primera respuesta no es satisfactoria, reformular el prompt o proporcionar más detalles puede ayudar a obtener una respuesta más precisa.
6	Recordar los límites de ChatGPT	ChatGPT tiene limitaciones. No puede proporcionar consejos médicos o legales precisos, no puede realizar tareas que requieran conocimiento del mundo real en tiempo real y puede tener problemas para responder preguntas que requieran un juicio ético o moral.
7	Feedback al modelo	Algunas implementaciones de modelos de IA permiten la posibilidad de dar feedback, lo que puede ayudar a mejorar las respuestas a lo largo del tiempo.

**Nota:** Elaboración propia a partir del uso de ChatGPT-4 (OpenAI, 2021).

#### Desarrollar una tabla de especificaciones

Una tabla de especificaciones, también conocida como matriz de prueba o tabla de planificación de pruebas, es una herramienta útil para diseñar exámenes y garantizar que cubran el contenido y las habilidades apropiadas (Wiggins, 2018). En términos simples, es una especie de mapa que define cuántas preguntas se deben incluir para cada tema y tipo de habilidad cognitiva, como bien lo indica el INEE (2019) "En este proceso, se hace una definición exhaustiva de los contenidos y habilidades involucrados, cuidando que no falte ni sobre ningún elemento central, ya que esta especificación será el referente para la construcción del instrumento." (p.11).

A ChatGPT-4 se le solicitó que elaborara una tabla de especificaciones genérica, donde pudiera sustituir conceptos clave como: grado, tema o área y habilidades cognitivas. Con esta simple instrucción generó lo siguiente (OpenAI, 2021):

Aquí hay un ejemplo general de cómo podrías pedirme que te ayude a crear una tabla de especificaciones:

Usuario: "Por favor, genera una tabla de especificaciones para un examen de ciencias de octavo grado que cubre los temas de la estructura atómica, la tabla periódica, los enlaces químicos y las reacciones químicas. Quiero que el examen evalúa tanto el conocimiento factual como el razonamiento aplicado."

A partir de esta solicitud, podría generar una tabla que identifica los temas (estructura atómica, tabla periódica, enlaces químicos, reacciones químicas), los tipos de habilidades cognitivas (conocimiento, comprensión, aplicación), y el número y tipo de preguntas para cada cruce de tema y habilidad.

Al observar que no había desarrollado del todo la petición, se le volvió a describir que a eso que había generado quitara los conceptos que se pueden sustituir para obtener resultados sistematizados y esto fue lo que resultó (OpenAI, 2021):

Usuario: "Por favor, genera una tabla de especificaciones para un examen de [nivel y asignatura] que cubre los temas de [tema 1, tema 2, tema 3, tema 4, etc.]. Quiero que el examen evalúa tanto [habilidad cognitiva 1, habilidad cognitiva 2, etc.]."

Donde:

[nivel y asignatura] puede ser cualquier combinación de nivel educativo y área de estudio (por ejemplo, "matemáticas de segundo grado", "ciencias sociales de la escuela secundaria", "biología universitaria", etc.).

[tema 1, tema 2, tema 3, tema 4, etc.] son los temas o unidades específicas que quieres que el examen cubra. Por ejemplo, en un curso de historia de los Estados Unidos, podrías tener temas como "la Guerra Civil", "la era de la Reconstrucción", "la Gran Depresión", etc.

[habilidad cognitiva 1, habilidad cognitiva 2, etc.] son las habilidades cognitivas que deseas que los estudiantes demuestran en el examen. Algunos ejemplos pueden ser "recuerdo de hechos", "comprensión", "análisis", "evaluación", etc. Recuerda que estas habilidades deben reflejar los objetivos de aprendizaje del curso y deben estar alineadas con los niveles de dificultad y complejidad apropiados para el nivel de los estudiantes.

### **Ejemplo aplicado para desarrollar una tabla de especificaciones**

Por ello, a lo que se prosiguió fue a la sustitución de los conceptos para saber si cumpliría con lo necesario. En este sentido se le pidió lo siguiente: Por favor, genera una tabla de especificaciones para un examen de [primer grado de primaria de matemáticas] que cubre los temas de [representación de números, adición y sustracción]. Quiero que el examen evalúe tanto [aplicación como evaluación] a partir de la Taxonomía de [Anderson y Krathwohl].

No se omitieron los corchetes con el fin de dejar visible lo sustituido. El resultado fue el siguiente, donde también se incluye la Tabla 3 ya que tiene la posibilidad de desarrollar tablas con una mejor precisión que la versión 3.5 (OpenAI, 2021):

A continuación, te presento una tabla de especificaciones para un examen de primer grado de primaria de matemáticas, que cubre los temas de representación de números, adición y sustracción. El examen evaluará las habilidades específicas dentro de cada dominio y el porcentaje aproximado de preguntas que se destinarán a cada habilidad en el examen.



## Examen de Matemáticas - Primer Grado de Primaria - Tabla de Especificaciones

**Tabla 3**

*Ejemplo de una tabla de especificaciones de matemática de primer grado de primaria*

Dominio	Habilidades Evaluadas	Nivel de Evaluación	Porcentaje de Preguntas
Representación de números	- Identificar y escribir números del 1 al 20	Aplicación	40%
	- Relacionar números con su representación en objetos	Evaluación	
	- Ordenar números de menor a mayor	Evaluación	
Adición	- Resolver sumas simples con números del 1 al 10	Aplicación	40%
	- Entender el concepto de suma como adición de objetos	Evaluación	
	- Aplicar la propiedad conmutativa de la adición	Evaluación	
Sustracción	- Resolver restas con números del 1 al 10	Aplicación	20%
	- Entender el concepto de resta como "quitar objetos"	Evaluación	
	- Reconocer la relación entre suma y resta	Evaluación	
Evaluación general	- Resolver problemas que involucren adición y sustracción	Evaluación	-

**Nota:** El porcentaje asignado a cada habilidad es aproximado y puede variar según la longitud del examen y los criterios de evaluación específicos del cuerpo docente encargado del examen.

**Fuente:** Elaboración propia a partir del uso de ChatGPT-4 (OpenAI, 2021).

Lo interesante de la Tabla 3 es que añade una evaluación general, realizando una inferencia de la necesidad de incurrir a una evaluación general. Posteriormente se le solicitó omitir para poder crear el examen final. Otra de las situaciones es que respondió que la taxonomía seleccionada define así a la aplicación y a la evaluación:

La Taxonomía de Anderson y Krathwohl es una revisión de la Taxonomía de Bloom, que clasifica las habilidades cognitivas en diferentes niveles. Los niveles más relevantes para este examen son:

**Aplicación:** Utilizar el conocimiento en situaciones concretas.

**Evaluación:** Hacer juicios y valoraciones basados en criterios y estándares establecidos.

### Elaboración de prompts para ítems

Una vez que se obtuvo la tabla de especificaciones, se procedió a preguntar la mejor manera para darle una indicación o prompt para elaborar ítems a partir de dicha tabla a lo cual indicó lo siguiente (OpenAI, 2021):

Un "prompt" claro y detallado podría incluir:

Dominio específico: Indica el dominio o tema de matemáticas sobre el cual deseas que se centre el ítem, como representación de números, adición o sustracción.

**Nivel de dificultad:** Especifica si deseas un ítem de fácil, moderado o alto nivel de dificultad. Esto puede depender del grado de los estudiantes, los objetivos de evaluación y el tiempo disponible para responder cada ítem.

**Habilidad a evaluar:** Describe la habilidad o concepto particular que deseas evaluar en el ítem. Por ejemplo, puede ser la identificación de números, el uso de la propiedad conmutativa de la adición o la comprensión del concepto de resta.

**Formato del ítem:** Indica si deseas un ítem de opción múltiple, respuesta corta, de completar el espacio en blanco, problemas matemáticos, o cualquier otro formato que consideres adecuado para evaluar la habilidad específica.

**Contexto o enunciado del ítem:** Proporciona un contexto o enunciado breve para el ítem que permita al estudiante aplicar la habilidad o concepto en un escenario concreto.

Por ejemplo, podrías decir:

- Dominio: Adición
- Nivel de dificultad: Moderado
- Habilidad a evaluar: Aplicar la propiedad conmutativa de la adición
- Formato del ítem: Opción múltiple

Sin embargo, después de intentar diversas formas de dar la indicación se optó por generar una sola indicación que contuviera lo siguiente: Elabora 20 ítems a partir de la tabla de especificaciones creada, considerando el dominio, la habilidad, el nivel de dificultad. Siendo ítems con tres opciones múltiples; señala la respuesta correcta. Esto fue lo que se obtuvo (OpenAI, 2021):

Elabora 20 ítems de opción múltiple a partir de la tabla de especificaciones creada. Cada ítem debe evaluar las habilidades y conceptos específicos de matemáticas en primer grado de primaria, considerando el dominio, la habilidad y el nivel de dificultad indicados en la tabla. Selecciona tres opciones para cada ítem y señala claramente la respuesta correcta. Asegúrate de proporcionar enunciados claros y concisos que permitan a los estudiantes demostrar su conocimiento y comprensión de los temas de representación de números, adición y sustracción.

Con el fin de generar una sistematización se podría sustituir por:

Elabora [] ítems de opción múltiple a partir de la tabla de especificaciones creada. Cada ítem debe evaluar las habilidades y conceptos específicos de [] en [] grado de [], considerando el dominio, la habilidad y el nivel de dificultad indicados en la tabla. Selecciona [] opciones para cada ítem y señala claramente la respuesta correcta. Asegúrate de proporcionar enunciados claros y concisos que permitan a los estudiantes demostrar su conocimiento y comprensión de los temas de representación de números, adición y sustracción.

Un ejemplo de examen a través de ChatGPT-4

Así, se obtuvieron un total de 20 ítems con fines de ejemplificar el tipo de preguntas que genera ChatGPT-4 con las indicaciones que se le dieron paso a paso según su propia lógica. Por otro lado, sí concuerda con el porcentaje dado a cada dominio: representación de números 40%, adición 40% y sustracción 20%. Cada uno de estos ítems, como se puede observar en la Tabla 4, se alinea con un dominio y una habilidad a evaluar, ya que así lo expresó el chat. Por otro lado, las opciones de respuesta fueron precisas, no hubo errores en cuestión de estructura.

No obstante, si se sigue utilizando una línea de chat por más de 4 horas se pueden comenzar a observar errores en la conversación, como repeticiones de otras instrucciones, como si ya no

podiera cambiar su forma de procesar la información si ya le disté una instrucción muy específica. Incluso se podría llegar a insistirle que no es lo que buscas y pedirle que olvide esas instrucciones; en todo caso es más conveniente abrir otro chat.

**Tabla 4**

*Ítems elaborados por ChatGPT-4*

Ítem	Dominio	Habilidad a Evaluar	Preguntas	Opciones	Respuesta
1	Representación de números	Identificar y escribir números del 1 al 20	Escribe el número que sigue después de 12.	a) 13, b) 14, c) 15	a) 13
2	Adición	Resolver sumas simples con números del 1 al 10	¿Cuánto es $4 + 2$ ?	a) 5, b) 6, c) 7	b) 6
3	Representación de números	Relacionar números con su representación en objetos	¿Cuál de las siguientes opciones representa la cantidad de peces en la imagen?	a) 3, b) 5, c) 8	c) 8
4	Adición	Resolver sumas simples con números del 1 al 10	¿Cuánto es $7 + 2$ ?	a) 8, b) 9, c) 10	b) 9
5	Representación de números	Ordenar números de menor a mayor	Ordena los siguientes números de menor a mayor: 14, 8, 12.	a) 8, 12, 14, b) 12, 8, 14, c) 8, 14, 12	a) 8, 12, 14
6	Sustracción	Resolver restas con números del 1 al 10	¿Cuánto es $6 - 3$ ?	a) 2, b) 3, c) 4	b) 3
7	Adición	Entender el concepto de suma como adición de objetos	Si tienes 4 manzanas y luego te dan 3 manzanas más, ¿cuántas manzanas tienes en total?	a) 6, b) 7, c) 8	c) 8
8	Sustracción	Entender el concepto de resta como "quitar objetos"	Si tienes 5 globos y das 2 globos a tu amigo, ¿cuántos globos te quedan?	a) 3, b) 4, c) 5	a) 3
9	Representación de números	Identificar y escribir números del 1 al 20	¿Cuál de las siguientes opciones representa el número 19?	a) 9, b) 15, c) 19	c) 19
10	Adición	Aplicar la propiedad conmutativa de la adición	Si tienes 2 pelotas azules y 5 pelotas rojas, ¿cuál de las siguientes operaciones es igual a la suma de todas las pelotas?	a) $2 + 5$ , b) $5 + 2$ , c) $2 - 5$	b) $5 + 2$

11	Adición	Resolver sumas simples con números del 1 al 10	¿Cuánto es $9 + 1$ ?	a) 8, b) 9, c) 10	c) 10
12	Representación de números	Ordenar números de menor a mayor	Ordena los siguientes números de menor a mayor: 6, 3, 9.	a) 3, 6, 9, b) 6, 3, 9, c) 3, 9, 6	a) 3, 6, 9
13	Adición	Resolver sumas simples con números del 1 al 10	¿Cuánto es $5 + 4$ ?	a) 7, b) 8, c) 9	c) 9
14	Sustracción	Resolver restas con números del 1 al 10	¿Cuánto es $10 - 2$ ?	a) 7, b) 8, c) 9	b) 8
15	Representación de números	Identificar y escribir números del 1 al 20	¿Cuál de los siguientes números es el número que sigue después de 16?	a) 17, b) 18, c) 19	a) 17
16	Adición	Resolver sumas simples con números del 1 al 10	¿Cuánto es $3 + 1$ ?	a) 2, b) 3, c) 4	c) 4
17	Representación de números	Ordenar números de menor a mayor	Ordena los siguientes números de menor a mayor: 2, 7, 5.	a) 2, 5, 7, b) 5, 2, 7, c) 2, 7, 5	a) 2, 5, 7
18	Sustracción	Resolver restas con números del 1 al 10	¿Cuánto es $7 - 2$ ?	a) 4, b) 5, c) 6	c) 6
19	Adición	Resolver sumas simples con números del 1 al 10	¿Cuánto es $2 + 3$ ?	a) 4, b) 5, c) 6	b) 5
20	Representación de números	Relacionar números con su representación en objetos	¿Cuál de las siguientes opciones representa la cantidad de manzanas en la imagen?	a) 4, b) 6, c) 8	b) 6

**Nota:** Elaboración propia a partir del uso de ChatGPT-4 (OpenAI, 2021).

Finalmente, se le solicitó al ChatGPT-4 que describiera el proceso de elaboración para este examen, dando una clara seguridad sobre su propio sistema "La flexibilidad y capacidad de razonamiento de ChatGPT-4 permitieron adaptar las indicaciones y la redacción de cada ítem para crear preguntas claras, comprensibles y adecuadas para el nivel de los estudiantes." (OpenAI, 2021). Con toda esta información, podemos desarrollar algunas conclusiones.

## DISCUSIÓN

Como se ha ilustrado en esta discusión, la IA ha sido una fuerza disruptiva en muchos ámbitos, incluyendo la educación y la evaluación educativa. Su implementación en estos campos ha sido motivada por la promesa de optimizar procesos y potenciar los logros de los estudiantes. Sin embargo, a pesar de sus potenciales beneficios, también se han levantado importantes

preocupaciones, particularmente en relación a la ética y el manejo de datos, el posible sesgo en los sistemas de IA y las implicaciones de estas tecnologías para la privacidad de los estudiantes (Dimitriadou & Lanitis, 2023).

En particular, el uso de ChatGPT para la generación automática de ítems de prueba presenta tanto oportunidades como desafíos. Por un lado, la capacidad de generar contenido de forma autónoma puede conducir a una mayor eficiencia y diversidad en las evaluaciones (Tilli et al., 2023). Sin embargo, también se han identificado problemas significativos, como la posibilidad de facilitar la trampa en exámenes y trabajos escritos, inconsistencias en el diseño de evaluaciones de aprendizaje y preocupaciones sobre la precisión, equidad y adecuación del contenido de aprendizaje proporcionado (Tilli et al., 2023). Además, se mencionan preocupaciones sobre la privacidad y cómo se manejan las conversaciones de los usuarios, así como la posibilidad de que ChatGPT proporcione información incorrecta o excesiva.

Aun así, algunos estudios también han señalado que ChatGPT puede ser de gran beneficio, particularmente en la reducción del tiempo requerido para la elaboración de ítems de exámenes, siempre que se le proporcionen preguntas bien estructuradas y la guía adecuada (Issayeva, 2023).

A la luz de estas deducciones, es importante seguir explorando y determinando las mejores prácticas para utilizar ChatGPT y otras herramientas de IA en la educación. Esto incluye no solo evaluar las afirmaciones críticas, sino también descubrir formas efectivas de aprovechar estas herramientas en el entorno educativo. Asimismo, es esencial abordar los desafíos éticos y regulatorios que presentan estas tecnologías, y garantizar que su implementación se realice de una manera que proteja los derechos e intereses de los estudiantes.

## CONCLUSIÓN

La elaboración de exámenes es una tarea fundamental pero desafiante para los profesores. Aquí es donde ChatGPT-4 puede ser una valiosa herramienta de apoyo. Basado en el análisis de las interacciones con el sistema y los resultados obtenidos, se pueden establecer algunas conclusiones:

**Especificidad:** La claridad y la especificidad son cruciales para obtener los mejores resultados de ChatGPT-4. Los usuarios deben formular prompts detallados y claros que indiquen con precisión lo que quieren lograr.

**Sistematización:** El sistema demostró ser capaz de crear una tabla de especificaciones a partir de una descripción generalizada, permitiendo la sustitución de conceptos clave para adaptarse a diferentes necesidades. Esta característica podría ser útil para sistematizar la elaboración de exámenes.

**Generación de ítems:** La capacidad de ChatGPT-4 para generar ítems a partir de las especificaciones también es notable. A través de un prompt bien estructurado, el sistema generó preguntas alineadas con los dominios y habilidades especificados. Esto puede ayudar a los profesores a diseñar evaluaciones de manera más eficiente.

**Límite de tiempo:** Durante las pruebas, se observó que, si se utiliza una línea de chat durante más de 4 horas, el sistema podría comenzar a repetir instrucciones o a cometer errores. Esto puede indicar que, para obtener los mejores resultados, podría ser más efectivo realizar sesiones más cortas o iniciar una nueva conversación después de cierto tiempo.


**Sistematización:** ChatGPT-4 demostró capacidad para adaptar indicaciones y redactar ítems de examen adecuados para diferentes niveles de estudiantes.

Aún queda mucho por hacer para diseñar mejores sistemas, pero apoyarse en estas herramientas puede ser funcional para ciertos procesos, sobre todo, en los que el investigador o diseñador se encuentra cansado. No obstante, también ello puede llevar a prácticas donde se podrá exigir más rendimiento a estos actores. Por último, este artículo forma parte de una serie de artículos en los cuales se explorarán los límites del uso de la herramienta en los procesos de elaboración de exámenes o pruebas.



## REFERENCIAS

- AlFarsi, G., Tawafak, R. M., ElDow, A., Malik, S. I., Jabbar, J., & Al Sideiri, A. (2021). Smart classroom technology in artificial intelligence: A review paper. In International Conference on Culture Heritage, Education, Sustainable Tourism, and Innovation Technologies (pp. 229–235). CESIT.
- Dimitriadou, E., & Lanitis, A. (2023). A critical evaluation, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learning Environments*, 10(12). <https://doi.org/10.1186/s40561-023-00231-3>
- Gregorcic, B., & Pendrill, A. (2023). ChatGPT and the frustrated Socrates. *Physics Education*, 58(3). <https://doi.org/10.1088/1361-6552/acc299>
- Instituto Nacional para la Evaluación de la Educación (INEE). (2019). Cuadernillo Técnico de Evaluación Educativa 4. Desarrollo de instrumentos de evaluación: pruebas. <https://www.inee.edu.mx/wp-content/uploads/2019/08/P2A354.pdf>
- Issayeva, L. (2023). What ChatGPT Brings Us For Item Writing. *Assess*. <https://assess.com/chatgpt-item-writing/>
- OpenAI (2021). ChatGPT [Language model]. <https://openai.com>. Acceso el [31 de julio de 2023].
- Rodríguez, E. (2023). GPT-4, guía a fondo: qué es, cómo funciona y qué novedades integra con respecto a ChatGPT 3.5. GenBeta. <https://www.genbeta.com/a-fondo/gpt-4-guia-a-fondo-que-como-funcionara-cuando-se-lanzara-que-novedades-integrara-respecto-a-chatgpt-3-5>
- Roose, K. (2023). How ChatGPT kicked off an A.I. arms race. *The New York Times*. <https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html>
- Russell, M., & Airasian, P. (2012). *Classroom Assessment: Concepts and applications*. Mc Graw Hill.
- Saini, M. K., & Goel, N. (2019). How smart are smart classrooms? A review of smart classroom technologies. *ACM Computing Surveys*, 52(6), 1–28. <https://doi.org/10.1145/3347448>
- Southgate, E., Blackmore, K., Pieschl, S., Grimes, S., McGuire, J., & Smithers, K. (2019). *Artificial intelligence and emerging technologies in schools*. Newcastle: University of Newcastle, Australia.
- Teijlingen, V., & Hundley, V. (2001). The importance of pilot studies. *Social Research Update*, 35(1), 1–4.
- Theall, M., & Franklin, J. L. (2010). Assessing Teaching Practices and Effectiveness for Formative Purposes. In K. J. Gillespie & D. L. Robertson (Eds.), *A Guide to Faculty Development*. Jossey Bass.
- Tlili, A., Shehata, B., Adarkwah, M. A., et al. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(15). <https://doi.org/10.1186/s40561-023-00237-x>
- Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. University of Chicago Press.
- Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. Jossey-Bass.

Todo el contenido de **LATAM Revista Latinoamericana de Ciencias Sociales y Humanidades**, publicados en este sitio está disponibles bajo Licencia [Creative Commons](#) .