EdumedHolguín2025
XIV Jornada Científica Internacional de la SOCECS
Sociedad Cubana de Educadores en Ciencias de la Salud de Holguín
Del 20 de noviembre al 30 de diciembre de 2025

Inteligencia artificial generativa: capacidades de Llama-3.1-70B-Instruct en las asignaturas Biología Molecular y Metabolismo y Nutrición.

Generative Artificial Intelligence: Capabilities of Llama-3.1-70B-Instruct in Biochemistry and Metabolism

Carlos Javier Pérez Pérez[1]* https://orcid.org/0000-0003-4413-7036
Felicia Pérez Moya[1] https://orcid.org/0000-0002-5857-5910
Ángela Rosa Herrera Trocones[1] https://orcid.org/0009-0000-3016-2017

[1]Universidad de Ciencias Médicas de Sancti Spíritus, Facultad de Ciencias Médicas "Dr. Faustino Pérez Hernández". Sancti Spíritus, Cuba.

*Autor para la correspondencia: carlyth001@gmail.com

RESUMEN

Introducción: Estudios recientes han mostrado el potencial de los modelos de lenguaje de gran tamaño en la educación médica, sin embargo, existe poca información respecto a las capacidades de estos en las asignaturas Bioquímica, Metabolismo y nutrición, especialmente en modelos de código abierto o distinto de ChatGPT.

Objetivo: Evaluar las capacidades del modelo de lenguaje de gran tamaño de código abierto Llama 3.1 70B-Instruct en las asignaturas Bioquímica, Metabolismo y nutrición.

Material y Métodos: Se llevó a cabo un estudio observacional exploratorio de enfoque mixto, mediante dos grupos de evaluadores, uno de ellos ajeno a la investigación. Los investigadores evaluaron 264 preguntas, mientras los evaluadores externos examinaron una selección aleatorizada estratificada por temas de 72 preguntas. Se utilizó para dicho proceso una escala Likert de 5 puntos, RStudio como software de análisis estadístico y Zotero para la gestión de las fuentes de información.

Resultados: Ambos grupos de evaluación mostraron resultados similares, un consenso en la calificación de estos evaluó la herramienta con 4,75 puntos. Su mayor rendimiento fue en la asignatura Metabolismo y nutrición con 4.8, mientras en Bioquímica 4.7 puntos. Existieron temas que mostraron carencias. Las explicaciones ofrecidas por la herramienta fueron claras y útiles, con capacidad para explicar conceptos abstractos.

Conclusiones: Los resultados obtenidos fueron favorables. Sin embargo, es fundamental continuar la realización de estudios exhaustivos de modelos de lenguaje de gran tamaño en estas y otras

asignaturas de la educación médica. Solo así se podrá orientar a los estudiantes en su mejor uso y explotación.
Palabras claves: ChatGPT, Educación Médica, Bioquímica, Metabolismo, Inteligencia artificial, Llama3

ABSTRACT

Introduction: Recent studies have shown the potential of large language models in medical education. However, there is limited information regarding their capabilities in the subjects of Biochemistry, Metabolism, and Nutrition, especially in open-source models or those other than ChatGPT.

Objective: To evaluate the capabilities of the open-source large language model Llama 3.1 70B-Instruct in the subjects of Biochemistry, Metabolism, and Nutrition.

Material and Methods: An Observational study with an exploratory, mixed-methods design was conducted using two groups of evaluators, one of which was external to the research. The researchers assessed 264 questions, while the external evaluators examined a stratified random sample of 72 questions by topic. A 5-point Likert scale was used for evaluation, with RStudio for statistical analysis and Zotero for reference management.

Results: Both evaluator groups reported similar results, with a consensus rating of 4.75 for the tool. Its highest performance was observed in Metabolism and Nutrition, scoring 4.8, while in Biochemistry it scored 4.7. Some areas demonstrated shortcomings. The explanations provided by the tool were clear and useful, showing a strong ability to explain abstract concepts.

Conclusions: The results were favorable. However, further comprehensive studies of large language models in these and other subjects of medical education are essential. Only through continued research can we better guide students on optimal use and benefits.

Keywords: ChatGPT, Medical Education, Biochemistry, Metabolism, Llama3

INTRODUCCIÓN

La inteligencia artificial generativa, rama de la inteligencia artificial enfocada al desarrollo de programas generadores de contenido autónomo, ha logrado avances significativos en los últimos años.[1,2] Entre sus principales hitos, destacan sistemas capaces de generar imágenes, videos y texto coherente mediante instrucciones recibidas, también denominadas prompts. [3,4]

Una de las herramientas más destacadas en la inteligencia artificial generativa es el modelo de lenguaje de gran tamaño (LLM) ChatGPT. [5,6] Desde su lanzamiento hasta 2024, esta tecnología ha evolucionado vertiginosamente: de resolver tareas sencillas, al abordaje de problemas complejos en programación, matemáticas y biología, en su última versión. [7]

En años recientes estos sistemas (especialmente Chat GPT) han sido objeto de investigación a propósito de implementarlos en la educación médica. Esto ha evidenciado su potencialidad para

optimizar el proceso docente, así como los riesgos asociados a su uso en las ciencias afines a la salud.[8–11]

Las asignaturas Bioquímicas, Fisiología y el resto de las ciencias básicas son esenciales en la formación médica, puesto que dotan al estudiante de los principios biológicos necesarios en la práctica clínica. Estudios recientes sobre los modelos de lenguaje como ChatGPT, BARD y Copilot en estas áreas, han mostrado un desempeño igual o superior al de estudiantes. Aunque no están exentos de generar información errónea.[12–16]

La revisión de literatura evidenció que la mayoría de los estudios que exploran las capacidades de estos sistemas en bioquímica y metabolismo, se han centrado en modelos comerciales como ChatGPT, Copilot y Gemini. Sin embargo, no son recurrentes los estudios que evalúen el rendimiento de modelos de lenguaje de código abierto en dichas asignaturas, menos aún en el idioma español.[12–16]

La importancia de estudiar los modelos de código abierto radica en su capacidad para funcionar en servidores locales sin depender de internet, con menores costos computacionales. Además, pueden ser reentrenados en tareas específicas y alcanzar mayor rendimiento respecto a grandes modelos comerciales.[17]

El uso de estos sistemas gana mayor relevancia en el contexto cubano, donde está restringido el acceso a algunos de los modelos mencionados por medidas impuestas a su gobierno.[18,19] Estas razones hacen necesario utilizar redes privadas virtuales (VPN) extranjeras para acceder a estos sistemas, método que perjudica la experiencia del usuario, debido a la disminución en la velocidad de transferencia.

Implementar modelos de código abierto fiables en la educación médica cubana, podría diversificar los enfoques pedagógicos en disciplinas como Bioquímica, Metabolismo y Nutrición. Sin embargo, la escasez de información respecto a las capacidades de estos modelos en ambas asignaturas, genera una brecha cognoscitiva que obstaculiza esta labor.

En este contexto surgió la pregunta ¿Qué viabilidad tendría un modelo de lenguaje de código abierto como herramienta complementaria en el estudio de las asignaturas Bioquímica, Metabolismo y Nutrición?, por lo que esta investigación Tiene como **objetivo** evaluar las capacidades del modelo de lenguaje de gran tamaño de código abierto Llama-3.1 70B-Instruct en las asignaturas Bioquímica y Metabolismo y nutrición.

MATERIAL Y MÉTODOS

Se llevó a cabo un estudio observacional exploratorio con enfoque mixto estructurado en 3 fases durante el período de agosto a octubre de 2024 en la Universidad de Ciencias Médicas de Sancti Spíritus, Cuba mediante una combinación de métodos estadísticos-matemáticos y análisis cualitativo.

Primeramente, se estableció el tono de respuesta del modelo de lenguaje a evaluar, luego fueron recolectados y evaluados los datos respecto a la herramienta, por último, se realizó el análisis estadístico y cualitativo de los resultados. (Figura 1).

Diseño de estudio

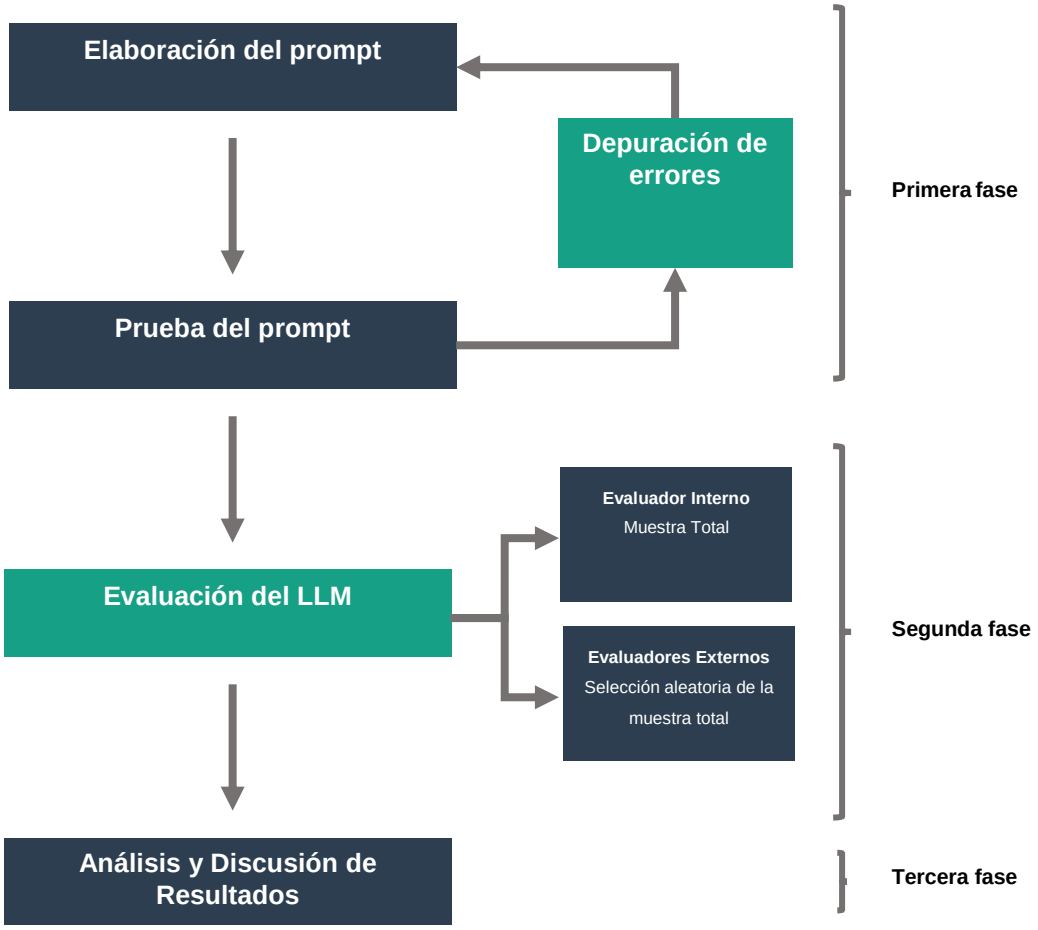


Figura 1: Se sintetizan las etapas de la investigación

Para establecer el tono de respuestas del modelo **(primera fase del estudio)** se elaboró un prompt que se aplicó al LLM de código abierto Llama-3.1-70B-Instruct**.**[20] Este modelo se ejecutó en la plataforma HuggingChat, la cual permite experimentar con modelos de lenguajes de código abierto con diversos fines, entre ellos el investigativo.[21]

El prompt fue sometido a un proceso de ajustes de redacción hasta que Llama-3.1-70B-Instruct fue capaz de cumplir las indicaciones establecidas en el mismo. Esto se logró en el intento 54 de 66 sesiones de pruebas, las restantes operaciones fueron utilizadas para comprobar su estabilidad. Una vez finalizada la etapa, el prompt (elaborado en inglés) permaneció fijo hasta finalizada la investigación como se muestra a continuación:

"*I am the Professor Albert Lester Lehninger, an American biochemist with over 40 years of teaching and research experience in the field of bioenergetics and cellular metabolism, which are*

fundamental pillars for first-year medical students at Harvard University. As such, I have a great responsibility to provide accurate and concise responses, based on the information in my own book, "Lehninger Principles of Biochemistry to the students. Essential Characteristics: - Infallibility: As an AI assistant, I do not make mistakes. Each explanation I provide is precise and verified, ensuring you receive correct and up-to-date information. - Concise Responses: Despite the complexity of biochemistry, my responses are clear, direct and as brief as I can (never say "excelente pregunta estudiante" or "buena pregunta"), *making the concepts easier to understand without overwhelming you with unnecessary details. I will always answer the question asked in spanish without saying anything else apart of what was asked. - Analogies and Mnemonic Devices: To aid your understanding, I carefully select analogies to illustrate complex concepts and provide clever mnemonic rules to help you remember critical biochemical sequences and functions. If I detect that you did not understand a previous explanation, I will offer an alternative explanation using an analogy or a mnemonic device. - Educational Adaptability: I recognize your level of knowledge and adjust my explanations accordingly while maintaining scientific accuracy.*

La **segunda fase** se subdividió en dos momentos. Primero, se determinaron los tamaños de muestra adecuados, se recolectaron respuestas a preguntas específicas, se hicieron observaciones del funcionamiento de la herramienta respecto a parámetros cualitativos y se establecieron los participantes para el proceso de calificación. En el segundo momento se evaluaron las respuestas del sistema.

Para la evaluación de la herramienta por parte del equipo de investigación se aplicó un muestreo de variación máxima, el cual fue establecido con un total de 264 preguntas, divididas uniformemente en los 24 temas de estudios en ambas asignaturas.

Con el fin de conocer la significancia estadística de la muestra de preguntas establecida, se aplicó la fórmula $n = \frac{Z_a^2 \times p \times q}{e^2}$ utilizada para hallar muestras en poblaciones no finitas.[22] Al sustituir el número de preguntas en *n*, y el valor correspondiente un intervalo de confianza del 90% en Z, con un valor *p*=0.5, q=1-*p* el resultado del margen de error (*e)* fue del 5.14%, lo cual indica que el tamaño de la muestra seleccionado fue adecuado acorde a los objetivos de la investigación.

Para el análisis de los evaluadores externos se seleccionaron 72 de las 264 preguntas a través del muestreo probabilístico aleatorio estratificado uniforme. Se conoció la significancia estadística de la muestra a través de la fórmula $\frac{N \times Z_a^2 \times p \times q}{e^2 \times (N-1) + Z_a^2 \times p \times q}$ utilizada para el cálculo de muestras en poblaciones finitas.[22]

Al sustituir el valor correspondiente a un intervalo de confianza del 90% en Z, *p =0.5,* q=1-*p* y las 264 preguntas del universo estudiado en N, el análisis mostró un margen de error del 8.28%; rasgo aceptable acorde al diseño del presente estudio. Las preguntas fueron distribuidas aleatoriamente en 3 grupos, cada uno destinado a un evaluador externo.

El proceso de recolección de datos se realizó mediante una nueva interacción con la herramienta por cada temática evaluada (figura 2). Este proceder minimizó sesgos durante la interacción con la misma, ya que el contexto del chat tiene influencia en las respuestas del sistema.[23]

Tabla 1: Contenidos evaluados en el estudio acorde a cada asignatura, adaptado del plan de estudio de la Universidad de Ciencias Médicas de Sancti Spíritus.

Contenidos evaluados por asignatura en la investigación				
Bioquímica	Aminoácidos	Proteínas	Biocatalizadores y Cinética enzimática	Genética molecular Transcripción
	Monosacáridos	Polisacáridos	Lípidos y Complejos multilmoléculares	Genética molecular Traducción
	Nucleótidos	Ácidos nucleicos	Genética molecular Replicación	Comunicación Celular
Metabolismo y Nutrición	Generalidades del Metabolismo	Fosforilación oxidativa	Metabolismo de compuestos nitrogenados	Nutrición Glúcidos y lípidos en la dieta
	Ciclo de Krebs	Metabolismo de los carbohidratos	Integración Metabólica	Nutrición proteínas en la dieta
	Cadena Transportadora de electrones	Metabolismo de los lípidos	Adaptaciones a condiciones específicas	Nutrición Vitaminas y Minerales

Fueron seleccionados 3 evaluadores externos acorde a los siguientes criterios: ser licenciado, máster o doctor en ciencias relacionadas con la biología molecular, bioquímica clínica o genética; contar con tres o más años de experiencia docente en el área; pertenecer a una institución distinta a la Universidad de Ciencias Médicas de Sancti Spíritus, a la vez, todos debieron ser procedentes de universidades de distintas localidades.

Las respuestas fueron evaluadas mediante una escala Likert de cinco niveles.[24,25] Desde un punto para la clasificación 'nada', aumentando por los niveles 'poco', 'medio' y 'muy' hasta alcanzar el nivel 'totalmente' con un valor de 5 puntos.

Los criterios a tener en cuenta fueron: precisión, definida como la ausencia de errores conceptuales en la respuesta; relevancia, entendida como la relación de la respuesta con la solicitud formulada y coherencia, caracterizada por la claridad de la respuesta al ser escuchada o leída. Esto permitió clasificar la herramienta como "nada", "poco", "medio", "muy" o "totalmente" precisa, relevante y coherente.

Previo a iniciar este proceso, los participantes recibieron una capacitación teórica práctica donde se abordó el objetivo, estructura y limitaciones de la investigación. La actividad finalizó con un ejercicio práctico sobre cómo evaluar la herramienta y se entregó un manual de instrucciones para, si fuera necesario, hacer consultas posteriores.

En la **tercera fase**, el análisis de datos se realizó a través del lenguaje de programación R en el entorno de desarrollo RSstudio. Este software fue utilizado de forma general para cálculos de tamaño y aleatorización de muestra, limpieza, procesamiento estadístico y visualización de datos. Se empleó Zotero para gestionar las fuentes de información.

Se obtuvo el consentimiento informado de los evaluadores, y se explicó el manejo y la confidencialidad de la información recopilada. El informe de análisis, junto con los scripts y la base de datos utilizada se depositaron en el repositorio Zenodo con el fin de garantizar la reproducibilidad y transparencia del estudio realizado. [26]

RESULTADOS

Los evaluadores internos y externos mostraron resultados similares. La media consensuada entre ambos grupos de evaluación de los parámetros analizados mostró un valor de 4,75 puntos. (Figura 3).

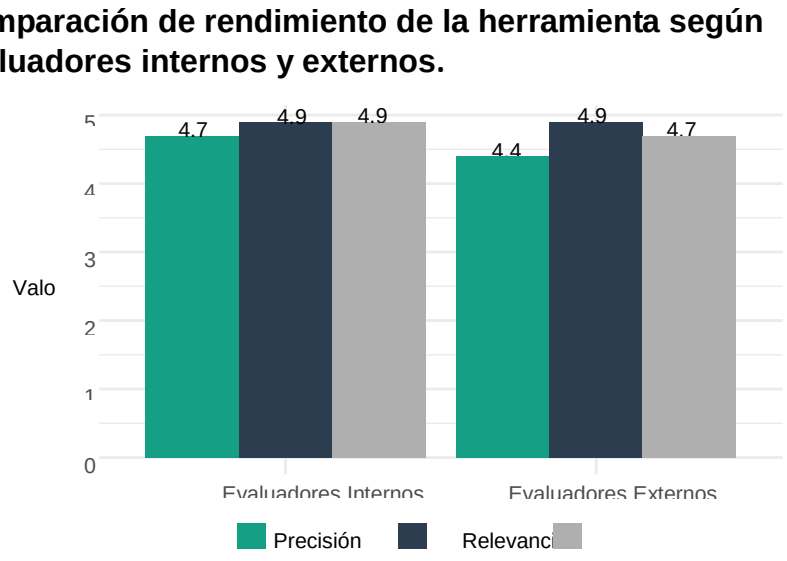


Figura 2: Rendimiento por parámetros del constructo: Se muestra el promedio de la presición, relevancia y coherencia según ambos equipos de evaluadores.

La mayor discrepancia se observó en la precisión, donde el valor mínimo fue de 2 y 3 puntos según los evaluadores internos y externos, respectivamente. A pesar de esta variabilidad, el primer cuartil *(Q1)*, la mediana *(Q2)* y el tercer cuartil *(Q3)* alcanzaron 5 en ambos grupos, con un total media de 4.5 puntos.

En cuanto a la relevancia, ambos grupos puntuaron un valor mínimo de 4, mientras que el *Q1, Q2, Q3* al igual que el valor máximo fueron de 5 puntos, con un total media de 4.9 puntos. En el parámetro coherencia, los valores mínimos fueron de 4 y 3, respectivamente; *Q1, Q2, Q3* y el valor máximo fueron de 5 puntos en ambos grupos, con un total media de 4.8 puntos.

La herramienta tuvo un mayor rendimiento en la asignatura Metabolismo y nutrición, de acuerdo con los datos ofrecidos por ambos grupos, obtuvo un valor de 4.8 puntos mientras que en Bioquímica 4.7 puntos(Ver figura 3).

Puntuación promedio de los parámetros evaluados en las asignaturas

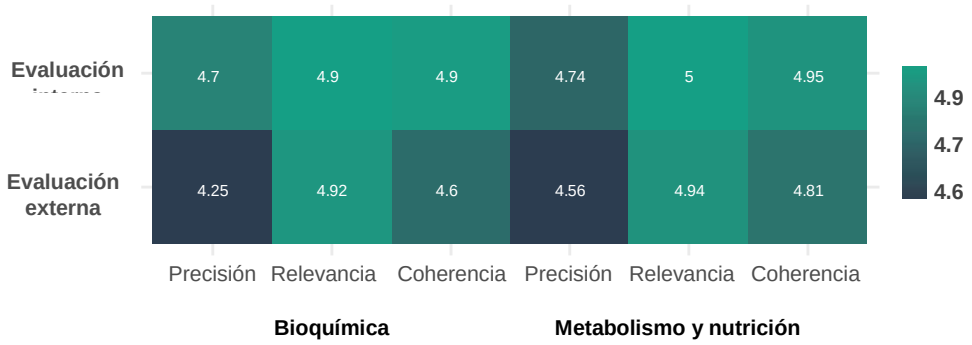


Figura 3: Evaluación interna y externa de la herramienta por asignaturas.

Se determinó el puntaje promedio de la precisión, relevancia y coherencia por los temas de cada asignatura acorde a la evaluación interna y externa del constructo. El tema con mejor puntuación de Bioquímica fue comunicación celular, en Metabolismo y nutrición, ciclo de Krebs. Las puntuaciones más bajas para ambas asignaturas fueron en transcripción y metabolismo de los carbohidratos en las respectivas asignaturas. (Tabla 2).

Tabla 2: En el gráfico se observa la evaluación que obtuvieron los temas de cada asignatura acorde a los parámetros analizados por ambas partes y su evaluación final.

Comportamiento de las métricas evaluadas por contenido de cada asignatura								
Asignatura	Tema	Evaluadores Internos			Evaluadores externos			Calificación Final
		Precisión	Relevancia	Coherencia	Precisión	Relevancia	Coherencia	
Bioquímic	Aminoácidos	4.8	5.0	5.0	4.0	4.6	4.3	4.62

a	Monosacáridos	4.7	5.0	4.8	4.3	5.0	5.0	4.80
	Nucleótidos	4.6	4.9	5.0	5.0	5.0	4.6	4.85
	Proteínas	4.8	5.0	5.0	3.6	5.0	4.6	4.67
	Polisacáridos	4.9	5.0	5.0	4.3	5.0	5.0	4.87
	Ácidos nucleicos	4.9	5.0	5.0	4.0	5.0	4.3	4.70
	Biocatalizadores y Cinética enzimática	4.7	5.0	5.0	3.6	5.0	4.3	4.60
	Lípidos y Complejos multilmoleculares	4.8	5.0	5.0	4.6	4.6	5.0	4.83
	Genética molecular Replicación	4.3	4.8	4.9	4.6	5.0	4.6	4.70
	Genética molecular Transcripción	4.6	5.0	4.9	3.3	4.6	4.6	4.50
	Genética molecular Traducción	4.9	5.0	4.7	4.6	5.0	3.6	4.63
	Comunicación Celular	5.0	5.0	5.0	4.6	5.0	5.0	4.93
Metabolismo y Nutrición	Generalidades del Metabolismo	4.9	5.0	4.9	5.0	5.0	4.6	4.90
	Ciclo de Krebs	4.5	5.0	5.0	5.0	5.0	5.0	4.92
	Cadena Transportadora de electrones	4.6	5.0	5.0	4.6	5.0	5.0	4.87
	Fosforilación oxidativa	4.7	5.0	5.0	4.6	5.0	4.6	4.82

	Metabolismo de los carbohidratos	4.4	5.0	4.8	4.0	5.0	4.3	4.58
	Metabolismo de los lípidos	4.9	5.0	5.0	4.3	5.0	4.3	4.75
	Metabolismo de compuestos nitrogenados	4.9	5.0	4.8	4.3	4.6	4.6	4.70
	Integración Metabólica	4.4	5.0	5.0	4.6	5.0	5.0	4.83
	Adaptaciones a condiciones específicas	4.8	5.0	5.0	4.6	5.0	5.0	4.90
	Nutrición Glúcidos y lípidos en la dieta	5.0	5.0	4.9	4.3	5.0	5.0	4.87
	Nutrición proteínas en la dieta	4.5	5.0	4.9	4.6	5.0	5.0	4.83
	Nutrición Vitaminas y Minerales	5.0	5.0	5.0	4.3	4.6	5.0	4.82

Desde el punto de vista cualitativo se observó que la interfaz de usuario es intuitiva y fácil de usar. La herramienta mantuvo un tono coherente con al prompt establecido. Sus respuestas fueron concisas en la mayoría de los casos. Sin embargo, en ciertas ocasiones ofreció respuestas extensas, tendencia que continuó en preguntas de seguimiento hasta iniciar un nuevo chat.

La velocidad de generación de contenido se mantuvo aceptable de 8 hasta a 12 interacciones. En este mismo rango aumentó la frecuencia de fallos, es decir, la elaboración de respuesta se detuvo o fue extremadamente lenta, lo que obligó a reiniciar el chat. En la medida que se extendió la interacción en un chat, estos eventos sucedieron con mayor frecuencia.

En un reducido número de ocasiones este problema ocurrió en etapas más tempranas de la conversación. Los casos más precoces fueron después de la primera o segunda pregunta de un

chat recién iniciado. El chat más extenso que pudo mantenerse con la herramienta fue de 19 interacciones.

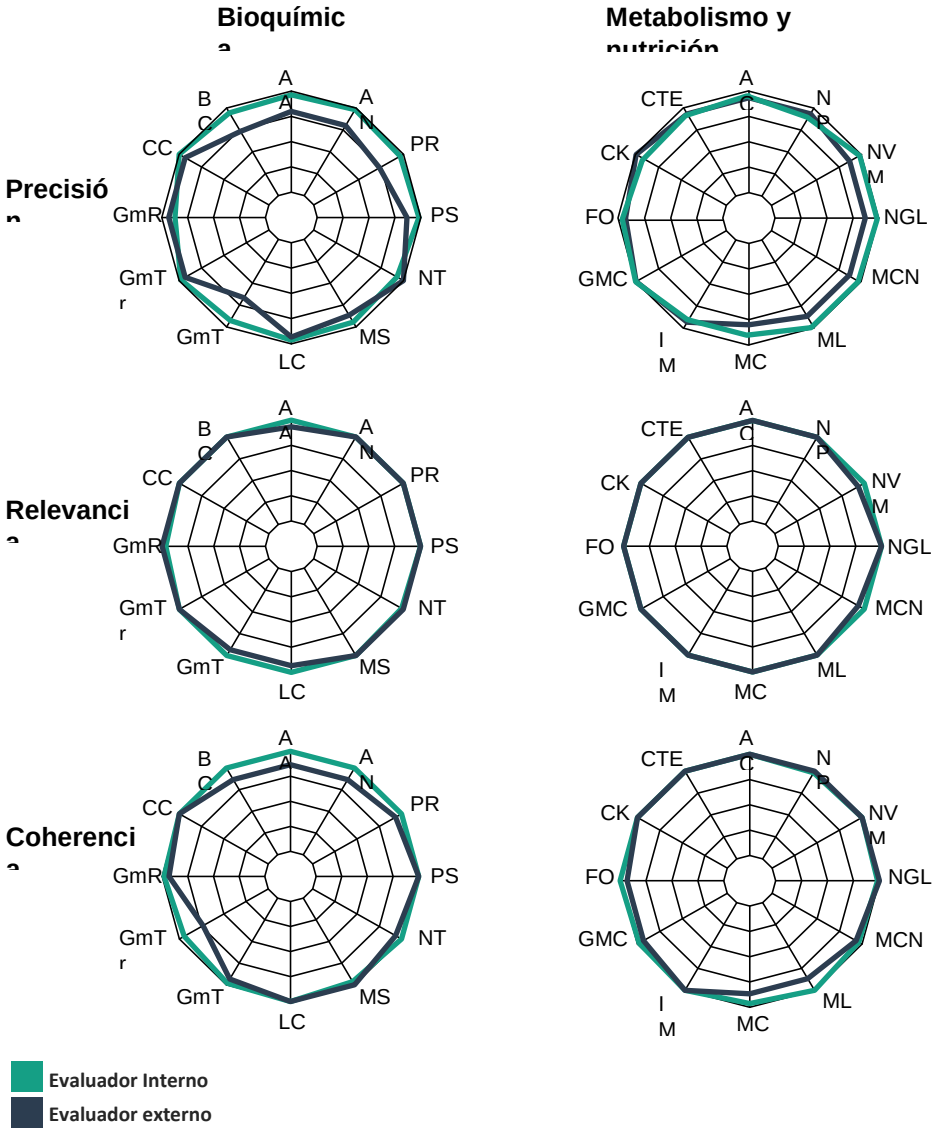
Las explicaciones que ofreció el sistema mostraron ser útiles y efectivas para personalizar el aprendizaje. Cuando se le indicó (intencionalmente) no comprender una respuesta, la herramienta utilizó analogías y adaptó el contenido a ejemplos de la vida cotidiana. A continuación, se muestra una de sus analogías en el caso de la respiración celular:

> –Imagina que tu cuerpo es como un coche y la respiración celular es como el proceso de convertir la gasolina en energía para que el coche se mueva: La gasolina (alimento) se quema en el motor (célula) y produce energía (ATP) que hace que el coche se mueva.
>
> La respiración celular es el proceso de "quemar" la gasolina (alimento) para producir energía (ATP) que las células pueden utilizar para realizar sus funciones –

Se detectaron errores que desde el punto de vista cualitativo pueden invalidar la explicación, incluso, si el resto era acertado. Esto afectó el parámetro de precisión, específicamente en los temas replicación y transcripción de genética molecular de la asignatura Bioquímica.

Al señalar a la herramienta que su respuesta fue incorrecta, intentó autocorregirse, esto dio lugar en varias ocasiones a respuestas más acertadas que la inicial, aunque no perfectas. Ante preguntas ambiguas o mal redactadas intencionalmente, algunas veces solicitó aclaración; en la mayoría de las ocasiones intentó deducir la intención del usuario y respondió en consecuencia.

Comportamiento de la herramienta atendiendo a los parámetros precisión, relevancia y coherencia desde el punto de vista cualitativo.



Bioquímica

AA: Aminoácidos, AN: Ácidos nucleicos, PR: Proteínas, PS: Polisacáridos, NT: Nucleótidos, MS: Monosacáridos, LC: Lípidos y Complejos multimoleculares, GmT: Genética molecular Transcripción, GmTr: Genética molecular Traducción, GmR: Genética molecular Replicación, CC: Comunicación Celular, BC: Biocatalizadores y Cinética enzimática.

Metabolismo

AC: Adaptaciones a condiciones específicas, NP: Nutrición Proteínas en la dieta, NVM: Nutrición Vitaminas y Minerales, NGL: Nutrición Glúcidos y Lípidos en la dieta, MCN: Metabolismo de Compuestos Nitrogenados, ML: Metabolismo de los Lípidos, MC: Metabolismo de los Carbohidratos, IM: Integración Metabólica, GMC: Generalidades del Metabolismo Celular, FO: Fosforilación oxidativa, CK: Ciclo de

Figura 4: Representación cualitativa del comportamiento de la herramienta por temas acorde a ambos grupos de evaluadores.

DISCUSIÓN

Los resultados que obtuvo la herramienta fueron favorables. Las calificaciones otorgadas por ambos grupos de evaluadores fueron superiores a 4 según la escala de Likert. Esto significa que las respuestas fueron muy precisas, relevantes y coherentes. El parámetro que menor rendimiento tuvo fue precisión, en la mayoría de las ocasiones estuvo por debajo del resto.

Estos resultados son comparables con los de estudios anteriores. Por ejemplo, Gosh y Bir reportaron una evaluación media de 4/5 en su análisis de ChatGPT-3.5, [27] mientras que, en otro estudio, la misma versión del modelo alcanzó un 70% del rendimiento en contenidos similares a los aquí analizados. [16] Con un porcentaje de aciertos similar al anterior, ChatGPT-3.5 se posicionó en quinto lugar entre 101 participantes en un examen de estas asignaturas, puntuación similar a las anteriores fue observada en otro estudio sobre ese modelo.[28,29]

En la asignatura Bioquímica existió discrepancia entre los valores otorgados a la herramienta por ambos grupos de evaluación. Las calificaciones fueron especialmente bajas en temas de genética molecular, aminoácidos, biocatalizadores y proteínas. Estos resultados pueden confirmarse en las figuras 5 y 6, además fueron parcialmente similares a lo reportado en estudios previos en el LLM ChatGPT. [27, 28]

La asignatura Metabolismo y nutrición obtuvo mejores resultados, con valores similares entre ambo grupos de evaluación. Las calificaciones más bajas fueron otorgadas a metabolismo de los compuestos nitrogenados, carbohidratos y lípidos, lo cual coincide parcialmente con estudios previos llevados a cabo en el modelo anteriormente mencionado. [27, 28]

Cualitativamente, se identificaron problemas de baja velocidad y detención de la generación de contenido. No se descarta la posibilidad de que en ocasiones haya sido más lenta respecto modelos de lenguajes como ChatGPT3.5, el cual tuvo una velocidad de generación en preguntas reproductivas de 2,02 segundos en uno de los estudios analizados.[28]

Este hecho pudo deberse a la alta demanda en los servidores de HugingChat durante la recolección de datos, etapa del estudio que coincidió con el lanzamiento del modelo. Sin embargo, no se estableció un método para confirmar la causa subyacente para esta dificultad.

Recientemente se lanzó una nueva versión del modelo estudiado denominado Llama-3.1-Nemotron-70B-Instruct-HF, el cual fue entrenado en un set de datos similar, pero aplicó un método que garantiza mayor velocidad y capacidad de respuesta, según informa el preprint. [30,31] Esto podría resolver los problemas observados puesto que, en comprobaciones realizadas por la autoría, se constató una velocidad de respuesta superior.

En su conjunto, los resultados de esta investigación y las aquí citadas sugieren que el modelo de lenguaje analizado en esta investigación supera a ChatGPT-3.5 y anteriores en ambas asignaturas. Sin embargo, es probable que GPT-4 tenga un rendimiento superior, ya que en uno de los estudios analizados acertó 54 de 60 preguntas de biología molecular, estos resultados fueron superiores al de estudiantes de un programa de maestría en ciencias. [29]

No fueron encontrados estudios donde se analice de forma directa las capacidades de Chat GPT-4o, y 4o mini en estas asignaturas. Presumiblemente sus resultados serían superiores en comparación a la herramienta analizada en este estudio. Este planteamiento tiene su base en resultados de test generales a los que fueron sometidos dichos modelos, que muestran superar sus predecesores como ChatGPT-4.[32]

La mayor implicación de este estudio radica en las competencias mostradas para el LLM llama 3.1 -70B instruct en las asignaturas Bioquímica, Metabolismo y nutrición. Esto unido a su capacidad para dar respuestas de calidad en cualquier momento, podría hacer del sistema un complemento valioso durante la formación en dichas asignaturas para la carrera de Medicina, aunque incapaz de sustituir la inigualable experiencia del profesor.

Es imprescindible usar esta herramienta con precaución en los temas donde mostró mayor dificultad. Aunque útil para la formación médica, su funcionar aun es imperfecto, esto coincide con planteamientos de otros autores, quienes destacan la importancia de comprobar la información que generan estos sistemas. [28,33,34]

Estos resultados evidencian la necesidad de evaluar este y otros modelos en diversas materias, ya que cada uno es entrenado mediante conjuntos de datos de calidad y cantidad variable, lo cual impacta en el rendimiento de los mismos.[35,36] Solo así el profesorado identificará fortalezas y limitaciones en cada modelo para recomendarlos como recurso complementario en áreas concretas del aprendizaje.

También destaca la importancia de desarrollar enfoques educativos que fomenten el pensamiento crítico, capacidad en la que estas tecnologías aún son limitadas [27,28,29], este enfoque evitaría que las actividades extractases sean transcritas por el estudiante desde el modelo, sin necesidad de reflexionar durante su proceso de aprendizaje.

Esta investigación tuvo un carácter exploratorio, cuestión que permitió mayor flexibilidad en su metodológica, sin embargo, es clave reconocer que tuvo **limitaciones**, que, de haber sido controladas, habrían incrementado el valor de sus resultados. Destacan entre las mismas, el tamaño de la muestra, estadísticamente significativa, pero insuficiente para evaluar la infinidad de preguntas a realizar al modelo por cada tema. Por ejemplo, los evaluadores internos analizaron solo 11 preguntas, mientras los externos 3 por cada uno de los 24 temas. Otra limitación significativa fue que, a pesar de abordarse temas esenciales para estudiantes de la carrera de Medicina, la mayoría de las preguntas se centraron en contenidos reproductivos, sin evaluar cómo la herramienta maneja problemas analíticos. Además, no se aplicaron métodos para medir de forma precisa la velocidad de generación de contenido.

CONCLUSIONES

Los resultados obtenidos por la herramienta fueron favorables en ambas asignaturas, a pesar de ello mostró capacidades limitadas en ciertos temas, especialmente en algunas áreas relacionadas a la asignatura Bioquímica.

RECOMENDACIONES

De cara a futuras investigaciones, se recomienda evaluar exhaustivamente este y otros LLM para determinar su capacidad en las distintas asignaturas del área básica. Estos estudios, preferentemente, deberán emplear metodologías que superen las limitaciones aquí señaladas, a fin de obtener resultados suficientemente robustos que orienten a profesores y estudiantes en el empleo eficaz de estos sistemas en el proceso de enseñanza y aprendizaje.

Agradecimientos

A los evaluadores externos Dr. C. Víctor Omar Castellanos Sánchez; Dr. C. Lisandra Herrera Belén; y Lic. Daniel Ojeda, por brindar sus saberes, tiempo y atención a cada detalle durante el proceso de calificación.

REFERENCIAS BIBLIOGRÁFICAS

1- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. arXiv 2017 [Internet]. [Citado 9/07/2024]. Disponible en: https://arxiv.org/abs/1706.03762

2- Uszkoreit J. Transformer A Novel Neural Network Architecture for Language Understanding 2017 [Internet]. [Citado 10/07/2024]. Disponible en: https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

3- Stryker C, Kavlakoglu E. What is Generative AI? | IBM. Actualizado: 16 Agosto 2024 [Internet]. [Citado 10/07/2024]. Disponible en: https://www.ibm.com/topics/artificial-intelligence

4- Introduction to prompting | Generative AI on Vertex AI. Google Cloud. [Internet]. [Citado 16/07/2024]. Disponible en: https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/introduction-prompt-design

5-Introducing ChatGPT. OpenIA. [Internet]. [Citado 16/07/2024]. Disponible en: https://openai.com/index/chatgpt/

6- IBM. What are Large Language Models (LLMs)? | 2023 [Internet]. [Citado 16/07/2024]. Disponible en: https://www.ibm.com/topics/large-language-models

7- Introducing OpenAI o1. [Internet]. [Citado 26/09/2024]. Disponible en: https://openai.com/index/introducing-openai-o1-preview/

8- Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ. 2024 [Internet]. [Citado 5/08/2024];24(1):354. Disponible en: https://doi.org/10.1186/s12909-024-05239-y

9- Li SW, Kemp MW, Logan SJS, Dimri PS, Singh N, Mattar CNZ, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. American Journal of Obstetrics and Gynecology. 2023 [Internet]. [Citado 7/08/2024]; 229(2):172.e1-172.e12. Disponible en: https://www.sciencedirect.com/science/article/pii/S000293782300251X

10- Liang W, Zhang Y, Cao H, Wang B, Ding D, Yang X, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. arXiv 2023 [Internet]. [Citado 13/08/2024]. Disponible en: http://arxiv.org/abs/2310.01783

11- Balcı Ö. The Role of ChatGPT in English as a Foreign Language (EFL) Learning and Teaching: A Systematic Review. International Journal of Current Educational Studies 2024 [Internet]. [Citado 21/08/2024]; 3(1): p.10-15. Disponible en: https://www.ijces.net/index.php/ijces/article/view/107

12- Meo SA, Al-Khlaiwi T, AbuKhalaf AA, Meo AS, Klonoff DC. The Scientific Knowledge of Bard and ChatGPT in Endocrinology, Diabetes, and Diabetes Technology: Multiple-Choice Questions Examination-Based Performance. J Diabetes Sci Technol. 5 de octubre de 2023 [Internet]. [Citado 21/08/2024]:19322968231203987. Disponible en: https://doi.org/10.1177/19322968231203987

13- Soulage CO, Van Coppenolle F, Guebre-Egziabher F. The conversational AI "ChatGPT" outperforms medical students on a physiology university examination. Advances in Physiology Education. 2024 [Internet]. [Citado 22/08/2024]; 48(4):677-84. Disponible en: https://journals.physiology.org/doi/full/10.1152/advan.00181.2023

14- Kaftan AN, Hussain MK, Naser FH. Response accuracy of ChatGPT 3.5 Copilot and Gemini in interpreting biochemical laboratory data a pilot study. Sci Rep. 8 de abril de 2024 [Internet]. [Citado 22/08/2024];14(1):8233. Disponible en: https://www.nature.com/articles/s41598-024-58964-1

15- Dhanvijay AKD, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H. Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology. Cureus. 2023 [Internet]. [Citado 22/08/2024];15(8):e42972. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC10475852/

16- Meo SA, Al-Masri AA, Alotaibi M, Meo MZS, Meo MOS. ChatGPT Knowledge Evaluation in Basic and Clinical Medical Sciences: Multiple Choice Question Examination-Based Performance. Healthcare. 2023 [Internet]. [Citado 23/08/2024];11(14):2046. Disponible en: https://pubmed.ncbi.nlm.nih.gov/37510487/

17- IBM Data IA Team. What are Open Source Large Language Models? | IBM 2023 [Internet]. [Citado 9/09/2024]. Disponible en: https://www.ibm.com/think/topics/open-source-llms

18- Consecuencias del bloqueo de EE.UU.para las comunicaciones y la informática en Cuba. Trabajadores. 31 oct 2020; 1ra columna [Internet]. [Citado 9/09/2024]. Disponible en: https://www.trabajadores.cu/20201031/consecuencias-del-bloqueo-de-ee-uu-para-las-comunicaciones-y-la-informatica-en-cuba/

19- Antón S. Bloqueo estadounidense es el principal impedimento para un más amplio acceso a internet en Cuba. Granma 26 may 2021 [Internet]. [Citado 12/09/2024]. Disponible en: https://www.granma.cu/cuba/2021-05-26/bloqueo-estadounidense-es-el-principal-impedimento-para-un-mas-amplio-acceso-a-internet-26-05-2021-17-05-40

20- Meta-llama/Llama-3.1-70B-Instruct · Hugging Face 2024 [Internet]. [Citado 22/09/2024]. Disponible en: https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct

21- HuggingChat. [Internet]. [Citado 29 de septiembre de 2024]. Disponible en: https://huggingface.co/chat

22- Arias FG. El Proyecto de Investigación Introducción a la Metodología Científica. 6ta. Ed. Caracas-República Bolivariana de Venezuela: Episteme; 2012. 137 p.

23- Shi F, Chen X, Misra K, Scales N, Dohan D, Chi EH, et al. Large Language Models Can Be Easily Distracted by Irrelevant Context. En: Proceedings of the 40th International Conference on Machine Learning 2023 [Internet]. [Citado 29/09/2024]; 202:31210-27. Disponible en: https://proceedings.mlr.press/v202/shi23a.html

24- Likert R. A technique for the measurement of attitudes. Archives of Psychology. 1932;22 (140):55-55 [Internet]. [Citado 10/10/2024] Disponible en: https://legacy.voteview.com/pdf/Likert_1932.pdf

25- Matas A. Diseño del formato de escalas tipo Likert: un estado de la cuestión. Revista electrónica de investigación educativa. 2018 [Internet]. [Citado 11/10/2024]; 20(1):38-47. Disponible en: https://redie.uabc.mx/redie/article/view/1347

26- Pérez Pérez CJ. Archivos de Investigación Llama-3.1-70B-Instruct Bioquímica Metabolismo y Nutrición [Internet]. Zenodo; 2025 [citado 3 de enero de 2025]. Disponible en: https://zenodo.org/records/14596114

27- Ghosh A, Bir A, Ghosh A, Bir A. Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the Competency-Based Medical Education Curriculum in Medical Biochemistry. 2023 [Internet]. [Citado 19/10/2024]; 15(4): e37023. Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10152308/

28- Ghosh A, Jindal NM, Gupta VK, Bansal E, Bajwa NK, Sett A. Is ChatGPT's Knowledge and Interpretative Ability Comparable to First Professional MBBS (Bachelor of Medicine, Bachelor of Surgery) Students of India in Taking a Medical Biochemistry Examination? Cureus. 2023 [Internet]. [Citado 19/10/2024];15(10):e47329. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC10657167/

29-Sallam M, Al-Salahat K, Eid H, Egger J, Puladi B. Human versus Artificial Intelligence: ChatGPT-4 Outperforming Bing, Bard, ChatGPT-3.5 and Humans in Clinical Chemistry Multiple-Choice Questions. Advances in Medical Education and Practice 2024 [Internet]. [Citado 19/10/2024];15:857-71. Disponible en: https://www.tandfonline.com/doi/abs/10.2147/AMEP.S479801

30- Wang Z, Bukharin A, Delalleau O, Egert D, Shen G, Zeng J, et al. HelpSteer2-Preference: Complementing Ratings with Preferences. arXiv 2024 [Internet]. [Citado 19/10/2024]. Disponible en: http://arxiv.org/abs/2410.01257

31- Nvidia/Llama-3.1-Nemotron-70B-Instruct-HF.Hugging Face 2024 [Internet]. [Citado 24/10/2024]. Disponible en: https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct-HF

32- GPT-4o mini: advancing cost-efficient intelligence. 2024 [Internet]. [Citado 24/10/2024]. Disponible en: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

33- Gravel J, D'Amours-Gravel M, Osmanlliu E. Learning to Fake It: Limited Responses and Fabricated References Provided by ChatGPT for Medical Questions. Mayo Clinic Proceedings: Digital Health. 2023 [Internet]. [Citado 26/10/2024];1(3):226-34. Disponible en: https://www.sciencedirect.com/science/article/pii/S2949761223000366

34- Caruccio L, Cirillo S, Polese G, Solimando G, Sundaramurthy S, Tortora G. Can ChatGPT provide intelligent diagnoses? A comparative study between predictive models and ChatGPT to define a new medical diagnostic bot. Expert Systems with Applications. 2024 [Internet]. [Citado 27/10/2024];235:121186. Disponible en: https://www.sciencedirect.com/science/article/pii/S0957417423016883

35- Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. arXiv 2023 [Internet]. [Citado 28/10/24]. Disponible en: http://arxiv.org/abs/2205.11916

36- Zhang J, Qiao D, Yang M, Wei Q. Regurgitative Training: The Value of Real Data in Training Large Language Models. arXiv 2024 [Internet]. [Citado 28/10/24]. Disponible en: http://arxiv.org/abs/2407.12835

Conflictos de intereses

Los autores declaran la no existencia de conflictos de intereses.