Tecnociencia, Vol. 28, N°2: 157-172
Julio-Diciembre 2026
ISSN L 2415-0940

Comparación reproducible de modelos de lenguaje en problemas textuales de mecánica clásica mediante un benchmark automatizado

Reproducible comparison of language models in textual problems of classical mechanics using an automated benchmark

Omayra Janeth Pérez Castro
Universidad de Panamá, Facultad de Ciencias Naturales, Exactas y Tecnología, Departamento de Física, Panamá. omayra.perez@up.ac.pa https://orcid.org/0000-0002-7080-5598

Bernardo Fernández García
Universidad de Panamá, Facultad de Ciencias Naturales, Exactas y Tecnología, Departamento de Física, Panamá. bernardo.fernandezg@up.ac.pa https://orcid.org/0000-0002-7947-3147

Luis Antonio Marín Díaz
Universidad de Panamá, Facultad de Ciencias Naturales, Exactas y Tecnología, Departamento de Física, Panamá. luis.marin-d@up.ac.pa https://orcid.org/0000-0001-8580-4128

Juan Manuel Rodríguez Cisneros
Universidad de Panamá, Facultad de Ciencias Naturales, Exactas y Tecnología, Departamento de Física, Panamá. juan-m.rodriguez-c@up.ac.pa https://orcid.org/0000-0002-2013-6752

Fecha de recepción: 30 de abril de 2026
Fecha de aceptación: 11 de mayo de 2026

DOI: https://doi.org/10.48204/j.tecno.v8n2.a10864

RESUMEN

Este estudio presenta un benchmark textual para comparar nueve modelos comerciales de lenguaje de gran escala en la resolución de problemas de Mecánica Clásica. Se empleó un total de 42 ítems auto-contenidos y sin dependencia explícita de diagramas, extraídos de materiales abiertos de MIT OpenCourseWare correspondientes a 2005 y 2016. Los modelos evaluados pertenecieron a tres proveedores: OpenAI, Anthropic y Google Gemini. Cada problema se presentó mediante un mismo prompt estructurado con respuesta final, resumen breve del razonamiento, ecuaciones, unidades, nivel de confianza. La corrida completa produjo 378 respuestas utilizables. La referencia de corrección se construyó en dos etapas: consenso numérico para 22 ítems y curación manual para 19 casos sin consenso; un ítem permaneció ambiguo por dependencia residual de figuras ausentes. El pipeline alcanzó 100% de éxito final, aunque 29 salidas truncadas requirieron

recuperación a nivel de parser. En acuerdo heurístico con la referencia curada, Gemini obtuvo el mejor rendimiento agregado por proveedor (43.7%; 55/126), mientras que gemini-3.1-pro-preview alcanzó el mejor resultado a nivel de modelo (61.9%; 26/42). Anthropic produjo las respuestas más extensas y lentas, y gemini-3-flash-preview fue el modelo más veloz. Los hallazgos muestran que los LLMs ya apoyan evaluaciones reproducibles en física sin reemplazar expertos.

PALABRAS CLAVE

Modelos de lenguaje, enseñanza de la física, mecánica clásica, benchmark, evaluación comparativa

ABSTRACT

This study presents a textual benchmark to compare nine large-scale commercial language models for solving Classical Mechanics problems. A total of 42 self-contained items, without explicit dependence on diagrams, were extracted from MIT OpenCourseWare open-source materials from 2005 and 2016. The evaluated models came from three vendors: OpenAI, Anthropic, and Google Gemini. Each problem was presented using the same structured prompt, including a final answer, a summary of the reasoning, equations, units, and confidence level. The complete run yielded 378 usable responses. The correction reference was built in two stages: numerical consensus for 22 items and manual curation for 19 cases lacking consensus; one item remained ambiguous due to residual dependence on missing figures. The pipeline achieved 100% final success, although 29 truncated outputs required parser-level recovery. In heuristic agreement with the curated reference, Gemini achieved the best aggregate performance by vendor (43.7%; 55/126), while gemini-3.1-pro-preview achieved the best result at the model level (61.9%; 26/42). Anthropic produced the longest and slowest responses, and gemini-3-flash-preview was the fastest model. The findings show that LLMs already support reproducible assessments in physics without replacing experts.

KEYWORDS

Large language models, physics education, classical mechanics, benchmark, comparative evaluation

INTRODUCCIÓN

En los últimos años, los grandes modelos de lenguaje (LLMs, por sus siglas en inglés) han pasado de ser herramientas de generación textual a convertirse en sistemas capaces de intervenir en tareas de explicación, tutoría y resolución de problemas. Ese desplazamiento ha reactivado la discusión sobre su valor en la educación en ciencias y sobre el alcance real de sus competencias lingüísticas y metalingüísticas (Beguš et al., 2025; Rodríguez Fernández et al., 2024).

En Física, el interés por estos sistemas no se limita a su fluidez verbal. Un modelo útil en este dominio debe interpretar enunciados, seleccionar principios relevantes, traducir relaciones físicas a expresiones matemáticas y sostener una explicación compatible con las leyes de la mecánica. Esa combinación exige algo más que solo usar lenguaje: requiere razonamiento

cuantitativo, control conceptual y vigilancia sobre unidades, condiciones iniciales y supuestos del problema (Ding et al., 2023; Zhao et al., 2024).

La resolución de problemas de Mecánica Clásica constituye un terreno especialmente exigente para evaluar estas capacidades. La literatura sobre didáctica de la Física ha mostrado que resolver problemas no equivale a aplicar fórmulas de forma automática: los problemas más ricos demandan análisis cualitativo, construcción de representaciones, selección de principios físicos y evaluación crítica de los resultados obtenidos (Gil Pérez & Valdés Castro, 1994; Navarro Altamar, 2006; Perales Palacios, 1993; Guisasola et al., 2015). Esa diferencia entre razonamiento novato y razonamiento experto resulta clave al analizar respuestas de LLMs, pues muchos de sus errores no son meramente aritméticos, sino estructurales o conceptuales.

Las evaluaciones disponibles sugieren que los LLMs pueden aproximarse al desempeño humano en ciertos ejercicios introductorios de Física, pero su rendimiento cae cuando los problemas exigen composición matemática, consistencia entre múltiples pasos o interpretación de condiciones implícitas (Kortemeyer, 2023; Dao et al., 2023; Sato, 2024). Más recientemente, han surgido benchmarks de mayor cobertura, como UGPhysics y PhysReason, que documentan avances relevantes pero también limitaciones persistentes en razonamiento físico multi-paso (Xu et al., 2025; Zhang et al., 2025).

Sin embargo, todavía hacen falta evaluaciones comparativas acotadas, reproducibles y metodológicamente homogéneas que controlen de manera estricta el prompt, la estructura de salida, el postprocesamiento y el criterio de puntuación, particularmente cuando se comparan proveedores comerciales bajo un mismo corpus disciplinar. Mientras UGPhysics y PhysReason privilegian la amplitud temática, el presente trabajo se concentra en un banco pequeño, homogéneo y exclusivamente textual de Mecánica Clásica, con énfasis en la trazabilidad del pipeline y en la comparabilidad directa entre respuestas.

El objetivo de este artículo es triple. Primero, documentar un pipeline reproducible para correr, parsear, puntuar y visualizar respuestas de múltiples LLMs en Mecánica Clásica. Segundo, comparar el desempeño de los modelos mediante una referencia curada explícita. Tercero, identificar en qué dimensiones del proceso experimental se ubican las principales fuentes de variación: formato de salida, estabilidad operativa, latencia, longitud de respuesta y acuerdo con la solución de referencia.

MATERIALES Y MÉTODOS

El estudio se articuló en torno a tres preguntas clave. La primera fue operacional: ¿pueden distintos LLMs comerciales completar un benchmark de Mecánica Clásica con un formato

estructurado y sin fallos duros de parseo? La segunda fue comparativa: ¿qué modelos muestran mayor acuerdo con una referencia de corrección curada bajo un mismo protocolo de inferencia? La tercera fue analítica: ¿cómo cambian esos resultados según proveedor, familia de modelo, latencia, extensión de la respuesta y área temática del problema?

Se realizó un estudio comparativo, descriptivo y reproducible con enfoque mixto. La unidad básica de análisis fue la respuesta generada por un modelo ante un problema de Mecánica Clásica. El diseño combinó métricas operativas de ejecución y parseo con una comparación heurística frente a una referencia de corrección curada. Esta estrategia permitió evaluar simultáneamente la robustez de la infraestructura y la calidad aparente de las respuestas. El trabajo corresponde a un benchmark de laboratorio computacional, no a una intervención educativa con estudiantes.

El corpus estuvo compuesto por 42 ítems provenientes de dos colecciones abiertas de MIT OpenCourseWare: Physics I: Classical Mechanics (Massachusetts Institute of Technology [MIT], 2005) y Classical Mechanics (MIT, 2016). Todos los problemas incluidos pertenecieron a una versión local text-only del benchmark y cumplieron tres criterios operativos: ser auto-contenidos, no requerir diagramas para su interpretación inmediata y estar listos para evaluación exclusivamente textual.

El pipeline partió de un archivo maestro con metadatos por ítem. La selección final retuvo únicamente problemas etiquetados como utilizables para LLM, auto-contenidos y preparados para modalidad textual. En esta etapa se excluyeron ejercicios con dependencia visual indispensable, formatos no textuales o enunciados excesivamente fragmentarios. El corpus no corresponde a un instrumento psicométricamente validado: todos los registros del archivo maestro estaban clasificados como open_oer_non_psychometric. Por ello, los resultados deben interpretarse como evidencia comparativa de desempeño en un banco de problemas abiertos de Mecánica Clásica, y no como una medición estandarizada del aprendizaje conceptual en el sentido de instrumentos como el Force Concept Inventory (Hestenes et al., 1992).

Predominaron los problemas numérico-simbólicos (40 de 42 ítems), mientras que solo dos combinaron explicación cualitativa con respuesta numérica. El área temática más representada fue cinemática (10 ítems), seguida por movimiento circular (5) y movimiento rotacional (4).

Tabla 1
Composición temática del corpus de evaluación

Área temática	**n**
Cinemática	10
Movimiento circular	5
Movimiento rotacional	4
Cantidad de movimiento e impulso	4
Restricciones y fuerzas resistivas	3
Transferencia de masa y velocidad relativa	3
Energía potencial	3
Choques	3
Rodadura	2
Trabajo y energía cinética	2
Energía	1
Momento angular	1
Leyes de Newton	1
Total	42

Nota. El corpus final estuvo compuesto por problemas abiertos de MIT OpenCourseWare adaptados a evaluación textual.

Se evaluaron nueve modelos comerciales, tres por proveedor: OpenAI (gpt-5.4, gpt-5.4-mini y gpt-5.4-nano), Anthropic (claude-opus-4-6, claude-sonnet-4-6 y claude-haiku-4-5) y Google Gemini (gemini-3.1-pro-preview, gemini-3-flash-preview y gemini-3.1-flash-lite-preview). Se muestran los identificadores exactamente como aparecieron, para que el experimento pueda rastrearse y verificarse. Esos identificadores reflejan lo que estaba disponible localmente en marzo de 2026.

Cada par modelo-ítem se evaluó una sola vez bajo configuración de producción fija, sin muestreo repetido ni búsqueda de la mejor respuesta entre múltiples corridas. El límite de salida fue de 1400 tokens por solicitud. OpenAI se consultó mediante Responses API con salida estructurada; Anthropic mediante Messages API con restricciones de salida JSON y parser tolerante a truncamientos; y Gemini mediante generateContent con application/json, candidateCount = 1 y temperatura cero. En Gemini 3.1 Pro se habilitó un presupuesto moderado de pensamiento, mientras que en los modelos Flash y Flash Lite se desactivó. Dado que los catálogos de modelos cambian con rapidez, la comparación debe leerse como una fotografía experimental fechada.

La latencia reportada en este artículo debe interpretarse como una métrica descriptiva del entorno local de ejecución y de las condiciones de red observadas el día del experimento. No se trata, por tanto, de una medición absoluta de rendimiento de infraestructura entre proveedores.

Tabla 2
Configuración experimental por proveedor

Proveedor	**Modelos**	**Estrategia de salida estructurada**
OpenAI	gpt-5.4; gpt-5.4-mini; gpt-5.4-nano	Responses API con salida estructurada y un solo intento por ítem
Anthropic	claude-opus-4-6; claude-sonnet-4-6; claude-haiku-4-5	Messages API con restricciones de salida JSON y parser tolerante a truncamientos
Gemini	gemini-3.1-pro-preview; gemini-3-flash-preview; gemini-3.1-flash-lite-preview	generateContent con application/json, temperature = 0 y candidateCount = 1

Nota. Todas las corridas se ejecutaron en marzo de 2026 con el prompt versionado physics-benchmark-json-v1 y un límite de 1400 tokens de salida.

Todos los modelos recibieron el mismo prompt base. Dicho prompt solicitó una única salida JSON con siete campos obligatorios: final_answer, reasoning_summary, equations_used, units, confidence, needs_human_review y review_note. La intención de este diseño fue evitar formatos libres difíciles de comparar y, al mismo tiempo, preservar suficiente información para revisar la estructura del razonamiento.

El prompt incluyó metadatos del ítem—identificador, título, broad_topic, subtopic y answer_format_guess— y una instrucción explícita para marcar needs_human_review cuando el problema fuese ambiguo, incompleto o dependiera de una figura ausente. Todas las ejecuciones quedaron registradas en archivos JSONL y CSV con metadatos de proveedor, modelo, tokens, latencia, motivo de finalización y texto crudo de respuesta.

El barrido produjo 378 respuestas para 9 modelos sobre 42 problemas. Aunque el parseo final alcanzó 100%, se registraron 29 respuestas truncadas que debieron recuperarse a nivel de parser. Estas recuperaciones se concentraron en claude-opus-4-6 (16), claude-sonnet-4-6 (9) y gemini-3.1-pro-preview (4).

La corrida se ejecutó en modo single-pass: no se realizaron reruns para mejorar respuestas individuales ni se permitió que un modelo viera respuestas de otro. El objetivo fue simular

una evaluación ciega, homogénea y trazable, más cercana a un benchmark comparativo que a un proceso iterativo de tutoría.

La referencia utilizada para puntuar las respuestas no se tomó de un solucionario oficial único. En su lugar, se construyó una clave de referencia curada en dos etapas. Primero, se identificaron agrupamientos de consenso exacto o numérico entre múltiples modelos. Cuando al menos tres respuestas coincidían tras normalización o dentro de una tolerancia relativa de 2%, esa solución se tomó como referencia provisional.

Segundo, los ítems sin consenso robusto se revisaron manualmente mediante derivación física razonada. Como resultado, 22 ítems quedaron clasificados como draft_numeric_consensus, 19 como curated_manual_reasoning y 1 como text_unanswerable_or_ambiguous. Ese último caso correspondió al ítem MIT05_PS02_Q02 "Geometry and Angles" (Physics I: Classical Mechanics, MIT OCW 2005, Problem Set 2, problema 2), un ejercicio cuyo enunciado, aun habiendo sido marcado como auto-contenido y text_only_ready en el archivo maestro, conservó una dependencia residual de información geométrica originalmente codificada en tres figuras del material fuente con ángulos numerados y marcas de ángulo recto.

El enunciado textual aislado no permitía reconstruir unívocamente esa configuración: en el archivo de respuestas, los nueve modelos coincidieron en marcar parsed_needs_human_review=True para este único ítem (el único caso de unanimidad en autoseñalización del estudio), y siete de ellos lo declararon explícitamente en parsed_review_note como "missing figures", "diagrams not provided" o equivalentes.

El caso ilustra un límite importante del razonamiento puramente textual: cuando un problema de mecánica codifica supuestos críticos en un diagrama (direcciones, signos, geometría de vínculos o convenciones de ejes), ningún LLM sin entrada visual puede recuperar esa información por inferencia lingüística, y la ambigüedad se traslada en bloque a la respuesta final.

Conviene aclarar, además, que la dependencia residual de figuras no se limitó a este único ejercicio: el conteo de menciones en parsed_review_note muestra que 19 ítems distintos recibieron al menos una queja explicita por figura ausente o geometría no especificada, aunque solo MIT05_PS02_Q02 resultó lo bastante ambiguo como para impedir la fijación de una respuesta de referencia. Por ello, este ítem se reportó como ambiguo y se excluyó del cómputo del acuerdo, pero se preservó en el corpus por su valor diagnóstico sobre las fronteras de la modalidad text-only.

Este procedimiento produjo una referencia operativa útil para la comparación entre modelos, pero no reemplaza una rúbrica independiente validada por un panel externo de expertos. En consecuencia, los resultados se interpretan aquí como acuerdo con una referencia curada, y no como exactitud absoluta.

Se calcularon tasas de éxito de API, parseo correcto, disponibilidad de respuesta final, autoseñalización de revisión humana, latencia media, tokens medios y acuerdo heurístico con la referencia curada. Este último indicador se definió como coincidencia exacta de texto, inclusión sustantiva de la respuesta canónica o coincidencia numérica dentro de tolerancia. Dado que no se utilizó un juez semántico externo en esta versión del análisis, el acuerdo heurístico debe leerse como una aproximación conservadora y trazable al desempeño comparado.

La tolerancia numérica empleada fue de 2% en términos relativos y 1e-6 en valor absoluto. Cuando la respuesta contenía varias magnitudes numéricas, el scorer exigió coincidencia posicional completa entre listas de valores. No se aplicaron pruebas inferenciales ni contrastes de hipótesis entre modelos, porque el propósito de esta versión del manuscrito es descriptivo-comparativo y el corpus no fue diseñado como muestra probabilística. En cambio, se privilegiaron tablas resumidas, tasas agregadas por proveedor y modelo y análisis por áreas temáticas, complementados con conteos absolutos cuando estos mejoraban la transparencia interpretativa.

RESULTADOS

En términos operativos, el pipeline mostró un comportamiento notablemente estable. Las 378 solicitudes completaron la llamada de API y terminaron integradas al CSV final, lo que implica 100% de éxito de ejecución y 100% de parseo utilizable. Este resultado es importante porque uno de los principales obstáculos para evaluar LLMs en dominios técnicos no es solo la calidad de la respuesta, sino la consistencia del formato de salida.

Sin embargo, la robustez final no debe confundirse con ausencia de inci dentes intermedios. El sistema tuvo que recuperar 29 respuestas truncadas: 16 de claude-opus-4-6, 9 de claude-sonnet-4-6 y 4 de gemini-3.1-pro-preview. En otras palabras, el uso de salidas estructuradas redujo la fragilidad del pipeline, pero no eliminó del todo los problemas de serialización cuando algunos modelos produjeron respuestas extensas o alcanzaron el límite de tokens.

A nivel agregado por proveedor, Gemini mostró el mayor acuerdo heurístico con la referencia curada (43.7%; 55/126), seguido por Anthropic (34.1%; 43/126) y OpenAI (12.7%; 16/126). Anthropic produjo las respuestas más largas (837.7 tokens de salida, en promedio) y también

la mayor latencia media (12.18 s). OpenAI fue el proveedor más rápido en promedio (4.48 s), pero también el que mostró menor acuerdo con la referencia.
Además, fue el proveedor que con mayor frecuencia marcó sus respuestas para revisión humana (31.0%), lo que sugiere una actitud más conservadora, aunque no necesariamente más acertada.

Tabla 3
Desempeño descriptivo por proveedor

Proveedor	Éxito API	Parseo correcto	Acuerdo heurístico	Coincidencia numérica	Latencia media (s)	Tokens salida	Auto-revisión
gemini	100.0%	100.0%	43.7%	42.9%	5.62	304.3	8.7%
anthropic	100.0%	100.0%	34.1%	34.1%	12.18	837.7	20.6%
openai	100.0%	100.0%	12.7%	11.1%	4.48	325.1	31.0%

Nota. El acuerdo heurístico corresponde a coincidencia exacta, coincidencia por inclusión o coincidencia numérica dentro de una tolerancia relativa del 2%.

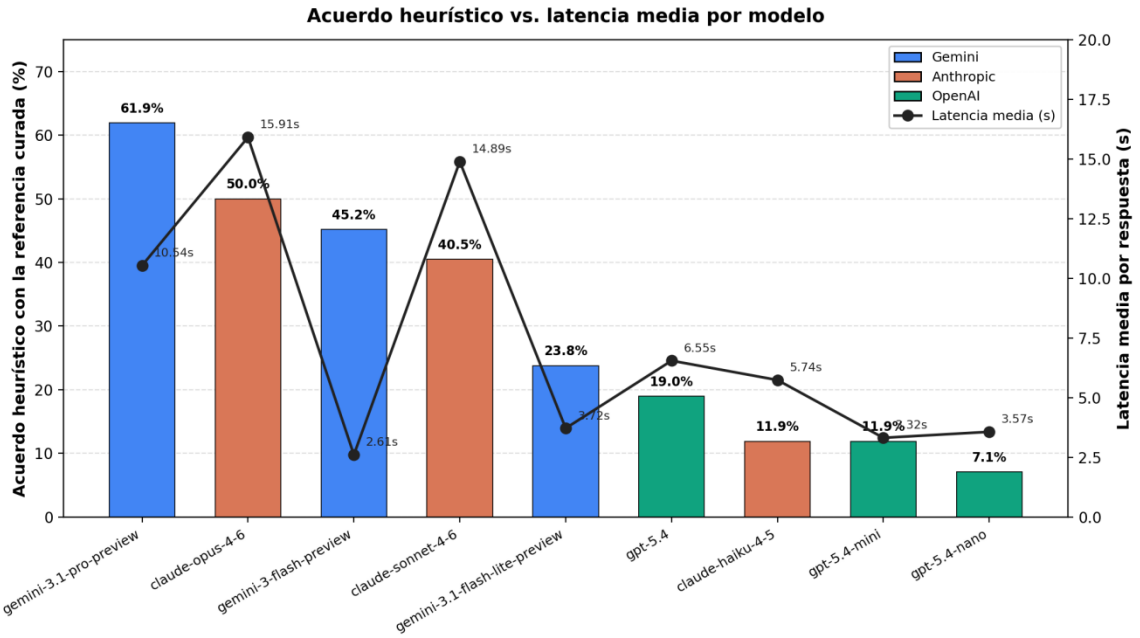
A nivel de modelo individual, el mejor rendimiento correspondió a gemini-3.1-pro-preview (61.9%; 26/42), seguido por claude-opus-4-6 (50.0%; 21/42). El modelo más veloz fue gemini-3-flash-preview con 2.61 s por respuesta, mientras que el más lento fue claude-opus-4-6 con 15.91 s. En la familia Anthropic se observó un patrón claro: mayor acuerdo heurístico se asoció con respuestas más extensas y mayor costo temporal. En OpenAI, la disminución de tamaño de modelo no se tradujo en mejor eficiencia epistémica; por el contrario, gpt-5.4-nano presentó el menor acuerdo y la mayor tasa de autoseñalización de revisión (61.9%).

Tabla 4
Desempeño descriptivo por modelo

Modelo	Proveedor	Acuerdo heurístico	Coincidencia numérica	Latencia media (s)	Tokens salida	Auto-revisión
gemini-3.1-pro-preview	gemini	61.9%	61.9%	10.54	268.5	9.5%
claude-opus-4-6	anthropic	50.0%	50.0%	15.91	938.1	19.0%
gemini-3-flash-preview	gemini	45.2%	42.9%	2.61	304.4	7.1%
claude-sonnet-4-6	anthropic	40.5%	40.5%	14.89	958.4	4.8%
gemini-3.1-flash-lite-preview	gemini	23.8%	23.8%	3.72	340.1	9.5%
gpt-5.4	openai	19.0%	16.7%	6.55	323.4	14.3%
claude-haiku-4-5	anthropic	11.9%	11.9%	5.74	616.6	38.1%
gpt-5.4-mini	openai	11.9%	9.5%	3.32	314.5	16.7%
gpt-5.4-nano	openai	7.1%	7.1%	3.57	337.3	61.9%

Nota. Los resultados corresponden a 42 ítems por modelo y se basan en agreement scoring heurístico contra una referencia curada.

Figura 1
Acuerdo heurístico vs. latencia media por modelo



Nota. Las barras representan el acuerdo heurístico con la referencia curada (eje izquierdo) y la línea negra la latencia media por respuesta en segundos (eje derecho). El color de cada barra indica el proveedor. Los modelos están ordenados de mayor a menor acuerdo. Elaboración propia a partir de los datos de la Tabla 4.

La Figura 1 sintetiza visualmente los hallazgos de la Tabla 4 y permite leer simultáneamente las dos dimensiones críticas del benchmark. En primer lugar, evidencia que el liderazgo de Gemini no es uniforme dentro del proveedor: gemini-3.1-pro-preview encabeza el ranking de acuerdo, pero gemini-3.1-flash-lite-preview cae al quinto lugar, lo que sugiere que la diferencia interna por familia es comparable a la diferencia entre proveedores.

La superposición de la línea de latencia muestra que el acuerdo no es una función monótona del tiempo de respuesta: gemini-3-flash-preview alcanza el tercer mejor acuerdo (45.2%) con la menor latencia del estudio (2.61 s), mientras que los modelos Anthropic obtienen acuerdos intermedios al costo de las latencias más altas (entre 14.89 s y 15.91 s). Finalmente, el bloque OpenAI se concentra en la zona de bajo acuerdo con latencias moderadas, lo que confirma que en este corpus la velocidad por sí sola no compensa el déficit de coincidencia con la referencia curada.

El análisis temático mostró diferencias relevantes entre proveedores. Gemini dominó varias áreas de cálculo y conservación, con sus mejores resultados en trabajo y energía cinética

(83.3%), choques (77.8%) y cantidad de movimiento e impulso. Anthropic se comportó relativamente mejor en problemas de energía potencial y en algunos ítems de movimiento rotacional, donde alcanzó 50.0% de acuerdo heurístico promedio.

Por contraste, OpenAI mantuvo un desempeño bajo en la mayoría de los ejes temáticos y solo se acercó a sus competidores en un subconjunto pequeño de trabajo y energía. También es relevante que el único ítem clasificado bajo la categoría broad_topic = energy produjo 0% de acuerdo heurístico en los tres proveedores. Este hallazgo puede interpretarse de dos maneras no excluyentes: el ítem introducía una formulación especialmente exigente para los modelos o el esquema de comparación heurística resultó demasiado estricto para una respuesta conceptualmente válida expresada con otra redacción.

DISCUSIÓN

Los resultados permiten sostener que la principal madurez observada en los LLMs evaluados no reside todavía en su precisión física global, sino en su capacidad para incorporarse a flujos de evaluación reproducibles. Con un prompt estructurado, control del formato de salida y un parser tolerante a truncamientos, fue posible ejecutar y consolidar 378 respuestas comparables sin fallos duros de parseo. Desde una perspectiva metodológica, este es un avance relevante para investigaciones futuras en evaluación automática de razonamiento científico.

No obstante, el rendimiento disciplinar siguió siendo heterogéneo. El mejor modelo del estudio no superó 61.9% de acuerdo con la referencia, y el mejor proveedor se situó en 43.7%. Estos valores distan de un nivel que permitiría tratar a los modelos como sustitutos confiables de una resolución experta en Mecánica Clásica. La diferencia entre modelos también sugiere que la escala o la verbosidad por sí solas no garantizan mejor desempeño: algunos sistemas fueron más rápidos y compactos, pero no necesariamente más correctos, y otros produjeron explicaciones largas sin mejorar de forma proporcional la calidad final.

Este patrón es coherente con evaluaciones previas y con benchmarks recientes de física, que muestran avances importantes en tareas introductorias, pero degradación cuando los problemas requieren razonamiento multi-paso, integración simbólico-numérica y control fino de condiciones implícitas (Kortemeyer, 2023; Xu et al., 2025; Zhang et al., 2025). Desde el punto de vista didáctico, el desempeño más prometedor aparece en problemas cuya solución puede representarse de manera algebraica o numérica con alta restricción formal. En cambio, los ítems con mayor carga interpretativa o formulaciones más abiertas siguen generando respuestas variables.

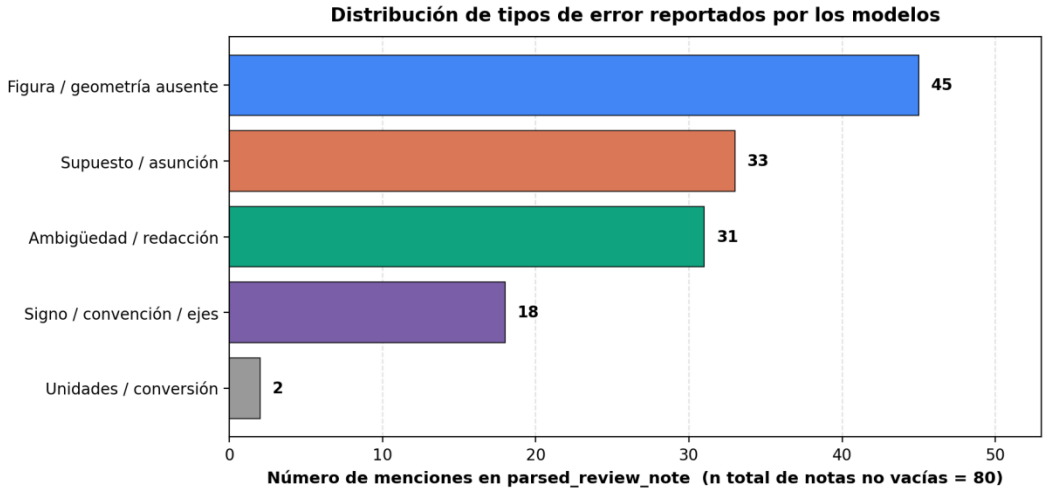
Una mirada cualitativa al subconjunto de respuestas no concordantes, apoyada en el conteo de menciones registradas en el campo parsed_review_note del archivo de resultados, permite

distinguir tipos de errores recurrentes. De las 378 respuestas, 80 incluyeron una nota de revisión no vacía; sobre ese subconjunto, las categorías siguientes no son excluyentes entre sí. El grupo más frecuente corresponde a quejas por dependencia residual de figuras o geometría no especificada (45 menciones, distribuidas en 19 ítems distintos): el modelo identifica el dominio pero declara explícitamente que el enunciado textual no fija ángulos, configuración de vínculos o relaciones espaciales necesarias. Un segundo grupo, casi de la misma magnitud (33 menciones), corresponde a errores de supuesto, en los que el modelo completa la información faltante con una asunción razonable pero no necesariamente coincidente con la convención de la fuente (por ejemplo, asumir un ángulo de 90° en una ranura en V cuyo ángulo no se especifica). Un tercer grupo (31 menciones) recoge errores asociados a ambigüedad o redacción corrupta del enunciado, que en algunos casos llevaron al modelo a proponer interpretaciones alternativas en la propia respuesta. Un cuarto grupo, más sutil (18 menciones), son errores de signo, dirección o convención de ejes (típicamente sentidos de fuerza, orientación del eje o sentido de giro que el enunciado no fija explícitamente).

Junto a estos, aparecen errores de planteamiento físico (selección de un principio inadecuado, como aplicar conservación de energía mecánica en presencia de fricción) y errores de ejecución algebraica o numérica sobre un planteamiento correcto, los cuales en su mayoría no quedaron registrados en parsed_review_note porque los modelos no los reconocieron como tales y por lo tanto deben inferirse por inspección de las respuestas. Errores explícitos de unidades fueron, en cambio, muy poco frecuentes (apenas 2 menciones), lo que sugiere que el ranking de modelos no estaría sesgado por la sensibilidad del scorer heurístico a las unidades. Mencion aparte merecen los errores de cierre por truncamiento: 29 respuestas alcanzaron el límite de tokens y debieron recuperarse a nivel de parser, concentradas en claude-opus-4-6, claude-sonnet-4-6 y gemini-3.1-pro-preview; críticamente, ninguna de las respuestas truncadas fue autoseñalizada por el modelo como necesidad de revisión, lo que indica que el truncamiento es un punto ciego de las propias respuestas y refuerza la utilidad de monitorear finish_reason además de parsed_needs_human_review.

En conjunto, esta distribución sugiere que el déficit dominante en el corpus no es aritmético sino representacional (reconstrucción geométrica y elección de convenciones), y que un juez semántico podría reclasificar parte de los casos de supuesto y convención sin alterar el orden general entre proveedores.

Figura 2
Distribución de tipos de error reportados por los modelos en parsed_review_note



Nota. Conteo de menciones por categoría sobre el subconjunto de 80 respuestas con parsed_review_note no vacía (de un total de 378). Las categorías no son mutuamente excluyentes: una misma nota puede mencionar varios tipos de error. Las menciones se identificaron mediante palabras clave en el texto crudo (figure/diagram/missing/geometry,assume,ambiguous/unclear/interpret/garbled,sign/direction/orient/axis/convention, unit/convert).

Tabla 5

Ítems con consenso de revisión (≥5 de 9 modelos marcaron needs_human_review)

benchmark_id	Título	broad_topic	Modelos que pidieron revisión	Fuente
MIT05_PS02_Q02	Geometry and Angles	kinematics	9 / 9	MIT 2005, PS2 Q2
MIT16_PS10_Q05	A Cylinder Rolling in a V-Groove	rotational_motion	7 / 9	MIT 2016, PS10 Q5
MIT16_PS02_Q04	Dragging two blocks	newtons_laws	5 / 9	MIT 2016, PS2 Q4
MIT16_PS08_Q03	Objects on a Ring	potential_energy	5 / 9	MIT 2016, PS8 Q3

Nota. Lista de los ítems donde al menos 5 de los 9 modelos marcaron parsed_needs_human_review = True. Solo MIT05_PS02_Q02 alcanzó unanimidad de 9/9 y fue el único clasificado como text_unanswerable_or_ambiguous en la curación manual.

La Figura 2 y la Tabla 5 hacen visible un patrón que el texto solo dejaba implicado: la concentración de las observaciones autocríticas de los modelos en problemas con

dependencia geométrica residual. De los cuatro ítems con consenso de revisión (≥5/9), tres

pertenecen a problemas de mecánica avanzada con configuraciones espaciales no triviales — ranura en V, bloques arrastrados, objetos sobre un anillo— donde, sin la figura original, queda ambigüa alguna pieza geométrica esencial (ángulo de la ranura, configuración de la polea, posición del punto de anclaje). Este consenso entre modelos heterogéneos en proveedor, escala y familia funciona como un indicador externo útil: cuando varios LLMs coinciden en pedir revisión humana sobre un mismo ítem, es un buen criterio operativo para auditar el enunciado en la siguiente iteración del corpus.

CONCLUSIONES

La comparación realizada muestra que los LLMs pueden integrarse de forma operativamente robusta en benchmarks reproducibles de Mecánica Clásica, siempre que se impongan salidas estructuradas y se controle cuidadosamente el postprocesamiento. En este plano, el pipeline desarrollado constituye un aporte metodológico útil para futuras investigaciones en evaluación automática de razonamiento físico y, de manera indirecta, para la educación científica asistida por inteligencia artificial.

En el plano disciplinar, sin embargo, los resultados son más sobrios. Gemini alcanzó el mejor rendimiento agregado y gemini-3.1-pro-preview lideró a nivel de modelo, pero incluso los mejores sistemas se mantuvieron en un rango de acuerdo moderado con la referencia curada. Por ello, la evidencia disponible apoya el uso de LLMs como herramientas auxiliares para exploración, apoyo tutorial o generación de borradores de solución, pero no como reemplazo del juicio experto ni de la validación docente en Física.

Como siguiente etapa, el estudio debería ampliarse con una clave de corrección validada por un panel de especialistas, un componente de judge-based scoring y una muestra que incorpore más ítems conceptuales y multimodales. Esa ampliación permitiría transformar el presente manuscrito desde un benchmark operativo sólido hacia una evaluación más robusta del razonamiento físico de los modelos.

REFERENCIAS BIBLIOGRÁFICAS

Beguš, G., Dąbkowski, M., & Rhodes, R. (2025). Large linguistic models: Investigating LLMs' metalinguistic abilities. IEEE Transactions on Artificial Intelligence, 6(12), 3453–3467. https://doi.org/10.1109/TAI.2025.3575745

Dao, X.-Q., Le, N.-B., Phan, X.-D., Ngo, B.-B., & Vo, T.-D. (2023). Evaluation of ChatGPT and Microsoft Bing AI chat performances on physics exams of Vietnamese National High School Graduation Examination. arXiv. https://doi.org/10.48550/arXiv.2306.04538

Ding, J., Cen, Y., & Wei, X. (2023). Using large language model to solve and explain physics word problems approaching human level. arXiv. https://doi.org/10.48550/arXiv.2309.08182

Gil Pérez, D., & Valdés Castro, P. (1994). La resolución de problemas de física: De los ejercicios de aplicación al tratamiento de situaciones problemáticas. Revista Española de Física, 8(3), 21–26.

Guisasola, J., Zuza, K., Garmendia, M., & Barragués, J. I. (2015). Resolver ejercicios no es fácil: El papel de la metodología científica en la resolución de problemas de física. Revista Brasileira de Ensino de Física, 37(3), 3508. https://doi.org/10.1590/S1806-11173731969

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. The Physics Teacher, 30(3), 141–158. https://doi.org/10.1119/1.2343497

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? Physical Review Physics Education Research, 19(1), 010132. https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

Massachusetts Institute of Technology. (2005). Physics I: Classical Mechanics. MIT OpenCourseWare. https://ocw.mit.edu/courses/8-01-physics-i-classical-mechanics-fall-2005/

Massachusetts Institute of Technology. (2016). Classical Mechanics. MIT OpenCourseWare. https://ocw.mit.edu/courses/8-01sc-classical-mechanics-fall-2016/

Navarro Altamar, S. L. (2006). La resolución de problemas planteada como un proceso de investigación hacia la enseñanza de la física. Prospectiva, 4(2), 60–65. https://www.redalyc.org/articulo.oa?id=496251108010

Perales Palacios, F. J. (1993). La resolución de problemas: Una revisión estructurada. Enseñanza de las Ciencias. Revista de Investigación y Experiencias Didácticas, 11(2), 170–178. https://doi.org/10.5565/rev/ensciencias.4533

Rodríguez Fernández, L., Fernández Mena, A. L., Torres Magaña, M. P., Rodríguez Magaña, M. A., & Rodríguez Fernández, M. A. (2024). Inteligencia artificial en la educación: Modelo de lenguaje de gran tamaño (LLM) como recurso educativo. Revista Ipsumtec, 7(2), 157–164. https://doi.org/10.61117/ipsumtec.v7i2.321

Sato, K. (2024). Exploring the educational landscape of AI: Large language models' approaches to explaining conservation of momentum in physics. arXiv. https://doi.org/10.48550/arXiv.2407.05308

Xu, X., Xu, Q., Xiao, T., Chen, T., Yan, Y., Zhang, J., Diao, S., Yang, C., & Wang, Y. (2025). UGPhysics: A comprehensive benchmark for undergraduate physics reasoning with large language models. Proceedings of Machine Learning Research, 267, 69849–69877.

Zhang, X., Dong, Y., Wu, Y., Huang, J., Jia, C., Fernando, B., Shou, M. Z., Zhang, L., & Liu, J. (2025). PhysReason: A comprehensive benchmark towards physics-based reasoning. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 16593–16615). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.811

Zhao, J., Tong, J., Mou, Y., Zhang, M., Zhang, Q., & Huang, X. (2024). Exploring the compositional deficiency of large language models in mathematical reasoning through trap problems. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 16361–16376). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.915