Imagen Diagnóstica
https://imagendiagnostica.org/

Imadi. 2025, **16**(1), 19–27.

Aumento de herramientas con calculadora obligatoria y mapeo de conceptos erróneos basado en solapamiento para el razonamiento cuantitativo odontológico con modelos de lenguaje de gran tamaño

José Antonio Rodríguez-Martínez*

Radiología e Imagen, Hospital General, SSA, Tlalpan, Ciudad de México, México

(Received: 21 January 2025. Accepted: 09 May 2025. Published online: 30 June 2025.)

Resumen

Antecedentes: Los modelos de lenguaje de gran tamaño (LLMs) han mostrado un rendimiento prometedor en preguntas de conocimiento y comprensión en odontología; sin embargo, los ítems cuantitativos (p. ej., dosificación, conversiones de unidades, estequiometría, razones/ppm y mezcla de materiales) siguen siendo una debilidad constante y un posible riesgo para la seguridad. Además, distintas versiones del modelo suelen compartir las mismas respuestas incorrectas, lo que sugiere modos de error estructurados, similares a conceptos erróneos, en lugar de fallos puramente aleatorios. **Métodos:** Proponemos un marco reproducible centrado en un benchmark seguro en términos de derechos de autor D_q y tres sistemas: un LLM base $\mathcal{M}$, un sistema aumentado con herramientas y uso obligatorio de calculadora $\mathcal{M}$+Calc que debe invocar una calculadora/solucionador simbólico para cada subpaso numérico, y $\mathcal{M}$+Calc+Remed, que amplía $\mathcal{M}$+Calc con diagnósticos de error informados por solapamiento para construir un "mapa de conceptos erróneos en odontología" y activar listas de verificación de validación dirigidas. Definimos exactitud, fiabilidad numérica (valor+unidades) y reducción por tipo de error, y describimos pruebas inferenciales mediante bootstrap pareado/pruebas de McNemar y calibración con teoría de respuesta al ítem (IRT). **Resultados:** Mediante un estudio de simulación Monte Carlo transparente para ilustrar la tubería de análisis, la obligatoriedad del uso de calculadora mejora la exactitud global de 73.7% a 82.0% y reduce los errores de ejecución de 11.1% a 3.6%; añadir remediación basada en solapamiento incrementa la exactitud a 85.4% y disminuye los errores conceptuales a 7.6%. **Conclusiones:** La metodología propuesta separa los fallos conceptuales de los fallos de ejecución, proporciona un procedimiento accionable para mapear conceptos erróneos y motiva una evaluación empírica con ítems cuantitativos odontológicos de autoría abierta para posibilitar una tutoría con IA más segura y una instrucción cuantitativa mejor dirigida.

Palabras clave: educación odontológica; modelos de lenguaje de gran tamaño; razonamiento cuantitativo; herramientas de cálculo; conversión de unidades; estequiometría; teoría de respuesta al ítem; taxonomía de errores; mapeo de conceptos erróneos; seguridad en IA

1. Introducción

El razonamiento cuantitativo en odontología abarca la dosificación de medicamentos y cálculos de concentración, verificaciones de dosis máxima de anestésicos locales, regímenes de fluoruro (p. ej., ppm), proporciones de materiales (p. ej., relaciones polvo–líquido) y química básica/estequiometría relevantes para la ciencia biomédica y de materiales dentales. Estos cálculos aparecen con frecuencia en los planes de estudio preclínicos y en evaluaciones estandarizadas de admisión/licenciamiento. Estudios recientes han informado que los LLMs pueden tener un rendimiento sólido en preguntas de conocimiento y comprensión, mientras muestran menor exactitud en ítems de análisis matemático y presentan alucinaciones o inestabilidad bajo distintas estrategias de prompting [1,5,6]. En contextos críticos para la seguridad, un modelo que "suena confiado" pero maneja mal conversiones de unidades o aritmética de dosificación puede resultar perjudicial.

Este manuscrito propone un programa de investigación completo para (i) *corregir* debilidades cuantitativas mediante aumento de herramientas con uso obligatorio de calculadora y (ii) *diagnosticar* errores residuales usando solapamiento entre modelos para derivar conglomerados tipo concepto erróneo. La idea clave es que muchos fallos cuantitativos son errores de ejecución (deslices aritméticos, manejo incorrecto de unidades, aplicación errónea de fórmulas) más que ausencia de conocimiento conceptual. El aumento con herramientas puede reducir errores de ejecución, mientras que los diagnósticos basados en solapamiento pueden identificar vacíos conceptuales persistentes que permanecen incluso cuando el cálculo es correcto.

*Corresponding author (resumenespepe@gmail.com)

Contribuciones.

1. Definimos un diseño de benchmark y un protocolo de evaluación para un conjunto de datos de odontología cuantitativa D_q, elaborado y revisado para ser seguro en términos de derechos de autor y clínicamente significativo.

2. Formalizamos tres sistemas—$\mathcal{M}$, $\mathcal{M}$+Calc y $\mathcal{M}$+Calc+Remed—y especificamos métricas de exactitud, fiabilidad numérica y reducción por tipo de error.

3. Introducimos el mapeo de conceptos erróneos informado por solapamiento como una forma fundamentada de transformar patrones de error compartidos en objetivos de remediación accionables.

4. Proporcionamos un estudio de simulación transparente que demuestra cómo se reportarían las ganancias esperadas y los diagnósticos, sin afirmar resultados a partir de ítems de examen propietarios.

2. Revisión de la literatura y formulación del problema

Los métodos de inteligencia artificial han avanzado rápidamente en odontología, con aplicaciones que abarcan imagenología, apoyo al diagnóstico y evaluación educativa [2–4]. Dentro de este panorama más amplio, estudios recientes que evalúan sistemas tipo ChatGPT en exámenes odontológicos y tareas de aprendizaje reportan un patrón consistente: buen rendimiento en ítems de conocimiento factual y comprensión, pero menor rendimiento en preguntas cuantitativamente intensivas [1]. Esta debilidad recurrente sugiere que reportar únicamente la exactitud agregada es insuficiente. En su lugar, los marcos de evaluación deben diseñarse para abordar explícitamente modos de fallo conocidos del razonamiento cuantitativo.

Esta necesidad se refuerza con la literatura más amplia sobre alucinación y fiabilidad en modelos de lenguaje de gran tamaño. La alucinación, la sobreconfianza y el razonamiento inestable siguen siendo preocupaciones centrales para el despliegue en dominios de alto riesgo [5]. Enfoques de prompting como chain-of-thought y razonamiento zero-shot pueden mejorar el rendimiento en algunos escenarios [9, 10], mientras que la decodificación por self-consistency puede aumentar la robustez al agregar múltiples trayectorias de razonamiento [11]. Sin embargo, en tareas cuantitativas odontológicas, la fluidez explicativa por sí sola no garantiza corrección. El principal cuello de botella suele ser la fiabilidad numérica, incluyendo ejecución aritmética, manejo de unidades y consistencia entre cálculos intermedios y respuestas finales.

Una dirección prometedora es el aumento con herramientas. Un creciente cuerpo de trabajo muestra que los LLMs pueden mejorar el razonamiento cuantitativo cuando se acoplan con herramientas externas como calculadoras, solucionadores simbólicos o entornos de ejecución de programas [12–15]. Estos estudios motivan un principio de diseño más fuerte orientado a la seguridad en entornos educativos y clínico-adyacentes: siempre que una respuesta requiera cálculo numérico, los subpasos aritméticos deben delegarse a un motor de cómputo confiable en lugar de ser generados puramente por el modelo de lenguaje. En este marco, la obligatoriedad del uso de calculadora no es solo una optimización de exactitud, sino una restricción de fiabilidad.

La medición educativa y la psicometría ofrecen una perspectiva complementaria para analizar errores persistentes del modelo. Los marcos clásicos y modernos de calibración de ítems, incluyendo Rasch y la teoría de respuesta al ítem (IRT), proporcionan herramientas fundamentadas para modelar dificultad y discriminación de ítems [18–20]. En paralelo, la investigación en educación química ha mostrado que los errores de los estudiantes suelen agruparse en torno a conceptos erróneos conceptuales estables, más que ser fallos aleatorios [21]. Extendemos esta idea al razonamiento cuantitativo odontológico utilizando *solapamiento de errores* entre variantes del modelo para inferir conglomerados tipo concepto erróneo. Bajo esta perspectiva, fallos repetidos compartidos por múltiples configuraciones del modelo pueden revelar puntos ciegos conceptuales latentes que requieren remediación dirigida, en lugar de ajuste genérico de prompts.

Motivados por estas líneas de trabajo, formalizamos el razonamiento dental cuantitativo como un problema estructurado de predicción y fiabilidad. Sea

$$D_q = \{(q_i, a_i, s_i)\}_{i=1}^{N}$$

un conjunto de datos de ítems cuantitativos en odontología, donde q_i es el texto de la pregunta (opcionalmente acompañado de campos numéricos estructurados), a_i es la respuesta de referencia (incluyendo valor numérico y unidad cuando corresponda), y $s_i \in \mathcal{S}$ es una etiqueta de habilidad como álgebra, probabilidad, estequiometría, conversión de unidades, o cálculos dentales de dosificación/materiales.

Para un sistema f que asigna a cada ítem q_i una respuesta predicha $\hat{a}_i = f(q_i)$, nuestro objetivo es optimizar el desempeño en cuatro dimensiones acopladas: corrección global en D_q, fiabilidad numérica (valor y unidad correctos simultáneamente), reducción de tipos de error clínicamente y educativamente significativos (p. ej., errores conceptuales, de ejecución y de unidades), y construcción de un mapa de conceptos erróneos informado por solapamiento que priorice fallos conceptuales persistentes para su remediación. Esta formulación desplaza la tarea desde un ejercicio de benchmark de una sola puntuación hacia un marco diagnóstico consciente de la fiabilidad.

Dentro de este marco, estudiamos tres configuraciones de sistema. La primera, $\mathcal{M}$, es un LLM base con prompting estandarizado. La segunda, $\mathcal{M}$+Calc, es un LLM aumentado con herramientas y restringido a invocar una calculadora o solucionador simbólico para cada subpaso numérico. La tercera, $\mathcal{M}$+Calc+Remed, extiende el sistema aumentado con herramientas con diagnósticos informados por solapamiento que activan listas de verificación estructuradas y prompts de remediación dirigidos cuando los patrones de error sugieren un comportamiento persistente tipo concepto erróneo. En conjunto, estos sistemas operacionalizan la hipótesis central de este estudio: mejorar el razonamiento cuantitativo odontológico requiere no solo una generación más fuerte, sino también una fundamentación computacional explícita y mecanismos de retroalimentación diagnóstica.

3. Metodología, análisis teórico y experimentos numéricos

Desarrollamos un marco de evaluación orientado a la fiabilidad para el razonamiento cuantitativo odontológico que integra diseño de benchmark, aumento de herramientas con uso obligatorio de calculadora, diagnósticos de error informados por solapamiento, descomposición teórica del error y un estudio de simulación. La metodología está diseñada para ser legalmente reproducible, diagnósticamente informativa y directamente extensible a futuras evaluaciones en el mundo real una vez que una cadena de herramientas implementada y un benchmark curado estén disponibles.

Un componente central es la construcción del benchmark de odontología cuantitativa D_q. Para garantizar tanto legalidad como reproducibilidad, D_q se ensambla a partir de tres fuentes y procedimientos: enunciados cuantitativos con licencia abierta o de dominio público (cuando estén disponibles) que eviten reproducir contenido de exámenes protegido por derechos de autor; ítems de nueva autoría preparados por docentes de odontología y bioestadísticos, con cada ítem acompañado de una etiqueta de habilidad, una solución trabajada y validación explícita de unidades; y pruebas piloto con estudiantes de odontología para identificar ambigüedades y obtener estimaciones preliminares de dificultad del ítem para una posterior calibración psicométrica usando modelos basados en IRT [19, 20]. Este diseño enfatiza la transparencia y respalda análisis posteriores tanto de exactitud como de estructura del error.

La Tabla 1 resume la cobertura prevista del benchmark a través de habilidades cuantitativas clave relevantes para la odontología. Los conteos deberían finalizarse mediante mapeo curricular, revisión por panel de expertos y análisis psicométrico piloto.

Table 1: Plan de cobertura del benchmark propuesto de odontología cuantitativa D_q. Los conteos se finalizan mediante mapeo curricular y revisión de expertos.

Etiqueta de habilidad	**Ejemplos**	**Relevancia dental (ilustrativa)**
Álgebra	ecuaciones lineales, proporciones	cálculos en consulta; escalado de mezclas
Probabilidad/estadística	probabilidad condicional, sensibilidad/especificidad	razonamiento diagnóstico, métricas de cribado
Estequiometría/química	molaridad, dilución, rendimiento de reacción	química de materiales, preparación de soluciones
Conversión de unidades	mg–g, mL–L, ppm, tiempo/temperatura	ppm de fluoruro, dosificación, proporciones de materiales
Dosificación/materiales dentales	verificación de dosis máxima, dilución, proporciones	anestesia local, regímenes de fluoruro, mezcla de cementos

La intervención central en la canalización de desarrollo del modelo es el aumento de herramientas con

uso obligatorio de calculadora. En lugar de permitir simplemente el uso opcional de herramientas, el sistema se restringe para delegar todos los subpasos numéricos a una calculadora o solucionador simbólico confiable. Este enfoque está conceptualmente alineado con trabajos previos sobre modelos de lenguaje aumentados con herramientas [12,13], pero aquí se operacionaliza como un requisito estricto de fiabilidad para tareas cuantitativas odontológicas. El objetivo es reducir fallos aritméticos y relacionados con unidades, preservando al mismo tiempo el papel del modelo en el análisis del problema, la planificación y la explicación.

El Algoritmo 1 describe el procedimiento de respuesta cuantitativa con uso obligatorio de calculadora para $\mathcal{M}$+Calc. La canalización primero analiza la pregunta en cantidades conocidas, variables objetivo y restricciones de unidades; descompone el problema en subpasos numéricos; normaliza unidades; ejecuta cada operación numérica mediante un motor de cómputo confiable; y finalmente valida magnitudes intermedias y unidades antes de componer la respuesta final.

Algorithm 1 Respuesta cuantitativa con uso obligatorio de calculadora ($\mathcal{M}$+Calc)

Entrada: pregunta q
Analizar q en (i) cantidades conocidas, (ii) objetivo desconocido, (iii) restricciones de unidades.
Descomponer en subpasos $\{g_k\}_{k=1}^{K}$ donde cada g_k es una operación numérica o aplicación de fórmula.
for $k = 1$ to K **do**
    Convertir las unidades a un sistema consistente (si es necesario).
    Llamar a la calculadora/solucionador simbólico para calcular $x_k \leftarrow \mathrm{Calc}(g_k)$.
    Validar unidades intermedias y magnitud (comprobaciones de plausibilidad).
end for
Componer la respuesta final $\hat{a}$ con valor numérico y unidad.
Salida: $\hat{a}$

Para ir más allá de la exactitud agregada, introducimos una taxonomía de errores y un marco de mapeo de conceptos erróneos informado por solapamiento. Los errores se clasifican en tres categorías amplias: errores conceptuales (elección de fórmula incorrecta, supuestos inválidos o mala comprensión del dominio), errores de ejecución (deslices aritméticos, sustitución incorrecta o errores de manipulación algebraica) y errores de unidades (respuestas numéricamente plausibles pero con unidades incorrectas o inconsistentes). Esta taxonomía permite análisis e intervención dirigidos, particularmente en escenarios donde un modelo puede parecer fluido pero aun así producir salidas cuantitativas inseguras.

Para identificar modos de fallo persistentes entre variantes del modelo, definimos una puntuación de solapamiento basada en dos sistemas $f^{(1)}$ y $f^{(2)}$, como diferentes checkpoints o versiones del modelo:

$$\mathrm{Overlap} = \frac{\sum_{i=1}^{N} \mathbb{I}[\hat{a}_i^{(1)} \neq a_i \wedge \hat{a}_i^{(2)} \neq a_i]}{\sum_{i=1}^{N} \mathbb{I}[\hat{a}_i^{(1)} \neq a_i \vee \hat{a}_i^{(2)} \neq a_i]}.$$

Esta cantidad es el solapamiento de Jaccard de los dos conjuntos de error. Un solapamiento alto indica que es más probable que los fallos surjan de debilidades estables y compartidas (estructuras candidatas tipo concepto erróneo) que de variabilidad aleatoria o sensibilidad al prompt.

El sistema mejorado con remediación $\mathcal{M}$+Calc+Remed utiliza esta taxonomía y la señal de solapamiento para activar intervenciones dirigidas. La Tabla 2 resume señales de detección representativas y acciones de remediación. En particular, errores conceptuales repetidos con solapamiento entre variantes activan listas de verificación centradas en conceptos y recuperación de principios, mientras que errores de ejecución y de unidades activan compuertas más estrictas de validación computacional y dimensional.

La evaluación se realiza utilizando un conjunto de métricas consciente de la fiabilidad. Reportamos la exactitud estándar,

$$\mathrm{Acc} = \frac{1}{N} \sum_{i} \mathbb{I}[\hat{a}_i = a_i],$$

así como la fiabilidad numérica (que exige tanto el valor correcto como la unidad correcta cuando corresponda), las proporciones por tipo de error (conceptual, de ejecución y de unidad) en un subconjunto anotado por expertos, y la robustez frente a perturbaciones del prompt, como paráfrasis o cambios de formato. Para comparaciones inferenciales, utilizamos intervalos de confianza bootstrap pareados y pruebas de McNemar para resultados de exactitud pareados, e incorporamos modelos IRT (2PL/3PL, según corresponda) para estimar

Table 2: Taxonomía de errores y acciones de remediación dirigidas utilizadas en $\mathcal{M}$+Calc+Remed.

Clase de error	**Señal de detección**	**Acción de remediación (ejemplos)**
Conceptual	familia de fórmulas incorrecta; solapamiento repetido en el concepto	imponer lista de verificación conceptual; solicitar análisis dimensional; recuperar principio relevante; re-derivar fórmula
Ejecución	aritmética intermedia inconsistente; discrepancia con la calculadora	exigir llamada a calculadora en cada paso; mostrar valores intermedios; volver a verificar sustituciones
Unidad	magnitud plausible pero unidad inconsistente	normalización estricta de unidades; análisis dimensional; compuerta final de validación de unidades

la dificultad de los ítems y comparar parámetros de habilidad a nivel de sistema [18–20]. Esta estrategia inferencial combinada respalda tanto comparaciones pareadas clásicas como una interpretación en términos de rasgo latente.

El análisis teórico aclara cómo las intervenciones propuestas apuntan a canales de error distintos. Sean E_C, E_E y E_U los eventos de error conceptual, de ejecución y de unidad, respectivamente. Si una respuesta final es correcta solo cuando ninguno de estos ocurre, entonces

$$\Pr(\text{Correct}) = 1 - \Pr(E_C \cup E_E \cup E_U).$$

Aplicando la cota de la unión se obtiene la desigualdad conservadora

$$\Pr(\text{Correct}) \geq 1 - (\Pr(E_C) + \Pr(E_E) + \Pr(E_U)) .$$

Esta descomposición hace explícita la lógica de intervención: se espera que la obligatoriedad del uso de calculadora reduzca $\Pr(E_E)$ directamente y $\Pr(E_U)$ de manera secundaria mediante normalización de unidades y verificaciones dimensionales, mientras que la remediación informada por solapamiento está destinada a reducir componentes persistentes de $\Pr(E_C)$. Así, el marco no trata todos los errores como homogéneos; asigna cada intervención al mecanismo de error que está diseñada para suprimir.

La puntuación de solapamiento también puede interpretarse teóricamente como un proxy de estructura conceptual latente. Supóngase que un subconjunto de ítems está influido por un factor de concepto erróneo no observado Z que incrementa la probabilidad de error conceptual en múltiples variantes del modelo. Entonces, a medida que aumentan la varianza de Z y las cargas específicas de cada variante sobre Z, los fallos compartidos entre variantes se vuelven más probables, y el solapamiento de errores observado debería aumentar en consecuencia. En la práctica, el agrupamiento basado en solapamiento proporciona por tanto una hipótesis falsable: los ítems que fallan conjuntamente entre variantes deberían estar enriquecidos en modos de fallo conceptual compartidos y deberían beneficiarse de forma desproporcionada de una remediación centrada en conceptos, en lugar de correcciones únicamente aritméticas.

Debido a que el contenido de exámenes propietarios no puede reproducirse y a que un estudio empírico completo requiere una pila de herramientas implementada y un benchmark curado D_q, incluimos una simulación Monte Carlo para demostrar la canalización de análisis prevista y el formato de reporte. Estas simulaciones explícitamente no constituyen afirmaciones sobre ningún modelo comercial en particular. Su propósito es metodológico: mostrar cómo se descompondrían las mejoras en reducciones por tipo de error, cómo se reportarían las estadísticas de solapamiento y cómo se presentarían las comparaciones entre sistemas en un marco reproducible.

En la simulación, generamos $N = 2000$ ítems en cinco categorías de habilidades con dificultad conceptual latente y complejidad de ejecución. Al sistema base $\mathcal{M}$ se le asignan propensiones no triviales a errores conceptuales, de ejecución y de unidad. El sistema con calculadora obligatoria $\mathcal{M}$+Calc reduce errores de ejecución y de unidad, mientras que $\mathcal{M}$+Calc+Remed reduce además errores conceptuales, especialmente en ítems más difíciles, imitando el efecto previsto de la remediación guiada por solapamiento. La Figura 1 proporciona un esquema del flujo completo de evaluación, desde la construcción del benchmark hasta el análisis inferencial y el mapeo de conceptos erróneos.

Los resultados globales se resumen en la Tabla 3. El patrón simulado concuerda con la hipótesis de diseño: $\mathcal{M}$+Calc mejora sustancialmente la exactitud con respecto a la línea base principalmente mediante reducciones

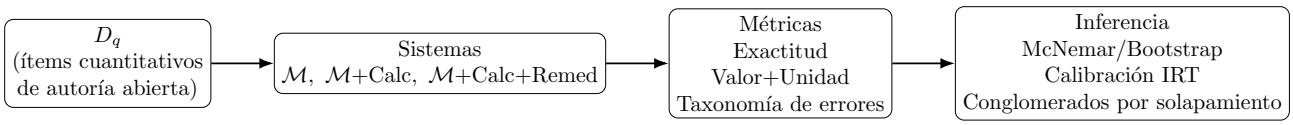

Figure 1: Canalización de evaluación propuesta: construcción del benchmark → evaluación del sistema → cálculo de métricas → análisis inferencial y mapeo de conceptos erróneos.

en errores de ejecución y de unidad, mientras que $\mathcal{M}$+Calc+Remed produce ganancias adicionales al reducir errores conceptuales. Cabe destacar que el modelo mejorado con remediación puede mostrar una pequeña redistribución de errores de ejecución en algunas simulaciones porque las intervenciones centradas en conceptos pueden alterar las trayectorias de solución y exponer ramas computacionales diferentes, lo que subraya el valor de la descomposición por tipo de error en lugar de depender solo de la exactitud.

Table 3: Desempeño global (simulación). "Error de unidad" denota una unidad incorrecta dada una computación numérica por lo demás correcta.

Sistema	**Exactitud (%)**	**Error conceptual (%)**	**Error de ejec. (%)**	**Error de unidad (%)**
$\mathcal{M}$ (línea base)	73.7	11.8	11.1	5.0
$\mathcal{M}$+Calc	82.0	10.9	3.6	3.8
$\mathcal{M}$+Calc+Remed	85.4	7.6	4.5	3.2

Los resultados por habilidad en la Tabla 4 ilustran además cómo los beneficios de la intervención pueden variar según el tipo de tarea. En esta simulación, el álgebra exhibe un alto rendimiento de línea base y ganancias modestas, mientras que probabilidad, estequiometría, conversión de unidades y tareas de dosificación/materiales dentales muestran mejoras mayores bajo aumento con herramientas y remediación. Este patrón heterogéneo es consistente con la expectativa de que los dominios que combinan interpretación conceptual con cálculo de múltiples pasos y gestión de unidades son los que más probablemente se beneficien del marco propuesto.

Table 4: Exactitud por categoría de habilidad (simulación).

Categoría de habilidad	$\mathcal{M}$	$\mathcal{M}$+Calc	$\mathcal{M}$+Calc+Remed
Álgebra	91.8%	92.0%	96.2%
Probabilidad	76.2%	85.5%	87.0%
Estequiometría	65.8%	76.0%	78.5%
Conversión de unidades	71.5%	80.0%	84.2%
Dosificación/materiales dentales	63.2%	76.8%	80.8%

Por último, simulamos dos variantes del modelo con errores correlacionados para ilustrar el reporte de solapamiento. En este ejemplo, el solapamiento de Jaccard de los dos conjuntos de error es 31.2%, y Pr(Variant 2 wrong | Variant 1 wrong) = 46.9%. Estas estadísticas demuestran cómo el solapamiento de errores entre variantes puede cuantificar estructura de fallo compartida más allá de las tasas marginales de error por sí solas, motivando así la remediación informada por solapamiento y los análisis de agrupamiento en futuros despliegues empíricos. La Tabla 5 presenta el resumen.

Table 5: Solapamiento de errores entre variantes (simulación).

Estadístico	**Valor**	**Interpretación**
Tasa de error de la Variante 1	26.6%	fracción incorrecta
Tasa de error de la Variante 2	25.9%	fracción incorrecta
Solapamiento de Jaccard de conjuntos de error	31.2%	errores compartidos / unión de errores
Pr(Variant 2 wrong | Variant 1 wrong)	46.9%	solapamiento condicional

4. Resultados, discusión y limitaciones

La simulación ilustrativa resalta un patrón que se espera que sea comprobable en futuros estudios empíricos sobre el benchmark D_q. El efecto más inmediato es una fuerte supresión de errores de ejecución bajo la obligatoriedad del uso de calculadora, con reducciones claras en errores aritméticos y de sustitución en comparación con el sistema de línea base (Tabla 3). Esto respalda la premisa central de diseño del marco: cuando los subpasos numéricos se delegan a un motor de cómputo confiable, una porción sustancial del fallo cuantitativo evitable puede eliminarse sin requerir que el propio modelo de lenguaje se vuelva intrínsecamente más preciso numéricamente.

Un segundo patrón notable es que las mayores ganancias absolutas ocurren en categorías clínica y educativamente relevantes como dosificación/materiales dentales y conversión de unidades (Tabla 4). Estos son precisamente los entornos en los que la normalización de unidades, las comprobaciones de consistencia dimensional y las restricciones explícitas de cómputo son más trascendentales. En términos prácticos, la simulación sugiere que el aumento con herramientas orientado a la fiabilidad puede proporcionar un beneficio desproporcionado en dominios donde errores numéricos o de unidades aparentemente pequeños pueden tener consecuencias posteriores desmedidas.

Al mismo tiempo, los resultados también indican brechas conceptuales residuales persistentes. Incluso después de la obligatoriedad del uso de calculadora, permanecen tasas no triviales de error conceptual, especialmente en problemas de estequiometría y de dosificación/materiales dentales que requieren selección correcta de fórmulas, interpretación del dominio o una formulación adecuada antes de realizar cualquier aritmética (Tabla 3). Esto refuerza la motivación para el mapeo de conceptos erróneos basado en solapamiento y la remediación conceptual dirigida: los mecanismos de corrección numérica por sí solos son necesarios pero no suficientes para un razonamiento cuantitativo odontológico robusto.

De forma más amplia, este estudio impulsa una agenda de investigación que desplaza la pregunta de evaluación desde "¿Qué tan bien puntúa un LLM?" hacia "¿Qué modos de fallo permanecen después de imponer un cómputo confiable, y cómo pueden mitigarse de forma segura?" En este marco, el aumento con herramientas no es simplemente una estrategia de optimización del rendimiento; funciona como una restricción orientada a la seguridad, alineada con preocupaciones de larga data sobre alucinación, sobreconfianza y fiabilidad en modelos de lenguaje [5]. Para tareas cuantitativas, los riesgos más relevantes incluyen deslices aritméticos, fallos en conversión de unidades y selección injustificada de fórmulas. La obligatoriedad del uso de calculadora y las compuertas de validación de unidades abordan directamente las dos primeras clases, mientras que los diagnósticos informados por solapamiento ofrecen una vía estructurada para identificar y priorizar debilidades conceptuales más profundas.

La perspectiva basada en solapamiento también puede tener valor más allá del benchmarking de modelos. Desde un punto de vista educativo, los conglomerados de error tipo concepto erróneo inferidos a partir de fallos compartidos entre variantes del modelo pueden ayudar a revelar cuellos de botella curriculares donde tanto estudiantes como sistemas de IA tienen dificultades, reflejando el papel de los inventarios de conceptos erróneos en educación química [21]. Para programas de odontología, tales conglomerados podrían informar módulos de remediación dirigidos, rediseño de ítems y evaluaciones más equilibradas. Para desarrolladores de modelos, esos mismos conglomerados pueden guiar ajuste fino específico del dominio, verificación aumentada por recuperación o estrategias de prompting conscientes del concepto que aborden déficits conceptuales recurrentes en lugar de errores aislados.

Deben enfatizarse varias limitaciones. En primer lugar, la evaluación empírica completa de $\mathcal{M}$, $\mathcal{M}$+Calc y $\mathcal{M}$+Calc+Remed depende de una canalización de herramientas implementada y de un benchmark curado de autoría abierta D_q; en consecuencia, este manuscrito presenta un protocolo completo y simulaciones ilustrativas, en lugar de resultados sobre materiales de examen propietarios. En segundo lugar, la inferencia de conceptos erróneos a partir del solapamiento depende inherentemente de la elección de variantes del modelo, prompts y condiciones de evaluación. Por tanto, el solapamiento de errores se interpreta mejor como una señal generadora de hipótesis y no como evidencia definitiva, y requiere revisión experta y validación del dominio. En tercer lugar, la obligatoriedad del uso de calculadora mitiga principalmente errores de ejecución y algunos errores de unidad, pero no garantiza un encuadre conceptual correcto; un modelo aún puede producir una respuesta numéricamente

consistente pero clínicamente incorrecta si los supuestos subyacentes o las fórmulas son erróneos. Por último, cualquier despliegue en el mundo real en entornos educativos debe abordar preocupaciones de gobernanza, incluyendo la sobredependencia del usuario y el posible uso indebido en contextos de evaluación. Por ello, las salvaguardas técnicas deben complementarse con políticas institucionales claras, límites de uso y prácticas de integración responsables.

5. Conclusión

Este estudio presenta un marco centrado en la fiabilidad para mejorar y evaluar el razonamiento cuantitativo en aplicaciones odontológicas con LLMs combinando aumento de herramientas con uso obligatorio de calculadora, validación explícita de unidades y diagnósticos de conceptos erróneos informados por solapamiento. En lugar de tratar el rendimiento como una única puntuación de exactitud, el enfoque propuesto descompone los errores en componentes conceptuales, de ejecución y relacionados con unidades, vinculando así intervenciones específicas con los modos de fallo que están destinadas a reducir. Dentro de este marco, la obligatoriedad del uso de calculadora funciona como una restricción orientada a la seguridad para subpasos numéricos, mientras que el análisis de errores basado en solapamiento proporciona un mecanismo fundamentado para identificar debilidades conceptuales persistentes que requieren remediación dirigida.

La simulación ilustrativa demuestra el comportamiento cualitativo esperado del marco: supresión sustancial de errores de ejecución bajo uso obligatorio de herramientas, ganancias marcadas en categorías clínicamente relevantes como dosificación/materiales y conversión de unidades, y errores conceptuales residuales persistentes que motivan remediación centrada en conceptos. Estos hallazgos, aunque simulados, muestran cómo los futuros estudios empíricos pueden reportar mejoras de manera transparente y diagnósticamente significativa en lugar de depender únicamente de puntuaciones agregadas de benchmark.

En conjunto, la metodología propuesta ofrece una base práctica y extensible para benchmarking legalmente reproducible (D_q), inferencia cuantitativa más segura y análisis de errores educativamente significativos en sistemas LLM orientados a odontología. El trabajo futuro debería implementar la canalización completa de herramientas, construir y validar el benchmark de autoría abierta con revisión de expertos y calibración psicométrica, y probar si la remediación informada por solapamiento mejora el rendimiento real de los modelos y la utilidad educativa en entornos auténticos de aprendizaje odontológico.

Referencias

[1] M. Dashti, S. Ghasemi, N. Ghadimi, *et al.* Performance of ChatGPT 3.5 and 4 on U.S. dental examinations: the INBDE, ADAT, and DAT. *Imaging Science in Dentistry*, 54 (2024) 271–275.

[2] F. Schwendicke, J. Krois. Data Dentistry: How Data Are Changing Clinical Care and Research. *Journal of Dental Research*, 101(1) (2022) 21–29.

[3] J. Sur, S. Bose, F. Khan, D. Dewangan, E. Sawriya, A. Roul. Knowledge, attitudes, and perceptions regarding the future of artificial intelligence in oral radiology in India: A survey. *Imaging Science in Dentistry*, 50(3) (2020) 193–198.

[4] K.-H. Yu, A. L. Beam, I. S. Kohane. Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2(10) (2018) 719–731.

[5] Z. Ji, N. Lee, R. Frieske, *et al.* Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12) (2023) Article 248.

[6] OpenAI. GPT-4 Technical Report. *arXiv:2303.08774*, 2023.

[7] L. Ouyang, J. Wu, X. Jiang, *et al.* Training language models to follow instructions with human feedback. *NeurIPS*, 35 (2022).

[8] T. B. Brown, B. Mann, N. Ryder, *et al.* Language models are few-shot learners. *NeurIPS*, 33 (2020) 1877–1901.

[9] J. Wei, X. Wang, D. Schuurmans, *et al.* Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35 (2022).

[10] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa. Large language models are zero-shot reasoners. *arXiv:2205.11916*, 2022.

[11] X. Wang, J. Wei, D. Schuurmans, *et al.* Self-consistency improves chain of thought reasoning in language models. *arXiv:2203.11171*, 2022.

[12] T. Schick, J. Dwivedi-Yu, R. Dessì, *et al.* Toolformer: Language models can teach themselves to use tools. *arXiv:2302.04761*, 2023.

[13] L. Gao, A. Madaan, S. Zhou, *et al.* PAL: Program-aided language models. *arXiv:2211.10435*, 2022.

[14] K. Cobbe, V. Kosaraju, M. Bavarian, *et al.* Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

[15] A. Lewkowycz, A. Andreassen, D. Dohan, *et al.* Solving quantitative reasoning problems with language models. *NeurIPS*, 35 (2022).

[16] D. Hendrycks, C. Burns, S. Kadavath, *et al.* Measuring mathematical problem solving with the MATH dataset. *arXiv:2103.03874*, 2021.

[17] M. Suzgun, N. Scales, N. Schärli, *et al.* Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv:2210.09261*, 2022.

[18] G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests.* Danish Institute for Educational Research, Copenhagen, 1960.

[19] S. E. Embretson, S. P. Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.
[20] R. J. de Ayala. *The Theory and Practice of Item Response Theory*. Guilford Press, New York, 2009.
[21] M. B. Nakhleh. Why some students don't learn chemistry: chemical misconceptions. *Journal of Chemical Education*, 69(3) (1992) 191–196.
[22] D. R. Mulford, W. R. Robinson. An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education*, 79(6) (2002) 739–744.
[23] E. Yüzbaşıoğlu. Attitudes and perceptions of dental students towards artificial intelligence. *Journal of Dental Education*, 85(1) (2021) 60–68.
[24] R. J. Weyant, S. L. Tracy, T. Anselmo, *et al.* Topical fluoride for caries prevention: Executive summary of evidence-based clinical recommendations. *Journal of the American Dental Association*, 144(11) (2013) 1279–1291.
[25] S. F. Malamed. *Handbook of Local Anesthesia*. 7th ed., Elsevier, 2019.
[26] A. J. Feilzer, A. J. de Gee, C. L. Davidson. Setting stress in composite resin in relation to configuration of the restoration. *Journal of Dental Research*, 71(10) (1992) 1697–1703.
[27] F. Schwendicke, J. Krois. Data-driven dentistry: How artificial intelligence is changing dentistry. *Journal of Dental Research*, 100(4) (2021) 331–339.
[28] E. Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books, 2019.