

DOCUMENTO PARA MEDIOS. Edición 1. Costa Rica. 28 de setiembre de 2026.

Publicado el 29 de setiembre de 2026. Autores: Sebastián Urbina Cañas y Roberto Echeverría Cardenas.

Hallazgos y guía de lectura del estudio Sextante

Este anexo reúne 24 hallazgos sobre seis hospitales privados de Costa Rica, con fuentes y notas de interpretación. Incluye la ficha técnica y una guía de conceptos para interpretar los resultados. Los resultados describen presencia y exactitud en la IA; no miden calidad clínica, decisiones de pacientes ni participación de mercado.

G1 a G5 identifican las tablas del anexo de datos y videos en PDF. Los porcentajes de menciones, respuestas y afirmaciones tienen bases distintas; los puntajes sobre 100 no son porcentajes. Las referencias al informe corresponden a la edición general en español del 28 de setiembre de 2026, de 15 páginas.

Concentración de menciones

1. Dos de seis hospitales concentran el 64,4 % de las menciones. CIMA concentra el 32,5 % y Clínica Bíblica el 31,9 %, para un total de **64,4 % de las menciones entre los seis hospitales**. Esta concentración describe cómo se distribuyen las menciones entre las instituciones del panel, sin equivaler a participación de mercado ni a preferencia de los pacientes.

Fuente: G2.

2. La concentración es todavía mayor al aparecer primero. Cuando el primer hospital nombrado es uno de los seis evaluados, **el 93,5 % de esas primeras menciones corresponde a CIMA o Clínica Bíblica**. Sobre el total de preguntas abiertas, ambos suman una tasa de primera mención del 72,2 %. Son dos bases distintas para entender la prominencia de esas instituciones.

Fuente: G2 e informe general, páginas 5 y 13.

3. Ser mencionado con frecuencia no asegura encabezar la respuesta. CIMA aparece en el **84,9 %** de las preguntas abiertas y Clínica Bíblica en el **83,5 %**, una presencia muy cercana. Sin embargo, sus tasas de primera mención son **49,4 % y 22,8 %**, respectivamente. La diferencia muestra por qué importa medir también el orden en que aparecen.

Fuente: Informe general, página 13.

4. La clasificación presenta una brecha de 19,8 puntos entre el segundo y el tercer puesto. En el índice principal, correspondiente a configuraciones con búsqueda web habilitada, CIMA obtiene **79,7 puntos**, Clínica Bíblica **76,7**, La Católica **56,9**, Metropolitano **50,8**, UNIBE **41,6** y San Rafael Arcángel **39,1**, sobre 100. La distancia entre posiciones ayuda a dimensionar la representación de cada hospital en este conjunto de respuestas.

Fuente: G1.

5. La Católica y Metropolitano aparecen en menos de la mitad de las preguntas abiertas. Sus tasas de mención espontánea son **47,7 % y 38,1 %**, respectivamente. Ambas instituciones forman parte de las opciones que presentan los modelos, pero con menor frecuencia que los dos primeros hospitales. Para sus equipos, el dato señala una diferencia concreta de exposición.

Fuente: Informe general, página 13.

Reconocimiento y exactitud

6. Los modelos pueden reconocer a UNIBE sin incluirlo espontáneamente entre las opciones. UNIBE aparece en el 5,4 % de las preguntas abiertas, pero alcanza 87,6 puntos sobre 100 en reconocimiento cuando se pregunta por su nombre. Son indicadores distintos: una descripción al preguntar por nombre y una mención en una consulta abierta muestran formas diferentes de presencia. El estudio no evaluó intervenciones para convertir una en otra.

Fuente: Informe general, página 13.

7. San Rafael Arcángel muestra una brecha similar entre reconocimiento y aparición espontánea. El hospital de Liberia figura en el 1,9 % de las preguntas abiertas y obtiene 85,3 puntos de reconocimiento directo. El resultado ilustra un desafío para la visibilidad de esta institución fuera de las búsquedas que ya incluyen su nombre; no permite generalizar a todos los hospitales de Guanacaste.

Fuente: Informe general, página 13.

8. CIMA registra la mayor proporción de afirmaciones verificables correctas del panel. El 99,2 % de 6.132 afirmaciones calificadas sobre el hospital es correcto. En una medida diferente, el 1,1 % de 4.027 respuestas calificadas contiene al menos un error. La distinción importa: una respuesta puede incluir varias afirmaciones.

Fuente: G3.

9. Más de una de cada cuatro respuestas calificadas sobre Metropolitano contiene algún error. La proporción alcanza el 28,7 % de 1.202 respuestas calificadas. Son errores de los modelos sobre datos institucionales, como acreditación, operación, ubicación o propiedad. El hallazgo permite orientar una revisión de la información pública que recibe quien consulta sobre el hospital.

Fuente: G3.

10. En La Católica, casi una de cada cuatro respuestas calificadas incluye información incorrecta. El estudio identifica al menos un error en el 23,5 % de 1.690 respuestas calificadas. Este indicador mide errores de la IA al describir la institución y no errores médicos ni resultados de atención del hospital.

Fuente: G3.

11. Una presencia reducida puede coexistir con información mayoritariamente correcta. UNIBE alcanza 93,7 % de afirmaciones correctas sobre una base de 508, y San Rafael Arcángel 93,1 % sobre 850. Junto con sus bajas tasas de aparición espontánea, estos datos muestran que frecuencia de aparición y exactitud describen dimensiones diferentes de la representación hospitalaria.

Fuente: G3.

Control y diferencias entre configuraciones

12. **Las nueve configuraciones describieron como abierto un hospital cerrado desde 2018.** El hospital se incluyó como control y fue presentado como operativo en alguna respuesta de **las nueve configuraciones evaluadas**. La frecuencia osciló entre **1,4 % y 5,9 % de las 1.568 respuestas de cada configuración**. El caso evidencia la necesidad de comprobar información tan básica como si un establecimiento sigue funcionando. El informe fuente no identifica al establecimiento de control; su identificación y los registros del caso deben solicitarse a la dirección técnica.

Fuente: G4 e informe, página 9.

13. **La frecuencia de errores cambia ampliamente entre configuraciones.** El porcentaje de respuestas calificadas con al menos un error va de **7,6 % en GPT-5.6 Luna con búsqueda habilitada**, sobre 1.322 respuestas calificadas, a **40,4 % en DeepSeek V4 Flash sin búsqueda**, sobre 997. Es una comparación dentro de las condiciones del estudio, no una clasificación universal de proveedores.

Fuente: G4.

14. **En GPT-5.6 Luna, la configuración con búsqueda presenta menos errores.** Las respuestas calificadas con al menos un error pasan de **20,1 % sin búsqueda a 7,6 % con búsqueda**, una diferencia de **12,5 puntos porcentuales**. Las bases son 617 y 1.322 respuestas calificadas, respectivamente. El resultado muestra cuánto puede variar la información según la configuración consultada. Las bases calificadas difieren entre condiciones; esta observación no demuestra que habilitar búsqueda cause una reducción equivalente.

Fuente: G4.

15. **Habilitar búsqueda web no elimina automáticamente los errores.** En Gemini 3.6 Flash, el indicador de respuestas calificadas con algún error es **21,4 % sin búsqueda y 22,8 % con búsqueda**, sobre 1.474 y 1.472 respuestas calificadas. La diferencia es pequeña y requiere cautela por la incertidumbre de la clasificación. El hallazgo relevante es que tener acceso a la web no garantiza que toda la información sea correcta.

Fuente: G4.

16. **La Católica y Metropolitano intercambian posiciones según el modo de consulta.** Sin búsqueda, Metropolitano obtiene **56,3 puntos** y La Católica **55,5**. Con búsqueda habilitada, La Católica alcanza **56,9** y Metropolitano **50,8**. La comparación muestra que una sola modalidad de consulta puede ofrecer una lectura incompleta de la presencia institucional. La ventaja de 0,8 puntos sin búsqueda es pequeña y no debe presentarse como concluyente frente a la incertidumbre del método.

Fuente: Informe, páginas 4 y 6.

17. **UNIBE presenta el mayor aumento del índice entre las condiciones comparadas.** Pasa de **37,3 puntos sin búsqueda a 41,6 con búsqueda**, una diferencia de **4,3 puntos**. Es una señal para investigar cómo cambia su representación al disponer de información web. La comparación no identifica por sí sola qué página o contenido explica la diferencia.

Fuente: Informe, páginas 4 y 6.

Variaciones por condición e idioma

18. **San Rafael Arcángel también obtiene un puntaje mayor con búsqueda habilitada.** Pasa de 35,1 a 39,1 puntos, una diferencia de 4,0 puntos. Para el hospital, este contraste ayuda a separar lo que responden los modelos desde su entrenamiento de lo que se observa cuando pueden consultar la web.

Fuente: Informe, páginas 4 y 6.

19. **Metropolitano registra el mayor descenso entre esas dos condiciones.** Su índice pasa de 56,3 puntos sin búsqueda a 50,8 con búsqueda, una diferencia de 5,5 puntos. El dato justifica revisar las respuestas y sus fuentes antes de decidir qué corregir. No demuestra que una acción específica del hospital haya causado la variación.

Fuente: Informe, páginas 4 y 6.

Nota de comparabilidad. Los hallazgos 20 a 23 proceden de G5. La fuente publica sus denominadores, pero no precisa la composición por configuraciones que explicaría su diferencia con G3. Se informan como un corte separado, sin agregarlos, compararlos directamente con G3 ni atribuir causas.

20. **La información sobre La Católica es menos exacta en inglés en el corte por idioma.** Las afirmaciones correctas representan 88,0 % en español y 79,1 % en inglés, una diferencia de 8,9 puntos porcentuales, sobre 1.494 y 1.917 afirmaciones calificadas. Es un hallazgo relevante para revisar la información que puede encontrar una persona que consulta en inglés sobre atención en Costa Rica.

Fuente: G5.

21. **UNIBE presenta una diferencia de 6,8 puntos porcentuales por idioma.** La proporción de afirmaciones correctas es 95,0 % en español y 88,2 % en inglés, sobre 337 y 398 afirmaciones calificadas. El resultado respalda la necesidad de evaluar la información institucional en ambos idiomas, especialmente cuando se busca atender a pacientes internacionales.

Fuente: G5.

22. **Metropolitano también muestra menor exactitud en inglés.** El corte por idioma registra 76,1 % de afirmaciones correctas en español y 72,2 % en inglés, sobre 1.341 y 1.508 afirmaciones calificadas. La diferencia es de 3,9 puntos porcentuales. El estudio identifica la brecha observada, pero no establece su causa ni cuantifica su impacto comercial.

Fuente: G5.

23. **La diferencia por idioma no es igual para todos los hospitales.** CIMA registra 99,1 % de afirmaciones correctas tanto en español como en inglés, y Clínica Bíblica 96,7 % en ambos idiomas, al nivel de redondeo publicado. Las bases son 4.493 y 5.902 afirmaciones para CIMA, y 3.950 y 5.247 para Bíblica. Esto refuerza la utilidad de analizar cada institución por separado.

Fuente: G5.

Presencia por servicios y cuadro de resultados

24. El puntaje general puede ocultar diferencias importantes entre procedimientos. En la matriz general de líneas de servicio, CIMA obtiene 77 puntos en las consultas sobre aumento mamario y 75 en reemplazo total de rodilla. Clínica Bíblica alcanza 73 en angioplastia electiva; La Católica, 55 en reemplazo de rodilla, y Metropolitano, 52 en manga gástrica. Son los valores redondeados publicados en esa matriz. Describen representación en IA por procedimiento y no resultados clínicos.

Fuente: Informe general, página 10.

Índice principal. Corresponde a las cinco configuraciones con búsqueda web habilitada. Combina presencia (40 %), exactitud y especificidad de la descripción (30 %) y reputación expresada por los modelos (30 %). El componente de exactitud del índice es distinto del porcentaje de afirmaciones verificables correctas.

Hospital	Índice sobre 100
Hospital CIMA San José	79,7
Hospital Clínica Bíblica	76,7
Hospital La Católica	56,9
Hospital Metropolitano	50,8
Hospital UNIBE	41,6
Hospital San Rafael Arcángel (Liberia)	39,1

Fuente: G1. Puntajes de representación en IA sobre 100. No son calificaciones de calidad clínica.

Dos unidades distintas para describir errores. Una respuesta puede contener varias afirmaciones. El porcentaje de respuestas con algún error y el porcentaje de afirmaciones correctas no son complementarios ni suman necesariamente 100.

Hospital	Respuestas con error %	N respuestas	Afirmaciones correctas %	N afirmaciones
CIMA San José	1,1	4.027	99,2	6.132
Clínica Bíblica	4,5	3.664	96,8	5.325
La Católica	23,5	1.690	80,6	2.048
Metropolitano	28,7	1.202	74,7	1.456
UNIBE	7,7	416	93,7	508
San Rafael Arcángel	10,0	588	93,1	850

Fuente: G3. N de respuestas se refiere a respuestas calificadas con afirmaciones verificables. N de afirmaciones se refiere a afirmaciones verificables calificadas. Una respuesta puede mencionar a más de un hospital; no se suman las bases por hospital para obtener el tamaño del corpus.

Ficha técnica y límites de interpretación

Alcance y fechas. 14.112 respuestas, nueve configuraciones de seis proveedores, español e inglés, generadas el 16 y el 17 de agosto de 2026. Banco de 162 ítems activos, fijado el 14 de agosto, que produjo 1.568 instancias por configuración con repeticiones. La relación detallada entre ítems e instancias no se desglosa en los archivos agregados.

Universo. CIMA San José, Clínica Bíblica, La Católica, Metropolitano, UNIBE y San Rafael Arcángel en Liberia. La dirección técnica seleccionó hospitales privados con presencia relevante en IA para esta edición. Es un panel definido, no un censo del sistema ni una muestra probabilística. No permite afirmar que otras instituciones nunca aparezcan.

Configuraciones. GPT-5.6 Luna, Gemini 3.6 Flash y Claude Sonnet 5 se evaluaron con y sin búsqueda. Sonar y Llama 4 Maverick, con búsqueda; DeepSeek V4 Flash, sin búsqueda. Tener búsqueda habilitada no implica que se usara en cada respuesta. Se evaluaron modelos mediante interfaces técnicas; las aplicaciones de consumo pueden incorporar personalización y otras funciones.

Verificación y codificación. Un modelo evaluador codificó las respuestas con un esquema fijo. Se contrastaron acreditación, operación de sedes, ubicación y propiedad con fichas basadas en fuentes oficiales. Cerca del 42 % de las afirmaciones extraídas, sobre precios, servicios o resultados, quedó fuera de esas tasas por no contar con una referencia verificable dentro del alcance.

Revisión institucional. Los hospitales pudieron revisar sus fichas del 7 al 21 de setiembre. Según la confirmación de la dirección técnica, ninguno corrigió datos. Esa ausencia de correcciones no demuestra que todos respondieran ni supone su aprobación del estudio.

Incertidumbre. La recodificación de respuestas idénticas coincidió entre el 72 % y el 76 % de las afirmaciones individuales. Los intervalos al 90 % proceden de 1.000 remuestreos y describen variación dentro de la muestra; no incorporan esa incertidumbre del evaluador. Los seis puestos se conservaron en esos remuestreos. Las diferencias pequeñas requieren cautela; no se garantiza el orden de una medición futura.

Condición de Anthropic con búsqueda. Se ejecutó mediante una herramienta para desarrolladores y parte de las respuestas conserva su encuadre. El informe señala que los registros afectados están identificados. Esa condición requiere cautela al comparar configuraciones. La condición sin búsqueda se regeneró en la API estándar.

Causas e impacto. Las diferencias por idioma o configuración describen observaciones. El estudio no determina su causa ni evalúa una intervención hospitalaria. No mide cuántos pacientes usan IA, si cambian de hospital, si sufren un daño o si una institución gana o pierde ingresos.

Fuentes y trazabilidad. Informe general de 15 páginas, edición en español del 28 de setiembre de 2026 (metodología, páginas 12 y 13; límites, página 14; reproducibilidad, página 15), y tablas G1 a G5 del anexo de datos y videos en PDF. El equipo conserva respuestas y codificaciones; las tablas publicadas contienen datos agregados, no todo el corpus.

Contacto para verificar una cifra o solicitar soporte de un caso: prensa@indicesextante.com.
Metodología: <https://indicesextante.com/metodologia>

Modelos de lenguaje y asistentes de inteligencia artificial

Qué es un LLM. La sigla inglesa corresponde a modelo de lenguaje de gran escala. Es un sistema entrenado para procesar lenguaje y generar respuestas a partir de una consulta y del contexto que recibe. La fluidez de una respuesta no demuestra por sí sola que sus afirmaciones sean correctas.

Modelo y asistente no son sinónimos. El modelo genera las respuestas. Un asistente es la aplicación que lo pone a disposición del usuario y puede añadir búsqueda, memoria, personalización u otras herramientas. Sextante consultó modelos mediante interfaces técnicas. Sus resultados no reproducen todas las experiencias posibles en aplicaciones de consumo.

Qué significa una configuración. Es una condición concreta de evaluación. Consultar el mismo modelo con búsqueda web habilitada o sin ella produce dos configuraciones. Esta edición comparó seis modelos en nueve configuraciones: cinco con búsqueda disponible y cuatro sin ella.

Proveedor	Modelo evaluado	Con búsqueda	Sin búsqueda
OpenAI	GPT-5.6 Luna	Sí	Sí
Google	Gemini 3.6 Flash	Sí	Sí
Anthropic	Claude Sonnet 5	Sí	Sí
Perplexity	Sonar	Sí	No
Meta	Llama 4 Maverick	Sí	No
DeepSeek	DeepSeek V4 Flash	No	Sí

Fuente: G4 e informe general, página 12. Las columnas indican las condiciones evaluadas. Una casilla «No» significa que esa condición no formó parte del estudio, no que el producto carezca de esa capacidad.

Con búsqueda web. El sistema puede consultar información externa al responder. Disponer de la herramienta no implica que la use en cada ocasión, que encuentre todas las fuentes relevantes o que las interprete correctamente.

Sin búsqueda web. La respuesta se genera sin consultar la web en esa condición. Actualizar una página pública no modifica de inmediato la información aprendida durante el entrenamiento del modelo. Conviene distinguir la disponibilidad de información nueva de su incorporación efectiva a una respuesta.

Qué se puede comparar. Las cifras describen el comportamiento observado en el protocolo y las fechas del estudio. No constituyen una clasificación universal de asistentes. La ejecución de Anthropic con búsqueda tiene una particularidad de herramienta que se explica en la ficha técnica y debe considerarse al comparar resultados.

Cómo leer los porcentajes y los puntajes

Tres unidades diferentes. Una respuesta es el texto generado ante una consulta. Dentro de ella puede haber menciones a varios hospitales y varias afirmaciones que se puedan verificar. Contar esas unidades responde a preguntas distintas.

Unidad	Qué se cuenta	Indicador de ejemplo
Mención	Apariciones de los hospitales del panel en las respuestas.	Participación en las menciones.
Respuesta	Textos generados que cumplen el alcance del indicador.	Respuestas calificadas con al menos un error.
Afirmación	Datos institucionales verificables extraídos de las respuestas.	Afirmaciones verificables correctas.

El denominador define el significado. El 64,4 % es la participación conjunta de CIMA y Clínica Bíblica en las menciones a los seis hospitales del panel. No significa que estén presentes en el 64,4 % de todas las respuestas, ni que reciban ese porcentaje de pacientes. Las menciones a hospitales ajenos al panel no forman parte de esa base.

Ejemplo de lectura, no resultado del estudio. Una respuesta hipotética contiene diez afirmaciones verificables y una es incorrecta. Esa respuesta tiene al menos un error, aunque nueve de sus diez afirmaciones sean correctas. Por eso el porcentaje de respuestas con error y el de afirmaciones correctas no son complementarios.

Reconocimiento y aparición espontánea. El primero examina cómo responde un modelo cuando se le da el nombre de un hospital. La segunda registra su aparición en preguntas abiertas. UNIBE obtiene 87,6 puntos de reconocimiento y aparece en el 5,4 % de las preguntas abiertas. Son escalas distintas; no se restan para calcular una brecha porcentual.

Un índice sobre 100 no es un porcentaje de respuestas. El índice principal combina presencia (40 %), especificidad y exactitud de la descripción (30 %) y reputación expresada por los modelos (30 %). Los 79,7 puntos de CIMA no significan que aparezca en el 79,7 % de las consultas ni constituyen una nota de calidad médica. La reputación del índice describe lo que expresan los modelos, no una encuesta a pacientes.

Qué dice la recodificación. Repetir la clasificación de respuestas idénticas produjo coincidencia en el 72 % al 76 % de las afirmaciones individuales. Esto mide consistencia del evaluador, no el porcentaje de datos verdaderos. Tampoco equivale a un margen de error de 24 a 28 puntos. Cerca del 42 % de las afirmaciones extraídas quedó fuera de las tasas de verificación; quedar fuera no significa ser falsa.

Qué cubren los intervalos. Los intervalos al 90 % se calcularon con 1.000 remuestreos, es decir, recombinando observaciones de la muestra para examinar su variación. No incorporan toda la incertidumbre del modelo evaluador ni garantizan el mismo resultado en otra fecha. Los detalles se conservan en la ficha técnica.

Fuente: G1, G2, G3 e informe general, páginas 12, 13 y 14. El ejemplo de diez afirmaciones es exclusivamente didáctico.