RESPOSTA DOS AUTORES AO PARECER

Artigo Avaliado CASTILLO BADILLA, Carlos Andrés. Modelo replicável de engenharia de prompts para recuperação de literatura científica com ChatGPT. Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, Florianópolis/SC, Brasil, v. 31, p. 1–30, 2026.

Avaliador 1: 1-No se define adecuadamente qué se entiende por “efectividad”. Dado que los resultados no se comparan con un resultado esperado, el término no se utiliza en su sentido tradicional en ciencias de la computación. Debe definirse con claridad o emplearse métricas más específicas. 2-La pregunta de investigación es demasiado amplia para el experimento diseñado. No se analiza el “cómo”, sino solamente el “si”. 3-Se utilizan términos poco claros, como “ChatGPT Plus (Investiga a fondo)”. Debe especificarse la versión exacta del modelo utilizado (por ejemplo, ChatGPT-4o, ChatGPT-5.2, etc.) para permitir replicabilidad. 4- El protocolo Zero-Hallucination (ZHP-7) no está explicado ni citado. 5- Aunque se afirma adherir a principios FAIR, no se entregan datos esenciales para verificación, como el conjunto de resultados de las 9 búsquedas y las fechas exactas de ejecución. 6- Variables como el “tiempo de respuesta” no están definidas con precisión, ni se justifica su relevancia. 7- No se define ni justifica el criterio de saturación muestral para las búsquedas. 8- No se diferencian claramente conceptos como alucinaciones, referencias inventadas o consistencia conceptual . 9 Se seleccionan 8 resultados como muestra cualitativa sin justificar su representatividad ni detallar el total de resultados obtenidos. 10-Respecto de las referencias: Algunas están inactivas al momento de la revisión (BRASIL, 2016). Otras seis no están citadas en el texto. Existe una referencia duplicada (CHILE 2021 y CHILE 2024).

Resposta: 1- Observación realizada: En este marco, resulta necesario precisar el sentido en que se emplea el término eficacia en el presente estudio. Este concepto no se utiliza como una comparación entre sistemas de recuperación ni como una evaluación frente a un estándar externo. Tampoco se pretende establecer el rendimiento del modelo en términos algorítmicos clásicos. En cambio, la eficacia se entiende como el desempeño operativo verificable de ChatGPT cuando es aplicado bajo un protocolo estructurado de ingeniería de prompts, considerando su capacidad para generar resultados bibliográficos sin referencias ficticias, con DOI o URL activos y con coherencia temática respecto del objetivo declarado. Esta definición se sustenta en indicadores observables como la validación manual de las referencias, la trazabilidad documental y la eficiencia temporal, en consonancia con el diseño exploratorio–descriptivo adoptado.
2- Observación realizada: ¿Permite la aplicación de un protocolo metodológico de ingeniería de prompts obtener resultados bibliográficos verificables y sin referencias ficticias en la recuperación de literatura científica?
3- Observación realizada: Estas variables se midieron a partir de un conjunto de búsquedas controladas realizadas en ChatGPT, modelo GPT-5.2 mediante suscripción ChatGPT Plus y ejecutado en la interfaz web oficial de OpenAI entre julio y agosto de 2025. Las búsquedas se ejecutaron aplicando el prompt metodológico del estudio como instrucción base dentro de ChatGPT y, en simultáneo, se utilizó la función “Investiga a fondo” como modo de ejecución para ampliar la exploración y estructuración de resultados, manteniendo constante el protocolo de ingeniería de prompts y la verificación manual de DOI/URL. La función “Investiga a fondo” corresponde a la herramienta denominada oficialmente Deep Research, incorporada por OpenAI en 2025 como un modo avanzado de exploración y síntesis de información. A diferencia de una consulta convencional,

esta funcionalidad permite una navegación extensiva y autónoma, con análisis estructurado de múltiples fuentes digitales, generación de informes organizados y vinculación directa a documentos originales verificables. En el presente estudio, su uso se limitó a complementar la ejecución del instrumento metodológico, sin sustituir en ningún momento la validación humana ni los controles de trazabilidad definidos en el diseño de investigación.

4- Observación realizada: El instrumento central consistió en un protocolo de ingeniería de prompts diseñado ad hoc (Apéndice A). Dicho protocolo contempló fases de identificación, verificación y análisis de resultados, junto con criterios de inclusión y exclusión claramente definidos. Con el propósito de minimizar la generación de referencias no verificables y asegurar la trazabilidad de los resultados, se desarrolló un esquema de verificación estructurado denominado Zero-Hallucination (ZHP-7), concebido específicamente para el presente estudio como mecanismo interno de control metodológico, sin pretensión de constituir un estándar externo ni un marco previamente validado en la literatura. Este esquema se integró al instrumento principal y permitió sistematizar el proceso de validación bibliográfica en cada búsqueda realizada. El procedimiento comprendió la identificación de afirmaciones y citas generadas por la herramienta, la normalización de referencias conforme a las normas editoriales vigentes, la validación técnica de DOI o URL, la verificación de accesibilidad documental, la comprobación de concordancia temática con los objetivos de investigación, la detección de inconsistencias formales o metadatos incompletos y una revisión final de coherencia documental. Su aplicación fue constante en las 9 búsquedas ejecutadas, funcionando como salvaguarda metodológica frente a posibles inconsistencias generadas por la herramienta de inteligencia artificial.

5- Observación realizada: Con el fin de asegurar la coherencia entre la declaración de adhesión a los principios FAIR (Findable, Accessible, Interoperable, Reusable) y la disponibilidad efectiva de información verificable, se estructuró y depositó el conjunto completo de resultados derivados de las nueve búsquedas controladas realizadas en este estudio. El dataset incluye la identificación de cada búsqueda (Search_ID), fase geográfica, fecha exacta de ejecución, modelo utilizado (GPT-5.2), función activada (Deep Research), versión del prompt aplicado, número de referencias inicialmente recuperadas, número de referencias validadas tras verificación manual, detección de duplicados, tiempos de respuesta y estado de validación técnica de cada DOI o URL.

6- Observación realizada: Con el fin de asegurar la coherencia entre la declaración de adhesión a los principios FAIR (Findable, Accessible, Interoperable, Reusable) y la disponibilidad efectiva de información verificable, se estructuró y depositó el conjunto completo de resultados derivados de las nueve búsquedas controladas realizadas en este estudio. El dataset incluye la identificación de cada búsqueda (Search_ID), fase geográfica, fecha exacta de ejecución, modelo utilizado (GPT-5.2), función activada (Deep Research), versión del prompt aplicado, número de referencias inicialmente recuperadas, número de referencias validadas tras verificación manual, detección de duplicados, tiempos de respuesta y estado de validación técnica de cada DOI o URL.En este estudio, el “tiempo de respuesta” se definió operacionalmente como el intervalo total, medido en segundos, entre la ejecución del prompt y la generación completa de la respuesta final por parte del sistema. Esta variable fue incorporada como indicador complementario de desempeño operativo, permitiendo observar posibles variaciones en la latencia asociadas a la complejidad de la búsqueda y al volumen de resultados generados.La inclusión de esta variable se justifica en el contexto bibliotecario universitario, donde la eficiencia temporal constituye un criterio operativo relevante para la gestión de servicios de información y apoyo a la investigación.

7- En cada zona se ejecutó una búsqueda base y dos replicaciones confirmatorias bajo idénticas condiciones instrumentales, variando exclusivamente la fecha de ejecución. Este procedimiento

permitió evaluar la estabilidad operativa del protocolo frente a posibles variaciones contextuales del modelo, manteniendo homogeneidad en los
criterios de inclusión, exclusión y estructura del prompt. La replicación controlada se utilizó como mecanismo de consistencia interna, orientado a observar la estabilidad en tres indicadores: el número de referencias únicas verificadas por búsqueda, la proporción de referencias con validación técnica completa (DOI o URL activo) y la coherencia conceptual respecto del objetivo declarado en el prompt. La estabilidad en estos indicadores permitió analizar el comportamiento del protocolo sin introducir modificaciones en el instrumento metodológico.Los criterios de inclusión contemplaron artículos revisados por pares en revistas Q1–Q4 y documentos ministeriales con DOI o URL institucional activo. Los criterios de exclusión descartaron preprints, prensa y repositorios no indexados, con el fin de mantener estándares de validez y trazabilidad acordes con las directrices editoriales internacionales.

8- Observación realizada: Con el propósito de delimitar conceptualmente las categorías analíticas empleadas, se establecieron definiciones operativas diferenciadas. En primer lugar, se entiende por alucinación la generación por parte del modelo de información factualmente incorrecta, inexistente o no verificable, producida como resultado de procesos predictivos sin respaldo empírico explícito, fenómeno ampliamente documentado en la literatura sobre modelos de lenguaje de gran escala (Rahman et al., 2026). En el ámbito académico, esta problemática ha sido descrita en relación con la generación de contenidos inexactos o no contrastables en entornos de investigación asistida por IA (Ruiz Mendoza; Pedroza Zúñiga, 2024; Chelli et al., 2024). En segundo lugar, se denomina referencia inventada a aquellas citas que presentan estructura formal académica —autor, año, título, revista— pero cuya existencia no puede corroborarse en bases indexadas o repositorios oficiales mediante DOI activo o URL institucional válida, fenómeno documentado específicamente en estudios sobre el uso de ChatGPT para la búsqueda de literatura científica (Blum, 2023). Esta categoría constituye un subconjunto específico de alucinaciones con implicancias éticas y metodológicas en la investigación científica. Finalmente, la consistencia conceptual se define como la concordancia temática entre la referencia recuperada y el objetivo específico declarado en el prompt, evaluada mediante revisión manual de pertinencia disciplinar y coherencia argumentativa. Esta variable es independiente de la variabilidad algorítmica del modelo y se orienta exclusivamente a determinar la alineación entre el resultado generado y el propósito de la búsqueda. El subconjunto de documentos analizados con fines interpretativos para reflejar una diversidad de enfoques metodológicos y áreas temáticas dentro del corpus recuperado. Los estudios seleccionados incluyen: estúdios comparativos de desempeño, revisiones conceptuales sobre IA en bibliotecología, guías metodológicas para el uso de IA en bibliotecas y trabajos regionales en Latinoamérica, específicamente de México y Brasil. Esta diversidad de estudios permite representar una gama de metodologías empleadas en el campo de la recuperación de literatura con IA.

9- Observación realizada: El corpus validado total estuvo compuesto por 130 referencias únicas. Con el propósito de reforzar la trazabilidad analítica del conjunto, se seleccionó un subconjunto cualitativo de ocho estudios (6,15 % del total), derivados de las nueve búsquedas ejecutadas. La selección fue de carácter intencional y no probabilístico, orientada a representar diversidad metodológica y disciplinar dentro del corpus recuperado, sin pretensión de representatividad estadística. Se incluyeron estudios comparativos experimentales centrados en evaluación de desempeño de modelos de lenguaje, revisiones conceptuales sobre inteligencia artificial en bibliotecología, guías metodológicas para la aplicación de IA en procesos de recuperación de información y trabajos regionales latinoamericanos.

10-Observación realizada: Enlaces actualizados. Se cotejaron todas las citas vs sus referencias y la cuadratura esta ok. Mantuve CHILE 2024.Se cambairón y actualizarón las las citas recomendadas.

Avaliador 2: 1-Se menciona el concepto de saturación, pero no se operacionaliza suficientemente. 2-El estudio se describe como cuantitativo, aunque incorpora análisis cualitativos. Se sugiere ajustar la caracterización metodológica. 3-Los resultados son transferibles en condiciones controladas, pero no generalizables a otros modelos. 4-Existe excesiva uniformidad en indicadores con resultados del 100 %. Se sugiere incorporar una síntesis visual de datos 5-La variable “eficiencia temporal” requiere mayor análisis 6-Reducir el tono normativo o personal en las secciones finales

Resposta: 1- Observación realizada: Se eliminó el término “saturación”, ya que el diseño no corresponde a un muestreo progresivo hasta saturación teórica, sino a un esquema de replicación temporal controlada orientado a evaluar estabilidad operativa del protocolo. La población objetivo se definió como la literatura científica y los documentos institucionales publicados entre 2020 y 2026, recuperables en bases indexadas como Scopus, Web of Science y SciELO, así como en repositorios oficiales ministeriales. Se realizaron nueve búsquedas controladas, distribuidas en tres niveles geográficos diferenciados: global (Norteamérica, Europa, Asia–Pacífico, África y Medio Oriente), regional latinoamericano (excepto Chile) y local chileno, con tres ejecuciones por cada fase. El diseño correspondió a un estudio piloto comparativo con replicación controlada.En cada zona se ejecutó una búsqueda base y dos replicaciones confirmatorias bajo idénticas condiciones instrumentales, variando exclusivamente la fecha de ejecución. Este procedimiento permitió evaluar la estabilidad operativa del protocolo frente a posibles variaciones contextuales del modelo, manteniendo homogeneidad en los criterios de inclusión, exclusión y estructura del prompt. La replicación controlada se utilizó como mecanismo de consistencia interna, orientado a observar la estabilidad en tres indicadores: el número de referencias únicas verificadas por búsqueda, la proporción de referencias con validación técnica completa (DOI o URL activo) y la coherencia conceptual respecto del objetivo declarado en el prompt. La estabilidad en estos indicadores permitió analizar el comportamiento del protocolo sin introducir modificaciones en el instrumento metodológico.Los criterios de inclusión contemplaron artículos revisados por pares en revistas Q1–Q4 y documentos ministeriales con DOI o URL institucional activo. Los criterios de exclusión descartaron preprints, prensa y repositorios no indexados, con el fin de mantener estándares de validez y trazabilidad acordes con las directrices editoriales internacionales.
2- Observación realizada: El enfoque metodológico correspondió a un diseño de predominio cuantitativo con componente interpretativo cualitativo complementario. La dimensión cuantitativa se orientó a la medición de indicadores operativos verificables, tales como precisión, trazabilidad, coherencia conceptual, eficiencia. Es importante precisar que los indicadores de precisión, trazabilidad y coherencia fueron operacionalizados mediante criterios dicotómicos verificables en términos temporales y de detección de alucinaciones en las referencias. Posteriormente, el componente cualitativo se incorporó con fines interpretativos, mediante análisis narrativo de calidad metodológica y transparencia en el uso de prompts dentro del corpus validado, con el propósito de contextualizar y explicar los patrones observados en los indicadores cuantitativos. Estas variables se midieron a partir de un conjunto de búsquedas controladas realizadas en ChatGPT, modelo GPT-5.2 mediante suscripción ChatGPT Plus y ejecutado en la interfaz web oficial de OpenAI en el mes de febrero de 2026. Las búsquedas se ejecutaron aplicando el prompt

metodológico del estudio como instrucción base dentro de ChatGPT y, en simultáneo, se utilizó la función “Investiga a
fondo” como modo de ejecución para ampliar la exploración y estructuración de resultados, manteniendo constante el protocolo de ingeniería de prompts y la verificación manual de DOI/URL.
3- Observación realizada: En cuanto al alcance de los resultados, es importante precisar que la evidencia obtenida se circunscribe al modelo GPT-5 utilizado bajo condiciones controladas y con aplicación estricta del protocolo metodológico diseñado. En consecuencia, los hallazgos no deben interpretarse como generalizables a otros modelos de inteligencia artificial generativa, cuyas arquitecturas, parámetros y mecanismos de actualización pueden producir comportamientos distintos. No obstante, el protocolo desarrollado presenta un carácter transferible, en la medida en que puede ser aplicado y validado en otros entornos tecnológicos, siempre que se mantengan criterios equivalentes de control, verificación documental y trazabilidad. Esta delimitación refuerza la prudencia metodológica del estudio y orienta futuras investigaciones comparativas entre modelos.
4- Observación realizada: Actualizado en resultados.
5- Observación realizada: El presente estudio busca contribuir a superar este vacío mediante la evaluación aplicada de ChatGPT con ingeniería de prompts en la recuperación de literatura científica, considerando criterios de eficiencia temporal, trazabilidad de referencias y coherencia epistemológica. Con ello, se pretende aportar lineamientos metodológicos replicables y pertinentes para fortalecer la Bibliotecología y las Ciencias de la Información en el contexto latinoamericano. En este marco, resulta necesario precisar el sentido en que se emplea el término eficacia en el presente estudio. Este concepto no se utiliza como una comparación entre sistemas de recuperación ni como una evaluación frente a un estándar externo. Tampoco se pretende establecer el rendimiento del modelo en términos algorítmicos clásicos. En cambio, la eficacia se entiende como el desempeño operativo verificable de ChatGPT cuando es aplicado bajo un protocolo estructurado de ingeniería de prompts, considerando su capacidad para generar resultados bibliográficos sin referencias ficticias, con DOI o URL activos y con coherencia temática respecto del objetivo declarado. Esta definición se sustenta en indicadores observables como la validación manual de las referencias, la trazabilidad documental y la eficiencia temporal, en consonancia con el diseño exploratorio–descriptivo adoptado.
6- Se estructuro la sección con un tono más formal y se dejó de forma más precisa y robusta.