Ingeniería y Competitividad
Vol 28 (2) 2026 doi: 10.25100/iyc.v28i2.15700

Vol 28 (2)

ISSN
0123-3033
e- 2027-8284

Artículo Original

Generación de reportes descriptivos usando inteligencia artificial generativa: Un mapeo sistemático

Generation of descriptive reports using generative artificial intelligence: A systematic mapping

Jhonfer Ruiz Figueroa[1] Hugo Armando Ordóñez Erazo[1] Roxana María Romero Luna[1]

[1] Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca. Popayán, Cauca, Colombia. Popayán, Colombia

Resumen

Introducción: Los modelos de lenguaje de gran tamaño (LLMs) han incrementado su uso entre científicos, estudiantes y docentes como herramientas de apoyo en actividades cotidianas.
Objetivo: Analizar el uso de tecnologías emergentes, como los LLMs, en la toma de decisiones a partir de conocimiento previamente procesado y extender el uso de generación de reportes descriptivos.
Metodología: Se realizó un mapeo sistemático que permitió identificar brechas y oportunidades en el uso de estos modelos.
Resultados: Los resultados evidencian que la mayoría de los autores proponen la gestión del conocimiento mediante enfoques como la generación aumentada por recuperación (RAG) y la inteligencia artificial explicable (XAI), con el fin de garantizar la fiabilidad de los textos generados. En la literatura se reporta el uso de modelos como ChatGPT-4, Llama y Gemini 2, destacando su evolución y capacidades en procesamiento de lenguaje natural.
Conclusiones: Aún existen barreras en el uso adecuado de los LLMs, por lo que se requieren investigaciones futuras orientadas a fortalecer la robustez y confiabilidad de los modelos en la generación de informes para la toma de decisiones.

Palabras clave: RLLM, XAI, RAG, Reportes, Toma de decisiones

Abstract

Introduction: Large Language Models (LLMs) have increased their use among scientists, students, and teachers as support tools for everyday activities.
Objective: To analyze the use of emerging technologies, such as LLMs, in decision-making based on previously processed knowledge and to extend the use of descriptive report generation.
Methodology: A systematic mapping was conducted to identify gaps and opportunities in the use of these models.
Results: The results show that most authors propose knowledge management approaches based on Retrieval-Augmented Generation (RAG) and Explainable Artificial Intelligence (XAI) to ensure the reliability of generated texts. The literature reports the use of models such as ChatGPT-4, Llama, and Gemini 2, highlighting their evolution and capabilities in natural language processing.
Conclusions: There are still barriers to the proper use of LLMs; therefore, future research is required to strengthen the robustness and reliability of these models in report generation for decision-making.

Keywords: LLM, XAI, RAG, Reporting, Decision making

¿Cómo citar?

Ruiz J, Ordóñez HA, Romero RM. Generación de reportes descriptivos usando inteligencia artificial generativa: Un mapeo sistemático. Ingeniería y Competitividad, 2026, 28(2)e-20315700

https://doi.org/10.25100/iyc.v28i2.15700

Recibido: 10/03/26
Revisado: 8/04/26
Aceptado: 14/05/26
Online: 21/05/26

Correspondencia

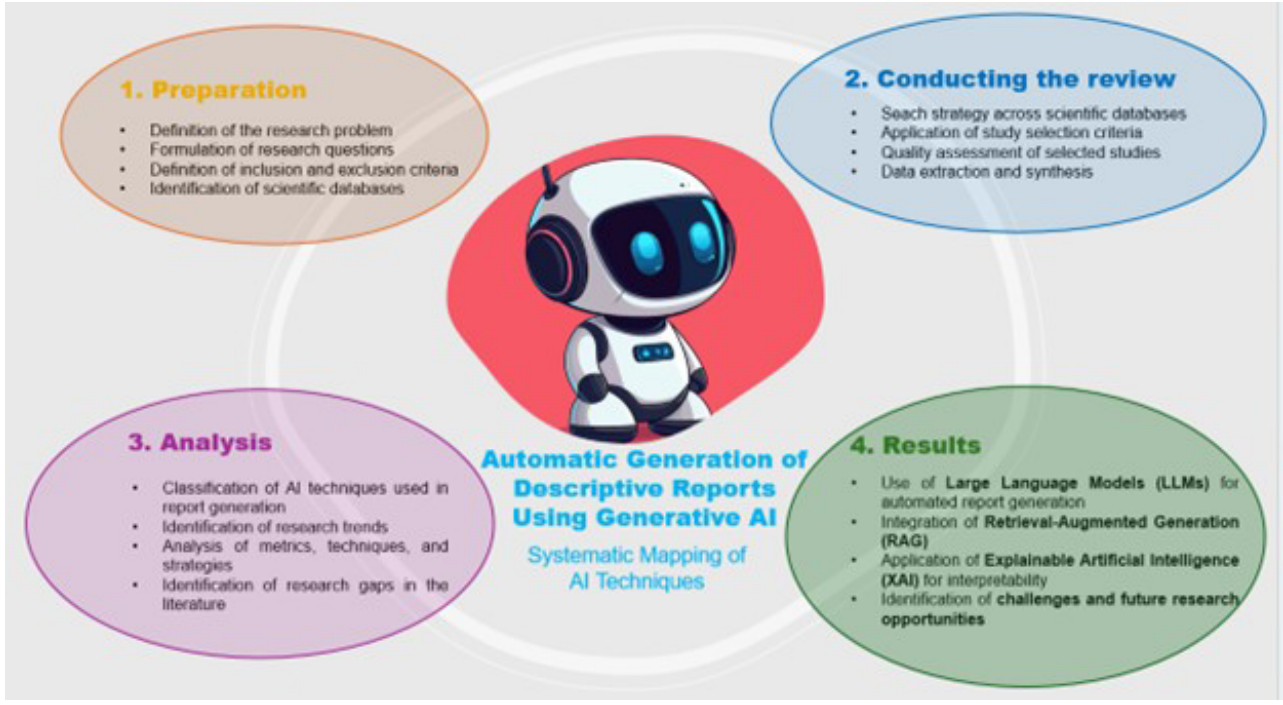
hugoordonez@unicauca.edu.co

¿Por qué se realizó el estudio?
Este trabajo permite organizar la literatura dispersa, identificar métricas, estrategias y limitaciones en la generación automática de reportes descriptivos basados en LLM y XAI. Además, se demuestra una ausencia de enfoques evaluativos de confiabilidad, evidenciando la falta de lineamientos estructurados para la integración de modelos explicables en contextos de toma de decisiones.

¿Cuáles fueron los hallazgos clave?
Los resultados más relevantes incluyen la identificación de altos niveles de precisión en dominios estructurados, específicamente en el sector salud, donde la información suele estar organizada mediante formatos, terminología controlada y variables claramente definidas. Estas características propician un entorno adecuado para el entrenamiento y la evaluación de los modelos LLM, permitiendo que los resultados sean consistentes y precisos en comparación con contextos abiertos o poco estructurados. Asimismo, se evidenció el uso recurrente de métricas como robustez, consistencia y acuerdo de razonamiento humano para evaluar la confiabilidad de los modelos. También, se encontró una escasa cantidad de herramientas integrales dirigidas a la generación automática de reportes explicables.

¿Qué aportan estos hallazgos?
Estos resultados constituyen una base para el desarrollo de futuras investigaciones orientadas a la creación de directrices estandarizadas, herramientas basadas en XAI y sistemas confiables para la generación automática de reportes descriptivos en diversos contextos.

Graphical Abstract



doi: 10.25100/iyc.v28i2.15700

Introducción

Los modelos de lenguaje de gran tamaño (LLM, por sus siglas en inglés) han despertado interés debido a su capacidad para generar texto en lenguaje natural y responder preguntas de forma coherente (1). Su implementación se basa en la arquitectura de transformadores, la cual permite el procesamiento en paralelo de secuencias completas, reduciendo el tiempo en comparación con las redes neuronales recurrentes (RNN, por sus siglas en inglés) (2). Estos modelos emplean esquemas de preentrenamiento auto supervisado combinado con ajustes posteriores supervisados, lo que ha permitido alcanzar resultados sobresalientes en diversas tareas de procesamiento de lenguaje natural (PNL, por sus siglas en inglés) (3).

Desde una perspectiva crítica, aún persisten desafíos relacionados con la confiabilidad del contenido generado. Entre estos se encuentran la generación de información incorrecta, la ampliación de sesgos y la falta de mecanismos claros de interpretación del proceso de inferencia. Estas condiciones impactan directamente la confianza de los textos producidos y restringen su aplicación en contextos sensibles como la elaboración de reportes descriptivos automatizados (3).

En respuesta a esta problemática se han desarrollado diversas técnicas, métodos y estrategias orientadas a potenciar la robustez y confiabilidad de los LLM. Entre ellas destacan los enfoques de inteligencia artificial explicable (XAI, por sus siglas en inglés) y métodos basados en perturbaciones (SHAP, por sus siglas en inglés), que buscan proporcionar interpretabilidad y transparencia en las predicciones generadas por los modelos (4). Sin embargo, estas contribuciones se encuentran dispersas en los estudios y carecen de una sistematización estructurada que permita identificar brechas, vacíos y oportunidades de investigación (5).

En este contexto el presente trabajo tiene como objetivo realizar un mapeo sistemático de la literatura publicada entre 2020 y 2025 sobre técnicas de confiabilidad aplicadas a LLM en la generación de reportes descriptivos. El mapeo permitirá identificar enfoques predominantes, limitaciones metodológicas y posibles lineamientos futuros para fortalecer la consistencia y confiabilidad en la elaboración automática de reportes descriptivos.

La sección 2 describe los materiales y métodos usados para la selección y análisis de los artículos incluidos en el mapeo sistemático. La sección 3 muestra los resultados y la discusión. Finalmente, la sección 4 expone las conclusiones y las oportunidades de investigaciones futuras derivadas del estudio.

Materiales y métodos

En este estudio, la metodología usada está en afinidad con el modelo propuesto en la literatura (6) (Ver Figura 1).

doi: 10.25100/iyc.v28i2.15700

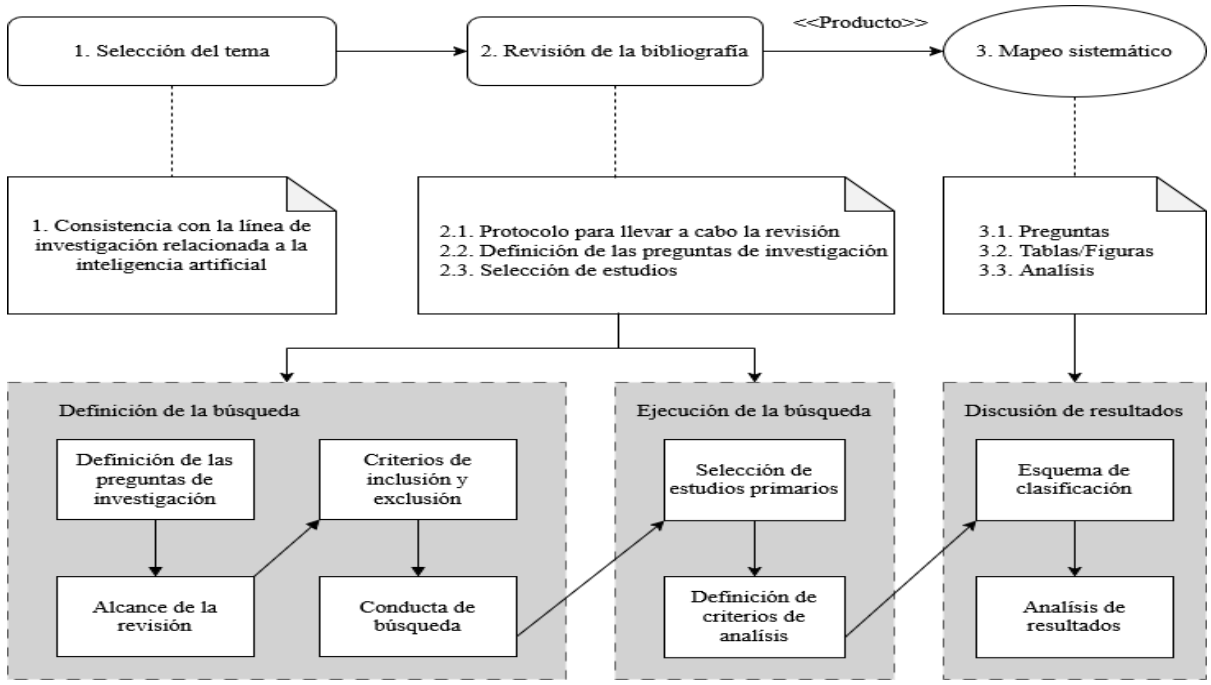


Figura 1. Proceso del Mapeo sistemático. Nota. Tomado de (6).

Los tres procesos principales se describen a continuación: 1. selección del tema; en este caso va relacionado a los reportes descriptivos con XAI, 2. Revisión bibliográfica, donde se especifica los filtros para obtener la literatura más relevante en el estudio y 3. el mapeo sistemático, donde se evidencia la metodología y estructura de selección de estudios relevantes aplicados. Todo lo anterior, es basado en las etapas descritas en (7).

Selección del tema

En esta etapa, se tiene en cuenta las siguientes características:

1. La selección del tema se fundamenta en la necesidad de analizar en la generación automática de reportes mediante LLM, integrando técnicas de interpretación como SHAP, debido a su relevancia en contextos de gobernanza.

2. La adaptabilidad de una herramienta en cualquier tipo de entorno o uso dentro de una entidad gubernamental, privada o administrativa para la elaboración de reportes automáticos.

Revisión bibliográfica

Uno de los pasos importantes es formular las preguntas de investigación para dirigir y enfocar el objetivo del estudio (Ver Tabla 1). En esta etapa se revisa el estado del arte con respecto al uso de XAI e IA en la generación de reportes descriptivos y determinar la literatura científica relevante para el estudio.

doi: 10.25100/iyc.v28i2.15700

Tabla 1. Preguntas de Investigación

Identificador	Preguntas de Investigación	Motivación
QIG	¿Cuál es el estado actual del conocimiento en el uso de modelos LLM con XAI en la generación automática de reportes?	Entender el estado actual del tema y los diferentes modelos usados para la generación de texto en NL.
QI1	¿Qué métricas y enfoques se utilizan para evaluar la calidad y confiabilidad de los reportes generados por LLM?	Conocer la eficacia de los métodos actuales para analizar y contextualizar en NL para la generación de reportes.
QI2	¿Cuáles son las métricas, técnicas y estrategias en el uso de modelos LLM con XAI y en reportes?	Identificar las métricas, técnicas y estrategias usadas que han sido efectivas en la generación de texto en NL.
QI3	¿Cuáles son las aplicaciones actuales en el uso de modelos LLM con XAI y en reportes?	Comprender la longitud del problema e identificar como se ha abordado, ya sea con estrategias especificas dentro del mismo contexto o, por el contrario, existen vacíos.

Nota. Elaboración propia.

El siguiente paso es establecer los criterios de inclusión y exclusión para determinar los textos especializados relevantes.

Criterios de inclusión

- Artículos publicados con una ventana de tiempo de 2022 a 2025.
- Artículos que den respuesta a las preguntas de investigación.
- Artículos relacionados a los LLM y los usos de PNL en la generación de reportes.

Artículos en inglés.

Criterios de exclusión

- Artículos duplicados considerando el más reciente.
- Artículos sin relación a los LLM, o PNL.
- Artículos incompletos.

En relación con las fuentes de información, se seleccionaron las bases de datos Science Direct, Springer Link, Google Scholar, y IEEE explore. Debido a su alta cobertura en áreas de IA, PNL e ingeniería.

doi: 10.25100/iyc.v28i2.15700

Science Direct y Springer Link: Fueron priorizadas por su gran volumen de textos especializados revisados por pares en ciencias computacionales y su reconocimiento en investigación académica de calidad.

Google Scholar: Se utilizó como una fuente complementaria para ampliar la cobertura y reducir el sesgo de indexación, lo que permitió abarcar literatura no incluida en bases de datos tradicionales.

IEEE explore: Se incluyó por su perspectiva especializada en ingeniería y tecnologías emergentes particularmente en IA aplicada.

De esta manera la combinación de estas fuentes permitió mantener la calidad, cobertura y actualidad en las publicaciones recolectadas. Por otra parte, el conjunto de fuentes bibliográficas Science Direct, Springer Link y IEEE explore, se llamarán fuentes indexadas.

Por último, se define los términos principales que permiten enfocar y delimitar la búsqueda de artículos relacionados a los LLM y la generación de reportes recopilados en la Tabla 2. Asimismo, las cadenas de búsqueda usadas considerando operadores lógicos como "AND" y "OR" para su construcción.

Tabla 2. Cadenas de Búsqueda

Términos principales	Cadena de búsqueda
LLM y NLP	(("Large language models" AND models) AND ("Report descriptive" AND Automation AND Visualization) AND (Natural))
LLM y Reportes Escritos	(("Large language models" AND models) AND ("Report descriptive" AND Automation AND Visualization))
LLM y XAI	(("Large language models" AND models) AND ("Report descriptive" AND Automation AND Visualization) AND ("Explainable Artificial Intelligence"))
	(("Large language models" AND models) AND ("Report descriptive" AND Automation AND Visualization) AND ("Explainable Artificial Intelligence" AND Natural))

Nota. Elaboración propia.

Estudios seleccionados

Para llegar a identificar los estudios primarios se aplicó los criterios de inclusión y exclusión para eliminar resultados duplicados o trabajos de investigación incompletos. Además, se incluyen dos filtros de revisión descritos a continuación:

doi: 10.25100/iyc.v28i2.15700

Primer filtro (1F): Revisión del título y el resumen.

Segundo filtro (2F): Las publicaciones que pasaron el primer filtro fueron sometidas a una lectura y análisis completo de su contenido.

Teniendo en cuenta todos los pasos anteriores se logró identificar los estudios primarios como se ilustra en el resumen de la Figura 2.

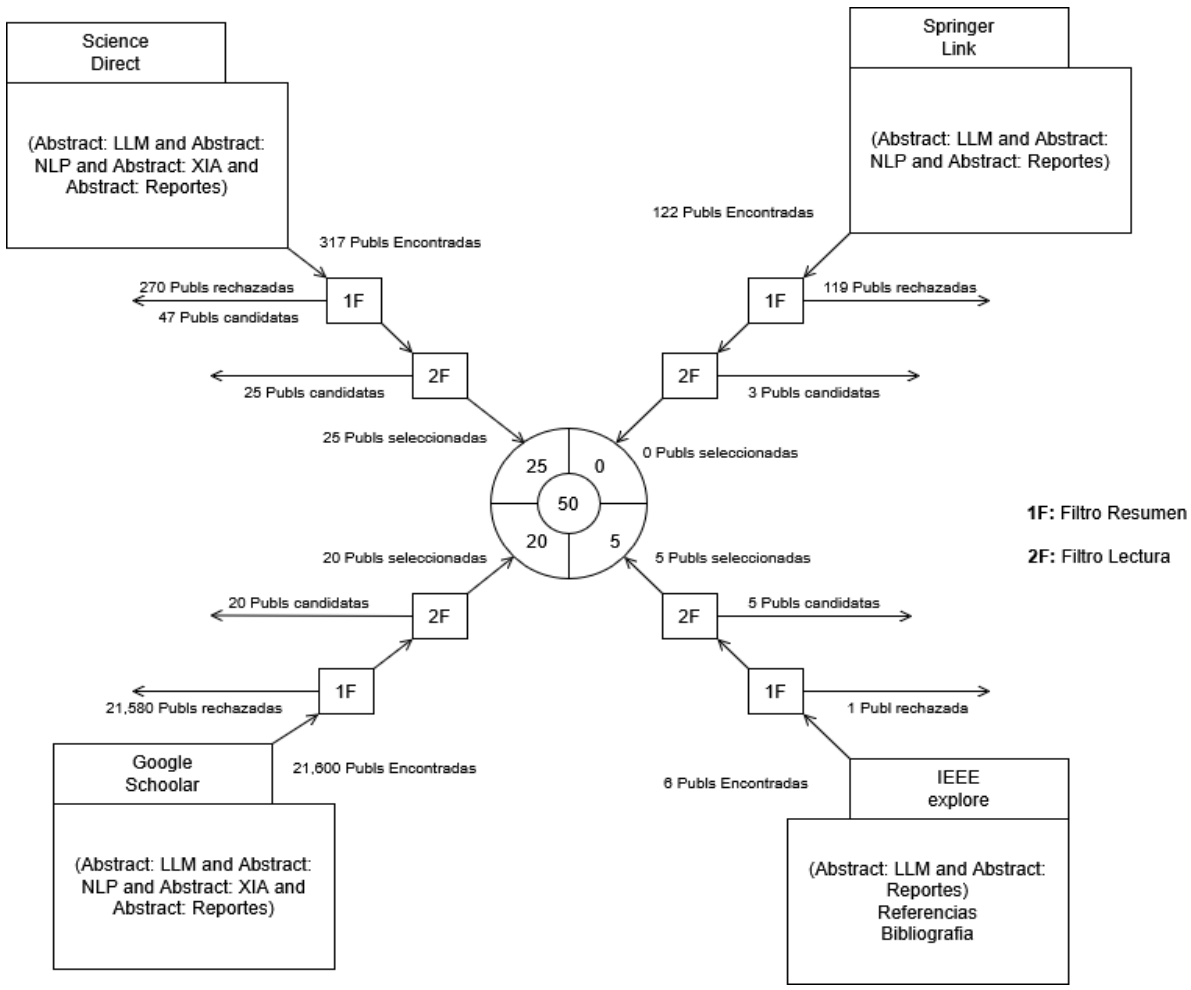


Figura 2. Gráfico de embudo para la búsqueda de estudios primarios en formato DCM.
Nota. Elaboración propia

El proceso mostrado en la figura 2 fue estructurado en dos etapas de filtrado. En primer lugar, se aplicaron los criterios de inclusión y exclusión junto con el primer filtro (1F) basado en la revisión del título y el resumen de las publicaciones identificadas. En conjunto con las fuentes indexadas (n = 55) y Google Scholar (n = 20), se obtuvo (n = 75) textos especializados para la siguiente fase. Por otro lado, se rechazaron para las fuentes indexadas (n = 390) y Google Scholar (n = 21580).

En el caso de Google Scholar, debido a su naturaleza y el gran volumen de trabajos (n = 21600) se implementó una estrategia de selección manual. En este proceso, se consideró los primeros 100 resultados ordenados por relevancia, dado que el motor de búsqueda prioriza publicaciones con mayor impacto, número de citaciones y coincidencia semántica con las cadenas de búsqueda.

doi: 10.25100/iyc.v28i2.15700

Posteriormente, se aplicó el segundo filtro (2F) y la segunda etapa, correspondiente a la lectura completa de los artículos seleccionados en la primera etapa. En esta fase, se excluyeron (n = 25) estudios por no cumplir con los criterios definidos, presentar información incompleta o no responder a las preguntas de investigación.

Para concluir, el proceso de selección permitió consolidar un total de (n = 50) estudios primarios, distribuidos en (n = 30) artículos provenientes de fuentes indexadas y (n = 20) artículos provenientes de la selección manual en Google Scholar. Este flujo de selección se presenta de manera detallada en la figura 3, siguiendo una adaptación del modelo PRISMA (8), lo que garantiza la trazabilidad y reproducibilidad del proceso.

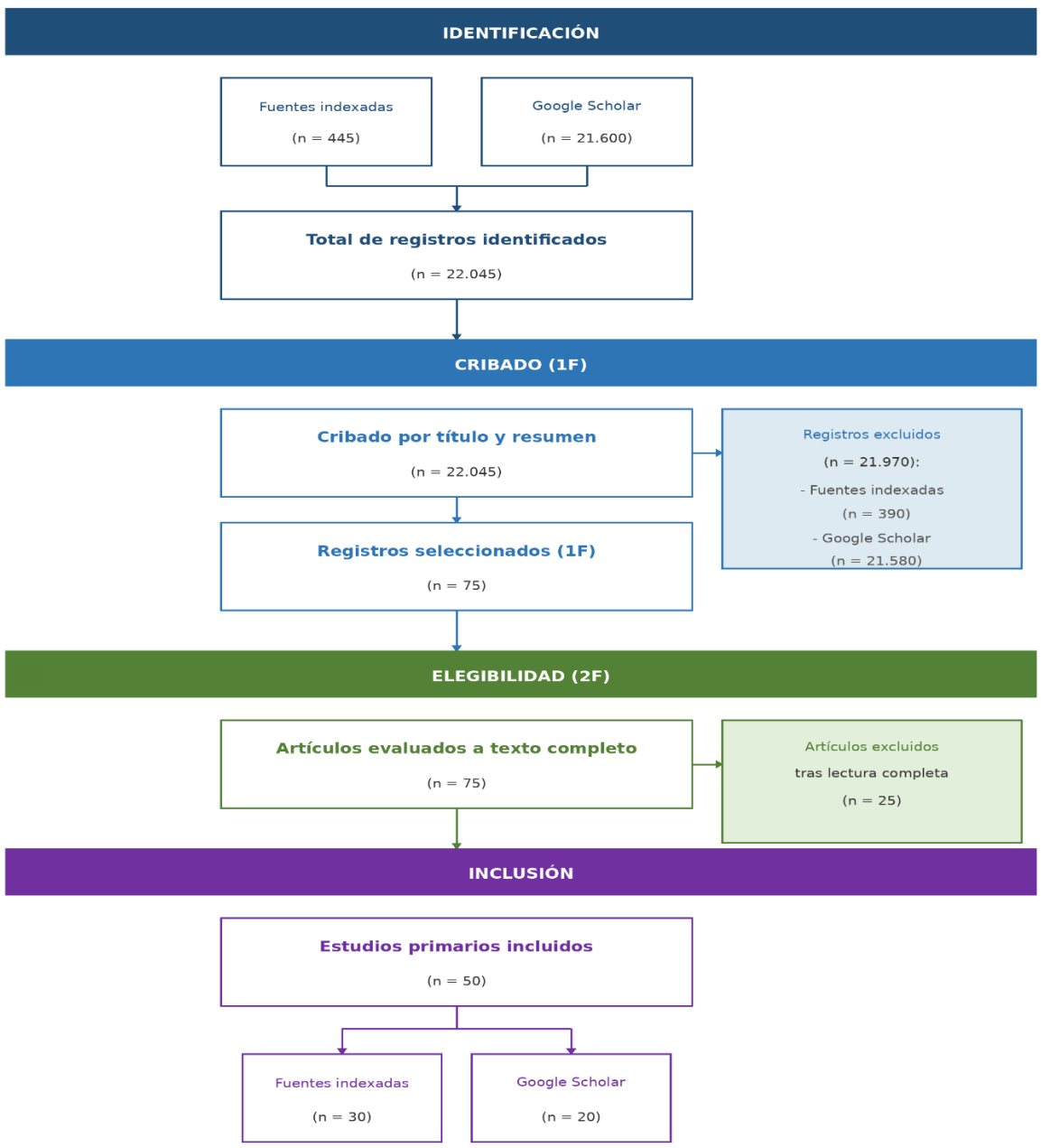


Figura 3. Gráfico del proceso usando el modelo PRISMA. Nota: Elaboración propia.

doi: 10.25100/iyc.v28i2.15700

Resultados y Discusión

La literatura que aborda problemas en la generación de reportes descriptivos a través de procesamiento de lenguaje natural y los modelos de lenguaje grande (LLM) es extensa, para poder facilitar el entendimiento del problema se dividieron en tres ejes principales: 1. Modelos LLM para generar reportes descriptivos, 2. Modelos LLM en el desarrollo de informes y 3. Modelos LLM y técnicas de aseguramiento de confianza.

Modelos LLM para generar reportes descriptivos

En los últimos años se han introducido esfuerzos significativos en la generación del lenguaje natural (NLG) para producir textos coherentes y legibles por humanos. No obstante, modelos clásicos como Seq2Seq tienden a generar respuestas genéricas y poco contextualizadas, lo que restringe su aplicabilidad en escenarios donde se requiere profundidad (9).

Diversos estudios han explorado aplicaciones de NLP y LLM en la generación de reportes descriptivos. En el ámbito de la salud, se han empleado técnicas de machine learning combinadas con modelos LLM, como el enfoque DSS – LLM, logrando procesar grandes volúmenes de información clínica, reducir tiempos de análisis y alcanzar precisiones entre 98.58% y 98.91% en diagnósticos tempranos (10). De manera complementaria, se han utilizado modelos basados en reglas y sistemas de comprensión del lenguaje natural (NLU) para automatizar procesos mediante esquemas de preguntas y respuestas entrenados en dominios específicos (11) (12).

Otros estudios han incorporado grafos de conocimiento (KG) y redes antagónicas (GAN) para adoptar la explicabilidad y confiabilidad de chatbots basados en LLM, asimismo apoyar la generación semiautomatizada de datos en procesos de automatización (13) (14). Del mismo modo, se ha evidenciado el uso de técnicas clásicas de NLP integradas en modelos actuales como Llama 2, aprovechando bases de conocimiento preexistente para agilizar el desarrollo de soluciones y mitigar problemas de acceso a la información (15).

En contextos con alta incertidumbre, se han propuesto enfoques híbridos que combinan LLM con redes bayesianas en conocimiento experto, incrementando la credibilidad de modelos que carecen de mecanismos formales de validación (16). En el ámbito farmacéutico y clínico, los LLM han demostrado una alta precisión diagnóstica para facilitar la toma de decisiones estratégias, simplificando la terminología médica y robustecer la interacción entre médico–paciente (17).

No obstante, estos avances presentan retos asociados a la privacidad, el consumo de recursos, la redundancia de información y la seguridad de los datos (18). Trabajos recientes analizan vulnerabilidades como inyección de prompts, envenenamiento de datos, entre otros; proponiendo estrategias de mitigación para fortalecer la robustez de los modelos (19) (20). De forma adicional, se han identificado una escasez de investigaciones centradas en la calidad y verificación de los KG utilizados junto con LLM, aunque se han reportado resultados prometedores en validación automática de declaraciones mediante modelos como llama 3 (21).

Finalmente, los LLM se han aplicado en la generación de reportes de incidencias en entornos industriales, alcanzando precisiones superiores al 90% en la extracción de información utilizando variantes como ChatGPT-3.5, lo que evidencia su potencial en escenarios operativos reales (22).

doi: 10.25100/iyc.v28i2.15700

Modelos LLM en el desarrollo de informes

El estado del arte muestra un creciente interés en el uso de LLM para el desarrollo de informes en distintos dominios. Diversas obras han evaluado su viabilidad en la generación de reportes estructurados, destacando adecuaciones significativas en eficiencia y reducción de tiempos frente a procesos manuales. Por ejemplo, el uso de ChatGPT-4 en la elaboración de informes de responsabilidad social corporativa (RSC) alcanzó una precisión del 87.14%, aunque reveló una tasa de inconsistencia del 32.87%, lo que subraya la necesidad de verificación humana (12).

En este contexto, se han integrado enfoques de generación aumentada por recuperación (RAG) para fortalecer la calidad de informes. Técnicas como ESGReveal lograron precisiones del 76.9% y tasas de extracción del 83.7% en el análisis de datos, evidenciando avances en la trazabilidad de la información (23). De igual manera, evaluaciones comparativas de distintos LLM en diagnósticos médicos, como la detección de osteoartritis, lograron sensibilidades de hasta 92.3% en modelos como ChatGPT-4, alcanzando niveles comparables a especialistas humanos, aunque se resalta la necesidad de validaciones por expertos (24).

Asimismo, los LLM se han aplicado en procesos de toma de decisiones mediante el análisis de grandes volúmenes de información provenientes de artículos periodísticos, informes gubernamentales, publicaciones sectoriales y redes sociales, combinados con técnicas de RAG (25). No obstante, estos trabajos identifican acotaciones recurrentes relacionados con la privacidad de los datos, la fiabilidad de los resultados y la presencia de errores médicos o informativos (26) (27) (28). Aunque los anteriores estudios provienen de dominios heterogéneos, permiten identificar patrones generales en el comportamiento de los LLM respecto a la generación de reportes.

Si bien los LLM pueden generar respuestas útiles a partir de instrucciones de usuario, su efectividad depende de gran medida del nivel de conocimiento del dominio por parte de quien los utiliza. En usuarios con experiencia, los modelos actúan como herramientas de apoyo; ahora bien, en contextos de bajo conocimiento, pueden introducir lagunas de información o reforzar premisas incorrectas (29).

Desde una perspectiva sistemática se han propuesto taxonomías para organizar el diseño y uso de LLM como soporte a la escritura en lenguaje natural, abordando problemas como la fragmentación de modelos y el uso de datos desactualizados. La integración de razonamiento avanzado y enfoques RAG ha permitido que los KG contribuyan a la actualización automática de información con menos márgenes de error (30) (31) (32).

Por último, los LLM se han aprovechado en la corrección técnica de textos especializados en áreas como la arquitectura, ingeniería y construcción, análogamente en la identificación de patrones estilísticos redundantes en textos generados por IA. Estas aproximaciones han dado lugar a nuevas categorizaciones para adaptar la generación automática de texto a estándares humanos de calidad (33) (34) (35) (36).

Modelos LLM y técnicas de aseguramiento de confianza

A medida que aumenta el uso de LLM también se incrementan las preocupaciones relacionadas con la fiabilidad y consistencia del conocimiento generado. En respuesta, la literatura ha propuesto

doi: 10.25100/iyc.v28i2.15700

diversas técnicas orientadas al aseguramiento de confianza, apoyadas principalmente en enfoques de XAI y mecanismo de comprobación automática.

En el ámbito educativo se han explorado modelos de aprendizaje basados en simulación integrados con LLM, los cuales permiten crear entornos realistas para el desarrollo de competencias docentes. Aunque estos enfoques muestran beneficios pedagógicos, persisten encares técnicos asociados a la latencia y a la comprensión de contextos complejos por parte de los modelos (37).

En panoramas industriales se han implementado estrategias de RAG con funciones aumentadas para ajustar procesos de tomas de decisiones complejas, como la recuperación de fallas de productos devueltos durante procesos de remanufactura. Estos planteamientos alcanzaron precisiones promedio entre 65% y 79%, demostrando su beneficio práctico pese a definiciones inherentes al modelo base usado (38).

Adicionalmente, se han propuesto métricas específicas para corroborar la eficacia de las técnicas de explicabilidad en LLM, entra las que destacan el acuerdo de razonamiento humano, la robustez, la consistencia y la contrastividad. Estas métricas permiten comparar modelos de manera objetiva y establecer bases para el desarrollo de arquitecturas confiables (3) (4) (13) (39) (40) (41) (42) (43) (44).

De forma complementaria diversas investigaciones han aplicado tácticas como ingeniería de prompts, RAG y la incorporación de KG para verificar la confiabilidad de la generación automática de reportes en dominios como la educación, la salud y la gestión del conocimiento. En este medio, se resalta la responsabilidad asociada al uso de modelos como ChatGPT-4 en procesos críticos de toma de decisiones, dada su creciente influencia en la generación y transmisión de conocimiento (45) (46) (47) (48).

Basado en los artículos se procede a analizar cada uno de ellos para identificar temas en común, relaciones entre los mismos y proporcionen una respuesta precisa a las preguntas de investigación planteadas (Ver Tabla 1). Teniendo en cuenta esto, se inicia con los siguientes pasos dentro del estudio.

Respuestas a las preguntas de investigación

QI1. ¿Qué métricas y enfoques se utilizan para evaluar la calidad y confiabilidad de los reportes generados por LLM?

La evaluación de la calidad y confiabilidad de los reportes generados se basa en un conjunto de métricas, cuantitativas y orientadas a la interpretabilidad y consistencia del modelo.

En primer lugar, se utilizan métricas cuantitativas tradicionales como la precisión, recall, f1-score y el área bajo la curva (AUC), particularmente en dominios estructurados como el ámbito clínico. Estas métricas permiten evaluar el desempeño del modelo en tareas de clasificación y predicción como indicadores base del rendimiento del sistema.

En segundo lugar, se identifican métricas orientadas a la confiabilidad y explicabilidad, tales como la consistencia, la robustez y el acuerdo de razonamiento humano. Estas métricas permiten valorar la estabilidad del modelo frente a variaciones en los datos y la alineación entre las respuestas

doi: 10.25100/iyc.v28i2.15700

generadas y la interpretación humana, siendo fundamental en la generación de reportes descriptivos.

Los estudios analizados muestran una perspectiva clara hacia altos niveles de precisión en dominios estructurados, como el sector de la salud, donde los modelos alcanzan valores superiores al 90%. Por el contrario, en ámbitos industriales y abiertos, la diferencia en el desempeño es mayor, con rangos entre 65% y 79%, tal como se evidencia en la figura 4. No obstante, en escenarios mas complejos, como la generación de reportes descriptivos, se requiere complementar estas métricas con enfoques que analicen la coherencia, la interpretabilidad y la confiabilidad del contenido generado.

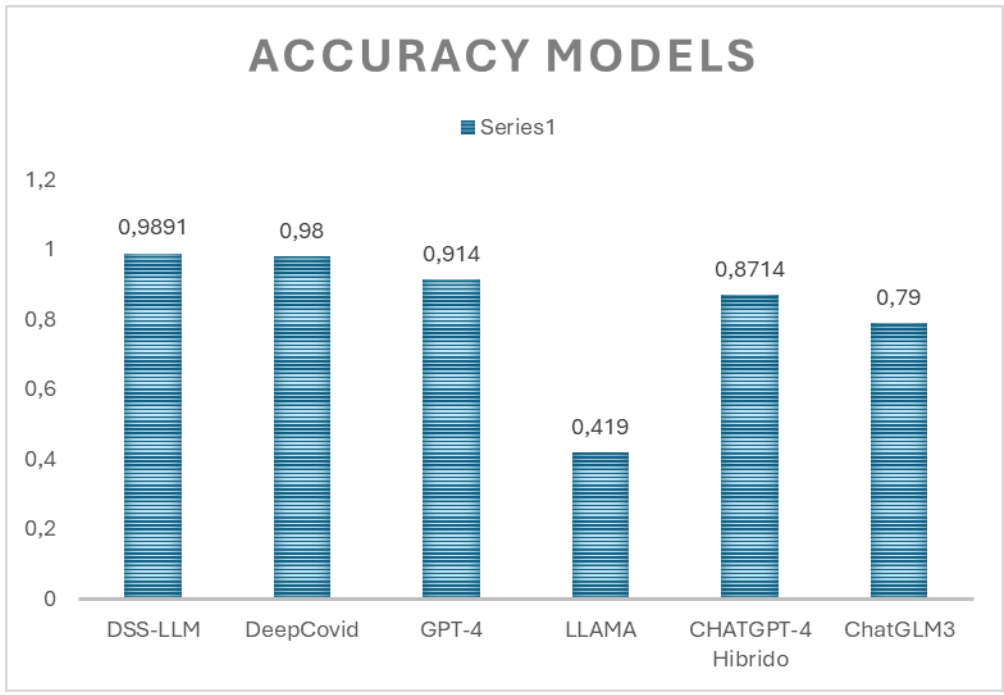


Figura 4. Gráfico de precisión de los modelos encontrados en la pregunta QI1. Nota. Elaboración propia

Por último, se observa que las métricas tradicionales como la precisión y el AUC predominan en dominios estructurados. Sin embargo, métricas más recientes como la consistencia y el acuerdo de razonamiento humano, se emplean para comprobar confiabilidad en la generación de reportes descriptivos.

QI2. ¿Cuáles son las métricas, técnicas y estrategias en el uso de modelos LLM con XAI y en reportes?

El mapeo sistemático identificó una variedad de métricas, técnicas y estrategias orientadas a incrementar la fiabilidad, interpretabilidad y estabilidad de los LLM en la generación automática de reportes. En lugar de orientaciones aisladas, los estudios convergen en tres dimensiones principales: evaluación explicable, dirección de la incertidumbre, y validación semántica.

En el ámbito de las métricas de evaluación explicable se destacan el acuerdo de razonamiento humano, la robustez, la consistencia y la contrastividad, aplicadas para comparar técnicas de explicabilidad sobre conjuntos de datos como Movie Reviews (IMDB) y Tweet Sentiment Extraction

doi: 10.25100/iyc.v28i2.15700

(TSE) (4). Estas métricas permiten la alineación entre el razonamiento del modelo y la interpretación humana.

En relación con las técnicas de explicabilidad se emplearon los SHAPStories y CFStories, diseñados para refinar la comprensión del usuario sobre predicciones generadas por LLM mediante narrativas explicativas (44). Asimismo, las relaciones metamórficas (MR) han sido planteadas como estrategias de validación empírica en múltiples LLM, evidenciando adopciones en estabilidad y coherencia semántica (49).

Con respecto a la gestión de la incertidumbre, se analizan mecanismos de incertidumbre verbalizada y no verbalizada para reducir vacíos interpretativos y mitigar percepciones erróneas derivadas de alucinaciones del modelo (50) (51).

Adicionalmente, en contextos periodísticos y de transparencia informativa se emplean métricas de coincidencia difusa, semántica y exacta para evaluar la precisión en la atribución de fuentes de modelos como ChatGPT-4, Claude y Gemini (52).

En conjunto, estos estudios consolidan un marco técnico para fortalecer la confianza, estabilidad y trazabilidad en la generación automática de reportes descriptivos basada en LLM.

QI3. ¿Cuáles son las aplicaciones actuales en el uso de modelos LLM con XAI y en reportes?

El análisis de las aplicaciones actuales en el uso de modelos LLM con XAI en la generación de reportes revela un panorama incipiente. La revisión sistemática identificó un número limitado de herramientas consolidadas, entre las cuales destaca LIDA (53), Un proyecto orientado a la generación de visualizaciones e infografías mediante el uso de modelos LLM como ChatGPT-4. Esta herramienta integra procesos que abarcan desde la interpretación de los datos hasta la generación de representaciones visuales, estableciendo un nicho especifico en la automatización de procesos analíticos ilustrativos.

Sin embargo, dentro de los estudios analizados la acotada presencia de herramientas muestra una brecha significativa en la investigación aplicada de LLM para la generación automatizada de reportes tanto descriptivos como visuales. Este vacío en el estado del arte sugiere la necesidad de desarrollar enfoques integrales que permitan articular la generación de contenido, la explicabilidad y la validación de resultados en un mismo entorno.

En este sentido, se identifica una coyuntura relevante para futuras investigaciones, esencialmente en el diseño de directrices y herramientas que faciliten la implementación de sistemas basados en LLM con capacidades de consistencias, orientados a contextos reales de gobernanza.

Como se presentó en las anteriores preguntas de investigación sé logró consolidar los métodos, los algoritmos y los modelos más precisos, que fueron: el DSS-LLM con una precisión de 98.91 %, DeepCovid y ChatGPT-4 en los diagnósticos de salud; evidenciado posibles candidatos para usar de referencia en cualquier investigación de generación de reportes descriptivos automáticos. Además, en el marco de métricas se tiene lo siguiente; acuerdo de razonamiento humano, robustez, consistencia, y contrastividad para medir la fiabilidad del modelo a usar y permitir aumentar la precisión y la rigurosidad del texto generado. De igual forma, sirven para validar o medir nuevas

doi: 10.25100/iyc.v28i2.15700

herramientas a la hora de evaluar y revisar la calidad del modelo en cuestiones de consistencia y coherencia en el texto generado. Asimismo, se encontró con la herramienta LIDA, donde se proporciona insumos en la interacción y creación de herramientas software en conjunto con los LLM. Incluso, se puede llegar a generar lineamientos y recomendaciones para tener en cuenta con este tipo de programas. En contraste, se debe emplear un apoyo en métricas sobre la autoatención y comprensión que tienen los modelos con respecto a lagunas de conocimiento generados por trabajar LLM con RAG de manera inadecuada. Esto sugiere que la integración de conocimiento externo destaca en la precisión, de igual manera, permite la trazabilidad del contenido generado.

En la Figura 5 se muestra la relevancia de artículos agrupados por año y temas compartidos. Entre más grande es la burbuja, mayor es el número de artículos con respecto al tema y el año de publicación. Si en la intersección no se encuentra una burbuja, significa que no hay artículos encontrados o seleccionados en ese punto.

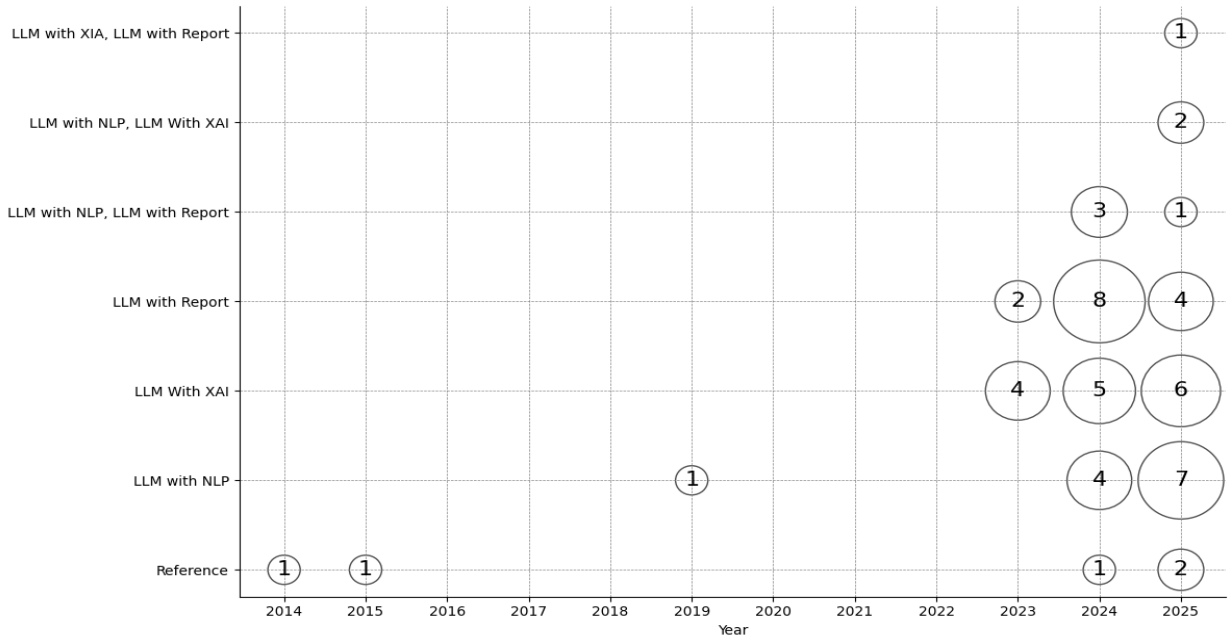


Figura 5. Gráfico de burbujas que representa el número de publicaciones por tema con respecto al año de publicación. Nota. Elaboración propia.

Por otro lado, se observa que la mayoría de los artículos son de los últimos años, para ser exactos 2023 al 2025. Esto es importante, ya que se infiere que la generación automática de reportes descriptivos usando XAI con los LLM ha sido tema de estudio durante los últimos años y tiene una tendencia a crecer en los próximos.

Discusión de los resultados

La discusión de los resultados identifica cuatro puntos entre tendencias y desafíos. En primer lugar, se evidencia que los LLM y la inteligencia artificial generativa han alcanzado rangos elevados de precisión en contextos estructurados, especialmente en dominios como la salud, donde modelos híbridos como ChatGPT-4 (24), Deepcovid (11) y DSS-LLM (10) reportan valores alrededor del 76.9% hasta 98.1%. Esto nos indica una buena relación entre los LLM y la creación de texto en

doi: 10.25100/iyc.v28i2.15700

los reportes usando NL. Sin embargo, a su vez pone un objetivo ambicioso para las herramientas que se tienen actualmente garantizando la confiabilidad y generalización de estos modelos en contextos más complejos y menos estructurados.

En segundo lugar, se encuentra una restricción relevante relacionada con la ausencia de lineamientos estandarizados para la validación y seguimiento de la robustez de los LLM. Si bien en los estudios propone diversas métricas como los acuerdos de razonamiento humano, la robustez, la consistencia y la contrastividad en el contexto de elaboración de reportes (4), a su vez otras aproximaciones como la percepción (51), las relaciones metamórficas (49), las coincidencias (52), estas se presentan de manera fragmentada. Esta disformidad dificulta y dilata la consolidación de un marco común de evaluación, lo que muestra la necesidad de futuros estudios orientados a integrar y validar métricas que permitan una comprobación consistente del modelo mediante el uso de técnicas de XAI (13).

En tercer lugar, los resultados evidencian una limitación en la integración de generación de reportes descriptivos y visuales dentro de un mismo entorno basado en LLM. A pesar de las modificaciones en las diferentes cadenas de búsqueda para abordar el tema desde una perspectiva holística, se identificó una escasa cantidad de estudios que desarrollen herramientas integrales que combinen generación textual, explicabilidad y visualización de datos (53). Esta situación pone de manifiesto una brecha en la investigación aplicada, específicamente en el diseño de soluciones completas orientadas a usuarios finales.

Finalmente, se encontró que hay un número reducido de herramientas consolidadas para el manejo de generación automática de textos especializados. Esta barrera puede estar asociada a la falta de directrices dentro de la integración de procesos de ingeniería de software con los LLM. Para abordar este dominio específico, se plantea como línea futura el desarrollo de soluciones que integren modelos LLM con técnicas de explicabilidad, permitiendo la generación de texto y la validación de su confiabilidad en un mismo entorno (4). Asimismo, aspectos hallados en la literatura como la sensibilidad inmediata, las alucinaciones y las características de varios modelos continúan presentando retos críticos, lo que refuerza la necesidad de incorporar enfoques basados en XAI y agentes inteligentes que desarrollen la trasparencia y fiabilidad de estos sistemas en ámbitos reales (51)

Conclusiones

El mapeo sistemático realizado a partir de 50 artículos relevantes identificó los desafíos y limitaciones en el uso de un modelo LLM para la generación automática de reportes fiables. Los resultados muestran una alta dispersión en metodologías, métricas y herramientas para esta temática, el cual dificulta la consolidación de buenas prácticas y condiciona su adopción en distintos ámbitos de aplicabilidad.

Si bien los modelos actuales han alcanzado altos niveles de precisión aún el camino es largo para lograr una generación de texto confiable y consistente en escenarios reales. De igual forma, se suma la falta criterios estandarizados para su evaluación, reforzando la necesidad de incluir marcos metodológicos unificados que integren métricas técnicas y enfoques de confiabilidad.

doi: 10.25100/iyc.v28i2.15700

Asimismo, la ausencia de herramientas consolidadas y la escasa integración con desarrollos y metodologías en la ingeniería de software aplicada a LLM, ha permitido identificar la falta de consenso de lineamientos técnicos entre estas dos temáticas. En ese sentido, la combinación de LLM con técnicas de XAI y su implementación en plataformas integradas se perfila como una estrategia clave para aumentar la confiabilidad.

En conjunto, estos hallazgos orientan hacia la formulación de lineamientos técnicos que favorezcan la robustez en la construcción de un LLM y la generación de reportes descriptivos automáticos, para permitir su aplicación efectiva en áreas críticas como la salud, el periodismo, la educación y los problemas del común en la sociedad.

Declaración sobre el uso de Inteligencia Artificial Generativa y Tecnologías Asistidas por IA en el Proceso de Redacción

Durante la preparación de este trabajo, los autores utilizaron ChatGPT (GPT-4) para asistir en el proceso de redacción y mejorar la legibilidad y el lenguaje. Después de utilizar esta herramienta, los autores revisaron y editaron el contenido según fue necesario y asumen plena responsabilidad por el contenido de la publicación.

Declaración de autoría de CrediT

Conceptualización - Ideas: Hugo Armando Ordóñez. Curación de datos: Hugo Armando Ordóñez. Análisis formal: Jhonfer Ruiz Figueroa. Investigación: Jhonfer Ruiz Figueroa. Metodología: Jhonfer Ruiz Figueroa. Gestión del proyecto: Hugo Armando Ordóñez. Recursos: Jhonfer Ruiz Figueroa. Software: Jhonfer Ruiz Figueroa. Supervisión: Roxana Maria Luna Romero. Validación: Roxana Maria Luna Romero. Redacción del borrador original - Preparación: Roxana Maria Luna Romero. Redacción - Revisión y edición - Preparación: Roxana Maria Luna Romero.

Financiación: no declara.

Conflicto de intereses: no declara. Aspecto ético: no declara.

Referencias

1. Lu Y, Aleta A, Du C, Shi L, Moreno Y. LLMs and generative agent-based models for complex systems research. Phys Life Rev. 2024;51:283-93.

https://doi.org/10.1016/j.plrev.2024.10.013

2. Izacard G, Grave E. Distilling Knowledge from Reader to Retriever for Question Answering. arXiv preprint arXiv:2012.04584 [Internet]. International Conference on Learning Representations, ICLR; 2022. Available from: http://arxiv.org/abs/2012.04584

3. Malhotra A, Jindal R. XAI Transformer based Approach for Interpreting Depressed and Suicidal User Behavior on Online Social Networks. Cogn Syst Res. 2024;84:101186.

https://doi.org/10.1016/j.cogsys.2023.101186

4. Mersha MA, Yigezu MG, Kalita J. Evaluating the effectiveness of XAI techniques for encoder-based language models. Knowl Based Syst. 2025;310:113042.

doi: 10.25100/iyc.v28i2.15700

https://doi.org/10.1016/j.knosys.2025.113042

5. Zohuri B, Behgounia F. Application of artificial intelligence driving nano-based drug delivery system. In: A Handbook of Artificial Intelligence in Drug Delivery [Internet]. Elsevier; 2023 [cited 2025 Sep 13]. p. 145-212.

https://doi.org/10.1016/B978-0-323-89925-3.00007-1

6. Campo Yule JE, Díaz Mage D alberto, Ordoñez HA. Técnicas de Machine Learning aplicadas al consumo de sustancias psicoactivas ilícitas: Un mapeo sistémico. Inge CuC. 2023;19(2):4.

https://doi.org/10.17981/ingecuc.19.2.2023.08

7. Petersen K, Vakkalanka S, Kuzniarz L. Guidelines for conducting systematic mapping studies in software engineering: An update. Inf Softw Technol. 2015;64:1-18.

https://doi.org/10.1016/j.infsof.2015.03.007

8. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;n71.

https://doi.org/10.1136/bmj.n71

9. Lin B. Reinforcement learning and bandits for speech and language processing: Tutorial, review and outlook. Expert Syst Appl. 2024;238:122254.

https://doi.org/10.1016/j.eswa.2023.122254

10. Zhou J, Li X, Xia Q, Yu L. Innovations in otolaryngology using LLM for early detection of sleep-disordered breathing. SLAS Technol. 2025;32:100278.

https://doi.org/10.1016/j.slast.2025.100278

11. Ravaut M, Zhao R, Phung D, Qin VM, Milovanovic D, Pienkowska A, et al. Targeting COVID-19 and Human Resources for Health News Information Extraction: Algorithm Development and Validation. JMIR AI. 2024;3:e55059.

https://doi.org/10.2196/55059

12. Yu D. Towards LLM-assisted move annotation: Leveraging ChatGPT-4 to analyse the genre structure of CEO statements in corporate social responsibility reports. English for Specific Purposes. 2025;78:33-49.

https://doi.org/10.1016/j.esp.2024.11.003

13. Kovari A. Explainable AI chatbots towards XAI ChatGPT: A review. Heliyon. 2025;11(2):e42077.

https://doi.org/10.1016/j.heliyon.2025.e42077

14. Tizaoui T, Tan R. Towards a benchmark dataset for large language models in the context of process automation. Digital Chemical Engineering. 2024;13:100186.

https://doi.org/10.1016/j.dche.2024.100186

15. Arslan M, Munawar S, Cruz C. Political Events using RAG with LLMs. Procedia Comput Sci. 2024;246:5027-35.

https://doi.org/10.1016/j.procs.2024.09.576

16. Rique T, Perkusich M, Gorgônio K, Almeida H, Perkusich A. Constructing the graphical structure of expert-based Bayesian networks in the context of software engineering: A systematic mapping study. Inf Softw Technol. 2025;177:107586.

doi: 10.25100/iyc.v28i2.15700

https://doi.org/10.1016/j.infsof.2024.107586

17. Chakraborty C, Bhattacharya M, Pal S, Chatterjee S, Das A, Lee SS. AI-enabled language models (LMs) to large language models (LLMs) and multimodal large language models (MLLMs) in drug discovery and development. J Adv Res. 2025;78:377-89.

https://doi.org/10.1016/j.jare.2025.02.011

18. Thomas J, Mudgal A, Liu W, Tahiraj N, Mohammed Z, Diddi D. Preserving Privacy, Increasing Accessibility, and Reducing Cost: An On-Device Artificial Intelligence Model for Medical Transcription and Note Generation.

https://doi.org/10.1101/2025.07.01.25330679

19. Yan B, Li K, Xu M, Dong Y, Zhang Y, Ren Z, et al. On protecting the data privacy of Large Language Models (LLMs) and LLM agents: A literature review. High-Confidence Computing. 2025;100300.

https://doi.org/10.1016/j.hcc.2025.100300

20. Ferrag MA, Alwahedi F, Battah A, Cherif B, Mechri A, Tihanyi N, et al. Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities. Internet of Things and Cyber-Physical Systems. 2025;5:1-46.

https://doi.org/10.1016/j.iotcps.2025.01.001

21. Adam D, Kliegr T. Traceable LLM-based validation of statements in knowledge graphs. Inf Process Manag. 2025;62(4):104128.

https://doi.org/10.1016/j.ipm.2025.104128

22. Nakamura M, Hayamizu S, Masanori H, Fuseya T, Iwamatsu H, Terada K. Causal Reasoning of Occupational Incident Texts Using Large Language Models. Procedia Comput Sci. 2024;246:820-9.

https://doi.org/10.1016/j.procs.2024.09.501

23. Zou Y, Shi M, Chen Z, Deng Z, Lei Z, Zeng Z, et al. ESGReveal: An LLM-based approach for extracting structured data from ESG reports. J Clean Prod. 2025;489:144572.

https://doi.org/10.1016/j.jclepro.2024.144572

24. Pagano S, Strumolo L, Michalk K, Schiegl J, Pulido LC, Reinhard J, et al. Evaluating ChatGPT, Gemini and other Large Language Models (LLMs) in orthopaedic diagnostics: A prospective clinical study. Comput Struct Biotechnol J. 2025;28:9-15.

https://doi.org/10.1016/j.csbj.2024.12.013

25. Arslan M, Mahdjoubi L, Munawar S. Driving sustainable energy transitions with a multi-source RAG-LLM system. Energy Build. 2024;324:114827.

https://doi.org/10.1016/j.enbuild.2024.114827

26. Jain T, Gao Y, Vanga S, Singla K. News Reporter: A Multi-lingual LLM Framework for Broadcast TV News. arXiv preprint arXiv:2410.07520 [Internet]. 2024. Available from: https://arxiv.org/pdf/2410.07520

27. Tonouchi Y, Nakai S, Nurakami K, Kataoka Y. Effectiveness of a Large Language Model-Based Feedback System for Case Report Writing in Novice Rehabilitation Staff Education: A Mixed-Methods Study. Rehabilitation [Internet]. Jxiv preprint; 2024. Available from: https://jxiv.jst.go.jp/index.php/jxiv/preprint/download/844/2450/2296

doi: 10.25100/iyc.v28i2.15700

28. Sacoransky E, Kwan BYM, Soboleski D. ChatGPT and assistive AI in structured radiology reporting: A systematic review. Curr Probl Diagn Radiol. 2024;53(6):728-37.

https://doi.org/10.1067/j.cpradiol.2024.07.007

29. Scanlon M, Breitinger F, Hargreaves C, Hilgert JN, Sheppard J. ChatGPT for digital forensic investigation: The good, the bad, and the unknown. Forensic Science International: Digital Investigation. 2023;46:301609.

https://doi.org/10.1016/j.fsidi.2023.301609

30. Gmeiner F, Yildirim N. Dimensions for Designing LLM-based Writing Support. In: In2Writing Workshop at CHI [Internet]. Hamburg, Germany: Association for Computing Machinery (ACM); 2023 [cited 2026 May 3]. Available from: https://www.frederic-otto.com/papers/DimensionsforDesigningLLM-basedWritingSupport.pdf

31. Hatem S, Khoriba G, Gad-Elrab MH, ElHelw M. Up To Date: Automatic Updating Knowledge Graphs Using LLMs. Procedia Comput Sci. 2024;244:327-34.

https://doi.org/10.1016/j.procs.2024.10.206

32. Jiang G, Shi X, Luo Q. Llm-collaboration on automatic science journalism for the general audience. arXiv preprint arXiv:2407.09756. 2024.

https://doi.org/10.48550/arXiv.2407.09756

33. Cruz-Castro L, Castelblanco G, Antonenko P. LLM-based system for technical writing real-time review in urban construction and technology. Proceedings of 60th Annual Associated Schools. 2024;5:130-8.

https://doi.org/10.29007/d9j3

34. Chakrabarty T, Laban P, Wu CS. Can AI writing be salvaged? Mitigating Idiosyncrasies and Improving Human-AI Alignment in the Writing Process through Edits. arXiv preprint arXiv:2409.14509 [Internet]. 2024. Available from:

https://doi.org/10.1145/3706598.3713559

35. Mazzone S, Harlan J, Xu LZ, Ow T. LLMs Enhance Emotional Expression While Maintaining Analytical Depth in News Writing. In: Scholar Space [Internet]. 2025.

https://doi.org/10.24251/HICSS.2025.278

36. Soós D. Who Wrote the Scientific News? Improving the Discernibility of LLMs to Human-Written Scientific News. Old Dominion University; 2024.

https://digitalcommons.odu.edu/computerscience_etds/178

37. Zheng L, Jiang F, Gu X, Li Y, Wang G, Zhang H. Teaching via LLM-enhanced simulations: Authenticity and barriers to suspension of disbelief. Internet High Educ. 2025;65:100990.

https://doi.org/10.1016/j.iheduc.2024.100990

38. Zhang H, Yan W, Hu H, Zhang X, Liu Q, Xia H, et al. An LLM-based knowledge and function-augmented approach for optimal design of remanufacturing process. Advanced Engineering Informatics. 2025;65:103206.

https://doi.org/10.1016/j.aei.2025.103206

39. Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Ser J Del, et al. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion. 2024;106:102301.

https://doi.org/10.1016/j.inffus.2024.102301

40. Haurogné J, Basheer N, Islam S. Vulnerability detection using BERT based LLM model with transparency obligation practice towards trustworthy AI. Machine Learning with Applications. 2024;18:100598.

https://doi.org/10.1016/j.mlwa.2024.100598

41. Brandsæter A, Glad IK. XAI in hindsight: Shapley values for explaining prediction accuracy. Expert Syst Appl. 2025;273:126845.

https://doi.org/10.1016/j.eswa.2025.126845

42. Weber L, Lapuschkin S, Binder A, Samek W. Beyond explaining: Opportunities and challenges of XAI-based model improvement. Information Fusion. 2023;92:154-76.

https://doi.org/10.1016/j.inffus.2022.11.013

43. Jouis G, Mouchère H, Picarougne F, Hardouin A. A methodology to compare XAI explanations on natural language processing. In: Benois-Pineau J, Bourqui R, Petkovic D, Quénot G, editors. Explainable Deep Learning AI [Internet]. Imprint Academic Press (libro 2023): Elsevier; 2023. p. 191-216.

https://doi.org/10.1016/B978-0-32-396098-4.00016-8

44. Martens D, Hinns J, Dams C, Vergouwen M, Evgeniou T. Tell me a story! Narrative-driven XAI with Large Language Models. Decis Support Syst. 2025;191:114402.

https://doi.org/10.1016/j.dss.2025.114402

45. Shimizu I, Kasai H, Shikino K, Araki N, Takahashi Z, Onodera M, et al. Developing Medical Education Curriculum Reform Strategies to Address the Impact of Generative AI: Qualitative Study. JMIR Med Educ. 2023;9.

https://doi.org/10.2196/53466

46. Erickson JS, Santos H, Pinheiro V, McCusker JP, McGuinness DL. LLM experimentation through knowledge graphs: Towards improved management, repeatability, and verification. J Web Semant. 2025;85:100853.

https://doi.org/10.1016/j.websem.2024.100853

47. Perera Molligoda Arachchige AS. Empowering radiology: the transformative role of ChatGPT. Clin Radiol. 2023;78(11):851-5.

https://doi.org/10.1016/j.crad.2023.08.006

48. Sarbaree Mishra. The age of explainable AI: improving trust and transparency in AI models. International Journal of Artificial Intelligence, Data Science, and Machine Learning. 2020;1.

https://doi.org/10.63282/3050-9262.IJAIDSML-V1I4P105

49. Chan PYP, Keung J, Yang Z. Effectiveness of symmetric metamorphic relations on validating the stability of code generation LLM. Journal of Systems and Software. 2025;222:112330.

https://doi.org/10.1016/j.jss.2024.112330

50. Xu Z, Song T, Lee YC. Confronting verbalized uncertainty: Understanding how LLM's verbalized uncertainty influences users in AI-assisted decision-making. Int J Hum Comput Stud. 2025;197:103455.

https://doi.org/10.1016/j.ijhcs.2025.103455

51. Rapp A, Di Lodovico C, Di Caro L. How do people react to ChatGPT's unpredictable behavior? Anthropomorphism, uncanniness, and fear of AI: A qualitative study on individuals' perceptions and understandings of LLMs' nonsensical hallucinations. Int J Hum Comput Stud. 2025;198:103471.

https://doi.org/10.1016/j.ijhcs.2025.103471

52. Vincent S, Wang P, Shi Z, Koka S, Fang Y. Measuring Large Language Models Capacity to Annotate Journalistic Sourcing.

https://doi.org/10.48550/arXiv.2501.00164

53. Dibia V. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. 2023

https://doi.org/10.18653/v1/2023.acl-demo.11

doi: 10.25100/iyc.v28i2.15700