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

Vol 28 (2)

ISSN
0123-3033
e- 2027-8284

Original Article

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

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

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

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

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

How to cite?

Ruiz J, Ordóñez HA, Romero RM. Generation of descriptive reports using generative artificial intelligence: A systematic mapping. Ingeniería y Competitividad, 2026, 28(2)e-30115700

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

Received: 10/03/26
Reviewed: 8/04/26
Accepted: 14/05/26
Online: 21/05/26

Correspondence

hugoordonez@unicauca.edu.co

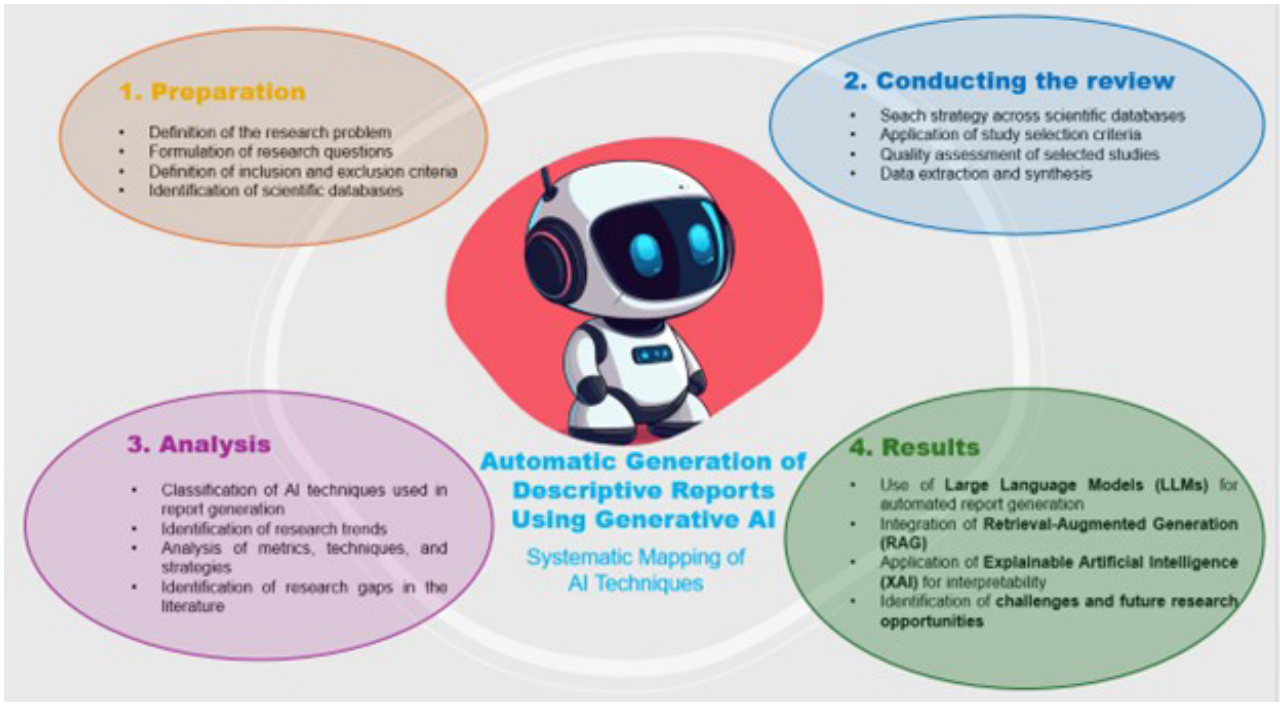
Spanish version

Why was the study conducted?
This work allows for the organization of scattered literature, the identification of metrics, strategies, and limitations in the automatic generation of descriptive reports based on LLM and XAI. Furthermore, it demonstrates a lack of evaluative approaches to reliability, highlighting the absence of structured guidelines for integrating explainable models into decision-making contexts.

What were the key findings?
The most relevant results include the identification of high levels of accuracy in structured domains, specifically in the healthcare sector, where information is typically organized using formats, controlled terminology, and clearly defined variables. These characteristics foster a suitable environment for training and evaluating LLM models, allowing for consistent and accurate results compared to open or unstructured contexts. Additionally, the study revealed the recurring use of metrics such as robustness, consistency, and human reasoning agreement to assess model reliability. Finally, it found a scarcity of comprehensive tools aimed at the automatic generation of explainable reports.

What do these findings contribute?
These results provide a basis for the development of future research aimed at creating standardized guidelines, XAI-based tools, and reliable systems for the automatic generation of descriptive reports in various contexts.

Graphical Abstract



doi: 10.25100/iyc.v28i2.15700

Introduction

Large Language Models (LLMs) have attracted significant interest due to their ability to generate natural language text and answer questions coherently (1). Their implementation is based on the transformer architecture, which enables parallel processing of entire sequences, reducing processing time compared to Recurrent Neural Networks (RNNs) (2). These models employ self-supervised pretraining schemes followed by supervised fine-tuning, enabling them to achieve outstanding performance across Natural Language Processing (NLP) tasks (3).

From a critical perspective, challenges related to the reliability of generated content persist. These include the generation of incorrect information, the amplification of biases, and the lack of clear mechanisms for interpreting the inference process. Such conditions directly affect trust in the generated texts and limit their application in sensitive contexts, such as the automated generation of descriptive reports (3).

In response to this issue, various techniques, methods, and strategies have been developed to enhance the robustness and reliability of LLMs. Among these, Explainable Artificial Intelligence (XAI) approaches and perturbation-based methods, such as SHAP (SHapley Additive exPlanations), stand out, as they aim to provide interpretability and transparency in model predictions (4). However, these contributions remain scattered across studies and lack a structured systematization that would allow the identification of research gaps, limitations, and opportunities (5).

In this context, the present study aims to conduct a systematic mapping of the literature published between 2020 and 2025 on reliability techniques applied to LLMs for the generation of descriptive reports. This mapping will make it possible to identify predominant approaches, methodological limitations, and potential future directions to strengthen consistency and reliability in the automated generation of descriptive reports.

Section 2 describes the materials and methods used for the selection and analysis of the articles included in the systematic mapping. Section 3 presents the results and discussion. Finally, Section 4 outlines the conclusions and future research opportunities derived from the study.

Materials and methods

In this study, the methodology used is aligned with the model proposed in the literature (6) (See Figure 1).

doi: 10.25100/iyc.v28i2.15700

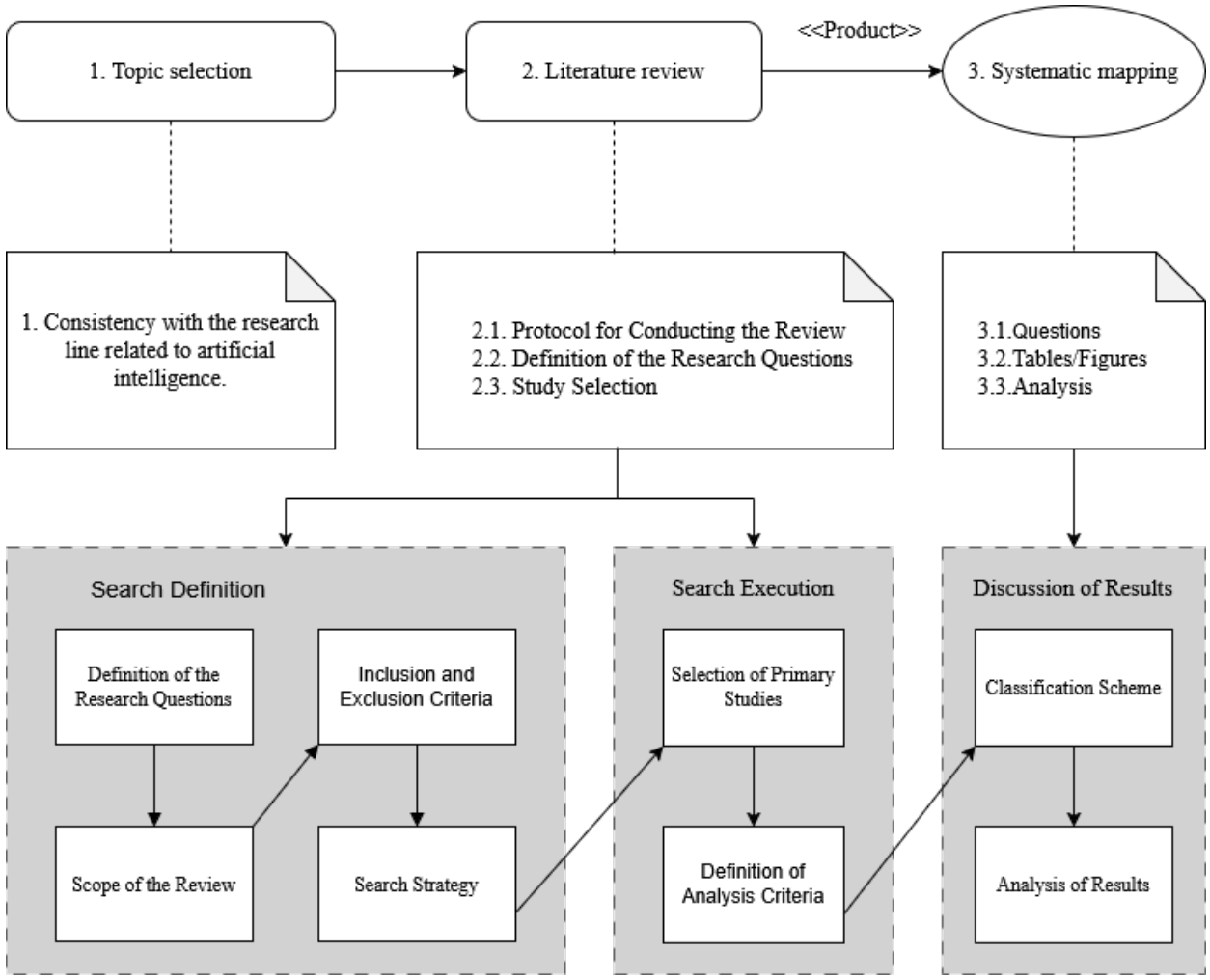


Figure 1. System Mapping Process Note. Adapted from (6).

The three main processes are described below: 1. Topic selection, which in this case is related to descriptive reports and XAI; 2. Literature review, where the filters used to obtain the most relevant literature for the study are specified, and 3. Systematic mapping, where the methodology and structure applied for selecting relevant studies are presented. All the above is based on the stages described in (7).

Topic Selection

At this stage, the following characteristics are considered:

1. The topic selection is based on the need to analyze the automatic generation of reports using LLMs, integrating interpretation techniques such as SHAP due to their relevance in governance contexts.

2. The adaptability of a tool to any type of environment or use within governmental, private, or administrative entities for the generation of automated reports.

Literature Review

One of the most important steps is formulating research questions to guide and focus on the objective of the study (see Table 1). At this stage, the state of the art regarding the use of XAI and AI

doi: 10.25100/iyc.v28i2.15700

in the generation of descriptive reports is reviewed, and the scientific literature relevant to the study is identified.

Table 1. Research Questions

Identifier	Research Questions	Motivation
QIG	What is the current state of knowledge regarding the use of LLMs models with XAI in the automatic generation of reports?	To understand the current state of the topic and the different models used for natural language (NL) text generation.
QI1	What metrics and approaches are used to evaluate the quality and reliability of reports generated by LLMs?	To understand the effectiveness of current methods for analyzing and contextualizing NL in report generation.
QI2	What metrics, techniques, and strategies are used in the application of LLMs models with XAI in reports?	To identify the metrics, techniques, and strategies that have proven effective in NL text generation.
QI3	What are the current applications of LLMs models with XAI in reports?	To understand the scope of the problem and identify how it has been addressed, whether through specific strategies within the same context or, on the contrary, whether research gaps still exist.

Note. Author own elaboration.

The next step is to establish the inclusion and exclusion criteria to determine the relevant specialized literature.

Inclusion Criteria

Articles published within the period from 2022 to 2025.

Articles that address the research questions.

Articles related to LLMs and the use of NLP in report generation.

Articles written in English.

Exclusion Criteria

Duplicate articles, considering only the most recent version.

Articles unrelated to LLMs or NLP.

Incomplete articles.

Regarding the information sources, the databases selected were ScienceDirect, SpringerLink, Google

doi: 10.25100/iyc.v28i2.15700

Scholar, and *IEEE Xplore* due to their extensive coverage in the fields of AI, NLP, and engineering.

ScienceDirect and SpringerLink: These were prioritized because of their large volume of peer-reviewed specialized literature in computer science and their recognition in high-quality academic research.

Google Scholar: This was used as a complementary source to broaden coverage and reduce indexing bias, allowing the inclusion of literature not found in traditional databases.

IEEE Xplore: This source was included due to its specialized focus on engineering and emerging technologies, particularly applied AI.

In this way, the combination of these sources ensured quality, coverage, and up-to-date publications in the collected literature. Furthermore, the bibliographic sources *ScienceDirect*, *SpringerLink*, and *IEEE Xplore* will hereafter be referred to as indexed sources.

Finally, the main terms used to focus and delimit the search for articles related to LLMs and report generation are defined in Table 2. Likewise, the search strings were constructed using logical operators such as "AND" and "OR."

Table 2. Search strings

Main Terms	Search String
LLMs y NLP	(("Large language models" AND models) AND ("Report descriptive" AND Automation AND Visualization) AND (Natural))
LLMs y Written Reports	(("Large language models" AND models) AND ("Report descriptive" AND Automation AND Visualization))
LLMs 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))

Note. Authors' own elaboration.

Selected studies

To identify the primary studies, the inclusion and exclusion criteria were applied to remove duplicate results and incomplete research works. In addition, two review filters were included, as described below:

doi: 10.25100/iyc.v28i2.15700

First filter (1F): Review of the title and abstract.

Second filter (2F): Publications that passed the first filter were subjected to a full reading and comprehensive content analysis.

Considering all the previous steps, the primary studies were successfully identified, as illustrated in the summary presented in Figure 2.

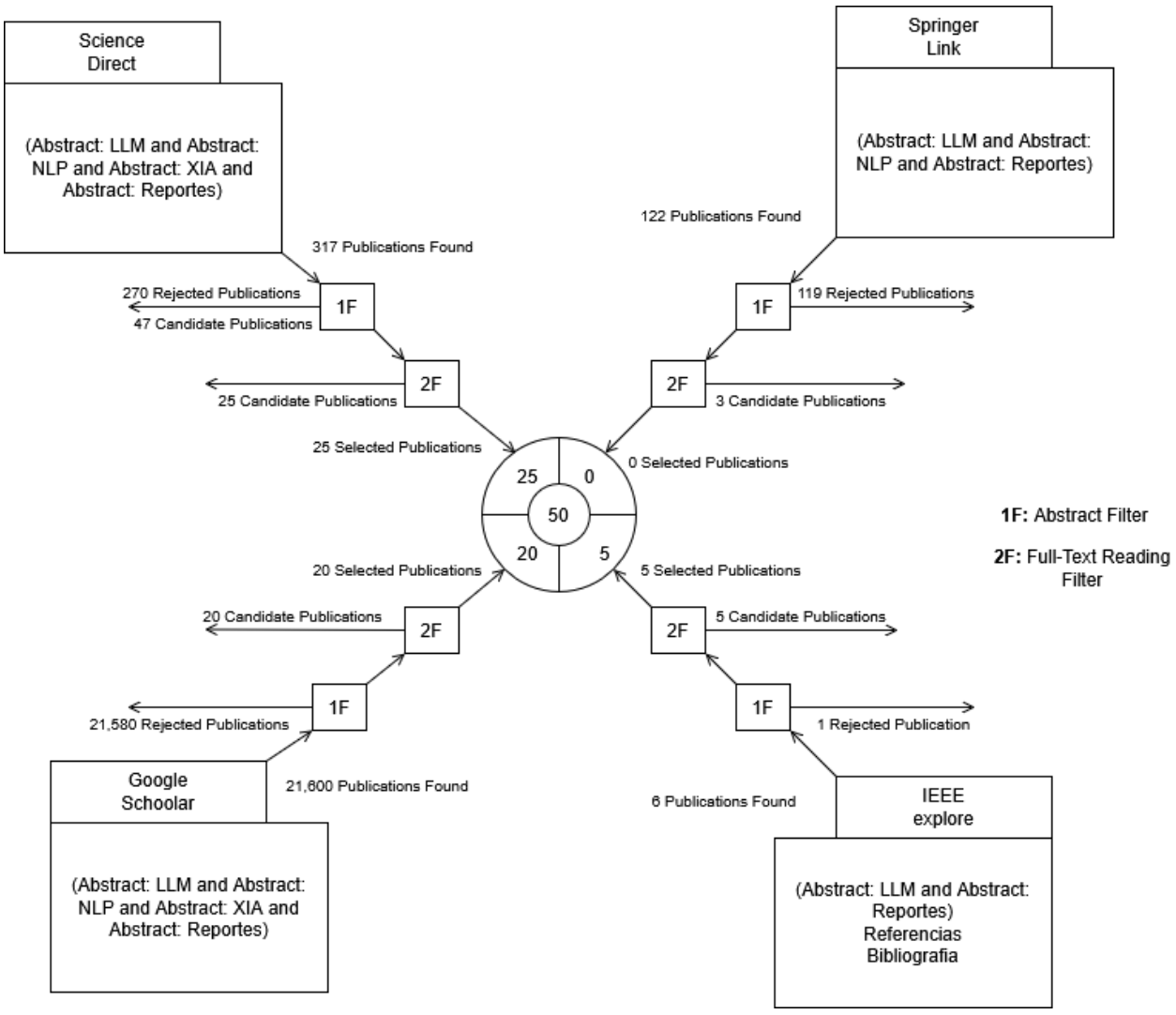


Figure 2. Funnel chart for the search for primary studies in DCM format. Note. Authors' own elaboration.

The process shown in Figure 2 was structured into two filtering stages. First, the inclusion and exclusion criteria were applied together with the first filter (1F), which was based on reviewing the title and abstract of the identified publications. Considering both the indexed sources (n = 55) and Google Scholar (n = 20), a total of (n = 75) specialized texts were obtained for the next phase. On the other hand, (n = 390) studies from indexed sources and (n = 21,580) from *Google Scholar* were rejected.

In the case of *Google Scholar*, due to its nature and the large volume of publications (n = 21,600), a manual selection strategy was implemented. In this process, the first 100 results sorted by relevance

doi: 10.25100/iyc.v28i2.15700

were considered, since the search engine prioritizes publications with greater impact, number of citations, and semantic similarity with the search strings.

Subsequently, the second filter (2F) and the second stage were applied, corresponding to the full reading of the articles selected in the first stage. During this phase, (n = 25) studies were excluded because they did not meet the defined criteria, contained incomplete information, or did not answer the research questions.

Finally, the selection process resulted in a total of (n = 50) primary studies, distributed into (n = 30) articles from indexed sources and (n = 20) articles obtained through the manual selection process in Google Scholar. This selection flow is presented in detail in Figure 3, following an adaptation of the PRISMA model (8), which ensures the traceability and reproducibility of the process.

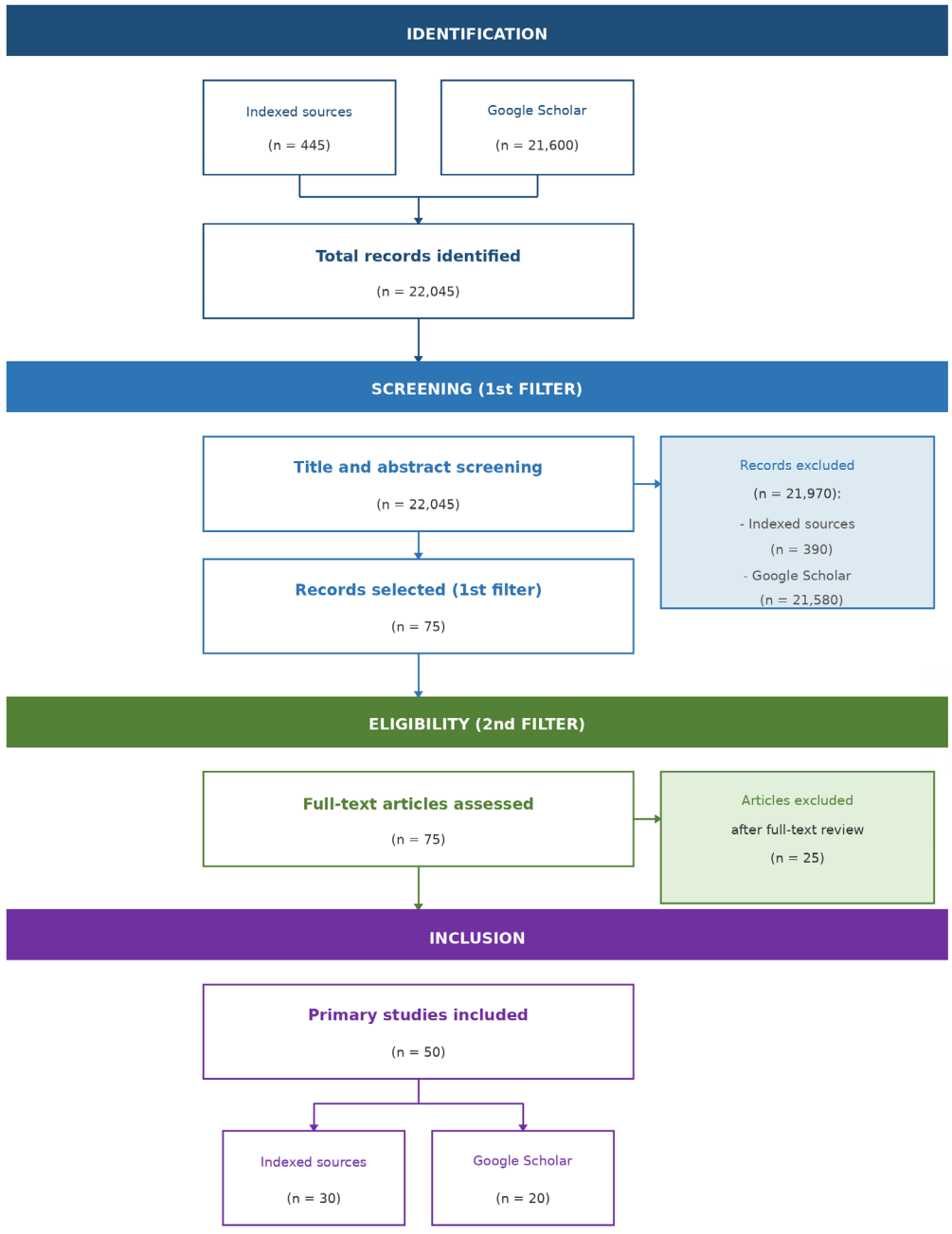


Figure 3. PRISMA model-based process flowchart. Note. Authors' own elaboration.

doi: 10.25100/iyc.v28i2.15700

Results and Discussion

The literature addressing problems related to the generation of descriptive reports through Natural Language Processing (NLP) and Large Language Models (LLMs) is extensive. To facilitate understanding of the difficulty, the studies were divided into three primary areas: 1. LLMs models for generating descriptive reports, 2. LLMs models in report development, and 3. LLMs models and trust assurance techniques.

LLMs Models for generating descriptive reports

In recent years, significant efforts have been made in Natural Language Generation (NLG) to produce coherent and human-readable texts. However, classical models such as Seq2Seq tend to generate generic and poorly contextualized responses, which limit their applicability in scenarios requiring deeper contextual understanding (9).

Several studies have explored NLP and LLMs applications in the generation of descriptive reports. In the healthcare domain, machine learning techniques combined with LLM-based models, such as the DSS–LLM approach, have been used to process large volumes of clinical information, reduce analysis times, and achieve accuracies between 98.58% and 98.91% in early diagnosis tasks (10). Complementarily, rule-based models and Natural Language Understanding (NLU) systems have been employed to automate processes through question-and-answer schemes trained on domain-specific datasets (11) (12).

Other studies have incorporated Knowledge Graphs (KGs) and Generative Adversarial Networks (GANs) to improve the explainability and reliability of LLM-based chatbots, as well as to support the semi-automated generation of data in automation processes (13) (14). Likewise, the use of classical NLP techniques integrated into modern models such as Llama 2 has been identified, leveraging preexisting knowledge bases to accelerate solution development and mitigate information access limitations (15).

In contexts characterized by high uncertainty, hybrid approaches combining LLMs with Bayesian networks and expert knowledge have been proposed, increasing the credibility of models that lack formal validation mechanisms (16). In pharmaceutical and clinical environments, LLMs have demonstrated high diagnostic accuracy to support strategic decision-making, simplify medical terminology, and strengthen physician–patient interaction (17).

Nevertheless, these advances also present challenges related to privacy, resource consumption, information redundancy, and data security (18). Recent studies have analyzed vulnerabilities such as prompt injection, and data poisoning, among others, recommending mitigation strategies to strengthen model robustness (19) (20). Additionally, a lack of research focused on the quality and verification of KGs used alongside LLMs has been identified, although promising results have been reported in the automatic validation of statements using models such as Llama 3 (21).

Finally, LLMs have also been applied to incident report generation in industrial environments, achieving accuracy above 90% in information extraction tasks using variants such as ChatGPT-3.5, demonstrating their potential in real-world operational scenarios (22).

doi: 10.25100/iyc.v28i2.15700

LLMs Models in report development

The state of the art demonstrates a growing interest in the use of LLMs for report development across different domains. Several studies have evaluated their feasibility in generating structured reports, highlighting significant improvements in efficiency and reductions in processing time compared to manual procedures. For example, the use of ChatGPT-4 in the preparation of Corporate Social Responsibility (CSR) reports achieved an accuracy of 87.14%, although it also revealed an inconsistency rate of 32.87%, emphasizing the need for human verification (12).

In this context, Retrieval-Augmented Generation (RAG) approaches have been integrated to strengthen report quality. Techniques such as ESGReveal achieved accuracies of 76.9% and extraction rates of 83.7% in data analysis tasks, demonstrating advances in information traceability (23). Similarly, comparative evaluations of different LLMs in medical diagnosis tasks, such as osteoarthritis detection, achieved sensitivities of up to 92.3% with models such as ChatGPT-4, reaching performance levels comparable to human specialists, although the need for expert validation remains highlighted (24).

Likewise, LLMs have been applied in decision-making processes through the analysis of large volumes of information obtained from news articles, government reports, sector publications, and social media, combined with RAG techniques (25). Nevertheless, these studies identify recurring limitations related to data privacy, result reliability, and the presence of medical or informational errors (26) (27) (28). Although these studies originate from heterogeneous domains, they allow the identification of general patterns in LLMs behavior regarding report generation.

While LLMs can generate useful responses based on user instructions, their effectiveness depends on the user's level of domain knowledge. For experienced users, the models function as support tools; however, in low-knowledge contexts, they may introduce information gaps or reinforce incorrect assumptions (29).

From a systematic perspective, taxonomies have been proposed to organize the design and use of LLMs as support tools for natural language writing, addressing issues such as model fragmentation and the use of outdated data. The integration of advanced reasoning capabilities and RAG-based approaches has enabled Knowledge Graphs (KGs) to contribute to automatic information updating with lower error margins (30) (31) (32).

Finally, LLMs have also been leveraged for the technical correction of specialized texts in areas such as architecture, engineering, and construction, as well as for the identification of redundant stylistic patterns in AI-generated texts. These approaches have led to new categorizations aimed at adapting automated text generation to human quality standards (33) (34) (35) (36).

LLMs Models and trust assurance techniques

As the use of LLMs increases, concerns regarding the reliability and consistency of generated knowledge also continue to grow. In response, the literature has proposed various techniques aimed at trust assurance, primarily supported by XAI approaches and automatic verification mechanisms.

doi: 10.25100/iyc.v28i2.15700

In the educational domain, simulation-based learning models integrated with LLMs have been explored, enabling the creation of realistic environments for the development of teaching competencies. Although these approaches demonstrate pedagogical benefits, technical challenges related to latency and the models' ability to understand complex contexts persist (37).

In industrial settings, RAG-based strategies with augmented functions have been implemented to improve complex decision-making processes, such as recovering failures in returned products during remanufacturing processes. These approaches achieved average accuracies between 65% and 79%, demonstrating practical benefits despite the inherent limitations of the underlying base models (38).

Additionally, specific metrics have been proposed to validate the effectiveness of explainability techniques in LLMs, including human reasoning agreement, robustness, consistency, and contrastiveness. These metrics enable objective model comparison and establish foundations for the development of trustworthy architectures (3) (4) (13) (39) (40) (41) (42) (43) (44).

Complementarily, several studies have applied strategies such as prompt engineering, RAG, and the incorporation of Knowledge Graphs (KGs) to verify the reliability of automated report generation in domains such as education, healthcare, and knowledge management. In this context, particular attention is given to the responsibility associated with the use of models such as ChatGPT-4 in critical decision-making processes, given their growing influence on the generation and dissemination of knowledge (45) (46) (47) (48).

Based on the reviewed articles, each study was analyzed to identify common themes, relationships among them, and elements that provide accurate answers to the proposed research questions (see Table 1). Considering this, the following steps of the study were conducted.

Answers to the research questions

QI1. What metrics and approaches are used to evaluate the quality and reliability of reports generated by LLMs?

The evaluation of the quality and reliability of generated reports is based on a set of quantitative metrics and approaches focused on interpretability and model consistency.

First, traditional quantitative metrics such as precision, recall, F1-score, and Area Under the Curve (AUC) are commonly used, particularly in structured domains such as the clinical field. These metrics make it possible to evaluate model performance in classification and prediction tasks, serving as baseline indicators of system performance.

Second, metrics focused on reliability and explainability have been identified, such as consistency, robustness, and human reasoning agreement. These metrics help assess the model's stability against variations in the data and the alignment between generated responses and human interpretation, which is essential in the generation of descriptive reports.

The analyzed studies show a clear tendency toward elevated levels of accuracy in structured domains, such as the healthcare sector, where models achieve values above 90%. In contrast, in industrial and open-domain environments, performance variation is greater, with ranges between

doi: 10.25100/iyc.v28i2.15700

65% and 79%, as illustrated in Figure 4. However, in more complex scenarios, such as descriptive report generation, it is necessary to complement these metrics with approaches that evaluate coherence, interpretability, and the reliability of the generated content.

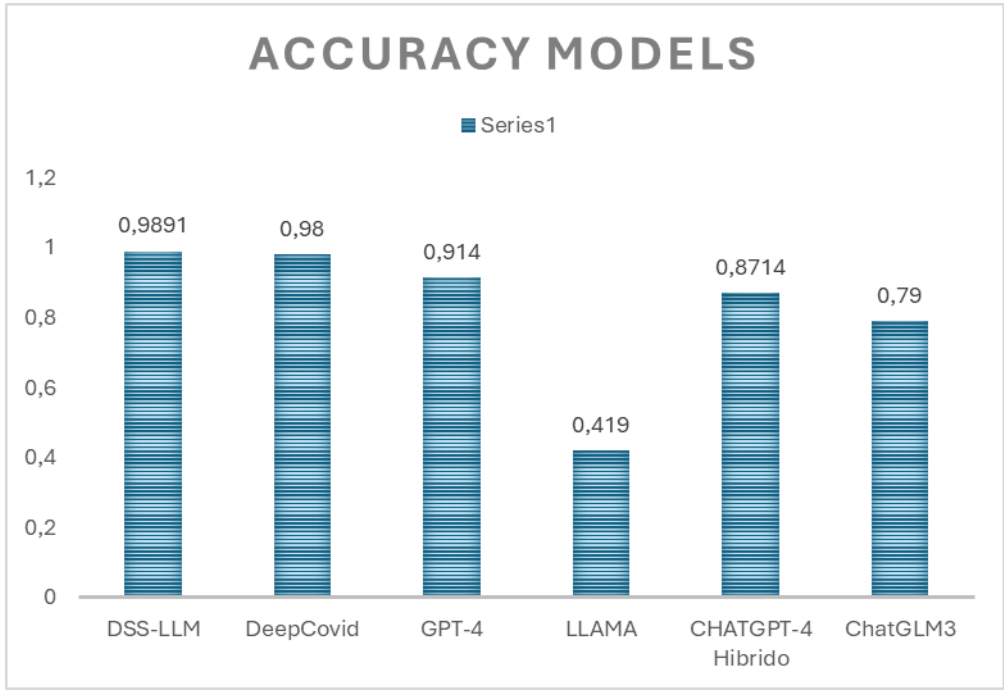


Figure 4. Accuracy graph of the models identified in research question QI1. Note. Authors' own elaboration

Finally, it is observed that traditional metrics, such as precision and AUC predominate in structured domains. However, more recent metrics such as consistency and human reasoning agreement are increasingly employed to validate reliability in descriptive report generation.

QI2. What metrics, techniques, and strategies are used in the application of LLMs models with XAI in reports?

The systematic mapping identified a variety of metrics, techniques, and strategies aimed at increasing the reliability, interpretability, and stability of LLMs in the automatic generation of reports. Rather than isolated approaches, the studies converge into three main dimensions: explainable evaluation, uncertainty management, and semantic validation.

In the domain of explainable evaluation metrics, human reasoning agreement, robustness, consistency, and contrastiveness stand out, being applied to compare explainability techniques on datasets such as Movie Reviews (IMDB) and Tweet Sentiment Extraction (TSE) (4). These metrics enable alignment between the model's reasoning and human interpretation.

Regarding explainability techniques, SHAPStories and CFStories were employed to improve user understanding of predictions generated by LLMs through explanatory narratives (44). Likewise, Metamorphic Relations (MRs) have been proposed as empirical validation strategies across multiple LLMs, demonstrating improvements in stability and semantic coherence (49).

doi: 10.25100/iyc.v28i2.15700

With respect to uncertainty management, verbalized and non-verbalized uncertainty mechanisms are analyzed to reduce interpretative gaps and mitigate misleading perceptions caused by model hallucinations (50) (51).

Additionally, in journalistic and information transparency contexts, fuzzy matching, semantic similarity, and exact matching metrics are used to evaluate the accuracy of source attribution in models such as ChatGPT-4, Claude, and Gemini (52).

Overall, these studies consolidate a technical framework aimed at strengthening trust, stability, and traceability in the automatic generation of descriptive reports based on LLMs.

QI3. What are the current applications of LLMs models with XAI in reports?

The analysis of current applications of LLMs models combined with XAI in report generation reveals an emerging landscape. The systematic review identified a limited number of consolidated tools, among which LIDA (53) stands out, a project focused on the generation of visualizations and infographics using LLMs such as ChatGPT-4. This tool integrates processes ranging from data interpretation to the generation of visual representations, establishing a specific niche in the automation of analytical and illustrative processes.

However, among the analyzed studies, the limited presence of tools reveals a significant gap in applied research on LLMs for the automated generation of both descriptive and visual reports. This gap in the state of the art suggests the need to develop comprehensive approaches capable of integrating content generation, explainability, and result validation within a single environment. In this sense, an important opportunity for future research is identified, particularly in the design of guidelines and tools that facilitate the implementation of LLM-based systems with consistency capabilities oriented toward real-world governance contexts.

As presented in the previous research questions, the study consolidated the methods, algorithms, and models with the highest reported accuracy, including DSS-LLM with an accuracy of 98.91%, as well as DeepCovid and ChatGPT-4 in healthcare diagnosis tasks. These models demonstrate strong potential as reference candidates for future research on automated descriptive report generation. Furthermore, regarding evaluation metrics, the study identified human reasoning agreement, robustness, consistency, and contrastiveness as key measures for assessing model reliability and improving the accuracy and rigor of generated text. Likewise, these metrics can be used to validate or evaluate new tools in terms of consistency and coherence in generated content.

Additionally, the tool LIDA was identified as an example that provides insights into the interaction and development of software tools integrated with LLMs. It may also support the development of guidelines and recommendations for the use of this type of system. In contrast, it remains necessary to incorporate support metrics focused on self-attention and comprehension capabilities, particularly regarding knowledge gaps generated by the improper use of LLMs combined with RAG approaches. This suggests that the integration of external knowledge contributes not only to improved accuracy but also to enhanced traceability of generated content.

doi: 10.25100/iyc.v28i2.15700

Figure 5 shows the relevance of articles grouped by year and shared topics. The larger the bubble, the greater the number of articles associated with a given topic and publication year. If no bubble is present at an intersection, it indicates that no articles were found or selected for that specific point.

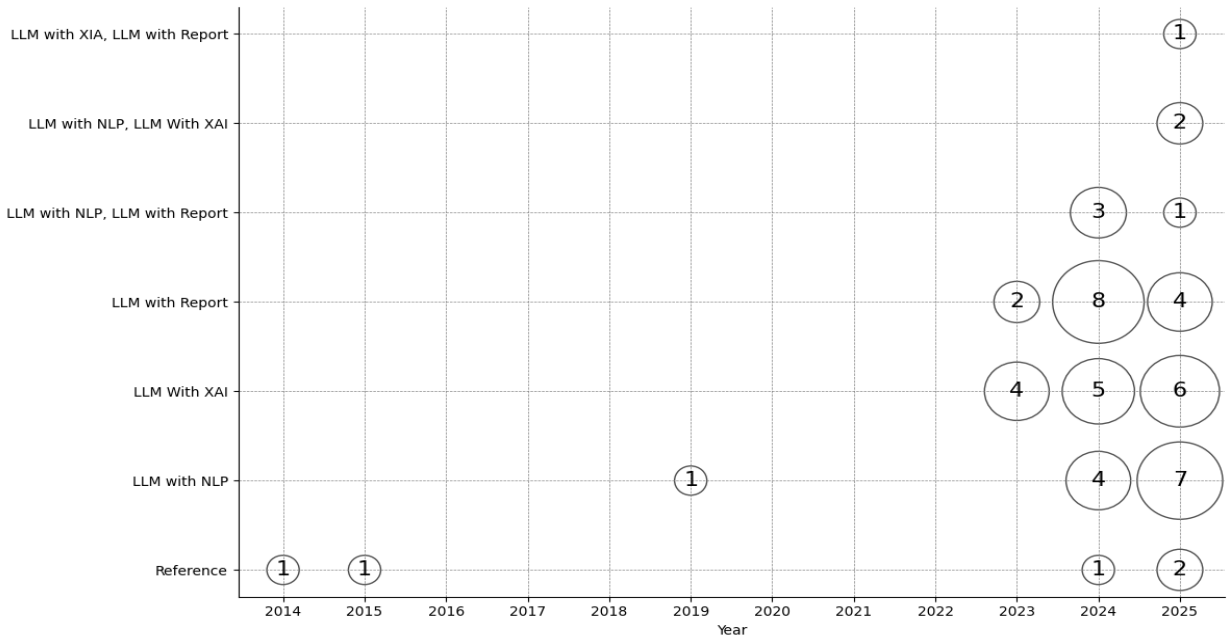


Figure 5. Bubble chart represents the number of publications by topic with respect to the year of publication. Note. Authors' own elaboration

On the other hand, it can be observed that most of the articles were published in recent years, specifically between 2023 and 2025. This is important, as it suggests that the automatic generation of descriptive reports using XAI with LLMs has become a relevant research topic in recent years and is likely to continue growing soon.

Discussion of results

The discussion of the results identifies four key points related to trends and challenges. First, LLMs and generative artificial intelligence have achieved high levels of accuracy in structured contexts, especially in domains such as healthcare, where hybrid models such as ChatGPT-4 (24), Deepcovid (11), and DSS-LLM (10) report values ranging from 76.9% to 98.1%. This indicates a strong relationship between LLMs and text generation for report creation using natural language (NL). However, it also establishes an ambitious objective for current tools in terms of guaranteeing the reliability and generalization of these models in more complex and less structured contexts.

Second, an important limitation was identified regarding the absence of standardized guidelines for validating and monitoring the robustness of LLMs. Although the studies propose several metrics, such as human reasoning agreement, robustness, consistency, and contrastiveness, in the context of report generation (4), other approaches such as perception-based evaluation (51), metamorphic

doi: 10.25100/iyc.v28i2.15700

relations (49), and matching techniques (52), are also presented. Nevertheless, these methods remain fragmented. This lack of uniformity complicates and delays the consolidation of a common evaluation framework, highlighting the need for future studies aimed at integrating and validating metrics that enable consistent model verification using XAI techniques (13).

Third, the results reveal a limitation in the integration of descriptive and visual report generation within the same LLM-based environment. Despite modifications made to the different search strings to address the topic from a holistic perspective, only a limited number of studies were identified that develop comprehensive tools combining text generation, explainability, and data visualization (53). This situation highlights a gap in applied research, specifically in the design of complete solutions oriented toward end users.

Finally, a limited number of consolidated tools for the automatic generation of specialized texts was identified. This barrier may be associated with the lack of guidelines for integrating software engineering processes with LLMs. To address this specific domain, a future research direction is proposed involving the development of solutions that integrate LLMs models with explainability techniques, enabling both text generation and reliability validation within the same environment (4).

Likewise, issues identified in the literature, such as prompt sensitivity, hallucinations, and the inherent characteristics of different models, continue to represent critical challenges, reinforcing the need to incorporate XAI-based approaches and intelligent agents capable of improving the transparency and reliability of these systems in real-world environments (51)

Conclusions

The systematic mapping conducted from fifty relevant articles identified the challenges and limitations associated with the use of LLMs for the automatic generation of reliable reports. The results reveal an important level of dispersion in methodologies, metrics, and tools related to this topic, which complicates the consolidation of best practices and limits their adoption across different application domains.

Although current models have achieved elevated levels of accuracy, there is still a long way to go before reliable and consistent text generation can be achieved in real-world scenarios. Likewise, the lack of standardized evaluation criteria reinforces the need to incorporate unified methodological frameworks that integrate both technical metrics and reliability-oriented approaches.

Furthermore, the absence of consolidated tools and the limited integration between LLMs developments and software engineering methodologies have highlighted the lack of consensus regarding technical guidelines between these two areas. In this sense, the combination of LLMs with XAI techniques and their implementation within integrated platforms emerges as a key strategy for increasing reliability.

Overall, these findings provide direction toward the formulation of technical guidelines that strengthen robustness in both LLMs construction and the automatic generation of descriptive reports, enabling their effective application in critical areas such as healthcare, journalism, education, and broader societal challenges.

doi: 10.25100/iyc.v28i2.15700

Declaration on the Use of Generative Artificial Intelligence and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used ChatGPT (GPT-4) to assist in the writing process and improve readability and language. After using this tool, the authors reviewed and edited the content as necessary and assumed full responsibility for the content of the publication.

CrediT authorship contribution statement

Conceptualization - Ideas: Hugo Armando Ordóñez . Data curation: Hugo Armando Ordóñez . Formal analysis: Jhonfer Ruiz Figueroa. Investigation: Jhonfer Ruiz Figueroa. Methodology: Jhonfer Ruiz Figueroa. Project Management: Hugo Armando Ordóñez. Resources: Jhonfer Ruiz Figueroa. Software: Jhonfer Ruiz Figueroa. Supervision: Roxana Maria Luna Romero. Validation: JRoxana Maria Luna Romero. Writing - original draft - Preparation: Roxana Maria Luna Romero. Writing - revision and editing -Preparation: Roxana Maria Luna Romero.

Financiación: does not declare.

Conflict of interest: does not declare. Ethical aspect: does not declare

References

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.

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.

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

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

doi: 10.25100/iyc.v28i2.15700

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.

doi: 10.25100/iyc.v28i2.15700

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