Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 67-79 recibido 21-11-2024 revisado 06-02-2025 aceptado 08-02-2025

Roadmap for Natural Language Generation: Challenges and Insights

Próximos Pasos en la Generación del Lenguaje Natural: Desafíos y Enfoques por Explorar

María Miró Maestre, Iván Martínez-Murillo, Tania J. Martin, Borja Navarro-Colorado, Antonio Ferrández, Armando Suárez Cueto, Elena Lloret
Department of Software and Computing Systems, University of Alicante, Spain
maria.miro@ua.es

Abstract: Generative Artificial Intelligence has experienced exponential growth largely due to the advent of Large Language Models (LLMs). This expansion is fueled by the impressive performance of deep learning methods used in Natural Language Processing (NLP) and its subfield, Natural Language Generation (NLG), which is the focus of this paper. Popular LLMs, such as GPT-4, Bard, and tools such as ChatGPT have set benchmarks for addressing various NLG tasks. This scenario raises critical questions regarding the future of NLG and its adaptation to emerging challenges in the LLM era. To explore these issues, the present paper reviews a representative sample of recent NLG surveys, thereby providing the scientific community with a research roadmap to identify NLG aspects that remain inadequately addressed and to suggest areas warranting further in-depth exploration in NLG.

Keywords: natural language generation, artificial intelligence, research gaps, generative models.

Resumen: La Inteligencia Artificial Generativa ha crecido exponencialmente debido, en gran medida, a la llegada de los Grandes Modelos del Lenguaje (LLMs). Esta expansión viene impulsada por el increíble rendimiento de los métodos de aprendizaje profundo utilizados en el Procesamiento del Lenguaje Natural (PLN) y su subcampo dedicado a la Generación del Lenguaje Natural (GLN), que supone el foco de este artículo. Algunos LLMs populares como GPT-4, Bard, y herramientas como ChatGPT se han convertido en referentes para abordar diversas tareas propias de la GLN. Este escenario plantea nuevas preguntas sobre el futuro de la GLN y su adaptación a los nuevos desafíos de la era de los LLMs. Para explorar estas cuestiones, el presente artículo analiza una muestra representativa de estudios recientes sobre la GLN, proporcionando así una hoja de ruta de investigación para identificar los aspectos de la GLN que siguen sin abordarse adecuadamente y proponer líneas de investigación que necesitan ser exploradas en profundidad para avanzar en la investigación en GLN.

Palabras clave: generación del lenguaje natural, inteligencia artificial, focos de investigación, modelos generativos.

1 Introduction

Natural Language Processing (NLP) is central to Artificial Intelligence (AI), as it facilitates more natural interactions between humans and machines. Despite NLP's recent surge in popularity, research in this field spans over 60 years. The inherent complexities involved in the understanding —Natural Language Understanding (NLU)— and the production of languages —Natural Language Generation (NLG)— are reflected in the relatively limited performance of semantic and pragmatic tasks, such as word sense disambiguation, coreference resolution, and intention detection.

Focusing on NLG, this subfield within NLP has changed drastically from when it was first studied in the end of the 1970s (McDonald, 2010). Originally, NLG architectures were a sequential pipeline of the macroplanning, microplanning and realization stages, known as modular architectures,

ISSN 1135-5948 DOI 10.26342/2025-74-5

with the standard architecture being the one proposed in (Reiter, 1994). Afterwards, these stages became more flexible, giving rise to a new approach known as planning perspectives, where the division of tasks was less strict than in modular architectures, enabling two or more different tasks to be combined and performed as one step.

Finally, task division was replaced by what is defined as global approaches. These architectures rely on statistical learning and perform the generation in just one stage. The major milestone within this group was the Transformer architecture, which achieved great results on NLG tasks (Topal, Bas, and van Heerden, 2021). Since then, several architectures based on Transformers have been proposed, with LLMs delivering better results and producing texts almost indistinguishable from texts written by humans.

Until very recently, AI systems focused on specific tasks, such as Question Answering, Description Generation, or Text Summarization. However, LLMs are trained over tons of information, making it possible for a single NLG system to address many applications, i.e., following a one-fits-all approach. This is the case, for instance, of ChatGPT, which was originally conceived as a chatbot, although it now provides solutions in natural language to a wide range of prompts (open questions, poetry generation, summaries, etc.). The popularity of these NLG tools, partly because of their versatility in the variety of tasks they solve, has placed AI research on the radar, in particular NLP.

Indeed, great advances have been made in NLP tasks thanks to neural models and the aforementioned LLMs (as machine translation, text classification, and text generation). The progress has been so great that some of these tasks can now be considered solved. The question arises as to how this will impact NLP and NLG going forward and how will their role shift in the face of recent advances in LLMs.

Languages are, however, more complex and ultimately LLMs are only specific models based mainly on contextual relationships between words. Indeed, new tasks or new NLU and NLG research lines are emerging, and others remain unsolved. Papers as (Church and Liberman, 2021) indicate some of the unsolved topics, such as syntactic parsing with Universal Dependencies, semantic compositionality or causality relationships.

The goal of this paper is to provide an analysis of several NLG survey papers published recently, exploring the unsolved research topics in NLG. Our work is presented as a NLG roadmap, detecting the areas requiring improvement and further in-depth research so NLG systems keep on evolving to face more complex tasks. We consider this to be of value to the research community in terms of revealing the key areas that need to be tackled in NLG going forward.

The paper is organized as follows. First, we describe the approach chosen to compile the surveys that structure this paper and analyse the selected NLG surveys in Section 2. This section also includes Table 1, where we list all the surveys analysed for the present research and show the linguistic and technical aspects covered in each survey. Then, Section 3 delves into an in-depth analysis on each of the nine research gaps we detected thanks to the previous survey analysis, remarking the main aspects to be addressed in each research gap. Subsequently, we define our proposal for a NLG research roadmap in Section 4, where we discuss the importance and possible consequences of setting aside each of the gaps detected via an Eisenhower matrix. Finally, Section 5 serves as a conclusion where we highlight the adequacy of the proposed roadmap and the many different tasks that need to be studied within the field.

2 *What Do Recent Surveys Say?*

Recapping surveys on the current state of NLG is essential for understanding and assessing developments in this evolving field. In this section, we outline our methodology for the survey compilation and examine the key findings and emerging trends in those surveys, providing comprehensive insights into the present and future directions of NLG research, with an emphasis on the content those systems can generate.

2.1 Survey Exploration

The methodology used to gather the NLG surveys was to first review the comprehensive set of NLG surveys from (Moreno, 2021), which organizes studies by their holistic or task-specific focus when addressing the NLG field chronologically. The surveys included a description of the task or domain they are devoted to. Then, we expanded this review to include recent publications, selecting surveys that address the chronological evolution of NLG systems, theoretical reviews of both traditional and neural models, as well as analyses of core techniques in NLG tasks, evaluation methods, and emerging issues.

Table 1 gathers data on the year the survey was published, and whether the survey includes the following: corpora, methods, and tools. The table also interprets the data, presenting some descriptive statistics to indicate research gaps and thereby, opportunities.

At a more macro level, the 19 papers analyzed can be grouped into three main categories. The first group focuses on providing an overview of the NLG field with an indication of future research directions. (Santhanam and Shaikh, 2019) provide a comprehensive overview of NLG approaches and suggest avenues for future research in open domain dialogue systems. (Gatt and Krahmer, 2018) explore developments in NLG since 2000, with a focus on data-driven techniques, vision-to-text generation, and the generation of artistic texts. (Dale, 2020) specifically examines commercial applications of NLG software, while also presenting an up-to-date overview and discussing challenges and limitations of using NLG in contexts such as non-English languages and highly technical domains. (Yu et al., 2022) present a comprehensive review of the work done in the field of knowledge-enhanced text generation. (Goyal, Kumar, and Singh, 2023) analyze the advancements in automated text generation over the past twelve years with a focus on the various automatic evaluation metrics and benchmark datasets and tools that exist in the field to provide researchers with a comprehensive resource. (Becker et al., 2024) address some of the main tasks within text generation to identify up to nine challenges common across all tasks, as well as the evaluation methodologies currently used for text generation systems, providing insights into future research directions for further exploration in the field. Finally, (Ignat et al., 2024) offers a comprehensive list of future directions that researchers can devote their studies to within the field of NLP. The survey categorizes open research questions into three primary areas –fundamental NLP, responsible NLP and applied NLP–encompassing a total of 14 research topics. In each of these categories, generation tasks are identified as playing a pivotal role in the advancement of the discipline, thereby offering a valuable comparative analysis for the research communities focused on both NLP and NLG.

The second group of papers provides a holistic overview of advancements in Neural Natural Language Generation (NNLG), a recent and growing research field. (Erdem et al., 2022) investigate recent developments and applications of NNLG from a multidimensional perspective, such as multimodality, multilinguality, controllability and learning strategies. (Tang, Guerin, and Lin, 2022) conduct a comprehensive survey of recent advancements in NNLG, categorising them into data construction, neural frameworks, training strategies, and evaluation metrics. (Lu et al., 2018) systematically survey NNLG, comparing properties of the models and their techniques through benchmarking experiments. (Topal, Bas, and van Heerden, 2021) focus on deep generative modelling for text generation, considering papers from 2015 onwards and evaluating approaches in different application domains. (Chandu and Black, 2020) offer a task-agnostic survey of modelling approaches in neural text generation, assisting researchers in positioning their work and identifying new challenges. (Iqbal and Qureshi, 2022) review various deep learning models used for text generation explaining the progress made in this area.

The third group concentrates on specific areas or tasks within NLG. (Perera and Nand, 2017) offer a detailed overview and classification of state-of-the-art approaches in NLG, particularly related to document planning, micro-planning, and surface realisation modules. (Kybartas and Bidarra, 2016) examine the automated versus manual authoring of plot and space components in story generation. (Gonçalo Oliveira, 2017) surveys intelligent poetry generators, focusing on languages, form and content features, techniques, reutilization of material, and evaluation. (Alabdulkarim, Li, and Peng, 2021) analyze machine learning approaches in story generation, addressing controllability, commonsense knowledge incorporation, reasonable character actions, and creative language generation. (Ji et al., 2023) provide a broad overview of the research progress and challenges in the hallucination problem in NLG, covering metrics, mitigation methods, and

Survey	Year	Corpora	Methods	Tools
Has It All Been Solved? Open NLP Research Questions Not Solved by Large Language Models (Ignat et al., 2024)	2024	✖	✔	✖
Text generation: A systematic literature review of tasks, evaluation, and challenges (Becker et al., 2024)	2024	✔	✔	✔
A Systematic survey on automated text generation tools and techniques: application, evaluation, and challenges (Goyal, Kumar, and Singh, 2023)	2023	✔	✔	✔
Survey of hallucination in natural language generation (Ji et al., 2023)	2023	✔	✔	✖
A survey of natural language generation (Dong et al., 2022)	2022	✔	✔	✔
A survey of knowledge-enhanced text generation (Yu et al., 2022)	2022	✔	✔	✔
Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning (Erdem et al., 2022)	2022	✔	✔	✖
Recent advances in neural text generation: A task-agnostic survey (Tang, Guerin, and Lin, 2022)	2022	✔	✔	✖
The survey: Text generation models in deep learning (Iqbal and Qureshi, 2022)	2022	✖	✔	✖
Exploring transformers in natural language generation: GPT, BERT, and XLNet (Topal, Bas, and van Heerden, 2021)	2021	✖	✔	✖
Positioning yourself in the maze of neural text generation: A task-agnostic survey (Chandu and Black, 2020)	2021	✖	✔	✖
Automatic story generation: Challenges and attempts (Alabdulkarim, Li, and Peng, 2021)	2021	✔	✔	✔
Natural language generation: The commercial state of the art in 2020 (Dale, 2020)	2020	✖	✖	✔
A survey of natural language generation techniques with a focus on dialogue systems - past, present and future directions (Santhanam and Shaikh, 2019)	2019	✔	✔	✔
Survey of the state of the art in natural language generation: Core tasks, applications and evaluation (Gatt and Krahmer, 2018)	2018	✔	✔	✔
Neural text generation: Past, present and beyond (Lu et al., 2018)	2018	✔	✔	✔
A survey on intelligent poetry generation: Languages, features, techniques, reutilization and evaluation (Gonçalo Oliveira, 2017)	2017	✖	✔	✔
Recent advances in natural language generation: A survey and classification of the empirical literature (Perera and Nand, 2017)	2017	✔	✔	✔
A survey on story generation techniques for authoring computational narratives (Kybartas and Bidarra, 2016)	2016	✖	✖	✔

Table 1: Analysis of NLG surveys.

task-specific advancements in the most common NLG tasks. (Dong et al., 2022) review NLG research, emphasizing data-to-text and text-to-text generation deep learning methods, as well as new applications, architectures, datasets, and evaluation challenges.

Overall, the 19 papers cover a wide range of topics in NLG, offering insights into commercial applications, knowledge integration, evaluation metrics, and specific tasks across various domains. They contribute to understanding the current state of the field and identifying future research directions.

3 Research Gaps to Explore

After the analysis of the surveys presented in Table 1, it can be argued that the excellent performance of LLMs in NLG tasks has revolutionized this discipline in an incredibly short time frame of five years (from 2019 with the emergence of GPT3 or T5 up to now). With this rapid development, we now face more complex tasks that require the input of further contextual knowledge and information modalities to achieve a performance that is actually comparable to a text written by a human. Consequently, this survey review serves as a starting point for identifying possible research gaps in NLG tasks, given the broad range of approaches from which these research issues are addressed. It is important to note that the research gaps discussed are based on the aspects related to the generation process to address specific tasks, rather than the methods used to evaluate NLG models.

Given the exponential growth that Generative AI methods have shown in the last few years, we believe these gaps bring a window of opportunities for researchers in the NLG discipline. By addressing them, we aim to ensure that LLMs cover complex aspects of language that would improve their overall performance for more demanding tasks.

3.1 Multimodality

Multimodality refers to the capacity of addressing different formats of input for language generation, such as text, data, audio, video, etc., (Erdem et al., 2022). This combined representation of different data formats represents an innovative approach to make NLG models improve their contextual knowledge, therefore boosting the addition of commonsense to the generated text, which constitutes one of the issues currently addressed in NLG. Indeed, many studies focus on multimodal input format. However, most of them tend to prioritize the information given in one of the modalities over the other (either data or text), therefore worsening the balance between the knowledge acquired from each input type (Erdem et al., 2022).

Given this inequality when processing the information contained in several modalities, we concluded that NLG systems need more multimodal training datasets to improve their performance. In this way, such systems would not miss the extra-linguistic information that may be detected by the combination of several information modalities. Moreover, an additional gap is to evaluate the knowledge balance between such formats to successfully solve some of the many emerging NLG tasks that make use of multimodal datasets, such as speech recognition, visual recognition, machine translation, etc.

3.2 Multilinguality

Multilinguality is another key issue not only for NLG systems but for NLP in general. The Internet has exacerbated the dominance of certain languages while others risk digital endangerment (Rehm and Way, 2023). An output of the survey review is that English is typically assumed as the “lingua franca” in NLG tasks. A clear example of this is found in the abscence of any mentions of the language chosen for the datasets in most surveys, from which we can infer that they are created in English. This reliance on English, used as a bridge in multilingual research where models are tested in more low-resource languages (see machine translation, text summarization, etc.), may overlook language-specific semantic nuances and hinder the generalizability of NLG architectures. Thus, future studies should focus on language-centric approaches and assess performance across different linguistic structures to study variances between languages and check if NLG models achieve the same performance.

Another drawback found in the analysis is that even high-resourced languages lack original datasets for key NLP tasks. Although Spanish ranks as the second most spoken language by native speakers in the world, and the third most used online, after English and Chinese (Vítores, 2024)(which are the most used languages in the surveys analyzed), most datasets on NLG specialized platforms like HuggingFace[1] are (semi)automatic translations of English, neglecting Spanish-specific semantic nuances. Moreover, the number of Spanish-language datasets on Hugging Face is 1789[2], which is relatively low compared to the 21520 datasets available in English. Consequently, another research gap in most current NLG surveys is the need for NLG systems oriented to high and low-resourced languages other than English.

3.3 Knowledge Integration and Controllable NLG

Neural models trained exclusively on a specific type of data, whether multimodal or not, possess constrained knowledge for generating the desired text. Including additional knowledge in neural models could enhance their performance and thereby, obtain a satisfactory output. Knowledge can be extracted from two different sources, internal and external (Yu et al., 2022). The former is obtained from the input text, such as keywords or linguistic features, and the latter is the knowledge that comes from outside sources, such as knowledge bases or external knowledge graphs.

Many studies have focused on two key steps involved in effectively integrating knowledge. The first step is concerned with obtaining helpful knowledge from different sources, and discarding what is irrelevant. The second step focuses on the successful understanding of knowledge and its incorporation into neural models. In our survey review, we have identified that, despite recent efforts that have led to significant progress in this area, there are still several gaps in effectively integrating knowledge into neural models. One key issue is the knowledge injected into the systems, which can become

[1] https://huggingface.co/datasets

[2] Data retrieved on February 7, 2025

outdated over time. These models typically rely on static knowledge. Additionally, if the knowledge retrieved is not relevant, it could introduce incorrect information, potentially leading to hallucinations.

Regarding controllable NLG, this topic arises from the need to control the final attributes of a text. The generation is guided by a control condition that can be, for example, stylistic (e.g., the emotion or intention of a text), or related to specific content (e.g., keywords or entities), as well as being based on demographic attributes of the speaker (Zhang et al., 2023)). There are two promising research lines on this topic: (1) proposing a unified framework to address the controllable generation task. Most of the research in this area has focused on specific tasks with specific conditions, so there lacks a global and unified framework. (2) Including additional commonsense knowledge so models generate texts according to a certain degree of fiction depending on the typology of the output, e.g. the degree of commonsense when writing a tale varies from the degree of commonsense needed to write a news article.

3.4 Hallucination

Hallucination is an issue present in state-of-the-art NLG tools. It occurs when a generated text seems to be fluent and natural, but its content is untrustworthy or illogical (Ji et al., 2023). Hallucinations can be intrinsic when a generated output differs from the source content, and extrinsic when a generated text cannot be corroborated by searching in the source text. Their origin can stem from two primary sources, i.e., the data, given the huge amounts of data models need to be trained, and the training and inference steps. As for the former, in the process of building datasets to train models many contradictions between the source and target can be introduced and consequently favor the appearance of hallucinations. Another problem is that duplicated data could bias the model to generate repeated data with more frequency. Regarding the latter, an inadequate training strategy can also introduce hallucinations. On the one hand, an encoder with a feeble understanding ability could learn wrong correlations of the training data. On the other hand, a decoder could focus on an erroneous part of the encoded input data, leading to hallucinations. Finally, the decoding strategy is also important because a strategy that increases the diversity of the generated output also increases the likelihood of hallucinations (Ji et al., 2023).

3.5 Explainability

Deep neural models, such as LLMs, have improved the effectiveness of NLG. Notwithstanding, these techniques have led indirectly to another social concern, which is explainability (Xu et al., 2019). Traditionally, NLG models were seen as *white box* systems where the decisions made by the models were guided by rules or decision trees. Consequently, these systems were inherently explainable. Since the development of deep neural models, improvements in performance have come at the cost of interpretability. These models, seen as *black box* systems, produce an output with no explanation of why the model has selected that result, or why it has arrived at a specific decision. As a result, it may trigger a lack of trust among users of these systems.

For these reasons, Explainable AI has become an interesting topic for the research community, and specifically the NLG field, to address. Ensuring that a system provides transparency as to how it arrives at decisions could help developers and users of systems. Explainability could help developers to detect data bias, identify mistakes made by the models, such as hallucinations, and improve these flaws. End users can also benefit when a system provides a decision as output because end users can understand why the system arrived at a decision and evaluate the trustworthiness of the steps taken. In this way, mistakes in reasoning can also be identified. Finally, explainability can be crucial in different socially impactful fields such as finance, medicine, or marketing. To sum up, although some advances have been made in this area, we still need more trustworthy and transparent NLG systems.

3.6 Narratives that Engage

LLMs can generate narrative texts creatively by producing stories with characters, sequential events, dialogues, etc.(Alabdulkarim, Li, and Peng, 2021). However, they struggle with more intricate narrative elements because of the impossibility to model these narrative components with only contextual relationships between words and sequential generation. For instance, while narratives depend on time-related events, they also require causal relationships between these events to maintain coherence, but LLMs struggle to model with simple word-to-word contextual relationships (Alabdulkarim, Li, and Peng, 2021). Additionally, crafting a compelling plot involves integrating conflict, suspense, and (possibly) a resolution (Alhussain and Azmi, 2021), but LLMs typically lack an overarching narrative framework to plan such elements effectively. Additionally, a narrative feature to capture the reader's attention is to generate suspense. This implies controlling what information is shown to the reader, where their attention is focused, the horizon of expectations, etc. None of these aspects are considered by LLMs. As a result, they generate boring narratives (Alabdulkarim, Li, and Peng, 2021; Alhussain and Azmi, 2021). They also fall short in developing authentic characters with the psychological depth and the relationships between them necessary to evoke reader empathy (Alabdulkarim, Li, and Peng, 2021). These are some of the main aspects of automatic narrative generation that LLMs are presently unable to manage. They are, therefore, open research topics in NLG that need complementary models. For some of these aspects, controlled generation (Alabdulkarim, Li, and Peng, 2021; Kybartas and Bidarra, 2016) is necessary, where a human decides how the narrative should be created.

3.7 Prompt Engineering and Beyond

Prompt engineering is the practice of optimizing textual input for generative AI (White et al., 2023). However, the flurry of interest in this field may not have a lasting impact, according to (Acar, 2023). The reason behind this is that as AI systems become more intuitive in understanding natural language, the need for meticulously crafted prompts is expected to decrease. New AI language models like GPT-4 also show promising results in generating effective prompts when asked, potentially rendering prompt engineering obsolete. Moreover, the effectiveness of prompts is often limited to specific algorithms, making them less applicable across different AI models and versions. As argued in (Acar, 2023), problem formulation is a more enduring and adaptable skill for leveraging the potential of generative AI. Problem formulation involves identifying, analyzing, and delineating problems. Well-formulated problems are crucial for achieving effective solutions, even when using sophisticated prompts. However, problem formulation is often overlooked and underdeveloped, with a disproportionate emphasis on problem-solving rather than problem formulation. Following (Acar, 2023), four key components of effective problem formulation are highlighted: diagnosis, decomposition, reframing, and constraint design. While prompt engineering is currently on everyone's radar, its lack of sustainability, versatility, and transferability restrict its long-term relevance. Emphasizing problem formulation over perfecting prompts enables a better understanding of problems and fosters effective collaboration with AI systems. Bearing this in mind, NLG could also consider wider approaches based on problem formulation which provide a platform for incorporating external knowledge and commonsense into generative AI.

3.8 Efficiency Issues

As reported by (Trabelsi et al., 2021), one of the disadvantages of LLMs is their high computation cost, causing constraints for both training and inference. This entails a processing limit on text length, as well as limits on access to updated data (e.g. ChatGPT's training data only goes up to 2021), which could be a serious handicap especially in NLG tasks. For example, LLMs have been successfully applied to Open-Domain Question Answering by generating answers to users' queries. However, the previous phase of compiling the passages of the relevant documents to extract the answer implies ad-hoc document retrieval, which is limited to the necessary processing of longer documents than LLMs allow (e.g. BERT cannot take input sequences longer than 512 tokens). In this way, the training of the LLM for this task is usually formed by triples such as "[document [CLS], query [SEP], passages [SEP]]" that frequently exceed 512 tokens.

To overcome this issue, several proposals have been developed, in which computational cost and memory complexity plays an important role. The common solution is to split the documents into smaller pieces of text, whether sentences or passages. However, as stated in (Kitaev, Kaiser, and Levskaya, 2020), ranking documents of length "L" using Transformers can require $\mathcal{O}(L^2)$ memory and time complexity (the authors reduce this complexity to $\mathcal{O}(L \cdot \log L)$), which renders these solutions unfeasible, even though the extraction of LLM-based document representation are run offline.

Therefore, decreasing memory complexity is an important research line in this area. For example, to reduce the dimension of the embeddings, vector compression methods have been proposed. Likewise, the combination of traditional bag-of-words (BOW) approaches (e.g. BM25) that filter the set of documents to a reduced set of passages, which are reranked using LLM-based semantic and relevance modes. Some researchers advocate for discarding these BOW approaches because they do not contain lots of important semantic information about documents. Thus, by proposing LLM embeddings to perform efficient retrieval based on the product quantization technique will assign for every document a real-valued codeword from the codebook or a binary code as in semantic hashing.

3.9 Ethical Concerns

LLMs can be a powerful tool to help humans in their daily life activities when used responsibly. However, given the large scale these models have acquired with their latest developments, several ethical considerations have emerged to preserve users' integrity, personal privacy and at the same time mitigate the wide range of societal biases that LLMs may reflect, which can come from very different sources (Hovy and Prabhumoye, 2021). Indeed, LLMs' potential has made researchers test their performance in increasingly specific tasks across professional disciplines which are not exempt from controversial decisions with serious consequences for humans. Within the legal setting, (Chen et al., 2019) raised a discussion about the limits of using NLP tools for legal decisions, as this work focused on the automatic prediction of prison terms via a dataset of records published by the Supreme People's Court of China. As for clinical NLG, the accuracy of the predictions that NLG architectures may provide cannot leave room for any mistake or doubt, as their generated information can have severe consequences for patients At the same time, legal concerns need to be considered within this professional field, as many studies need to feed their models with patients' medical records in order to learn clinical predictions, although by getting such data they may run the risk of interfering with the personal privacy of patients (Thirunavukarasu et al., 2023). Another current issue LLMs are coming up with is the existence of gender bias in either the data those models are trained with orin the information generated by those models. Language is a reflection of society, and when LLMs reflect these societal biases, they perpetuate harmful stereotypes for people belonging to different social groups (Vashishtha, Ahuja, and Sitaram, 2023). NLP research has already addressed the societal biases automatically reproduced in linguistic processing systems. Unfortunately, very little work has been done when approaching gender bias from the NLG perspective (Garimella et al., 2021). Given this lack of research on language generation biases, we believe that it is of high importance to address this issue by also considering the several grammatical structures used in different languages for detecting biases. The reason for this is that languages differ in the structures used to express a particular human genre given their cultural and societal context (Vashishtha, Ahuja, and Sitaram, 2023), so different approaches would have to be tested to mitigate this NLP issue. In this work, we only mentioned some of the tasks in which ethical issues may come up when automatically processing information, but these same concerns could be applied to many other research fields, as it has already been done in the area of news processing and how they deal with dis- and misinformation (Dulhanty et al., 2019), as well as the ethical consequences of using crowdworkers to so labelling and evaluation tasks within NLP research (Shmueli et al., 2021). In summary, such is the awareness of the ethical considerations that NLP researchers need to include in their work that the European Union already published the document "Ethics Guidelines for Trustworthy Artificial Intelligence" in 2019 (High-Level Expert Group on AI, 2019). This document, which includes sections for both the creation and evaluation of trustworthy AI, serves as a model of how researchers should develop technologies to preserve human integrity and mitigate untrustworthy information. Consequently, we believe that future studies on the creation of models with an ethical approach will be beneficial for both the NLP research community and society so we benefit from their potential preserving their trust.

4 Roadmap for Natural Language Generation

After describing some of the existing research gaps that need to be addressed on the NLG discipline, in this section we aim to devise a roadmap for NLG research. To accomplish so, we will discuss which of the previously identified gaps should be prioritized, by paying close attention to the guidelines stated by the current laws and guidelines on AI published by the European Union with the AI Act[3]. Moreover, it is worth mentioning some of the National AI Strategies that are currently being developed under different countries as France [4], Germany[5] or Spain (Ministerio para la Transformación Digital y de la Función Pública, 2024), which also determine some of next steps to be made within Generative AI research so as to consider the ethical implications and future challenges of doing research within this area. Then, we will classify the previously identified gaps in terms of their impact and potential consequences of setting them aside in line with the guidelines stated in those regulations and initiatives.

To this end, we designed an Eisenhower matrix (Bratterud et al., 2020) to represent which NLG research gaps could be addressed first according to four degrees of importance: important-urgent, important-non-urgent, unimportant-urgent and unimportant-non-urgent (Maksymov and Tryus, 2022). The Eisenhower matrix is shown in Figure 1, with urgency represented on the X-axis and importance on the Y-axis.

Level 1, characterized by high urgency and high importance, includes the gap on the hallucination issue, which poses risks as it could be exploited by malicious users to spread misinformation or manipulate content unfaithfully[6]. Another critical gap included in this quadrant is ethical concerns, as generating biased information could lead to unethical outcomes and hinder the development of AI systems that serve all users equitably[7]. Finally, we also incorporate efficiency issues, as advancements in LLMs have led to increased performance, but they also demand greater resources and energy. Finding the synergy between consumption and performance is therefore essential (Ministerio para la Transformación Digital y de la Función Pública, 2024).

Level 2, focused on high importance but low urgency, encompasses explainability and multimodality. Explainability helps clarify the reasoning processes within NLG models, aiding in the detection of potential errors in decision-making[8]. Multimodality is also valuable, as robust systems with multiple input and output formats can improve accessibility for diverse users (Ministerio para la Transformación Digital y de la Función Pública, 2024), potentially reducing the need of different models as one single model could achieve different inputs and outputs, thus enhancing efficiency.

Level 3, which focuses on gaps with high urgency but lower importance, includes knowledge integration and controllable NLG. Controlling the outputs of NLG models and integrating external knowledge can produce more reliable and accurate responses depending on users' culture and social context[9], therefore reducing the possibility of the NLG model to deviate from the user's purpose.

Finally, Level 4 features low urgency and low importance gaps, including multilinguality, prompt engineering, and narrative engagement. This quadrant involves those gaps which aim to enhance the user experience when interacting with NLG models, which is also of large importance for the creation of inclusive NLG systems that can generate outputs depending on the users' background. However, we included them in the last quadrant as, arguably, the preceding gaps affect the core performance and societal impact of any NLG system, regardless of the language in which the system is trained, the prompt we use to communicate with the system or more special needs as in the textual typology

[3] Official Journal version of 13 June 2024 available on: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

[4] https://www.info.gouv.fr/actualite/25-recommandations-pour-lia-en-france

[5] https://www.ki-strategie-deutschland.de/

[6] https://artificialintelligenceact.eu/es/recital/133/

[7] https://artificialintelligenceact.eu/es/recital/8/

[8] https://artificialintelligenceact.eu/es/recital/27/

[9] https://artificialintelligenceact.eu/es/recital/44/

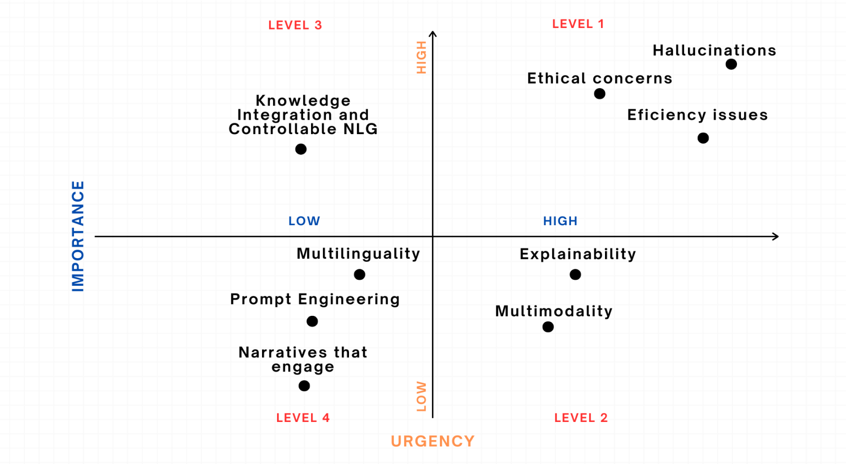

Figure 1: Urgency - importance matrix of the research gaps detected on NLG.

of narrative. Therefore, we believe those gaps need to be addressed not immediately, but in the long run.

Overall, we believe this distribution of research gaps in the matrix shown in Figure 1 matches the considerations estipulated in both the EU AI Act and the Spanish Government's *Estrategia de Inteligencia Artificial 2024*, therefore becoming a valuable roadmap of which research gaps should be address in the NLG discipline to ensure the evolution of such systems to improved versions based on governments' guidelines.

5 Conclusions

This paper outlined the state of the art in NLG by analyzing current key research lines derived from the gaps identified in the survey review. An analysis of 19 of the most recent surveys in the field identified the crucial areas that are being addressed in the context of NLG. The analysis also sheds some light on other unsolved and important problems to tackle. Indeed, although Generative AI and LLMs are capable of solving many NLG tasks by following a one-fits-all approach, they still have a lot of room for improvement to generate reliable and top quality texts.

The resulting roadmap on NLG research areas that need further in-depth studies focuses on the gaps concerning LLMs in the areas of multimodality, multilinguality, knowledge integration and controllable NLG, hallucination, explainability, creating engaging narratives, prompt engineering, their efficiency and several ethical concerns. By describing their complexities and possible consequences supported by some of the most recent governmental reports on AI, we consider that this survey can help researchers in the NLG field to identify potential research topics to address and draw a roadmap that guides NLG along its future path. The authors also emphasize that the identified gaps were those detected at the time of conducting this research. As the field of NLG continues to advance, new research gaps and opportunities will emerge. Therefore, it is crucial to continuously identify these limitations.

Acknowledgments

This research is part of the R&D projects “CORTEX: Conscious Text Generation” (PID2021-123956OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by “ERDF A way of making Europe”; “CLEAR.TEXT: Enhancing the modernization public sector organizations by deploying Natural Language Processing to make their digital content CLEARER to those with cognitive disabilities” (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and “European Union NextGenerationEU/PRTR”; and project “NL4DISMIS: Natural Language Technologies for dealing with dis- and misinformation with grant reference (CIPROM/2021/021)” funded by the Generalitat Valenciana. Moreover, it has been also partially funded by the Ministry of Economic Affairs and Digital Transformation and “European Union NextGenerationEU/PRTR” through the “ILENIA” project (grant number 2022/TL22/00215337) and “VIVES” subproject (grant number 2022/TL22/00215334).

References

Acar, O. A. 2023. AI prompt engineering isn't the future. Harvard Business Review. `https://hbr.org/2023/06/ai-prompt-engineering-isnt-the-future`.

Alabdulkarim, A., S. Li, and X. Peng. 2021. Automatic story generation: Challenges and attempts. In N. Akoury, F. Brahman, S. Chaturvedi, E. Clark, M. Iyyer, and L. J. Martin, editors, *Proceedings of the Third Workshop on Narrative Understanding*, pages 72–83, Virtual, June. Association for Computational Linguistics.

Alhussain, A. I. and A. M. Azmi. 2021. Automatic story generation: A survey of approaches. *ACM Computing Surveys (CSUR)*, 54(5):1–38.

Becker, J., J. P. Wahle, B. Gipp, and T. Ruas. 2024. Text generation: A systematic literature review of tasks, evaluation, and challenges. *arXiv preprint arXiv:2405.15604*.

Bratterud, H., M. Burgess, B. T. Fasy, D. L. Millman, T. Oster, and E. Sung. 2020. The sung diagram: Revitalizing the Eisenhower matrix. In *Diagrammatic Representation and Inference: 11th International Conference, Diagrams 2020, Tallinn, Estonia, August 24–28, 2020, Proceedings 11*, pages 498–502. Springer.

Chandu, K. R. and A. W. Black. 2020. Positioning yourself in the maze of neural text generation: A task-agnostic survey. *arXiv preprint arXiv:2010.07279*.

Chen, H., D. Cai, W. Dai, Z. Dai, and Y. Ding. 2019. Charge-based prison term prediction with deep gating network. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367, Hong Kong, China, November. Association for Computational Linguistics.

Church, K. and M. Liberman. 2021. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, 4:625341.

Dale, R. 2020. Natural language generation: The commercial state of the art in 2020. *Natural Language Engineering*, 26(4):481–487.

Dong, C., Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang. 2022. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38.

Dulhanty, C., J. L. Deglint, I. B. Daya, and A. Wong. 2019. Taking a stance on fake news: Towards automatic disinformation assessment via deep bidirectional transformer language models for stance detection. *arXiv preprint arXiv:1911.11951*.

Erdem, E., M. Kuyu, S. Yagcioglu, A. Frank, L. Parcalabescu, B. Plank, A. Babii, O. Turuta, A. Erdem, I. Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Garimella, A., A. Amarnath, K. Kumar, A. P. Yalla, N. Anandhavelu, N. Chhaya, and B. V. Srinivasan. 2021. He is very intelligent, she is very beautiful? On mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545.

Gatt, A. and E. Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Gonçalo Oliveira, H. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In J. M. Alonso, A. Bugarín, and E. Reiter, editors, *Proceedings of the 10th International Conference on Natural Language Generation*, pages 11–20, Santiago de Compostela, Spain, September. Association for Computational Linguistics.

Goyal, R., P. Kumar, and V. Singh. 2023. A systematic survey on automated text generation tools and techniques: Application, evaluation, and challenges. *Multimedia Tools and Applications*, 82(28):43089–43144.

High-Level Expert Group on AI. 2019. Ethics guidelines for trustworthy AI. Technical report, European Commission. `https://digital-strategy.`

ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Hovy, D. and S. Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and linguistics compass*, 15(8):e12432.

Ignat, O., Z. Jin, A. Abzaliev, L. Biester, S. Castro, N. Deng, X. Gao, A. E. Gunal, J. He, A. Kazemi, M. Khalifa, N. Koh, A. Lee, S. Liu, D. J. Min, S. Mori, J. C. Nwatu, V. Perez-Rosas, S. Shen, Z. Wang, W. Wu, and R. Mihalcea. 2024. Has it all been solved? open NLP research questions not solved by large language models. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8050–8094, Torino, Italia. ELRA and ICCL.

Iqbal, T. and S. Qureshi. 2022. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*, 34(6):2515–2528.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Kitaev, N., Ł. Kaiser, and A. Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451.*

Kybartas, B. and R. Bidarra. 2016. A survey on story generation techniques for authoring computational narratives. *IEEE Transactions on Computational Intelligence and AI in Games*, 9(3):239–253.

Lu, S., Y. Zhu, W. Zhang, J. Wang, and Y. Yu. 2018. Neural text generation: Past, present and beyond. *arXiv preprint arXiv:1803.07133.*

Maksymov, A. and Y. Tryus. 2022. Combined method of solving time management tasks and its implementation in the decision support system. In *International Scientific-Practical Conference "Information Technology for Education, Science and Technics"*, pages 131–146. Springer.

McDonald, D. D. 2010. Natural language generation. In *Handbook of natural language processing*. Chapman and Hall/CRC.

Ministerio para la Transformación Digital y de la Función Pública. 2024. Estrategia de inteligencia artificial 2024. Technical report, Ministerio para la Transformación Digital y de la Función Pública.

Moreno, M. V. 2021. *A Discourse-Aware Macroplanning Approach for Text Generation and Beyond*. Ph.D. thesis, Universitat d'Alacant/Universidad de Alicante.

Perera, R. and P. Nand. 2017. Recent advances in natural language generation: A survey and classification of the empirical literature. *Computing and Informatics*, 36:1–32.

Rehm, G. and A. Way. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Springer Nature.

Reiter, E. 1994. Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the Seventh International Workshop on Natural Language Generation*, pages 163–170.

Santhanam, S. and S. Shaikh. 2019. A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. *arXiv preprint arXiv:1906.00500.*

Shmueli, B., J. Fell, S. Ray, and L.-W. Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. *arXiv preprint arXiv:2104.10097.*

Tang, C., F. Guerin, and C. Lin. 2022. Recent advances in neural text generation: A task-agnostic survey. *arXiv preprint arXiv:2203.03047.*

Thirunavukarasu, A. J., D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Topal, M. O., A. Bas, and I. van Heerden. 2021. Exploring transformers in natural language generation: GPT, BERT, and XLNet. *arXiv preprint arXiv:2102.08036.*

Trabelsi, M., Z. Chen, B. D. Davison, and J. Heflin. 2021. Neural ranking models for document retrieval. *Information Retrieval Journal*, 24(6):400–444.

Vashishtha, A., K. Ahuja, and S. Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318, Toronto, Canada, July. Association for Computational Linguistics.

Vítores, D. F. 2024. El español: Una lengua viva. Informe 2024. *Centro Virtual Cervantes.*

White, J., Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382.*

Xu, F., H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, proceedings, part II 8*, pages 563–574. Springer.

Yu, W., C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, and M. Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38.

Zhang, H., H. Song, S. Li, M. Zhou, and D. Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.