Survey of Hallucination in Natural Language Generation

ZIWEI JI, NAYEON LEE, RITA FRIESKE, TIEZHENG YU, DAN SU, YAN XU, ETSUKO ISHII, YE JIN BANG, ANDREA MADOTTO, and PASCALE FUNG,
Hong Kong University of Science and Technology

Natural Language Generation (NLG) has improved exponentially in recent years thanks to the development of sequence-to-sequence deep learning technologies such as Transformer-based language models. This advancement has led to more fluent and coherent NLG, leading to improved development in downstream tasks such as abstractive summarization, dialogue generation, and data-to-text generation. However, it is also apparent that deep learning based generation is prone to hallucinate unintended text, which degrades the system performance and fails to meet user expectations in many real-world scenarios. To address this issue, many studies have been presented in measuring and mitigating hallucinated texts, but these have never been reviewed in a comprehensive manner before.

In this survey, we thus provide a broad overview of the research progress and challenges in the hallucination problem in NLG. The survey is organized into two parts: (1) a general overview of metrics, mitigation methods, and future directions, and (2) an overview of task-specific research progress on hallucinations in the following downstream tasks, namely abstractive summarization, dialogue generation, generative question answering, data-to-text generation, and machine translation. This survey serves to facilitate collaborative efforts among researchers in tackling the challenge of hallucinated texts in NLG.

CCS Concepts: • **Computing methodologies** → **Natural language generation**; **Neural networks;**

Additional Key Words and Phrases: Hallucination, intrinsic hallucination, extrinsic hallucination, faithfulness in NLG, factuality in NLG, consistency in NLG

ACM Reference format:
Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages.
https://doi.org/10.1145/3571730

1 INTRODUCTION

Natural Language Generation (NLG) is one of the crucial yet challenging sub-fields of **Natural Language Processing (NLP)**. NLG techniques are used in many downstream tasks such as summarization, dialogue generation, **Generative Question Answering (GQA)**, data-to-text

Authors' address: Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, Room 2602A, Center for Artificial Intelligence Research (CAiRE), Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong; emails: {zjiad, nyleeaa}@connect.ust.hk, rita.frieske@ust.hk, {tyuah, dsu, yxucb, eishii, yjbang, amadotto}@connect.ust.hk, pascale@ece.ust.hk.

0360-0300/2023/03-ART248 $15.00
https://doi.org/10.1145/3571730

generation, and **Machine Translation (MT)**. Recently, the rapid development of NLG has captured the imagination of many thanks to the advances in deep learning technologies, especially Transformer [138]-based models like BART [75], GPT-2 [105], and GPT-3 [13]. The conspicuous development of NLG tasks attracted the attention of many researchers, leading to an increased effort in the field.

Alongside the advancement of NLG models, attention toward their limitations and potential risks has also increased. Some early works [52, 146] focus on the potential pitfalls of utilizing the standard likelihood maximization based objective in training and decoding of NLG models. They discovered that such likelihood maximization approaches could result in *degeneration*, which refers to generated output that is bland, incoherent, or gets stuck in repetitive loops. Concurrently, it has been discovered that NLG models often generate text that is nonsensical, or unfaithful to the provided source input [109, 113, 139]. Researchers started referring to such undesirable generation as *hallucination* [90].[1]

Hallucination in NLG is concerning because it hinders performance and raises safety concerns for real-world applications. For instance, in medical applications, a hallucinatory summary generated from a patient information form could pose a risk to the patient. It may provoke a life-threatening incident for a patient if the instructions of a medicine generated by MT are hallucinatory. Hallucination can also lead to potential privacy violations. Carlini et al. [17] demonstrate that **Language Models (LMs)** can be prompted to recover and generate sensitive personal information from the training corpus (e.g., email address, phone/fax number, and physical address). Such memorization and recovery of the training corpus is considered a form of hallucination because the model is generating text that is not "faithful" to the source input content (i.e., such private information does not exist in the source input).

Currently, there are many active efforts to address hallucination for various NLG tasks. Analyzing hallucinatory content in different NLG tasks and investigating their relationship would strengthen our understanding of this phenomenon and encourage the unification of efforts from different NLG fields. However, to date, little has been done to understand hallucinations from a broader perspective that encompasses all major NLG tasks. To the best of our knowledge, existing surveys have only focused on specific tasks like abstractive summarization [57, 90] and translation [70]. Thus, we present a survey of the research progress and challenges in the hallucination problem in NLG and offer a comprehensive analysis of existing research on the phenomenon of hallucination in different NLG tasks, namely abstractive summarization, dialogue generation, GQA, data-to-text generation, and **Neural Machine Translation (NMT)**. We mainly discuss hallucination of the uni-modal NLG tasks that have textual input sources upon which the generated text can be assessed. We also briefly summarize hallucinations in multi-modal settings such as visual-language tasks [1, 10] and speech-to-text tasks [119, 124]. This survey can provide researchers a high-level insight derived from the similarities and differences of different approaches. Furthermore, given the various stages of development in studying hallucination from different tasks, the survey can assist researchers in drawing inspiration on concepts, metrics, and mitigation methods.

Organization of this Survey. The remainder of this survey is organized as follows. Sections 2 through 6 provide an overview of the hallucination problem in NLG by discussing its definition

[1]The term *hallucination* first appeared in Computer Vision in the work of Baker and Kanade [3] and carried more positive meanings, such as superresolution [3], image inpainting [35], and image synthesizing [160]. Such hallucination is something we take advantage of rather than avoid in Computer Vision. Nevertheless, recent works have started to refer to a specific type of error as "hallucination" in image captioning [113] and object detection [62], which denotes non-existing objects detected or localized incorrectly at their expected position. The latter conception is similar to "hallucination" in NLG.

and categorization, contributors, metrics, and mitigation methods, respectively. The second part of our survey discusses the hallucination problem associated with specific NLG tasks: abstractive summarization in Section 7, dialogue generation in Section 8, GQA in Section 9, data-to-text generation in Section 10, NMT in Section 11, and other tasks in Section 12. Finally, we conclude the whole survey in Section 13.

2 DEFINITIONS

In the general context outside of NLP, hallucination is a psychological term referring to a particular type of perception [38]. Blom [11] defines hallucination as "a percept, experienced by a waking individual, in the absence of an appropriate stimulus from the extracorporeal world." Simply put, a hallucination is an unreal perception that feels real. The undesired phenomenon of "NLG models generating unfaithful or nonsensical text" shares similar characteristics with such psychological hallucinations—explaining the choice of terminology. Hallucinated text gives the impression of being fluent and natural despite being unfaithful and nonsensical. It appears to be grounded in the real context provided, although it is actually hard to specify or verify the existence of such contexts. Similar to psychological hallucination, which is hard to tell apart from other "real" perceptions, hallucinated text is also hard to capture at first glance.

Within the context of NLP, the preceding definition of hallucination, *the generated content that is nonsensical or unfaithful to the provided source content* [37, 90, 100, 168], is the most inclusive and standard. However, there do exist variations in definition across NLG tasks, which will be further described in the later task-specific sections.

2.1 Categorization

Following the categorization from previous works [30, 57, 90], there are two main types of hallucinations, namely intrinsic hallucination and extrinsic hallucination. To explain the definition and categorization more intuitively, we give examples of each category of hallucinations for each NLG downstream task in Table 1 and expand on the two main types of hallucinations next:

(1) *Intrinsic hallucinations*: The generated output that contradicts the source content. For example, in the abstractive summarization task from Table 1, the generated summary "The first Ebola vaccine was approved in 2021" contradicts the source content "The first vaccine for Ebola was approved by the FDA in 2019."

(2) *Extrinsic hallucinations*: The generated output that cannot be verified from the source content (i.e., output can neither be supported nor contradicted by the source). For example, in the abstractive summarization task from Table 1, the information "China has already started clinical trials of the COVID-19 vaccine" is not mentioned in the source. We can neither find evidence for the generated output from the source nor assert that it is wrong. Notably, the extrinsic hallucination is not always erroneous because it could be from factually correct external information [90]. Such factual hallucination can be helpful because it recalls additional background knowledge to improve the informativeness of generated text. However, extrinsic hallucination is still treated with caution because its unverifiable aspect of the additional information increases the risk from a factual safety perspective.

2.2 Task Comparison

The previous section is about the definition and categorization of hallucination commonly shared by many NLG tasks. Yet there are some task-specific differences.

For the abstractive summarization, data-to-text, and dialogue tasks, the main difference is in what serves as the "source" and the level of tolerance toward hallucinations. The source in

Table 1. Examples of Each Category of Hallucinations for Each Task

Task	Sub-Task	Type	Source	Output
Abstractive Summarization [99]		Intrinsic	The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.	The first Ebola vaccine was approved in **2021**.
		Extrinsic		**China** has already started clinical trials of the COVID-19 vaccine.
Dialogue	Task-Oriented [82]	Intrinsic	inform (NAME = pickwick hotel, PRICERANGE = moderate)	the hotel named pickwick hotel is in a **high** price range.
		Extrinsic		the pickwick hotel **in san diego** is a moderate price range
	Open-Domain [116]	Intrinsic	Dialog History: Speaker 1: What do you think about Murray? Speaker 2: I think Murray is a great player he just needs to stay healthy in order to compete more. Who do you like best?	Speaker 1: I like **Roger Nadal** the most!
		Extrinsic	Knowledge: In tennis, the term Big Four refers to the quartet of men's singles players comprising Roger Federer, Rafael Nadal, Novak Djokovic, and Andy Murray.	Speaker 1: I like Djokovic. He has played **in the top ten singles players of the world**.
GQA [96]		Intrinsic	Question: dow jones industrial average please?	Answer: The Dow Jones Industrial Average(DJIA) is an index of **30 major U.S.stock indexes**. The DJIAwas created in 1896 by Charles Dow, a statistician and editor of The Wall Street Journal. The index was first published in the mid-1880s, and **was last updated on May 26,1896.**
		Extrinsic	Question: definition of sadduction	Answer: Sadducees. noun. **The definition of a sadducee is a person who acts in a deceitful or duplicitous manner. An example of a Sadduceee is a politician who acts deceitfully in order to gain political power**.1 a member of a Jewish sect that was active during the Second Temple.
Data2Text [149]		Intrinsic	TEAM: Rockets; CITY: Houston; WIN: 18; LOSS: 5; PTS: 108; FG_PCT: 44; BLK: 7	The Houston Rockets **(18-4)** defeated the Denver Nuggets (10-13) 108-96 on Saturday.
		Extrinsic	TEAM: Nuggets; CITY: Denver; WIN: 10; LOSS: 13; PTS: 96; FG_PCT: 38; BLK: 7	**Houston has won two straight games and six of their last seven.**
Translation [168]		Intrinsic	迈克周四去书店。 (Michael went to the bookstore on Thursday.)	**Jerry didn't go** to the bookstore.
		Extrinsic	迈克周四去书店。 (Michael went to the bookstore on Thursday.)	Michael **happily** went to the bookstore **with his friend.**

In the Data2Text task: H/A, H/A, home/away; MIN, minutes; PTS, points; REB, rebounds; AST, assists; BLK, blocks; FG_PCT, field goals percentage.

abstractive summarization is the input source text that is being summarized, whereas the source in data-to-text is non-linguistic data, and the source(s) in the dialogue system is dialogue history and/or the external knowledge sentences. Tolerance toward hallucinations is very low in both the summarization [99] and data-to-text tasks [100, 144, 145] because it is essential to provide faithful generation. In contrast, the tolerance is relatively higher in dialogue systems because the desired characteristics are not only faithfulness but also user engagement, especially in open-domain dialogue systems [56, 59]. For the GQA task, the exploration of hallucination is at its early stage, so there is no standard definition or categorization of hallucination yet. However, we can see that the GQA literature mainly focuses on "intrinsic hallucination" where the source is the world knowledge [77]. Last, unlike the aforementioned tasks, the categorizations of hallucinations in NMT vary within the task. Most relevant literature agrees that translated text is considered a hallucination when the source text is completely disconnected from the translated target [70, 93, 109].

2.3 Terminology Clarification

Multiple terminologies are associated with the concept of hallucination. We provide clarification of the commonly used terminologies *hallucination*, *faithfulness*, and *factuality* to resolve any confusion. *Faithfulness* is defined as staying consistent and truthful to the provided source—an antonym to "hallucination." Any work that tries to maximize faithfulness thus focuses on minimizing hallucination. For this reason, our survey includes all those works that address the faithfulness of machine generated outputs. *Factuality* refers to the quality of being actual or based on fact. Depending on what serves as the "fact," "factuality" and "faithfulness" may or may not be the same. Maynez et al. [90] differentiate "factuality" from "faithfulness" by defining the "fact" to be the world knowledge. In contrast, Dong et al. [25] use the source input as the "fact" to determine the factual correctness, making "factuality" indistinguishable from "faithfulness." In this article, we adopt the definition from Maynez et al. [90] because we believe having such distinction between source knowledge and world knowledge provides a more clear understanding.

Note that the judging criteria for what is considered faithful or hallucinated (i.e., the definition of hallucination) can differ across tasks. More details of these variation definitions will be provided in the later task-specific sections.

3 CONTRIBUTORS TO HALLUCINATION IN NLG

3.1 Hallucination from Data

The main cause of hallucination from data is source-reference divergence. This divergence happens as an artifact of heuristic data collection or is inevitably contained in data due to the nature of some NLG tasks. When a model is trained on data with this divergence, the model can be encouraged to generate text that is not necessarily grounded and not faithful to the provided source.

Heuristic Data Collection. When collecting large-scale datasets, some works heuristically select and pair real sentences or tables as the source and target [69, 149]. As a result, the target reference may contain information that cannot be supported by the source [100, 143]. For instance, when constructing WIKIBIO [69], a dataset for generating biographical notes based on the infoboxes of Wikipedia, the authors took the Wikipedia infobox as the source and the first sentence of the Wikipedia page as the target ground-truth reference. However, the first sentence of the Wikipedia article is not necessarily equivalent to the infobox in terms of the information they contain. Indeed, Dhingra et al. [22] points out that 62% of the first sentences in WIKIBIO have additional information not stated in the corresponding infobox. Such mismatch between source and target in datasets can lead to hallucination.

Another problematic scenario is when duplicates from the dataset are not properly filtered out. It is almost impossible to check hundreds of gigabytes of text corpora manually. Lee et al. [71] show that duplicated examples from the pre-training corpus bias the model to favor generating repeats of the memorized phrases from the duplicated examples.

Innate Divergence. Some NLG tasks by nature do not always have factual knowledge alignment between the source input text and the target reference, especially those that value diversity in generated output. For instance, it is acceptable for open-domain dialogue systems to respond in chit-chat style, subjective style [108], or with a relevant fact that is not necessarily present in the user input, history, or provided knowledge source—this improves the engagingness and diversity of the dialogue generation. However, researchers have discovered that such dataset characteristic leads to inevitable extrinsic hallucinations.

3.2 Hallucination from Training and Inference

As discussed in the previous section, source-reference divergence existing in dataset is one of the contributors of hallucination. However, Parikh et al. [100] show that the hallucination problem still occurs even when there is very little divergence in the dataset. This is because there is another contributor of hallucinations—training and modeling choices of neural models [109, 113, 139].

Imperfect Representation Learning. The encoder has the role of comprehending and encoding input text into meaningful representations. An encoder with a defective comprehension ability could influence the degree of hallucination [100]. When encoders learn wrong correlations between different parts of the training data, it could result in erroneous generation that diverges from the input [2, 36, 78, 134].

Erroneous Decoding. The decoder takes the encoded input from the encoder and generates the final target sequence. Two aspects of decoding contribute to hallucinations. First, decoders can attend to the wrong part of the encoded input source, leading to erroneous generation [134]. Such wrong association results in generation with facts mixed up between two similar entities [30, 121]. Second, the design of the decoding strategy itself can contribute to hallucinations. Dziri et al. [30] and Lee et al. [73] illustrate that a decoding strategy that improves the generation diversity, such as top-p sampling, is positively correlated with increased hallucination. Lee et al. [73] show that deliberately added "randomness" from sampling-based decoding increases the unexpected nature of the generation and the higher chance of containing hallucinated content.

Exposure Bias. Regardless of decoding strategy choices, the exposure bias problem [7, 107], defined as the discrepancy in decoding between training and inference time, can be another contributor to hallucination. It is common practice to train the decoder with teacher-forced MLE training, where the decoder is encouraged to predict the next token conditioned on the ground-truth prefix sequences. However, during the inference generation, the model generates the next token conditioned on the historical sequences previously generated by itself [142]. Such a discrepancy can lead to increasingly erroneous generation, especially when the target sequence gets longer.

Parametric Knowledge Bias. Pre-training of models on a large corpus is known to result in the model memorizing knowledge in its parameters [112]. This so-called parametric knowledge helps improve the performance of downstream tasks but also serves as another contributor to hallucinatory generation. Large pre-trained models used for downstream NLG tasks are powerful in providing generalizability and coverage, but Longpre et al. [86] have discovered that such models prioritize parametric knowledge over the provided input. In other words, models that favor gen-

Table 2. Evaluation Metrics and Mitigation Methods for Each Task

	Category	Task	Works
Automatic Metrics	Statistical	Dialogue	Shuster et al. [121]
		Data2Text	Dhingra et al. [22], Wang et al. [145]
		Translation	Martindale et al. [89]
	Model-based	Abstractive Summarization	Durmus et al. [26], Kryscinski et al. [67], Nan et al. [95], Wang et al. [140], Gabriel et al. [39], Goodrich et al. [46], Pagnoni et al. [99], Zhou et al. [168], Falke et al. [32], Laban et al. [68], Mishra et al. [92], Scialom et al. [117]
		Dialogue	Balakrishnan et al. [4], Honovich et al. [54], Li et al. [82], Dziri et al. [31], Gupta et al. [50], Santhanam et al. [116]
		GQA	Sellam et al. [118],* Zhang et al. [164],* Durmus et al. [26],* Wang et al. [140],* Su et al. [125]
		Data2Text	Dušek and Kasner [28], Liu et al. [85], Wiseman et al. [149], Filippova [37], Rebuffel et al. [111], Tian et al. [134]
		Translation	Kong et al. [65], Lee et al. [70], Parthasarathi et al. [101], Tu et al. [136], Feng et al. [36], Garg et al. [42], Raunak et al. [109], Zhou et al. [168]
		Task-Agnostic	Goyal and Durrett [48], Liu et al. [84], Zhou et al. [168]
Mitigation Method	Data-Related	Abstractive Summarization	Cao et al. [16], Gunel et al. [49], Nan et al. [95], Zhu et al. [170]
		Dialogue	Honovich et al. [54], Shen et al. [120], Shuster et al. [121], Wu et al. [151], Santhanam et al. [116]
		GQA	Bi et al. [9], Fan et al. [33], Yin et al. [157]
		Data2Text	Liu et al. [85], Nie et al. [98], Parikh et al. [100], Wang [143], Nie et al. [97], Rebuffel et al. [110]
		Translation	Junczys-Dowmunt [60], Lee et al. [70], Raunak et al. [109], Briakou and Carpuat [12]
	Modeling and Inference	Abstractive Summarization	Huang et al. [55], Li et al. [78], Song et al. [123], Zhao et al. [165], Aralikatte et al. [2], Cao et al. [14], Cao and Wang [15], Chen et al. [18]
		Dialogue	Balakrishnan et al. [4], Dziri et al. [30], Li et al. [82], Rashkin et al. [108]
		GQA	Fan et al. [33], Krishna et al. [66], Li et al. [77], Su et al. [125], Nakano et al. [94]
		Data2Text	Liu et al. [85], Tian et al. [134], Wang et al. [144, 145], Xu et al. [155], Filippova [37], Rebuffel et al. [110], Su et al. [127], Xiao and Wang [152], Puduppully and Lapata [104]
		Translation	Feng et al. [36], Lee et al. [70], Weng et al. [148], Xu et al. [154], Li et al. [81], Raunak et al. [109], Wang and Sennrich [142], Bengio et al. [7], Goyal et al. [47], Zhou et al. [168]

*The hallucination metrics are not specifically proposed for GQA, but they can be adapted for that task.

erating output with their parametric knowledge instead of the information from the input source can result in the hallucination of excess information in the output.

4 METRICS MEASURING HALLUCINATION

Recently, various studies have illustrated that most conventional metrics used to measure the quality of writing are not adequate for quantifying the level of hallucination [22, 26]. It has been shown that **State-of-the-Art (SOTA)** abstractive summarization systems, evaluated with metrics such as ROUGE, BLEU, and METEOR, have hallucinated content in 25% of their generated summaries [32]. A similar phenomenon has been shown in other NLG tasks, where it has been discovered that traditional metrics have a poor correlation with human judgment in terms of the hallucination problem [22, 26, 54, 66]. Therefore, there are active research efforts to define effective metrics for quantifying hallucination, as summarized in Table 2. FRANK [99] surveys the faithfulness metrics for summarization and compares these metrics' correlations with human judgments. To assess the example-level accuracy of metrics in diverse tasks, TRUE [53] reports their area under the ROC curve (ROC AUC) in regard to hallucinated example detection.

4.1 Statistical Metric

One of the simplest approaches is to leverage lexical features (n-grams) to calculate the information overlap and contradictions between the generated and the reference texts—the higher the mismatch counts, the lower the faithfulness and thus the higher the hallucination score.

Given that many traditional metrics leverage the target text as the ground-truth reference (ROUGE, BLEU, etc.), Dhingra et al. [22] build upon this idea and propose PARENT,[2] a metric that can also measure hallucinations using *both* the source and target text as references. Particularly, PARENT n-gram lexical entailment matches generated text with both the source table and target text. The F1-score that combines the precision and recall of the entailment reflects the accuracy in the table-to-text task. The source text is additionally used because it is not guaranteed that the output target text contains the complete set of information available in the input source text.

It is common for NLG tasks to have multiple plausible outputs from the same input, which is known as one-to-many mapping [126]. In practice, however, covering all possible outputs is too expensive and almost impossible. Thus, many works simplify the hallucination evaluation setup by relying on the source text as the sole reference. Their metrics just focus on the information referred by input sources to measure hallucinations, especially intrinsic hallucinations. For instance, Wang et al. [145] propose PARENT-T, which simplifies PARENT by only using table content as the reference. Similarly, Knowledge F1 [121]—a variant of unigram F1—has been proposed for **Knowledge-Grounded Dialogue (KGD)** tasks to measure the overlap between the model's generation and the knowledge used to ground the dialogue during dataset collection.

Furthermore, Martindale et al. [89] proposed a BVSS (bag-of-vectors sentence similarity) metric for measuring sentence adequacy in NMT, which only refers to the target text. This statistical metric helps determine whether the MT output has a different amount of information than the translation reference. Although simple and effective, one potential limitation of the lexical matching is that it can only handle the lexical information. Thus, it fails to deal with syntactic or semantic variations [118].

4.2 Model-Based Metric

Model-based metrics leverage neural models to measure the hallucination degree in the generated text. They are proposed to handle more complex syntactic and even semantic variations. The model-based metrics comprehend the source and generated texts and detect the knowledge/content mismatches. However, the neural models can be subject to errors that can propagate and adversely affect the accurate quantification of hallucination.

4.2.1 Information Extraction Based. It is not always easy to determine which part of the generated text contains the knowledge that requires verification. **Information Extraction (IE)**-based metrics use IE models to represent the knowledge in a simpler relational tuple format (e.g., *subject, relation, object*), then verify against relation tuples extracted from the source/reference. Here, the IE model is identifying and extracting the "facts" that require verification. In this way, words containing no verifiable information (stopwords, conjunctions, etc.) are not included in the verification step.

For example, ground-truth reference text ``Brad Pitt was born in 1963'' and generated text ``Brad Pitt was born in 1961'' will be mapped to the relation triples (Brad Pitt, born-in, 1963) and (Brad Pitt, born-in, 1961), respectively [46]. The mismatch between the dates (1963≠1961) indicates that there is hallucination. One limitation associated with this approach is the potential error propagation from the IE model.

[2]Note that PARENT is a general metric like ROUGE and BLEU, not only constrained to hallucination.

4.2.2 Question Answering Based. This approach implicitly measures the knowledge overlap or consistency between the generation and the source reference. This is based on the intuition that similar answers will be generated from a same question if the generation is factually consistent with the source reference. It is already put in use to evaluate hallucinations in many tasks, such as summarization [26, 117, 140], dialogue [54], and Data2Text generation [111].

The **Question Answering (QA)**-based metric that measures the faithfulness of the generated text is consists of three parts. First, given a generated text, a **Question Generation (QG)** model generates a set of question-answer pairs. Second, a QA model answers the generated questions given a ground-truth source text as the reference (containing knowledge). Last, the hallucination score is computed based on the similarity of the corresponding answers.

Similar to the IE-based metrics, the limitation of this approach is the potential error that might arise and propagated from either the QG model or the QA model.

4.2.3 Natural Language Inference Metrics. There are not many labeled datasets for hallucination detection tasks, especially at the early stage when the hallucination problem starts to gain attention. As an alternative, many works leverage the **Natural Language Inference (NLI)** dataset to tackle hallucinations. Note that NLI is a task that determines whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) given a "premise." These metrics are based on the idea that only the source knowledge reference should entail the entirety of the information in faithful and hallucination-free generation [28, 31, 32, 54, 57, 67, 68, 92]. More specifically, NLI-based metrics define the hallucination/faithfulness score to be the entailment probability between the source and its generated text, also known as the percentage of times generated text entails, neutral to, and contradicts the source.

According to Honovich et al. [54], NLI-based approaches are more robust to lexical variability than token matching approaches such as IE-based and QA-based metrics. Nevertheless, as illustrated by Falke et al. [32], off-the-shelf NLI models tend to transfer poorly to the abstractive summarization task. Thus, there is a line of research in improving and extending the NLI paradigm specifically for hallucination evaluation purposes [31, 32]. Apart from generalizability, Goyal and Durrett [48] point out the potential limitation of using sentence-level entailment models, namely their incapability to pinpoint and locate which parts of the generation are erroneous. In response, the authors propose a new dependency-level entailment and attempt to identify factual inconsistencies in a more fine-grained manner.

4.2.4 Faithfulness Classification Metrics. To improve upon NLI-based metrics, task-specific datasets are constructed to improve from the NLI-based metrics. Liu et al. [84] and Zhou et al. [168] constructed syntactic data by automatically inserting hallucinations into training instances. Santhanam et al. [116] and Honovich et al. [54] construct new corpora for faithfulness classification in dialogue responses. They manually annotate the Wizard-of-Wikipedia dataset [24], a KGD dataset, by judging whether each response is hallucinated.

Faithfulness-specific datasets can be better than NLI datasets because entailment or neutral labels of NLI datasets and faithfulness are not equivalent. For example, the hypothesis "Putin is U.S. president" can be considered to be either neutral to or entailed from the premise "Putin is president." However, from the faithfulness perspective, the hypothesis contains unsupported information "U.S.," which is deemed to be hallucination.

4.2.5 LM-Based Metrics. These metrics leverage two LMs to determine if each token is supported or not: an unconditional LM is only trained on the targets (ground-truth references) in the dataset, whereas a conditional language model LM_x is trained on both source and target data. It is assumed that the next token is inconsistent with the input if unconditional LM gets a smaller loss

than conditional LM_x during forced-path decoding [37, 134]. We classify the generated token as hallucinatory if the loss from LM is lower. The ratio of hallucinated tokens to the total number of target tokens $|y|$ can reflect the hallucination degree.

4.3 Human Evaluation

Due to the challenging and imperfect nature of the current automatic evaluation of hallucinations in NLG, human evaluation [116, 121] is still one of the most commonly used approaches. There are two main forms of human evaluation: (1) scoring, in which human annotators rate the hallucination level in a range, and (2) comparing, in which human annotators compare the output texts with baselines or ground-truth references [129].

5 HALLUCINATION MITIGATION METHODS

Common mitigation methods can be divided into two categories, in accordance with two main contributors of hallucinations: data-related methods and modeling and inference methods. We summarize these methods for each NLG downstream task in Table 2.

5.1 Data-Related Methods

5.1.1 Building a Faithful Dataset. Considering that noisy data encourage hallucinations, constructing faithful datasets manually is an intuitive method, and there are various ways to build such datasets. One way is employing annotators to write clean and faithful targets from scratch given the source [41], which may lack diversity [100]. Another way is employing annotators to rewrite real sentences on the web [100], or targets in the existing dataset [143]. Basically, the revision strategy consists of three stages: (1) phrase trimming: removing phrases unsupported by the source in the exemplar sentence; (2) decontextualization: resolving co-references and deleting phrases dependent on context; and (3) syntax modification: making the purified sentences flow smoothly. Meanwhile, other works [39, 54] leverage the model to generate data and instruct annotators to label whether these outputs contain hallucinations or not. This approach is typically used to build diagnostic evaluation datasets; however, it has the potential to build faithful datasets. Although less costly than building from scratch, it still requires tons of manpower and resources. Overall, building faithful datasets is task specific and lacks generalization.

5.1.2 Cleaning Data Automatically. To alleviate semantic noise issues, another approach is to find information that is irrelevant or contradictory to the input from the existing parallel corpus and then filter or correct the data. This approach is suitable for the case where there is a low or moderate level of noise in the original data [37, 98].

Some works [85, 109, 120] have dealt with the hallucination issue at the instance level by using a score for each source-reference pair and filtering out hallucinated ones. This corpus filtering method consists of several steps: (1) measuring the quality of the training samples in terms of hallucination utilizing the metrics described previously, (2) ranking these hallucination scores in descending order, and (3) selecting and filtering out the untrustworthy samples at the bottom. Instance-level scores can lead to a signal loss because divergences occur at the word level—that is, parts of the target sentence are loyal to the source input, whereas others diverge [110].

Considering this issue, other works [27, 98] correct paired training samples, specifically the input data, according to the references. This method is mainly applied in the data-to-text task because structured data are easier to correct than utterances. This method consists of two steps: (1) utilizing a model to parse the **Meaning Representation (MR)**, such as attribute-value pairs, from original human textual references, and (2) using the MR extracted from the reference to correct the input MR through slot matching. This method will enhance the semantic consistency between input and output without abandoning a part of the dataset.

5.1.3 Information Augmentation. It is intuitive that augmenting the inputs with external information will obtain a better representation of the source, because the external knowledge, explicit alignment, extra training data, and so forth can improve the correlation between the source and target and help the model learn better task-related features. Consequently, a better semantic understanding helps alleviate the divergence from the source issue. Examples of the augmented information include entity information [85], extracted relation triples from source documents [16, 55] obtained by Fact Description Extraction, pre-executed operation results [97], synthetic data generated through replacement or perturbation [18, 70], and retrieved external knowledge [9, 33, 49, 121, 170].

These methods enforce a stronger alignment between inputs and outputs. However, they will bring challenges due to the gap between the original source and augmented information, such as the semantic gap between an ambiguous utterance and a distinct MR of structured data, and the format discrepancy between the structured knowledge graph and natural language.

5.2 Modeling and Inference Methods

5.2.1 Architecture.

Encoder. The encoder learns to encode a variable-length sequence from input text into a fixed-length vector representation. As we mentioned in Section 5.1.3, hallucination appears when the models lack semantic interpretation over the input. Some works have modified the encoder architecture to make it more compatible with input and learn a better representation. For example, Huang et al. [55] and Cao et al. [16] propose a dual encoder, consisting of a sequential document encoder and a structured graph encoder to deal with the additional knowledge.

Attention. The attention mechanism is an integral component in neural networks that selectively concentrates on some parts of sequences while ignoring others based on dependencies [138]. To encourage the generator to pay more attention to the source, Aralikatte et al. [2] introduce a short circuit from the input document to the vocabulary distribution via source-conditioned bias. Krishna et al. [66] employ sparse attention to improve the model's long-range dependencies in the hope of modeling more retrieved documents so as to mitigate the hallucination in the answer. Wu et al. [151] adopt inductive attention, which removes potentially uninformative attention links by injecting pre-established structural information to avoid hallucinations.

Decoder. The decoder is responsible for generating the final output in natural language given input representations [138]. Several works modified the decoder structures to mitigate hallucination, such as the multi-branch decoder [110], uncertainty-aware decoder [152], and dual decoder, consisting of a sequential decoder and a tree-based decoder [123], and the constrained decoder with lexical or structural limitations [4]. Based on the observation that the "randomness" from sampling-based decoding, especially near the end of sentences, can lead to hallucination, Lee et al. [73] propose to iteratively reduce the "randomness" through time. These decoders improve the possibility of faithful tokens while reducing the possibility of hallucinatory ones during inference by figuring out the implicit discrepancy and dependency between tokens or restricted by explicit constraints. Since such decoders may have more difficulty generating fluent or diverse text, there is a balance to be struck between them.

5.2.2 Training.

Planning/Sketching. Planning is a common method to control and restrict what the model generates by informing the content and its order [103]. As shown in Figure 1(a), Planning can be a separate step in a two-step generator [18, 85, 104, 127, 144], which is prone to progressive amplification of the hallucination problem, or it can be injected into the end-to-end model

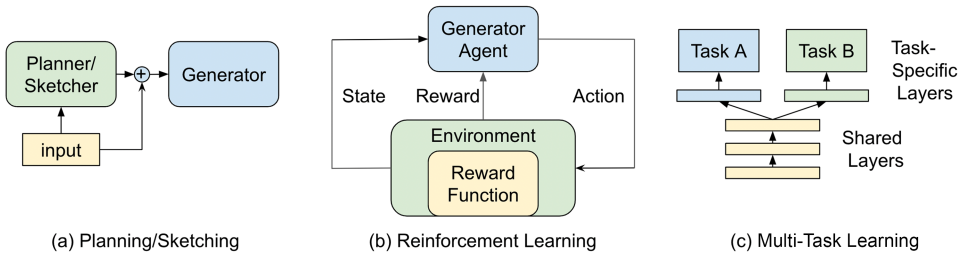

Fig. 1. The frameworks of training methods.

during generation [155]. Sketching has a similar function to planning and can also be adopted for handling hallucinations [144]. The difference is that the skeleton is treated as a part of the final generated text. Although providing more controllability, such methods also need to strike a balance between faithfulness and diversity.

Reinforcement Learning. As pointed out by Ranzato et al. [107], word-level maximum likelihood training leads to the problem of exposure bias. Some works [55, 65, 82, 91, 127] adopt Reinforcement Learning (RL) to solve the hallucination problem, which utilizes different rewards to optimize the model (shown in Figure 1(b)). The purpose of RL is for the agent to learn an optimal policy that maximizes the reward that accumulates from the environment [137]. The reward function is critical to RL, and if properly designed, it can provide training signals that help the model accomplish its goal of hallucination reduction. For example, Li et al. [82] propose a slot consistency reward that is the cardinality of the difference between the generated template and the slot-value pairs extracted from input dialogue act. Improving the slot consistency can help reduce the hallucination phenomenon of missing or misplacing slot values in generated templates. Mesgar et al. [91] attain a persona consistency sub-reward via an NLI model to reduce the hallucinations in personal facts. Huang et al. [55] use a combination of ROUGE and the multiple-choice cloze score as the reward function to improve the faithfulness of summarization outputs. The cloze score is similar to the QA-based metric, measuring how well a QA model can address the questions by reading the generated summary (as context), where the questions are automatically constructed from the reference summary. As the preceding examples show, some RL reward functions for mitigating hallucination are inspired by existing automatic evaluation metrics. Although RL is challenging to learn and converge due to the extremely large search space, this method has the potential to obtain the best policy for the task without an oracle.

Multi-Task Learning. Multi-task learning is also utilized for handling hallucinations in different NLG tasks. As shown in Figure 1(c), in this training paradigm, a shared model is trained on multiple tasks simultaneously to learn the commonalities of the tasks. The hallucination problem may be derived from the reliance of the training process on a single dataset, leading to the fact that the model fails to learn the actual task features. By adding proper additional tasks along with the target task during training, the model can suffer less from the hallucination problem. For example, Weng et al. [148] and Garg et al. [42] incorporate a word alignment task into the translation model to improve the alignment accuracy between the input and output, and thus faithfulness. Li et al. [78] combine an entailment task with abstractive summarization to encourage models to generate summaries entailed by and faithful to the source. Li et al. [77] incorporate rationale extraction and the answer generation, which allows more confident and correct answers and reduces the hallucination problem. The multi-task approach has several advantages, such as data efficiency improvement, overfitting reduction, and fast learning. It is crucial to choose which tasks should be learned jointly, and learning multiple tasks simultaneously presents new challenges of design and optimization [20].

Controllable Generation. Current works treat the hallucination level as a controllable attribute to keep the hallucination in outputs at a low level. Controllable generation techniques such as controlled re-sampling [108] and control codes that can be provided manually [37, 108, 151] or predicted automatically [151] are leveraged to improve faithfulness. This method may require some annotated datasets for training. Considering that hallucination is not necessarily harmful and may bring some benefits, controllable methods can be further adapted to change the degree of hallucination to meet the demands of different real-world applications.

Other general training methods such as regularization [61, 70, 93] and loss reconstruction [81, 142, 145] have also been proposed to tackle the hallucination problem.

5.2.3 Post-Processing. Post-processing methods can correct hallucinations in the output, and this stand-alone task requires less training data. Especially for noisy datasets where a large proportion of the ground truth references suffer from hallucinations, modeling correction is a competitive choice to handle the hallucination problem [18]. Cao et al. [14], Chen et al. [18], Dong et al. [25], and Dziri et al. [30] follow a generate-then-refine strategy. Although the post-processing correction step tends to result in ungrammatical texts, this method allows researchers to utilize SOTA models that perform best in respect of other attributes, such as fluency, and then correct the results specifically for faithfulness by using small amounts of automatically generated training data.

6 FUTURE DIRECTIONS

In this section, we point out the remaining challenges and potential directions in the metrics design and mitigation method.

6.1 Future Directions in Metrics Design

Fine-Grained Metrics. Most of the existing hallucination metrics measure intrinsic and extrinsic hallucinations together as a unified metric. However, it is common for a single generation to have both types and a number of hallucinatory sub-strings. Fine-grained metrics that can distinguish between the two types of hallucinations will provide richer insight to researchers.

To implement a fine-graded metric, the first step would be to identify the exact location of the hallucinatory sub-strings correctly. However, some metrics, such as those that are QA based, cannot identify the individual hallucinatory sub-strings. Improvements in this aspect would help improve the quality and explainability of the metrics. The next step would be to categorize the detected hallucinatory sub-strings. The hallucinatory sub-string will be intrinsic if it is wrong or nonsensical, and extrinsic if it is non-existing in the source context. Future work that explores an automatic method of categorization would be beneficial.

Fact-Checking. The factual verification of extrinsic hallucinations requires fact-checking against world knowledge, which can be time consuming and laborious. Leveraging an automatic fact-checking system for extrinsic hallucination verification is thus other future work that requires attention. Fact-checking consists of the knowledge evidence selection and claim verification sub-tasks, and the following are the remaining challenges associated with each sub-task.

The main research problem associated with the evidence selection sub-task is how to retrieve evidence from the *world* knowledge. Most of the literature leverages Wikipedia as the knowledge source [72, 132], which is only a small part of world knowledge. Other literature attempts to use the whole web as the knowledge source [88]. However, this method leads to another research problem—"how to ensure the trustworthiness of the information we use from the web" [44]. Source-level methods that leverages the meta-information of the web source (e.g., web traffic, PageRank, or URL structure) have been proposed to deal with this trustworthiness issue [5, 102]. Addressing the

aforementioned issues to allow evidence selection against world knowledge will be an important future research direction.

For the verification subtask, verification models perform relatively well if given correct evidence [74]. However, it has been shown that verification models are prone to adversarial attacks and are not robust to negation, numerical, or comparative words [133]. Improving this weakness of verification models would also be crucial because the factuality of a sentence can easily be changed by small word changes (i.e., changes in negations, numbers, and entities).

Generalization. Although we can see that the source and output text of different tasks are in various forms, investigating their relationship and common ground and proposing general metrics to evaluate hallucinations are worth exploring. Task-agnostic metrics with cross-domain robustness could help the research community build a unified benchmark. It is also important and meaningful to build open source platforms to collaborate and standardize the evaluation metrics for NLG tasks.

Incorporation of Human Cognitive Perspective. A good automatic metric should correlate with human evaluation. Humans are sensitive to different types of information. For instance, proper nouns are usually more important than pronouns in the generated text. Mistakes concerning named entities are striking to human users, but automatic metrics treat them equally if not properly designed. To address this issue, new metrics should be designed from the human cognitive perspective. The human ability to recognize salient information and filter the rest is evident in scenarios where the most important facts need to be determined and assessed. For instance, when signing an agreement, a prospective employee naturally skims the document to look at the entries with numbers first. In this way, humans classify what they believe is crucial.

Automatic check-worthy detection has the potential to be applied to improve the correlation with human judgment. Implementing the automatic human-like judgment mentioned earlier can further mitigate hallucination and improve NLG systems.

6.2 Future Directions in Mitigation Methods

General and Robust Data Pre-Processing Approaches. Since the data format varies between downstream tasks, there is still a gap for data processing methods between tasks, and currently, no universal method is effective for all NLG tasks [76]. Data pre-processing might result in grammatical errors or semantic transformation between the original and processed data, which can negatively affect the performance of generation. Therefore, we believe that general and robust data pre-processing methods can help mitigate the hallucinations in NLG.

Hallucinations in Numerals. Most existing mitigation methods do not focus on the hallucination in numerals. However, the correctness of numerals in text such as date, quantities, and scalars are important for readers [131, 163, 165]. For example, given the source "The optimal oxygen saturation (SpO_2) in adults with COVID-19 who are receiving supplemental oxygen is unknown. However, a target SpO_2 of 92% to 96% seems logical, considering that indirect evidence from patients without COVID-19 suggests that an SpO_2 of <92% or >96% may be harmful,"[3] the summary "The target SpO_2 range for patients with COVID-19 is 29–69%" includes wrong numbers, which could be fatal. Currently, some works [98, 131, 163] point out that using commonsense knowledge can help gain better numeral representation. And Zhao et al. [165] alleviate numeral hallucinations by re-ranking candidate-generated summaries based on the verification score of quantity entities. Therefore, we believe that explicitly modeling numerals to mitigate hallucinations is a potential direction.

[3] https://www.covid19treatmentguidelines.nih.gov/management/critical-care/oxygenation-and-ventilation.

Extrinsic Hallucination Mitigation. Although many works on mitigating hallucinations have been published, most do not distinguish between intrinsic and extrinsic hallucination. Moreover, the main research focus has been on dealing with intrinsic hallucination, whereas extrinsic hallucination has been somewhat overlooked as it is more challenging to reduce [57]. Therefore, we believe it is worth exploring different mitigation methods for intrinsic and extrinsic hallucinations, and relevant methods in fact-checking can be potentially used for this purpose.

Hallucination in Long Text. Many tasks in NLG require the model to process long input texts, such as multi-document summarization and GQA. We think adopting existing approaches to a Longformer [6]-based model could help encode long inputs. Meanwhile, a part of dialogue systems needs to generate long output text, in which the latter part may contradict history generation. Therefore, reducing self-contradiction is also an important future direction.

Reasoning. Misunderstanding facts in the source context will lead to intrinsic hallucination and errors. To help models understand the facts correctly requires reasoning over the input table or text. Moreover, if the generated text can be reasoned backward to the source, we can assume it is faithful. There are some reasoning works in the area of dialogue [21, 43], but few in reducing hallucinations. Moreover, tasks with quantities, such as logical table-to-text generation, require numerical reasoning. Therefore, adding reasoning ability to the hallucination mitigation methods is also an interesting future direction.

Controllability. Controllability means the ability of models to control the level of hallucination and strike a balance between faithfulness and diversity [30, 113]. As mentioned in Section 3, it is acceptable for chit-chat models to generate a certain level of hallucinatory content as long as it is factual. Meanwhile, for the abstractive summarization task, there is no agreement in the research community about whether factual hallucinations are desirable or not [90]. Therefore, we believe controllability merits attention when exploring hallucination mitigation methods.

7 HALLUCINATION IN ABSTRACTIVE SUMMARIZATION

Abstractive summarization aims to extract essential information from source documents and to generate short, concise, and readable summaries [158]. Neural networks have achieved remarkable results on abstractive summarization. However, Maynez et al. [90] observe that neural abstractive summarization models are likely to generate hallucinatory content that is unfaithful to the source document. Falke et al. [32] analyze three recent abstractive summarization systems and show that 25% of the summaries generated from SOTA models have hallucinated content. In addition, Zhou et al. [168] mention that even if a summary contains a large amount of hallucinatory content, it can achieve a high ROUGE score. This has encouraged researchers to actively devise ways to improve the evaluation of abstractive summarization, especially from the hallucination perspective.

In this section, we review the current progress in automatic evaluation and the mitigation of hallucination, and list the remaining challenges for future work.

7.1 Hallucination Definition in Abstractive Summarization

The definition of hallucination in abstractive summarization follows Section 2 and previous definitions [57, 90]: "a summary is hallucinated if it has any spans not supported by the input document." *Intrinsic hallucination* refers to output content that contradicts the source, whereas *extrinsic hallucination* refers to output content that the source cannot verify. For example in Table 1, the source document is "The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started." An

example of intrinsic hallucination is "The Ebola vaccine was rejected by the FDA in 2019" because "rejected" contradicts "approved." And an example of extrinsic hallucination is "China has already started clinical trials of the COVID-19 vaccine," because this statement is not mentioned in the given source. We can neither find evidence of it from the input article nor assert that it is wrong.

7.2 Hallucination Metrics in Abstractive Summarization

Existing metrics for hallucination in abstractive summarization are mainly model based. Following Huang et al. [57], we divide the hallucination metrics into two categories: (1) unsupervised metrics and (2) semi-supervised metrics. Existing hallucination metrics evaluate both intrinsic and extrinsic hallucinations together because it is difficult to automatically distinguish them.

7.2.1 Unsupervised Metrics. Given that hallucination is a newly emerging problem, there are only a few hallucination-related datasets. Therefore, researchers have proposed to adopt other datasets to build unsupervised hallucination metrics.

IE-Based Metrics. IE-based metrics leverage IE models to extract knowledge as relation tuples (*subject, relation, object*) from both the generation and knowledge source to analyze the factual accuracy of the generation [46]. However, IE models are not 100% reliable yet (making errors in the identification of the relation tuples). Therefore, Nan et al. [95] propose an entity-based metric relying on the Named-Entity Recognition model, which is relatively more robust. Their metric builds on the assumption that there will be a different set of named entities in the gold and generated summary if there exists hallucination.

NLI-Based Metrics. The NLI (a.k.a. textual entailment) model can be utilized to measure hallucination based on the assumption that a faithful summary will be entailed by the gold source. However, Falke et al. [32] discover that models trained on NLI datasets cannot transfer well to abstractive summarization tasks, degrading the reliability of NLI-based hallucination metrics. To overcome this issue, they release collected annotations as additional test data. Mishra et al. [92] find that the low performance of NLI-based metrics is mainly caused by the length of the premises in NLI datasets being shorter than the source documents in abstractive summarization. Thus, they propose to convert multiple-choice reading comprehension datasets into long-premise NLI datasets automatically. The results indicate that long-premise NLI datasets help the model achieve a higher performance. In addition, Laban et al. [68] introduce a simple but efficient method called $SUMMAC_{Conv}$ by applying NLI models to sentence units segmented from a document rather than the whole document.

QA-Based Metrics. QA-based metrics measure the knowledge overlap or consistency between summaries and the source documents based on the intuition that QA models will achieve similar answers if the summaries are factually consistent with the source documents. QA-based metrics such as FEQA [26], QAGS [140], and QuestEval [117] follow three steps to obtain a final score: (1) a QG model generates questions from the summaries, (2) a QA model obtains answers from the source documents, and (3) they calculate the score by comparing the set of answers from source documents and the set of answers from summaries. The results show that these reference-free metrics have substantially higher correlations with human judgments of faithfulness than the baseline metrics. Gabriel et al. [39] further analyze the FEQA and find that the effectiveness of QA-based metrics depends on the question. They also provide a meta-evaluation framework that includes QA metrics.

7.2.2 Semi-Supervised Metrics. Semi-supervised metrics are trained on the synthetic data generated from summarization datasets. Trained on these task-specific corpora, models can judge whether the generated summaries are hallucinatory. Kryscinski et al. [67] propose a weakly

supervised model named *FactCC* for evaluating factual consistency. The model is trained jointly for three tasks: (1) checking whether the synthetic sentences remain factually consistent, (2) extracting supporting spans in the source documents, and (3) extracting inconsistent spans in the summaries, if any exist. Zhou et al. [168] fine-tune pre-trained language models on synthetic data with automatically inserted hallucinations to detect the hallucinatory content in summaries. The model can classify whether spans in the machine-generated summaries are faithful to the article. This method shows higher correlations with human factual consistency evaluation than the baselines.

7.3 Hallucination Mitigation in Abstractive Summarization

7.3.1 Architecture Methods. Researchers have made modifications to the architecture design of the models to reduce hallucinated content in the summaries.

Encoder. Zhu et al. [170] use an explicit Graph Neural Network (GNN) to encode the fact tuples extracted from source documents. In addition to an explicit graph encoder, Huang et al. [55] further design a multiple-choice cloze test reward to encourage the model to better understand entity interactions. Moreover, Gunel et al. [49] use external knowledge from Wikipedia to make knowledge embeddings, and the results show improved factual consistency.

Decoder. Song et al. [123] incorporate a sequential decoder and a tree-based decoder to generate a summary sentence and its syntactic parse. Aralikatte et al. [2] introduce the Focus Attention Mechanism, which encourages decoders to generate tokens similar or topical to the source documents. The results show the efficiency of these methods to generate more faithful summaries.

Encoder-Decoder. Cao et al. [16] extract fact descriptions from the source text and apply a dual-attention seq2seq framework to force the summaries to be conditioned on both source documents and the extracted fact descriptions. Li et al. [78] propose an entailment-aware encoder and decoder with multi-task learning that incorporates the entailment knowledge into abstractive summarization models.

7.3.2 Training Methods. Aside from architecture modification, some works improved the training approach to reduce hallucination. Cao and Wang [15] introduce a contrastive learning method to train summarization models. The positive training data are reference summaries, whereas the negative training data are automatically generated hallucinatory summaries, and the contrastive learning system is trained to distinguish between them. Tang et al. [130] propose another contrastive fine-tuning strategy, named *CONFIT*, that can improve the factual consistency and overall quality of dialogue summaries.

7.3.3 Post-Processing Methods. Some works carry out post-editing to reduce the hallucination of generated summaries, which are viewed as draft summaries. Dong et al. [25] propose SpanFact, a pair of factual correction models that use knowledge learned from QA models to correct the spans in the generated summaries. Similarly, Cao et al. [14] introduce a post-editing corrector module to identify and correct hallucinatory content in generated summaries. The corrector module is trained on synthetic data that are created by adding a series of heuristic transformations to reference summaries. Zhao et al. [165] present HERMAN, a system that learns to recognize quantities (dates, amounts of money, etc.) in the generated summary and verify their factual consistency with the source text. According to the quantity hallucination score, the system chooses the most faithful summary where the source text supports its quantity terms from the candidate-generated summaries. Chen et al. [18] introduce a contrast candidate generation model replacing the named entities in the generated summaries with ones in the source, and the contrast candidate selection model selecting the best candidate as the final output summary.

7.4 Future Directions in Abstractive Summarization

Factual Hallucination Evaluation. Factual hallucinations contain information not found in source content, although it is factually correct. In the summarization task, this kind of hallucination could lead to better summaries. However, there is little work focused on evaluating factual hallucination explicitly. Fact-checking approaches could be potentially used in this regard.

Extrinsic Hallucination Mitigation. There has been little research on extrinsic hallucinations, as it is more challenging to detect and mitigate content based on world knowledge. We believe it is worth exploring extrinsic hallucination in terms of evaluation metrics and mitigation methods.

Hallucination in Dialogue Summarization. In conversational data, the discourse relations between utterances and co-references between speakers are more complicated than from, say, news articles. For example, Zhong et al. [166] show that 74% of samples in the QMSum dataset consist of inconsistent facts. We believe exploring the hallucination issue in dialogue summarization is an important and special component of research into hallucination in abstractive summarization.

8 HALLUCINATION IN DIALOGUE GENERATION

Dialogue generation is a task that automatically generates responses according to user utterances. The generated responses are required to be fluent, coherent, and consistent with the dialogue history. This task can be divided into two sub-tasks. *Task-oriented dialogue generation* aims to complete a certain task according to a user query in a specific domain, such as restaurant booking and hotel recommendation. *Open-domain dialogue generation* aims to establish a multi-turn, long-term conversation with users while providing an engaging experience.

8.1 Hallucination Definition in Dialogue Generation

The hallucination problem also exists in the dialogue generation task. It is important to note that a dialogue system is expected either to provide users with the required information or an engaging response without repeating utterances from the dialogue history. Thus, the tolerance for producing proper "hallucination" from the dialogue history is relatively higher.

The definition of hallucination in this task can be adopted from the general definition as follows. In *intrinsic hallucination*, the generated response is contradictory to the dialogue history or the external knowledge sentences. In *extrinsic hallucination*, the generated response is hard to verify with the dialogue history or the external knowledge sentences. In the following sections, the hallucination problem in open-domain and task-oriented dialogue generation tasks will be separately discussed according to the their natures.

8.2 Open-Domain Dialogue Generation

Although the term *hallucination* seems to have newly emerged in the NLP field, a related behavior, *inconsistency*, of neural models has been widely discussed. This behavior has been pointed out as a shortcoming of generation-based approaches for open-domain chatbots [114]. Two possible types of inconsistency occur in open-domain dialogue generation: (1) self-inconsistency [147, 159] or incoherence [8, 29] among the system utterances, such as when the system contradicts its previous utterance, and (2) external inconsistency with an external source, such as factually incorrect utterances. Some have recently started to call the second type *hallucination* [115]. Self-consistency can be considered as an intrinsic hallucination problem, whereas the external inconsistency involves both intrinsic and extrinsic hallucinations, depending on the reference source.

As mentioned earlier, a certain level of hallucination may be acceptable in open-domain chit-chat as long as it does not involve severe factual issues. It is almost impossible to verify factual correctness since the system usually lacks a connection to external resources. With the

introduction of the KGD task [24, 169] providing an external reference, however, there have been more active discussions of hallucination in open-domain dialogue generation.

8.2.1 Self-Consistency. In end-to-end generative open-domain dialogue systems, the inconsistency among system utterances has been pointed out as the bottleneck to human-level performance [139]. We often observe an inconsistency in the answers to semantically similar yet not identical questions. For example, a system may answer "What is your name?" and "May I ask your name ?" with different responses. Persona consistency has been the center of attention [79, 161] and it is one of the most obvious cases of self-contradiction regarding the character of the dialogue system. "Persona" is defined as the character that a dialogue system plays during a conversation, and can be composed of identity, language behavior, and an interaction style [79] (see Section 8.2.2).

8.2.2 External Consistency. An open-domain dialogue system should also generate persona-consistent and informative responses corresponding so as to user utterances to further engage with the user during conversation. In this process, an external resource containing explicit persona information or world knowledge is introduced into the system to assist the model generation process.

The PersonaChat datasets have accelerated research into persona consistency [23], where each conversation is accompanied by persona descriptions such as "I like to ski." By conditioning on the persona description, a chit-chat model is expected to acquire an ability to generate a more persona-consistent response. Lately, the application of NLI [80, 122] or RL frameworks [91] have been investigated. Although these methods on PersonaChat have been successful, further investigation of approaches that do not rely on the given persona descriptions is necessary because such descriptions are not always available, and covering all aspects of a person is impossible.

In addition, KGD in the open-domain requires the model to generate informative responses with the help of an external knowledge graph or knowledge corpus [24, 169] and considers the external knowledge sentences as part of the source. Hallucination in conversations, which is also considered as a factual consistency problem, has raised much research interest recently [30, 108, 116, 121]. Most KGD works tackle the hallucination problem when responses contain information that contradicts (intrinsic) or cannot be found in the provided knowledge input (extrinsic). Since world knowledge is enormous and ever-changing, the extrinsic hallucination may be factual but hard to verify. Dziri et al. [30] further adopt the similar hallucination definition to the knowledge graph grounded dialogue task, where intrinsic hallucination indicates misusing either the subject or object of the knowledge triple, and extrinsic hallucination indicates that there is no corresponding valid knowledge triple in the gold reference knowledge. Recently, there have been some attempts to generate informative responses only with the help of the implicit knowledge inside large pre-trained language models instead [156] during the inference time. Under this setting, the study of extrinsic hallucination is of great value but still poorly investigated.

8.2.3 Hallucination Metrics. For generation-based dialogue systems, especially open-domain chatbots, the hallucination evaluation method remains an open problem [114]. As of now, there is no standard metric. Therefore, chatbots are usually evaluated by humans on factual consistency or factual correctness [116, 151]. We also introduce some automatic statistical and model-based metrics as a reference, which will be described in more detail in the following.

Variants of F1 Metrics. KFI (Knowledge F1) measures the overlap between the generated responses and the gold knowledge sentences to which the human referred for conversation during dataset collection [121]. KF1 is only available for datasets with labeled ground-truth knowledge. The authors further propose RF1 (Rare F1), which only considers the infrequent words in the dataset when calculating F1 to avoid influence from the common uni-grams.

Model-Based Metric. Recently, several works have proposed model-based evaluation metrics for measuring consistency, such as using NLI [29, 147], training learnable evaluation metrics [159], or releasing an additional test set for coherence [8]. For the KGD task, Dziri et al. [31] propose the BEGIN benchmark, which consists of samples taken from Dinan et al. [24] with additional human annotation and a new classification task extending the NLI paradigm. Honovich et al. [54] present a trainable metric Q^2 for the KGD task, which also applies NLI. It is also noteworthy that Gupta et al. [50] propose datasets that can benefit fact-checking systems specialized for dialogue systems. The Conv-FEVER corpus [116] is a factual consistency detection dataset, which was created by adapting the Wizard-of-Wikipedia dataset [24]. It consists of both factually consistent and inconsistent responses and can be used to train a classifier to detect factually inconsistent responses with respect to the knowledge provided.

8.2.4 Mitigation Methods. The hallucination issue can be mitigated by data pre-processing, which includes introducing extra information into the data. Shen et al. [120] propose a measurement based on seven attributes of the dialogue quality, including self-consistency. Based on this measurement, the untrustworthy samples that get lower scores are filtered out from the training set to improve the model performance in terms of self-consistency (i.e., intrinsic hallucination). Shuster et al. [121] conduct a comprehensive investigation on the KGD task where a retriever is introduced to the system for knowledge selection. The experimental results show that retrieval helps substantially in improving performance and reducing the hallucination in conversations without sacrificing conversational ability on KGD tasks. Rashkin et al. [108] introduce a set of control codes and concatenate them with dialogue inputs to reduce the hallucination by forcing the model to be more aware of how the response relies on the knowledge evidence in the response generation.

Some researchers have also tried to reduce hallucinated responses during generation by improving dialogue modeling. Different from data pre-processing, addressing the hallucination issue from the modeling aspect does not require heavy human labor for annotation, but it only provides distant supervision on hallucination mitigation. Wu et al. [151] apply inductive attention into Transformer-based dialogue models, and potentially uninformative attention links are removed with respect to a piece of pre-established structural information between the dialogue context and the provided knowledge. Instead of improving the dialogue response generation model itself, Dziri et al. [30] present a response refinement strategy with a token-level hallucination critic and entity-mention retriever without further training the original dialogue model. The former module is designed to label the hallucinated entity mentioned in the generated responses, whereas the retriever is trained to retrieve more faithful entities from the provided knowledge graph.

8.3 Task-Oriented Dialogue Generation

A task-oriented dialogue system is often composed of several modules: a natural language understanding module, a dialogue manager, and an NLG module [40]. Intrinsic hallucination can occur between the dialogue manager and NLG, where a dialogue act such as `recommend(NAME=`*peninsula hotel*`, AREA=`*tsim sha tsui*`)` is transformed into a natural language representation "the hotel named *peninsula hotel* is located in *tsim sha tsui* area" [4, 82].

8.3.1 Hallucination Metrics. Some works introduce hallucination-specific automatic metrics in addition to traditional metrics like BLEU and human evaluation. Li et al. [82] use the slot error rate, which is computed by $(p + q)/N$, where N represents the total number of slots extracted by another model in the dialogue act. Here, p stands for the missing slots in the generated template, and q is the number of redundant slots. However, Balakrishnan et al. [4] introduce a novel metric called *tree accuracy*, which determines whether the prediction's tree structure is identical to that of the input MR.

8.3.2 Mitigation Methods. While Balakrishnan et al. [4] adopt tree-structured semantic representations and add constraints on decoding, Li et al. [82] frame an RL problem to which they apply a bootstrapping algorithm to sample training instances and then leverage a reward related to slot consistency. Recently, there has emerged another line of research in task-oriented dialogue, which is to build a single end-to-end system rather than connecting several modules [87, 150]. As discussed in previous sections of this article, there is potential for such end-to-end systems to produce extrinsic hallucinations, yet this remains less explored. For example, a model might generate a response with an entity that appears out of nowhere. In the example of hotel recommendation in Hong Kong given earlier, a model could generate a response such as "the hotel named *raffles hotel* is located in *central* area,"[4] which cannot be verified from the knowledge base of the system.

8.4 Future Directions in Dialogue Generation

Exploring longer memory of dialogue is a future direction for solving self-contradiction and the hallucination problem. One possible reason for self-contradiction is that current dialogue systems tend to have a short memory of dialogue history [114]. First, common dialogue datasets provide several turns of conversation, which are not long enough to assess a model's ability to deal with a long context. To tackle this, Xu et al. [153] introduce a dataset consisting of more than 40 utterances per episode on average. Second, they often truncate dialogue history into fewer turns to fit into models such as Transformer-based architectures, which makes it difficult for a model to memorize the past. In addition to the works on dialogue summarization, it would be beneficial to apply other works that aim to grasp the longer context but do not focus on dialogue generation [6].

In addition, fact-checking is a future direction in dealing with the hallucination problem in dialogue systems [50]. Dialogue fact-checking involves verifiable claim detection, which is an important line in distinguishing hallucination-prone dialogue, and evidence retrieval from an external source. This fact-checking in the dialogue system could be utilized not only as an evaluation metric for facilitating factual consistency but also to model such a system.

9 HALLUCINATION IN GQA

GQA aims to generate an abstractive answer rather than extract an answer to a given question from provided passages [34, 77]. It is an essential task since many of the everyday questions that humans deal with and pose to search engines require in-depth explanations [63] (e.g., why/how...?). The answers usually are long and cannot be directly extracted from existing phrase spans.

Normally, a GQA system involves searching an external knowledge source for information relevant to the question. Then it generates the answer based on the retrieved information [66]. In most cases, no single source (document) contains the answer, and multiple retrieved documents will be considered for answer generation. Those documents may contain redundant, complementary, or contradictory information. Thus, hallucination is common in the generated answers.

The hallucination problem is one of the most important challenges in GQA. Since an essential goal of a GQA system is to provide factually correct answers given the question, hallucination in the answer will mislead the user and damage the system performance dramatically.

9.1 Hallucination Definition in GQA

As a challenging yet underexplored task, there is no standard definition of hallucination in GQA. However, almost all works on GQA [34, 66, 94, 125] involve a human evaluation process, in which the *factual correctness* measuring the faithfulness of the generated answer can be seen as a measurement of the hallucination—that is, the more faithful the answer is, the less hallucinated content

[4]Raffles Hotel is a hotel located in Downtown Core, Singapore.

it contains. The most recent such work [77] uses the term *semantic drift*, which indicates how the answer drifts away from a correct one during generation, and this can also be seen as a specific definition of hallucination in GQA.

In line with the general categorization of hallucination in Section 2.1, we give two concrete hallucination examples in GQA in Table 1. The sources of both questions are Wikipedia web pages. For the first question, "dow jones industrial average please?," the generated answer "index of 30 major U.S. stock indexes" conflicts with the statement "of 30 prominent companies listed on stock exchanges in the United States" from Wikipedia. So we categorize it as an *intrinsic hallucination*. For the second example, the sentences "The definition of a Sadducee is a person who acts in a deceitful or duplicitous manner. An example of a Sadduceee is a politician who acts deceitfully in order to gain political power" in the generated answer can not be verified from the source documents, and thus we categorize it as an *extrinsic hallucination*.

9.2 Hallucination-Related Metrics in GQA

Although most works in GQA use automatic evaluation metrics such as ROUGE and F1 to measure the quality of the answer, these n-gram overlap-based metrics are not a meaningful way to evaluate hallucination due to their poor correlation with human judgments [66]. However, almost all GQA-related works involve human evaluation as a complement to automatic evaluation. Normally, human annotators will be asked to assign a score indicating the faithfulness of the answer, which can also be viewed as a measurement of hallucination. However, the human evaluation metrics usually come from a small sample of the data.

Semantic overlap [118], a learned evaluation metric based on BERT that models human judgments, could be considered a better measurement of hallucination for GQA. *Factual correctness* can also be a way to measure hallucination in GQA. Zhang et al. [164] propose to explicitly measure the factual correctness of a generated text against the reference by first extracting facts via an IE module. Then they define and measure the factual accuracy score to be the ratio of facts in the generation text equal to the corresponding facts in the reference. *Factual consistency*, which measures the faithfulness of the generated answer given its source documents, can be employed as another way to measure hallucination in GQA. Durmus et al. [26] and Wang et al. [140] propose an automatic QA-based metric to measure faithfulness in summary, leveraging the recent advances in machine reading comprehension. They first use a QG model to construct question-answer pairs from the summary, then a QA model is applied to extract short answer spans from the given source document for the question. The extracted answers that do not match the provided answers indicate unfaithful information in the summary. Although these metrics were first proposed in summarization, they can be easily adopted in GQA to measure hallucinations in the generated long-form answer.

The most recent work [125] proposes to estimate the faithfulness of the generated long-form answer via *zero-shot short answer recall* on extractive QA datasets. They first generate long-form answers for questions from two extractive QA datasets, both of which contain large-scale question-answer pairs, then they measure the ratio of golden short answer span contained in the generated long answer as an estimation of faithfulness of the generated long answer. Although the idea is similar to the factual consistency metric in summarization work [26], and also matches with our intuition to some extent, its correlation with human evaluation on faithfulness has not been verified.

9.3 Hallucination Mitigation in GQA

Unlike conditional text generation tasks such as summarization, or data-to-text generation, in which the source documents are provided and normally related to the target generation, the hallucination problem in GQA is more complicated. Generally speaking, it might come from two sources:

(1) the incompetency of the retriever, which retrieves documents irrelevant to the answer, and (2) the *intrinsic* and *extrinsic* hallucination in the conditional generation model itself. Normally these two parts are interconnected and cause hallucinations in the answer.

Early works on GQA mostly tried to improve the faithfulness of the answer by investigating reliable external knowledge sources or incorporating multiple information sources. Yin et al. [157] propose Neural Generative Question Answering (GENQA), an end-to-end model that generates answers to simple factoid questions based on the knowledge base, whereas Bi et al. [9] propose the Knowledge-Enriched Answer Generator (KEAG) to generate a natural answer by integrating facts from four different information sources, namely questions, passages, vocabulary, and knowledge. Nevertheless, these methods rely on the existence of high-quality, relevant resources that are not easily available.

Recent works focus more on the conditional generation model. Fan et al. [33] construct a local knowledge graph for each question to compress the information and reduce redundancy from the retrieved documents, which can be viewed as an early trial to mitigate hallucination. Li et al. [77] propose Rationale-Enriched Answer Generator (REAG), in which they add an extraction task to obtain the rationale for an answer at the encoding stage, and the decoder is expected to generate the answer based on both the extracted rationale and original input. The recent work [66] employs a Routing Transformer (RT), a sparse attention-based Transformer-based model that employs local attention and mini-batch k-means clustering for long-range dependence, as the answer generator in the hope of modeling more retrieved documents to mitigate the hallucination in the answer. Su et al. [125] propose a framework named *RBG* (*R*ead *B*efore *G*enerate) to jointly models answer generation with machine reading. They augment the generation model with fine-grained, answer-related salient information predicted by the MRC module, to enhance answer faithfulness. Such methods can exploit and utilize the information in the original input better, whereas they require the extra effort of building models to extract that information.

Most recently, Lin et al. [83] proposed a benchmark, which comprises 817 questions that span 38 categories, to measure the truthfulness of a language model in the QA task. This work investigates the performances of GPT-3 [13], GPT-Neo/J [141], GPT-2 [105], and a T5-based model [106]. The results suggest that simply scaling up the model is less promising than fine-tuning it in terms of improving truthfulness since larger models are better at learning the training distribution from web data and thus tend to produce more imitative falsehoods. In another recent work, Nakano et al. [94] fine-tuned GPT-3 to answer long-form questions with a web-browsing environment, which allows the model to navigate the web as well as use human feedback to optimize answer quality directly using imitation learning [58]. Although this method seems promising, it also hinges on how that feedback is processed.

9.4 Future Directions in GQA

Since GQA is challenging yet underexplored, many possible directions could be explored to improve the answer quality and mitigate hallucination. First, better automatic evaluation metrics are needed to measure hallucination. The previously mentioned metrics, such as the semantic overlap between generated and ground-truth answers, the faithfulness of generated answers, and the factual consistency between answers and source documents, only consider one aspect of hallucination. Metrics that can consider all factors related to hallucination (e.g., semantic overlap, faithfulness, or factual consistency) could be designed. Second, datasets with hallucination annotations should be proposed since none of the current GQA dataset has that information. Another possible direction to mitigate hallucination in the answer is improving the performance of the models. We need better retrieval models that retrieve relevant information according to queries and generation models that can synthesize more accurate answers from multi-source documents.

10 HALLUCINATION IN DATA-TO-TEXT GENERATION

Data-to-text generation is the task of generating natural language descriptions conditioned on structured data, such as tables [100, 149], and knowledge graphs [41]. Although this field has been recently boosted by neural text generation models, it is well known that these models are prone to hallucinations [149] because of the gap between structured data and text, which may cause semantic misunderstanding and erroneous correlation. Moreover, the tolerance of hallucination is very low when this task is applied to the real world, such as in the case of patient information table description, and analysis of experimental results tables in a scientific report. Recent years have seen a growth of interest in hallucinations in data-to-text generation, and researchers have proposed works from the aspect of evaluation and mitigation.

10.1 Hallucination Definition in Data-to-Text Generation

The definition and categories of hallucination in data-to-text generation follow the descriptions in Section 2. We follow the general hallucination definition in this task. First, there are *intrinsic hallucinations*, in which the generated text contains information that is contradicted by the input data [98]. For example, in Table 1, "The Houston Rockets (18-4)" uses the information "[TEAM: Rockets, CITY:Houston, WIN:18, LOSS: 5]" in the source table. However, "(18-4)" is contradicted by "[LOSS: 5]" and it should be "(18-5)". Second, there are *extrinsic hallucinations*, in which the generated text contains extra information irrelevant to the input [22, 98]. For example, in Table 1, "Houston has won two straight games and six of their last seven" is not mentioned in the source table [143].

10.2 Hallucination Metrics in Data-to-Text Generation

Statistical. PARENT [22] measures the accuracy of table-to-text generation by aligning n-grams from the reference description R and generated texts G to the table T. And it is the average F-score by combining the entailment precision and recall. Wang et al. [145] modify PARENT and denote this table-focused version as PARENT-T. Different from PARENT, which evaluates the i-th instance (T_i, R_i, G_i), PARENT-T ignores the reference description R and evaluates each instance (T_i, G_i).

IE Based. Liu et al. [85] estimate hallucination with two entity-centric metrics: table record coverage (the ratio of covered records in a table) and hallucinated ratio (the ratio of hallucinated entities in text). This metric first uses entity recognition to extract the entities of input and generated output, then aligns these entities by heuristic matching strategies, and finally calculates the ratios of faithful and hallucinated entities separately. Moreover, there are some general post hoc IE-based metrics that could be applied to hallucination evaluation, such as Slot Error Rate (SER) [155], Content Selection (CS), Relation Generation (RG), and Content Ordering (CO) [143, 149].

QA Based. Data-QuestEval [111] adapts QuestEval [117] from summarization into data-to-text generation. First, a *textual QG model* is trained on a textual QA dataset. For each sample (structured data, textual descriptions), the *textual QG model* generates synthetic problems based on the descriptions. The structured data, textual descriptions (answers), and synthetic questions make up a synthetic QG/QA dataset to train *synthetic QA/QG models*. Then, the *synthetic QG* model generates questions based on the textual description to be evaluated. The *synthetic QA* model then generates answers based on a synthetic question and the structured input data. Finally, BERTScore [162] measures the similarity between the generated answer and description, indicating faithfulness.

NLI Based. Dušek and Kasner [28] recognize the textual entailment between the input data and the output text for both omissions and hallucinations with an NLI model. This work measures the semantic accuracy in two directions: check omissions by inferring whether the input fact is

entailed by the generated text and check hallucinations by inferring the generated text from the input.

LM Based. Filippova [37] and Tian et al. [134] base their work on the intuition that when an unconditional LM, only trained on the targets, gets a smaller loss than a conditional LM_x, trained on both sources and targets, the token is predicted unfaithfully. Thus, they calculate the ratio of hallucinated tokens to the total target length to measure the hallucination level.

10.3 Hallucination Mitigation in Data-to-Text Generation

Data-Related Methods. Several clean and faithful corpora are collected to tackle the challenges from data infidelity. TOTTO [100] is an open-domain faithful table-to-text dataset, where each sample includes a Wikipedia table with several highlighted cells and a description. To ensure that targets exclude hallucinations, the annotators revise existing Wikipedia candidate sentences and clear the parts unsupported by the table. Moreover, RotoWire-FG (Fact-Grounding) [143] is a purified and enlarged and enriched version of RotoWire [149] generating NBA game summaries from score tables. Annotators trim the hallucination part in target texts and extract the mapped table records as content plans to better align input tables and output summaries.

For data processing, OpAtt [97] designs a gating mechanism and a quantization module for the symbolic operation to augment the record table with pre-calculated results. Nie et al. [98] utilize a language understanding module to improve the equivalence between the input MR and the reference utterance in the dataset. They train natural language understanding model with an iterative relabeling procedure. First, they train the model on original data, parse the MR by model inference, train the model on new paired data with high confidence, and then repeat the preceding processes. Liu et al. [85] select training instances based on faithfulness ranking. Finer-grained than the preceding instance-level method, Rebuffel et al. [110] label tokens according to co-occurrence analysis and sentence structure through dependency parsing in the pre-processing step to explicate the correspondence between the input table and the text. Generally, the data-related methods are appropriate when the training dataset is noisy.

Modeling and Inference Methods. Planning and skeleton generation are common methods to improve the faithfulness to the input in data-to-text tasks. Liu et al. [85] propose a two-step generator with a separate text planner augmented by auxiliary entity information. The planner predicts the plausible content plan based on the input data. Then, given the preceding input data and the content plan, the sequence generator generates the text. Similarly, Plan-then-Generate [127] also consists of a content planner and a sequence generator. In addition, this work adopts a structure-aware RL training to generate output text following the generated content plan faithfully. Puduppully and Lapata [104] first induce a macro plan consisting of multiple sequences of entities and events from the input table and its corresponding multi-paragraph long document. The predicted macro plan then serves as the input to an encoder-decoder model for surface realization. SANA [144] is a skeleton-based two-stage model that includes skeleton generation to select key tokens from the source table and edit-based generation to produce texts via iterative insertion and deletion operations. In contrast to the preceding two-step model using planning or skeleton, AGGGEN [155] is an end-to-end model that jointly learns to plan and generate at the same time. This architecture with a hidden Markov model and Transformer encoder-decoder reintroduces explicit sentence planning stages into neural systems by aligning facts in the target text to input representations.

Other modeling methods have also been proposed to mitigate the hallucination problem. Conjecturing that hallucinations can be caused by inattention to the source, Tian et al. [134] propose a confidence score and a variational Bayes training framework to learn the score from data. Wang et al. [145] introduce a table-text optimal-transport matching loss and an embedding similarity loss

to encourage faithfulness. The hallucination degree can also be treated as a controllable factor in generating texts. In the work of Filippova [37], the hallucination degree of each training sample is estimated and converted into a categorical value that is a part of the inputs as a controlled setting. This approach does not require the dismissal of any input or modification of the model structure.

To mitigate hallucinations at the inference step, Rebuffel et al. [110] propose a Multi-Branch Decoder that leverages word-level alignment labels between the input table and paired text to learn the relevant parts of the training instance. These word-level labels are gained through dependency parsing during the pre-processing step. The branches separately integrate three co-dependent control factors: content, hallucination, and fluency. UABS (Uncertainty-Aware Beam Search) [152] is an extension to beam search to reduce hallucination. Considering that the hallucination probability is positively correlated with predictive uncertainty, this work adds a weighted penalty term in the beam search that is able to balance the predictive probability and uncertainty. This approach is task-agnostic and can also be applied to other tasks, such as image captioning.

These various types of methods do not necessarily conflict and can collaborate to solve the hallucination problem in data-to-text generation.

10.4 Future Directions in Data-to-Text Generation

Given the challenges brought by the discrepancy between structure data and natural text, and the low fault tolerance in the data-to-text task, there are several potential directions worth exploring in terms of hallucination.

First, numbers contain information about scales and are common and crucial in the data-to-text generation [128]. It is frequent to have errors in numbers, which results in hallucinations and infidelity. This is a serious problem for data-to-text generation, yet models rarely give special consideration to the numbers found in the table or text [131]. The current automatic metrics of hallucinations also do not specifically treat numbers. This indiscriminate treatment contradicts findings in cognitive neuroscience, where numbers are known to be represented differently from lexical words in a different part of the brain [45]. Thus, considering or highlighting numbers when mitigating and assessing hallucinations is worth exploring. This requires the generative model to learn a better numerical presentation and capture scales, which will reduce the hallucinations caused by the misunderstanding of numbers.

Moreover, the logical data-to-text generation task requires logical inference, calculation, and comparison, which is challenging and causes hallucinations more easily. Thus, reasoning (including numerical reasoning), which is usually combined with graph structures [19], is another direction to improve the accuracy of entity relationships and alleviate hallucinations.

11 HALLUCINATIONS IN NMT

NMT is the task of generating translation of the source language into the target language via inference, given parallel data samples for training. Compared to **Statistical Machine Translation (SMT)**, the output of NMT is usually quite fluent and of human-level quality, which creates the danger of misinforming users when there are hallucinations [89].

11.1 Hallucinations Definition and Categories in NMT

The problem of hallucination was identified with the deployment of the first NMT models. Early work comparing SMT and NMT systems [64], without explicitly using the term *hallucination*, mentioned that NMT models tend to "sacrifice adequacy for the sake of fluency" especially when evaluated with out-of-domain test sets. Following further development of NMT, most relevant research papers agree that translated text is considered a hallucination when it is completely disconnected from the source [70]. The categorization of hallucination in NMT is unlike that in any other NLG

Table 3. Categories and Examples of Hallucinations in MT by Zhou et al. [168] and Raunak et al. [109]

Category	Source	Correct Translation	Hallucinatory Translation
Intrinsic	迈克周四去书店。	Mike goes to the bookstore on Thursday.	Jerry doesn't go to the bookstore on Thursday.
Extrinsic	迈克周四去书店。	Mike goes to the bookstore on Thursday.	Mike happily goes to the bookstore on Thursday with his friend.
Detached	Das kann man nur feststellen, wenn die kontrollen mit einer großen intensität durchgeführt werden.	This can only be detected if controls undertaken are more rigorous.	Blood alone moves the wheel of history, i say to you and you will understand, it is a privilege to fight.
Oscillatory	1995 das produktionsvolumen von 30 millionen pizzen wird erreicht.	1995 the production reached 30 million pizzas.	The US, for example, has been in the past two decades, but has been in the same position as the US, and has been in the United States.

tasks and uses various terms that are often overlapping. To maintain consistency with other NLG tasks, in this section we use the intrinsic and extrinsic hallucination categories applied to the NMT task by Zhou et al. [168]. After a formal definition, we will describe other identified types of hallucinations and hallucination categories mentioned in the relevant literature.

Intrinsic and Extrinsic Hallucinations. Following the idea that hallucinations are outputs that are disconnected from the source, Zhou et al. [168] suggest categorizing the hallucinatory content based on the way the output is disconnected. First, there are *intrinsic hallucinations*, in which translations contain incorrect information compared to information present in the source. In Table 3, "Jerry doesn't go," since the original name in the source is "Mike" and the verb "to go" is not negated. Second, there are *extrinsic hallucinations*, in which translations produce additional content without any regard to the source. In Table 3, "happily" and "with his friend" are the two examples of the hallucinatory content since they are added without any apparent connection to the input.

Other Categories and Types of Hallucinations. Raunak et al. [109] propose an alternative categorization of hallucinations. They divide hallucinations into hallucinations under perturbations and natural hallucinations. *Hallucinations under perturbation* are those that can be observed if a model tested on the perturbed and unperturbed test set returns drastically different content. Their work on hallucinations under perturbation strictly follows the algorithm proposed by Lee et al. [70]; see Section 11.2.2 on the entropy measure. And *natural hallucinations* are created with a connection to the noise in the dataset and can be further divided into detached and oscillatory, where *detached* hallucinations mean that a target translation is semantically disconnected from a source input, and *oscillatory* hallucinations mean those that are decoupled from the source by manifesting a repeating n-gram. Tu et al. [136] and Kong et al. [65] analyze this phenomenon under the name *overtranslation*—that is, a repetitive appearance of words that were not in the source text. Conversely, undertranslation is skipping the words that need to be translated [136]. Finally, abrupt jumps to the end of the sequence and outputs that remain mostly in the source language are also examples of hallucinatory content [70].

11.2 Hallucination Metrics in NMT

The definition of hallucinations in MT tends to be qualitative and subjective, and thus researchers often identify hallucinated content manually. Most detrimentally, the appearance of hallucinations is found not to affect the BLEU score of the translated text [168]. There are also several notable efforts to automatize and quantify the search for hallucinations using statistical methods.

11.2.1 Statistical Metrics. Martindale et al. [89] propose identifying sentence adequacy using the BVSS metric. This metric indicates that the information is lost because the reference contains

more information than the MT output, or that there is additional information inserted in the case when the MT output contains more information than the reference.

11.2.2 Model-Based Metrics.

Auxiliary Decoder. Faithfulness refers to the amount of source meaning that is faithfully expressed in the translation, and it is used interchangeably with the term *adequacy* [36, 135]. Feng et al. [36] propose adding another "evaluation decoder" apart from the standard translation decoder. In their work, faithfulness is based on word-by-word translation probabilities and is calculated in the evaluation module along with translation fluency. The loss returned by the evaluation module helps adjust the probability returned by the translation module.

Entropy Measure. In scenarios where the ground truth of a translation is not available, an entropy measure of the average attention distribution can be used to detect hallucinations. Tu et al. [136] and Garg et al. [42] show that hallucinations are visible in attention matrices. When the model outputs correct translation, the attention mechanism attends to the entire input sequence throughout decoding. However, it tends to concentrate on one point when the model outputs hallucinatory content. The entropy is calculated on the average attention weights when the model does or does not produce hallucinations during testing. For comparison, a clean test set is used along with the purposefully perturbed one, which is created to incite hallucinations (test sets featuring multiple repetitions). The mean entropy returned by hallucinatory models diverges from the mean of the models that do not produce hallucinations spontaneously [70].

Token-Level Hallucination Detection. Zhou et al. [168] propose a method for detecting hallucinated tokens within a sentence, making the search more fine-grained. They use a synthetic dataset that is created by adding noise to the source data—more specifically, it is generated by a language model with certain tokens of correct translations masked. Tokens in synthetic data are labeled as hallucinated (1) or not (0). Then the authors compute the hallucination prediction loss between binary labels and the tokens from the hallucinated sentence. This work further employs a word alignment-based method and overlap-based method as baselines for hallucination.

Similarity-Based Methods. Zhou et al. [168] use an unsupervised model that extracts alignments from similarity matrices of word embeddings and predicts the target token as hallucinated if it is not aligned to the source. Parthasarathi et al. [101] calculate faithfulness by computing similarity scores between perturbed source sentence and target sentence after applying the same perturbation.

Overlap-Based Methods. Zhou et al. [168] predict that the target token is hallucinated if it does not appear in the source. Since the target and source are two different languages, the authors use the density matching method for bilingual synonyms from Zhou et al. [167]. Kong et al. [65] suggest the Coverage Difference Ratio (CDR) as the metric to evaluate adequacy, which is especially successful in finding cases of undertranslation. It is estimated by comparing source words covered by generated translation with human translations. The overlap-based methods for detecting hallucinations are heuristics based on the assumption that all translated words should appear in the source. However, this is not always the case, such as when paraphrasing or using synonyms. Using word embeddings as similarity-based methods helps avoid such simplifications and allows more diverse, synonymous translations.

Approximate Natural Hallucination Detection. Raunak et al. [109] propose ANH (Approximate Natural Hallucination) detection based on the fact that hallucinations often occur as oscillations (repeating n-grams) and the lower unique bigram count indicates a higher appearance of

oscillatory hallucinations. Furthermore, the ANH detection method searches for repeated targets in the translation output. Their method finds translation above a certain n-gram threshold and searches for repeated targets in the output translation, following the assumption that if hallucinations are often incited by aligning unique sources to the same target, then repeating targets will also appear during the inference [136].

11.3 Hallucination Mitigation Methods in NMT

Hallucinations in MT are hard to discover for a person who is not fluent in the target language, and thus they can lead to many possible errors, or even dangers. Out of all the NLG tasks, NMT engines such as Google in the English-speaking internet and Baidu in the Sinosphere are probably the most widely accessible to netizens. Consequently, there is a big interest in improving NMT performance, also by mitigating hallucinations. This section compiles data-related and model-related methods of mitigating hallucinations in NMT.

11.3.1 Data Related. Data augmentation appears to be one of the most common methods for removing hallucination. Lee et al. [70] and Raunak et al. [109] suggest addition of perturbed sentences. Furthermore, perturbation, where the insertions of most common tokens are placed at the beginning of the sentence, seems to be the most successful in hallucination mitigation. A disadvantage of this method is the need to understand different types of hallucinations produced by the model to apply a correct augmentation method. Corpus filtering is a method of mitigating hallucinations caused by the noise in the dataset by removing the repetitive and mismatching source and target sequences [109]. Junczys-Dowmunt [60] implements a cross-entropy data filtering method for bilingual data, which uses cross-entropy scores calculated for noisy pairs according to two translation models trained on the clean data. The scores that suggest disagreement between sentence pairs from two models are subsequently penalized.

Whereas Lee et al. [70], Raunak et al. [109], and Junczys-Dowmunt [60] define noise as mismatched source and target sentences, Briakou and Carpuat [12] analyze the influence of fine-grained semantic divergences on NMT outputs. The authors consequently propose a mitigation method for fine-grained divergences based on semantic factors. The tags are applied to each source and target sentence to inform about the position of divergent tokens. Factorizing divergence not only helps mitigate hallucinations but also improves the overall performance of the NMT. This shows that tagging small semantic divergences can provide useful information for the network during training.

11.3.2 Modeling and Inference. Overexposure bias is a common problem in NMT, amplified by the teacher-forcing technique used in sequence-to-sequence models. The models are trained on the ground truth, but during inference, they attend to the past predictions, which can be incorrect [65, 107]. To mitigate this problem, Wang and Sennrich [142] propose substituting MLE as a training objective with minimum risk training. Scheduled sampling is a classic method of mitigating overexposure bias first proposed by Bengio et al. [7]. Based on that method, Goyal et al. [47] create a differentiable approximation to greedy decoding that shows a good performance in the NMT task. Xu et al. [154] propose further improvement of the scheduled sampling algorithm by optimizing the probability of source and target word alignments. This improvement helps address the issue of flexibility in word order between a source and target language when performing scheduled sampling.

Zhou et al. [168] propose a method of improving self-training of NMT based on hallucination detection. They create hallucination labels (see Section 11.2.2) and then discard losses of tokens predicted as hallucinations, which is known as token loss truncation. This is similar to the method proposed by Kang and Hashimoto [61], the latter for full sentences in the summarization task.

Furthermore, instead of adjusting losses, Zhou et al. [168] mask the hidden states of the discarded losses in the decoder in a procedure called *decoder HS masking*. Experimental results show both a translation quality improvement in terms of BLEU and also a large reduction in hallucination. The token loss truncation method shows good results in the low-resource languages scenario.

Another method to mitigate the impact of noisy datasets is TERM (Tilted Empirical Risk Minimization), a training objective proposed by Li et al. [81]. Lee et al. [70] mention that techniques such as dropout, L2E regularization, and clipping tend to decrease the number of hallucinations. Last, several authors propose methods of improving phrase alignment that are helpful both in increasing translation accuracy and identifying content that did not appear in the source translation [42, 148].

11.4 Future Directions in NMT

The future work on hallucinations in NMT is to define hallucinations in a quantifiable manner—that is, to specify a cut-off value between translation error and hallucinated content using a particular metric. Martindale et al. [89] propose a threshold between fluency and adequacy that is the closest to this ideal. They, however, do not concentrate on hallucinated content as such, and thus fluent but inadequate sentences may not always indicate hallucinations but also other types of translation errors. Balakrishnan et al. [4] mention constrained decoding as a method to mitigate hallucinations in dialogue systems, but it could also be applied in NMT. Hokamp and Liu [51] use constrained decoding to incorporate specific terminology into MT, but the preceding methods can be re-purposed to mitigate hallucinations.

Another direction for future work on hallucinations is improving existing methods of searching for hallucinatory content that are computationally expensive [109] or require the creation of an additional perturbed test set [70]. Similarly, for mitigation of lack of faithfulness and fluency, the method proposed by Feng et al. [36] requires the creation of a one-to-many architecture (one encoder and two decoders), which is also computationally expensive. Future directions would therefore include simplification of existing hallucination evaluation methods, applying them to different architectures like CNNs and transformers, and possibly conducting research on finding simpler hallucination search methods.

12 HALLUCINATION IN OTHER TASKS

Besides the uni-modal NLG tasks we discussed previously, hallucination also occurs in other tasks, such as **Vision-Language (VL)** tasks and speech-to-text tasks. The hallucination research in the multi-modal field is still in an early stage, and thus methods to measure and mitigate hallucination remain open questions. This section briefly introduces the recent relevant work and trends in these tasks.

In image captioning, object hallucination is the current mostly explored problem, which is defined as models generating captions containing non-existent or inaccurate objects from the input image. To automatically measure object hallucination, CHAIR [113] is proposed to calculate what proportion of object words generated are actually in the image according to the ground-truth captions. As an extension to beam search, the UABS [152] (previously mentioned in Section 10.3) can be applied in image captioning to reduce hallucination. Furthermore, Biten et al. [10] hypothesize that the main cause of object hallucination is the systematic co-occurrence of particular object categories in images and propose three data augmentation methods to make the co-occurrence statistics matrix more uniform. For other VL tasks, Alayrac et al. [1] show that a large VL model prompted with questions may hallucinate answers that seem reasonable if given the text only but are wrong or unverifiable given the additional visual input.

The topic of hallucinations in speech is currently underrepresented. Serai et al. [119] describe the hallucination in ASR (Automatic Speech Recognition) and present a model predicting hallucinated

word sequences. They utilize the hallucinations as data augmentation to improve the robustness of the ASR model. Despite the requirement for faithfulness in speech translation [124], few recent studies have tackled the hallucination issue.

13 CONCLUSION

In this survey, we provide the first comprehensive overview of the hallucination problem in NLG, summarizing existing evaluation metrics, mitigation methods, and the remaining challenges for future research. Hallucination is an artifact of NLG and is of concern because they appear fluent and can therefore mislead users. In some scenarios and tasks, hallucination can cause harm. We survey various contributors to hallucination such as noisy data, erroneous parametric knowledge, incorrect attention mechanism, and inappropriate training strategy. We show that there are two categories of hallucinations, namely intrinsic hallucination and extrinsic hallucination, and they need to be treated differently with diverse mitigation strategies. Hallucination is relatively easy to detect in abstractive summarization and NMT against the evidence in the source. For dialogue systems, it is important to balance diversity and consistency in responses. Hallucination is detrimental to GQA's performance, but research on mitigation methods is still quite preliminary in this area. For data-to-text generation, hallucination arises from the discrepancy between the input and output format. Most methods to mitigate hallucinations in NMT either aim to reduce dataset noise or alleviate exposure bias. There remain many challenges ahead in identifying and mitigating hallucinations in NLG, and it is our hope that research in this area can benefit from this survey.

REFERENCES

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, et al. 2022. Flamingo: A visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198* (2022).

[2] Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 6078–6095.

[3] S. Baker and T. Kanade. 2000. Hallucinating faces. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*. 83–88.

[4] Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. 2019. Constrained decoding for neural NLG from compositional representations in task-oriented dialogue. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[5] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

[6] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150* (2020).

[7] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15)*. 1171–1179.

[8] Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? Targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*. 4164–4173.

[9] Bin Bi, Chen Wu, Ming Yan, Wei Wang, Jiangnan Xia, and Chenliang Li. 2019. Incorporating external knowledge into machine reading for generative question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 2521–2530.

[10] Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE, Los Alamitos, CA.

[11] Jan Dirk Blom. 2010. *A Dictionary of Hallucinations*. Springer.

[12] Eleftheria Briakou and Marine Carpuat. 2021. Beyond noise: Mitigating the impact of fine-grained semantic divergences on neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 7236–7249.

[13] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*. Curran Associates, 1877–1901.

[14] Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

[15] Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6633–6649.

[16] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[17] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, et al. 2021. Extracting training data from large language models. In *Proceedings of the 30th USENIX Security Symposium*. 2633–2650.

[18] Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*. 5935–5941.

[19] Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. 2020. Logic2Text: High-fidelity natural language generation from logical forms. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2096–2111.

[20] Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796* (2020).

[21] Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1406–1416.

[22] Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4884–4895.

[23] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, et al. 2020. The second conversational intelligence challenge (ConvAI2). In *The NeurIPS'18 Competition*. Springer, 187–208.

[24] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations*.

[25] Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 9320–9331.

[26] Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the ACL*. 5055–5070.

[27] Ondřej Dušek, David M. Howcroft, and Verena Rieser. 2019. Semantic noise matters for neural natural language generation. In *Proceedings of the 12th International Conference on Natural Language Generation*. 421–426.

[28] Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*. 131–137.

[29] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3806–3812.

[30] Nouha Dziri, Andrea Madotto, Osmar R. Zaiane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2197–2214.

[31] Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The BEGIN benchmark. In *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics, 1–12.

[32] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2214–2220.

[33] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 4186–4196.

[34] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

[35] Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. 2016. Image inpainting through neural networks hallucinations. In *Proceedings of the 2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop*. IEEE, Los Alamitos, CA.

[36] Yang Feng, Wanying Xie, Shuhao Gu, Chenze Shao, Wen Zhang, Zhengxin Yang, and Dong Yu. 2020. Modeling fluency and faithfulness for diverse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[37] Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 864–870.

[38] William Fish. 2009. *Perception, Hallucination, and Illusion*. Oxford University Press.

[39] Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 478–487.

[40] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. 2–7.

[41] Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

[42] Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 4453–4462.

[43] Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. CIDER: Commonsense inference for dialogue explanation and reasoning. *arXiv:2106.00510* (2021).

[44] Alexandru L. Ginsca, Adrian Popescu, and Mihai Lupu. 2015. Credibility in information retrieval. *Foundations and Trends in Information Retrieval* 9, 5 (2015), 355–475.

[45] Silke M. Göbel and Matthew F. S. Rushworth. 2004. Cognitive neuroscience: Acting on numbers. *Current Biology* 14, 13 (2004), R517–R519.

[46] Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 166–175.

[47] Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2017. Differentiable scheduled sampling for credit assignment. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 366–371.

[48] Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 3592–3603.

[49] Beliz Gunel, Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2020. Mind the facts: Knowledge-boosted coherent abstractive text summarization. *arXiv:2006.15435* (2020).

[50] Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. DialFact: A benchmark for fact-checking in dialogue. *arXiv preprint arXiv:2110.08222* (2021).

[51] Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 1535–1546.

[52] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations*.

[53] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2nd DialDoc Workshop on Document-Grounded Dialogue and Conversational Question Answering*.

[54] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 7856–7870.

[55] Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[56] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems* 38, 3 (2020), Article 21, 32 pages.

[57] Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839* (2021).

[58] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys* 50, 2 (2017), Article 21, 35 pages.

[59] Ziwei Ji, Yan Xu, I.-Tsun Cheng, Samuel Cahyawijaya, Rita Frieske, Etsuko Ishii, Min Zeng, Andrea Madotto, and Pascale Fung. 2022. VScript: Controllable script generation with visual presentation. *arXiv:2203.00314* (2022).

[60] Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the 3rd Conference on Machine Translation: Shared Task Papers*. 888–895.

[61] Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 718–731.
[62] Osman Semih Kayhan, Bart Vredebregt, and Jan C. van Gemert. 2021. Hallucination in object detection—A study in visual part verification. In *Proceedings of the 2021 IEEE International Conference on Image Processing*. IEEE, Los Alamitos, CA, 2234–2238.
[63] Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question answering with diverse answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 421–433.
[64] Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the 1st Workshop on Neural Machine Translation*. 28–39.
[65] Xiang Kong, Zhaopeng Tu, Shuming Shi, Eduard Hovy, and Tong Zhang. 2019. Neural machine translation with adequacy-oriented learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6618–6625.
[66] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*. 4940–4957.
[67] Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 9332–9346.
[68] Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics* 10 (2022), 163–177.
[69] Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.
[70] Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation. In *Proceedings of the International Conference on Learning Representations*.
[71] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499* (2021).
[72] Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-Tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the 3rd Workshop on Fact Extraction and VERification (FEVER'20)*. 36–41.
[73] Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *arXiv preprint arXiv:2206.04624* (2022).
[74] Nayeon Lee, Chien-Sheng Wu, and Pascale Fung. 2018. Improving large-scale fact-checking using decomposable attention models and lexical tagging. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*. 1133–1138.
[75] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
[76] Bohan Li, Yutai Hou, and Wanxiang Che. 2021. Data augmentation approaches in natural language processing: A survey. *arXiv preprint arXiv:2110.01852* (2021).
[77] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, and Songfang Huang. 2021. Addressing semantic drift in generative question answering with auxiliary extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 942–947.
[78] Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1430–1441.
[79] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.
[80] Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! Making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4715–4728.
[81] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. 2020. Tilted empirical risk minimization. In *Proceedings of the International Conference on Learning Representations*.
[82] Yangming Li, Kaisheng Yao, Libo Qin, Wanxiang Che, Xiaolong Li, and Ting Liu. 2020. Slot-consistent NLG for task-oriented dialogue systems with iterative rectification network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 97–106.

[83] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).

[84] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704* (2021).

[85] Tianyu Liu, Xin Zheng, Baobao Chang, and Zhifang Sui. 2021. Towards faithfulness in open domain table-to-text generation from an entity-centric view. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 13415–13423.

[86] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*. 7052–7063.

[87] Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 1468–1478.

[88] Amr Magdy and Nayer Wanas. 2010. Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. 103–110.

[89] Marianna J. Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*. 233–243.

[90] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[91] Mohsen Mesgar, Edwin Simpson, and Iryna Gurevych. 2021. Improving factual consistency between a response and persona facts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. 549–562.

[92] Anshuman Mishra, Dhruvesh Patel, Aparna Vijayakumar, Xiang Lorraine Li, Pavan Kapanipathi, and Kartik Talamadupula. 2021. Looking beyond sentence-level natural language inference for question answering and text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*. 1322–1336.

[93] Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas*. 151–164.

[94] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* (2021).

[95] Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. 2727–2733.

[96] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv:1611.09268* (2016).

[97] Feng Nie, Jinpeng Wang, Jin-Ge Yao, Rong Pan, and Chin-Yew Lin. 2018. Operation-guided neural networks for high fidelity data-to-text generation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*.

[98] Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2673–2679.

[99] Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*. 4812–4829.

[100] Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 1173–1186.

[101] Prasanna Parthasarathi, Koustuv Sinha, Joelle Pineau, and Adina Williams. 2021. Sometimes we want ungrammatical translations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 3205–3227.

[102] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. 2173–2178.

[103] Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[104] Ratish Puduppully and Mirella Lapata. 2021. Data-to-text generation with macro planning. *Transactions of the Association for Computational Linguistics* 9 (2021), 510–527.

[105] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[106] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67.

[107] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR'16)*.

[108] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 704–718.

[109] Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologys (NAACL-HLT'21)*. 1172–1183.

[110] Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2022. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery* 36 (2022), 318–354.

[111] Clément Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. 2021. Data-QuestEval: A reference-less metric for data-to-text semantic evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

[112] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

[113] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

[114] Stephen Roller, Y.-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, et al. 2020. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442* (2020).

[115] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. 300–325.

[116] Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456* (2021).

[117] Thomas Scialom, Paul-Alexis Dray, Patrick Gallinari, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

[118] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7881–7892.

[119] Prashant Serai, Vishal Sunder, and Eric Fosler-Lussier. 2022. Hallucination of speech recognition errors with sequence to sequence learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 890–900.

[120] Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. 2021. Identifying untrustworthy samples: Data filtering for open-domain dialogues with Bayesian optimization. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*. 1598–1608.

[121] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 3784–3803.

[122] Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8878–8885.

[123] Kaiqiang Song, Logan Lebanoff, Qipeng Guo, Xipeng Qiu, Xiangyang Xue, Chen Li, Dong Yu, and Fei Liu. 2020. Joint parsing and generation for abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[124] Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7409–7421.

[125] Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! Faithful long form question answering with machine reading. *arXiv:2203.00343* (2022).

[126] Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[127] Yixuan Su, David Vandyke, Sihui Wang, Yimai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, 895–909.

[128] Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

[129] Yanli Sun. 2010. Mining the correlation between human and automatic evaluation at sentence level. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10)*.

[130] Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2021. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. *arXiv preprint arXiv:2112.08713* (2021).

[131] Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. Representing numbers in NLP: A survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*. 644–656.

[132] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 809–819.

[133] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Evaluating adversarial attacks against multiple fact verification systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 2944–2953.

[134] Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. 2020. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv:1910.08684* (2020).

[135] Zhaopeng Tu, Yang Liu, Lifeng Shang, Xiaohua Liu, and Hang Li. 2017. Neural machine translation with reconstruction. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

[136] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 76–85.

[137] Victor Uc-Cetina, Nicolas Navarro-Guerrero, Anabel Martin-Gonzalez, Cornelius Weber, and Stefan Wermter. 2021. Survey on reinforcement learning for language processing. *arXiv preprint arXiv:2104.05565* (2021).

[138] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[139] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. In *Proceedings of the ICML Deep Learning Workshop*.

[140] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[141] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. Retrieved November 23, 2022 from https://github.com/kingoflolz/mesh-transformer-jax.

[142] Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics*. 3544–3552.

[143] Hongmin Wang. 2019. Revisiting challenges in data-to-text generation with fact grounding. In *Proceedings of the 12th International Conference on Natural Language Generation*. 311–322.

[144] Peng Wang, Junyang Lin, An Yang, Chang Zhou, Yichang Zhang, Jingren Zhou, and Hongxia Yang. 2021. Sketch and refine: Towards faithful and informative table-to-text generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 4831–4843.

[145] Zhenyi Wang, Xiaoyang Wang, Bang An, Dong Yu, and Changyou Chen. 2020. Towards faithful neural table-to-text generation with content-matching constraints. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[146] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. In *Proceedings of the International Conference on Learning Representations*.

[147] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3731–3741.

[148] Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. 2020. Towards enhancing faithfulness for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

[149] Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2253–2263.

[150] Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. *arXiv preprint arXiv:1901.04713* (2019).

[151] Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, et al. 2021. A controllable model of grounded response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 14085–14093.

[152] Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.
[153] Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567* (2021).
[154] Weijia Xu, Xing Niu, and Marine Carpuat. 2019. Differentiable sampling with flexible reference word order for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19)*. 2047–2053.
[155] Xinnuo Xu, Ondrej Dušek, Verena Rieser, and Ioannis Konstas. 2021. AggGen: Ordering and aggregating while generating. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.
[156] Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, and Pascale Fung. 2022. Retrieval-free knowledge-grounded dialogue response generation with adapters. In *Proceedings of the 2nd DialDoc Workshop on Document-Grounded Dialogue and Conversational Question Answering*.
[157] Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2016. Neural generative question answering. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. 2972–2978.
[158] Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*. 5892–5904.
[159] Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021. D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2502–2516.
[160] Hongguang Zhang, Jing Zhang, and Piotr Koniusz. 2019. Few-shot learning via saliency-guided hallucination of samples. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA.
[161] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2204–2213.
[162] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations*.
[163] Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 4889–4896.
[164] Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5108–5120.
[165] Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2237–2249.
[166] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, et al. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*. 5905–5921.
[167] Chunting Zhou, Xuezhe Ma, and Graham Neubig Di Wang. 2019. Density matching for bilingual word embedding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'19)*. 1588–1598.
[168] Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 1393–1404.
[169] Kangyan Zhou, Shrimai Prabhumoye, and Alan W. Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
[170] Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2021. Enhancing factual consistency of abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'21)*. 718–733.

Received 11 March 2022; revised 17 October 2022; accepted 8 November 2022