Can We Catch the Elephant? A Survey of the Automatic Hallucination Evaluation on Natural Language Generation

Anonymous ACL submission

Abstract

Hallucination in Natural Language Generation (NLG) presents a significant challenge, often underestimated despite recent advances in model fluency and grammatical correctness. As text generation systems evolve, hallucination evaluation has become increasingly critical, yet current methodologies remain complex and varied, lacking clear organization. In this paper, we conduct a comprehensive survey on Automatic Hallucination Evaluation (AHE) techniques. We systematically categorize existing approaches based on the proposed evaluation pipeline: datasets and benchmarks, evidence collection, and comparison mechanisms. Our work aims to clarify these diverse approaches, highlighting limitations and suggesting avenues for future research to improve the reliability and safety of NLG models.

1 Introduction

Hallucination in Natural Language Generation (NLG) typically refers to situations where the generated text is inconsistent with or unsupported by the source input or external knowledge. Like an elephant in the room, this problem has existed since the beginning of NLG but often ignored in the early stage. As text generation models continue to evolve, technologies like Large Language Models (LLMs) have achieved grammatical correctness and fluency nearly indistinguishable from human writing. Consequently, hallucination has gradually surfaced and attracted increased attention. The automatic evaluation of hallucinations is important as it effectively drives the advancement of LLMs to be more reliable and safe. In this paper, we conduct a comprehensive survey on the process of Automatic Hallucination Evaluation (AHE) methods, which gives the current advancements made in catching hallucinations and shows future directions.

The concept of hallucination originally referred to grammatically correct but semantically inaccurate content based on source input (Lee et al., 2018). This was commonly observed in tasks like Summarization and Neural Machine Translation (NMT), where the source information is usually well-defined. The breakthrough came with the advent of LLMs like ChatGPT (OpenAI, 2022). Many NLG tasks can be effectively performed by prompting LLMs with designed instructions (Ouyang et al., 2022). However, their responses occasionally contain hallucinations that are unfaithful or factually incorrect, posing significant challenges for accurate evaluation.

Faithfulness and factuality are two concepts that are closely related when describing hallucinations and can be prone to confusion in some circumstances. In this paper, we add prefixes to both of them for better understanding by introducing **Source Faithfulness (SF)** and **World Factuality (WF)**. SF measures the degree to which the generated output accurately reflects and is consistent with the source input. SF has a limited scope, as there are specific sources that can be used to substantiate and verify the generated text. WF, on the other hand, assesses whether the generated output aligns with general world knowledge and facts. WF is a more expansive and challenging problem as it goes beyond the specific source and considers the broader context of common sense and established knowledge, which is more difficult to collect and encode comprehensively. Recent studies have recognized the critical importance of addressing and measuring the SF and WF of generated text.

Assessing from SF or WF aspects means the evaluators refer to different source information, which is closely tied to specific tasks. For example, in NMT, generated translations detached from the source text are considered unfaithful (Dale et al., 2023a). In summarization, summaries usually should be faithful to the source document, but some also argue that certain hallucinations can align with external facts (Dong et al., 2022;

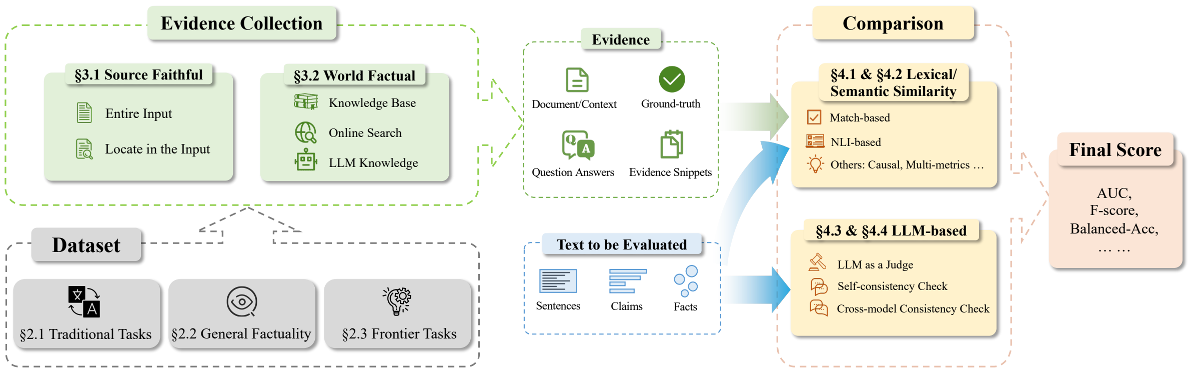

Figure 1: Automatic Hallucination Evaluation (AHE) methods typically follow a pipeline that includes dataset construction, evidence collection, and comparison between the generated output and reference evidence, resulting in a final score that reflects the level of hallucination.

Cao et al., 2022). In tasks involving LLMs, hallucinations exhibit greater diversity, occasionally encompassing both SF and WF issues simultaneously. Apart from these, LLMs face unique difficulties, such as updating world information and handling false-premise questions (Vu et al., 2023; Kasai et al., 2024).

Previous works have some introductions on methods for LLM hallucination evaluation (Huang et al., 2023; Zhang et al., 2023c; Ji et al., 2023; Huang et al., 2021), but they have neither categorized the existing benchmarks nor systematically summarized the processes of the evaluators, nor have they conducted a comparative analysis of the methods at different steps. In contrast, this paper comprehensively introduces AHE methods by following the structure of the proposed pipeline, as illustrated in Figure 1 and Figure 2. It begins with an overview of **Datasets and Benchmarks**, which is the first step and foundation of AHE (see § 2). This is followed by a discussion of **Evidence Collection**, which identifies WF/SF evidence for hallucination evaluations(see § 3). Then, this paper details how evaluators use the evidence for **Comparison** to get the quantitative evaluation results (see § 4). Although not all AHE methods fully implement each of these steps, this standardized pipeline methodology helps us understand the underlying connections between different approaches and their evolution from the pre-LLM era to the post-LLM era. We also present Table 1 and Table 2 for all the methods surveyed in this paper, including key aspects discussed in the following sections. Finally, following the pipeline, this paper summarizes the current state of research on AHE, outlining existing challenges and suggesting potential directions for future investigation.

2 Dataset and Benchmark

This section introduces datasets and benchmarks developed for evaluating model hallucination. Of the evaluators surveyed, 56.1% present their datasets or benchmarks for evaluation. The evolution has shifted from task-specific methods to general factuality assessments, with recent works focusing on more practical and diverse domains, adapting design patterns to various usage scenarios.

2.1 Task-specific

Task-specific datasets, though not designed for hallucination research, inherently exhibit relevant phenomena, making them suitable for hallucination evaluation. For summarization task, many works manually evaluate the model-generated summaries and publish the annotations. On the news datasets Xsum and CNN/DM, Maynez et al. (2020) publish XSumFaith with hallucination types (intrinsic or extrinsic) at the span positions, CoGenSumm (Falke et al., 2019) gives annotation on CNN/DM dataset, and QAGS (Wang et al., 2020) annotates each sentence with a binary label of SF on both datasets. Polytope (Huang et al., 2020) provides both SF and WF annotations to measure both extractive and abstractive summarization.

However, the binary classes of texts can be difficult to determine. FRANK (Pagnoni et al., 2021) collects annotation based on a more fine-grained defined typology of factual errors. Similarly, for dialogue summarization task, FactEval (Wang et al., 2022) includes hallucination error during annotating and RefMatters (Gao et al., 2023) further refines the error categories by combining content-based and form-based factual errors. Additionally, Devaraj et al. (2022) categorize 3 types of factual errors for data collected from Newsela (Xu et al., 2015) and Wikilarge (Zhang and Lapata, 2017) for text simplification. In dialogue generation, DialogueNLI (Welleck et al., 2018) provides three-type labels of the entailment of sentence pairs.

Besides annotating existing generated summaries, data augmentation serves as an additional method for creating training data. Falsesum (Utama et al., 2022) automates the augmentation process and can control the intrinsic and extrinsic errors in summaries. Task-specific data annotation and augmentation methods are advancing toward greater detail, automation, and scalability. As LLMs evolve, the task boundaries become increasingly blurred, suggesting that future datasets should align with more comprehensive domains.

2.2 General Factuality

Beyond task-specific datasets, some studies have shifted their focus toward more generalized evaluations to assess LLMs' ability to avoid hallucinations. This process is usually carried out through multiple turns of Questions and Answers (QA).

Within knowledge-grounded dialogue, Q^2 (Honovich et al., 2021) gives an annotated dataset of factual consistency with respect to a given knowledge. FACTOR (Muhlgay et al., 2023) follows the error types from FRANK (Pagnoni et al., 2021) and performs a multi-choice factual evaluation task with the help of Wikipedia, news, and expert-curated QA datasets. Also with the help of Wikipedia, PHD (Yang et al., 2023) focuses on passage-level entity-centric knowledge, and HaluEval (Li et al., 2023) verifies hallucinations in ChatGPT. The truthfulness of LLMs extends beyond mere knowledge to encompass other behaviors, where TruthfulQA (Lin et al., 2022) highlights the trade-off between truthfulness and informativeness in LLMs, stating that hedging is better than providing wrong answers. The evaluation of hallucinations in LLMs focuses more on WF accuracy. As a result, large-scale common knowledge sources, such as Wikipedia, are often used to support the construction of evaluation datasets.

2.3 Frontiers

Recent advancements have increasingly focused on AHE across multiple diverse and critical aspects.

Long Context/Generation Despite recent advancements in LLMs enabling them to handle long texts better, hallucination evaluation in a long context or generation remains a challenge. BAMBOO (Dong et al., 2023) includes the hallucination detection task to its multi-task long context benchmark, and FactScore (Min et al., 2023) provides long-form biographies sampled from Wikipedia and breaks the generated text into fine-grained atomic facts with each assigned a binary label.

Domain-specific Hallucinations in specialized fields such as medicine or law can lead to serious consequences, and constructing relevant datasets is particularly needed. MedHalt (Pal et al., 2023) gathers seven medical datasets to a benchmark for LLMs' hallucination evaluation. Magesh et al. (2024) provide references for QA in the law field, including legal questions from five aspects.

Non-English Languages Numerous Chinese LLMs have also emerged along with the trend and hallucination is also a crucial problem. UHGEval (Liang et al., 2023) hallucination dataset is generated by Chinese LLMs in news domain, while ChineseFactEval (Wang et al., 2023a) covers areas in daily life and specifically includes the modern Chinese history. Similarly, inspired by TruthfulQA (Lin et al., 2022), HalluQA (Cheng et al., 2023) summarizes the question patterns and combines them with Chinese culture, and categorizes hallucinations into imitative falsehoods and factual errors. Another Chinese-English benchmark ANAH (Ji et al., 2024) prompts the model to annotate hallucination for each sentence. Other than Chinese, multilingual datasets such as HalOmi (Dale et al., 2023b) can help evaluate hallucinations in different languages and distinguish them between translation errors.

Fact Reasoning Hallucination in LLM reasoning can be complex due to the muli-step process. Laban et al. (2023) build a benchmark SUMMEDITS, which provides a three-step protocol for inconsistency detection benchmark creation and implements it in a 10-domain benchmark.

Fresh Fact As the world is constantly changing, a critical question arises: how can we assess whether LLMs possess dynamic knowledge? The following benchmarks concentrate on constructing time-sensitive datasets to enable the evaluation of LLMs' capacity to incorporate up-to-date information. FreshQA (Vu et al., 2023) includes questions about current events and also inputs with false premises to the LLMs. RealTimeQA (Kasai et al., 2024) tests on both open- and closed-book QA systems. KoLA (Yu et al., 2023) uses both Wikipedia and continuously collected emerging news and novels for evaluation. ERBench (Oh et al., 2024) leverages the benefits of databases for easy updates through an entity-relationship model. To facilitate real-world applications, ToolBH (Zhang et al., 2024) evaluates the hallucination tendencies of LLMs by examining both depth and breadth across various scenarios and tasks.

2.4 Evaluate the Evaluators

Furthermore, for evaluating the evaluators themselves, SummEval (Fabbri et al., 2021), SummaC (Laban et al., 2022), Dialsummeval (Gao and Wan, 2022), and AGGREFACT (Tang et al., 2023) focus on summarization hallucination evaluation or detection. In the domain of dialogue generation, Wizard of Wikipedia (Dinan et al., 2018), CI-ToD (Qin et al., 2021), BEGIN (Dziri et al., 2022b), FaithDial (Dziri et al., 2022a) and TopicalChat (Gopalakrishnan et al., 2023) facilitate the measurement of consistency in evaluators. RealHall (Friel and Sanyal, 2023) is a benchmark for evaluation methods and contains both closed- and open-domain hallucinations, corresponding to SF and WF. FELM (Zhao et al., 2024) expands to diverse domains: science, math, recommendation, and reasoning. TRUE (Honovich et al., 2022) and BEAMetrics (Scialom and Hill, 2021) also can evaluate metrics across a series of NLG tasks.

In general, datasets and benchmarks from various field have emerged to better evaluate hallucinations. Despite the abundance of datasets, many suffer from limited data size and a one-to-one correspondence between datasets and evaluation methods. Future dataset construction should focus on integration from multiple sources, standardization, and maintaining both quality and quantity.

3 Evidence Collection

Datasets and benchmarks provide the foundation for AHE. Large-scale automation for evidence collection is essential to achieve AHE. In this section, we explore methods that do not rely on ground-truth references. For SF evaluation, evidence is directly derived from the source input or contextual information, whereas for WF evaluation, it is typically sourced from external or model knowledge.

3.1 SF Evidence

To determine the faithfulness of the generated text, the source input can be utilized in two ways: as an entire reference or by locating relevant evidence within it.

Entire Input as Evidence Utilizing the entire input as evidence implies that the evaluation process does not involve extracting specific sentences or spans. For tasks such as text summarization or simplification with long input, Maskeval (Liu et al., 2022) gets the token importance weights by concatenating the output and source text to fine-tune a masked language model. For NMT task, the input and output typically have approximately the same length and convey the same information. So it is natural for NMT evaluators to use the input as the comparison object (Guerreiro et al., 2023; Dale et al., 2023a). While this approach is straightforward and effective, it also has significant flaws that encompass much irrelevant information.

Locate Evidence in the Input To avoid information redundancy in evidence collection, more recent methods employ strategies to identify relevant evidence, specifically targeting content that either supports or contradicts the output text. One widely adopted approach for evaluating summarization tasks is Question Generation and Question Answer (QG-QA). A common framework is extracting QA pairs from the summary, using QA models to retrieve answers from the document, and checking consistency, such as FEQA (Durmus et al., 2020)and QAGS (Wang et al., 2020). In this context, the answer derived from the document serves as evidence to validate the summary answer. Because the summary should contain key information from the document, QuestEval (Scialom et al., 2021) trains a question weighter to label important questions. QAFactEval (Fabbri et al., 2022) further explores the use of abstractive QA models, but finding no significant difference in performance between extractive and abstraction QA approaches. This suggests that QA capability is not the primary bottleneck in the task. For answer selection, Fabbri et al. (2022) demonstrate that selecting noun phrase chunks as answers yields better performance than entities. These evidences are usually words or short spans, a more comprehensive approach involves dividing the context into segments (Zha et al., 2023) or representing the core content of the source input as a semantic graph (Ribeiro et al., 2022).

3.2 WF Evidence

Retrieving evidence from external sources is more challenging due to the difficulty in determining search boundaries, identifying connections, and extracting critical information [1].

External Knowledge Base (KB) Leveraging the external KBs offers a comprehensive reservoir of world knowledge. The main challenge is to accurately identify and extract relevant information from this extensive data pool. Among the KBs utilized, Wikipedia is the most commonly employed, with others such as YAGO, KGAP, and UMLS also being used (Feng et al., 2023). The format of knowledge extraction can vary, including entities (Yang et al., 2023), triplets (Feng et al., 2023), or fine-defined atomic facts (Min et al., 2023). Domain-wide KBs like PubMed are also essential for biomedical information retrieval (Pal et al., 2023). When multiple pieces of evidence are available, identifying the related ones is also a significant step before making the judgement (Wang et al., 2024).

Online Search While KBs can only provide static information, utilizing tools such as search engines can help access dynamic and up-to-date information. FacTool (Chern et al., 2023) decomposes the sentences into checkable atomic claims used for online searches. HaluAgent (Cheng et al., 2024) also combines smaller LLMs with search tools to retrieve evidences. Before searching, Factcheck-GPT (Wang et al., 2023c) incorporates a check-worthiness selection module for each claim.

LLM as Knowledge Base LLMs have massive learned knowledge while training, powerful LLMs can serve as KBs. In a closed-book setting, the LLM generates answers solely based on the parameteratic knowledge, without relying on any external KBs. Moreover, LLMs can be injected with more knowledge by fine-tuning and retrieving (Ovadia et al., 2023; Chen et al., 2024b). UFO (Huang et al., 2024b) introduces a fact verification framework that incorporates multiple sources of evidence, including knowledge from LLMs. Similarly, CONNER (Chen et al., 2023a) utilizes LLMs to generate related knowledge as evidence for evaluation. RefChecker (Hu et al., 2024a) applies LLMs' knowledge to solve the zero-context hallucination detection. These approaches are applied to knowledge-intensive tasks, such as open-domain question answering and knowledge-grounded dialogue.

The effectiveness of evidence derived from fixed sources, such as SF evidence and ones based on static KBs, is largely determined by extraction accuracy. Online search, while offering extensive coverage, can suffer from information loss due to the lengthy search pipeline, and the effectiveness of online search often depends on the quality of search recommendations. Reliance on LLMs for evidence retrieval may lead to the issue of "lying to verify lies", as LLMs themselves can suffer from hallucinations. The manner in which this evidence is utilized, specifically, how it is compared with the generated text directly determines the evaluation outcome.

[1]The retrieval-augmented phase of the Retrieval-Augmented Generation (RAG) framework follows a process similar to the methods discussed in this section.

4 Comparison

Various approaches have been proposed to compare the generated text with corresponding ground truths or collected evidence. These range from model-free methods to more advanced techniques that employ multiple models for judgment. While certain methods leverage the evidence to compute this similarity, others operate independently of the evidence, instead relying on the knowledge encoded within the model itself. In this section, we categorize the comparison methods into distinct groups and present an overview of the corresponding scoring metrics alongside the associated approaches.

4.1 Lexical Similarity

Lexical similarity refers to the measurement of the closeness or similarity between two pieces of text based on their word usage. Traditional n-gram methods like ROUGE (Lin, 2004) measure n-gram overlap between texts but show weak correlation with human evaluations (Maynez et al., 2020). Therefore, the methods discussed below represent statistical metrics grounded in the definition of facts instead of n-grams.

Exact Match (EM) EM score is based on the definition of facts. $Fact_{acc}$ (Goodrich et al., 2019) defines the fact schemas as triplet tuples (entity-relation-entity), and then the score is calculated by comparing the schema between the ground-truths and generated text. Maskeval (Liu et al., 2022) evaluates on the token level, and combines masked LM weights with EM scores.

QG-QA Answer Match In the context of QG-QA approaches, some answers are relatively short, such as entities or informative text segments. Within this framework, the similarity between system-generated outputs and source-derived answers can be quantitatively assessed through lexical overlap. For summarization task, FEQA (Durmus et al., 2020), QAGS (Wang et al., 2020) and QuestEval (Scialom et al., 2021) use F1-score to compare the answers. MQAG (Manakul et al., 2023a) computes the statistical distance (e.g. KL-Div) of answers over automatically generated multiple-choice questions,

QA Benchmark Answer Match To assess the hallucination level of LLMs, many of the benchmarks introduced in Sec. 2 are typically framed in QA tasks. While the focus of these benchmarks may differ, they all provide ground-truth answers for evaluation. One line of research involves prompting LLMs to generate answers to the given questions and subsequently evaluating their performance using EM scores (Kasai et al., 2024; Oh et al., 2024). Another line of research involves using multiple-choice tasks (Lin et al., 2022; Kasai et al., 2024; Oh et al., 2024; Dong et al., 2023), where accuracy or F-score are computed as the final performance metrics.

4.2 Semantic Similarity

The approaches presented in this section diverge from the lexical similarity, as they are not based on the word matching score. Instead, these methods exploit the semantic meaning of text, either by assessing the entailment likelihood between the generated text and the source evidence or leveraging from more diverse perspectives

Data-augmentation NLI One way to measure semantic similarity involves evaluating the degree of entailment using a NLI model, wherein the predicted likelihood is utilized as a measure of the entailment score. Among these methods, data augmentation is a widely adopted technique to enhance the performance of NLI models. Building positive and negative samples is an effective way to improve model ability to distinguish them. Positive data is usually built by paraphrasing or backtranslation (Kryscinski et al., 2020; Wang et al., 2022). For negative data, FactCC and FactCCX (Kryscinski et al., 2020) achieve this through word swapping and noise injection. And FactPush (Steen et al., 2023) further augments negative samples by appending random phrases. Alternatively, FactKB (Feng et al., 2023) augments the training data with external triplet knowledge, which can improve the model's ability of knowledge understanding.

Semantic-structure NLI With more focus on the encoding processes, some studies leverage sentence or document structure to construct semantic representations. For example, DAE (Goyal and Durrett, 2020) applies the entailment model on the dependency level of a sentence, specifically focusing on the relationship between the head and tail of a dependency arc. Expanding on this, FactGraph (Ribeiro et al., 2022) improves discourse understanding by encoding semantic structures as graphs for both the input and output.

NLI for Answer Match Beyond using NLI models solely for text entailment checking, studies (Fabbri et al., 2022; Honovich et al., 2021) within the QG-QA pipeline have demonstrated that leveraging NLI models for answer similarity checking is an effective approach. These works further highlight that QA-based and NLI-based metrics can provide complementary insights.

Other Methods The aforementioned NLI methods focus on evaluating similarity within a binary classification framework. However, hallucinations can be assessed from a broader range of perspectives, allowing for more nuanced evaluation. CoCo (Xie et al., 2021) introduces counterfactual data to measure the causal effects between source documents and generated summaries. AlignScore (Zha et al., 2023) builds an alignment model utilizing a LM and 3 individual linear layers as the 3-way classification (aligned, contradict, neutral), binary classification (aligned, not-aligned), and regression (score $\in [0, 1]$) heads.

In addition to employing a single metric for evaluation, several studies have explored the aggregation of multiple metrics in a collaborative manner to provide a more comprehensive assessment. WeCheck (Wu et al., 2023) introduces a weak supervision learning paradigm that builds upon existing metrics, utilizing a combination of NLI datasets for initialization and noise-aware fine-tuning to develop a target metric model. Similarly, STARE (Himmi et al., 2024) combines signals from internal model-based and external detectors to improve hallucination detection on NMT task. Other than using the off-the-shelf methods, ExtEval (Zhang et al., 2023b) identifies five broad categories of unfaithfulness issues in extractive summarization that cannot be fully addressed by entailment models, with each category being assessed through a specific sub-metric.

4.3 LLM as a Judge

In this section, we introduce approaches that leverage LLMs as evaluators for hallucination evaluation. The core premise of this approach is that LLMs possess parametric knowledge acquired during training and can be prompted to complete various tasks (Li et al., 2024).

The evaluation process involves first providing the LLM with the evaluation criteria and task description, followed by supplying the task inputs for judgment. The feasibility of ChatGPT as an effective evaluator is specifically examined by Wang et al. (2023b), demonstrating its potential for building evaluators with or without reference inputs. For specific tasks, SCALE (Lattimer et al., 2023) focuses on long-form dialogue, segmenting lengthy source documents into chunks and assessing the level of support provided by each text snippets. Chen et al. (2023b) experiments the few-shot and zero-shot scenarios to evaluate summarization task. Expanding to a broader range of tasks, GPTScore (Fu et al., 2023) and G-Eval (Liu et al., 2023) both offer multi-faceted evaluation frameworks that include consistency as a key metric. Chain-of-thoughts (CoT) also can enables the reasoning capabilities of LLMs (Liu et al., 2023; Friel and Sanyal, 2023; Akbar et al., 2024), as it provides transparency by outlining the intermediate steps involved in judging and improves the complex and nuanced judgments.

4.4 Consistency Cross Check

The evaluators discussed above primarily focus on comparing the target text with either extracted evidence or the broader context. However, when assessing LLMs, an alternative approach is to examine the consistency of the LLM's output. The underlying premise is that a model with lower generation uncertainty is likely to demonstrate higher confidence in producing hallucination-free content. This method can be categorized into two distinct approaches: self-consistency check and cross-model consistency check.

Self-consistency Check This approach assumes that an LLM will show self-consistency if it possesses relevant knowledge. Based on this, SelfCheckGPT (Manakul et al., 2023b) employs a zero-resource hallucination detection framework by evaluating the consistency of multiple sampled responses. InterrogateLLM (Yehuda et al., 2024) measures consistency by reconstructing the input query from generated responses and comparing it to the original. To evaluate LLMs' world knowledge, KoLA (Yu et al., 2023) develops a self-contrast metric by contrasting two completions generated by the same model and gets the similarity score.

In addition to examining the generated text, the semantic information retained within the internal states can also assist in the judgment process. Based on multiple generations, EigenScore (Chen et al., 2024a) leverages eigenvalues of responses' covariance matrix to measure self-consistency. Another line of research does not rely on multiple generations from the model but instead utilizes the difference between internal states and outputs. LLM-Check (Sriramanan et al.) employs internal attention kernel maps, hidden activations, and output prediction probabilities to assess hallucinations, while Lookback-Lens (Chuang et al., 2024) uses attention maps to detect contextual hallucinations. EGH (Hu et al., 2024b) models the distributional distance between embeddings and gradients of regular conditional and unconditional outputs through Taylor expansion. Likewise, PHR (Jesson et al., 2024) estimates hallucination rates by evaluating response log probabilities from conditional generative models.

Cross-model Consistency Check Although self-inconsistency in LLMs is often associated with hallucinations, self-consistency does not inherently ensure factual accuracy in generated content. Therefore, SAC^3 (Zhang et al., 2023a) includes verifier LMs to perform cross-checking, and considers both question inputs and answer outputs when measuring semantic consistency.

When ground truth or evidence is available, evaluation typically involves measuring lexical or semantic similarity, where the NLI models can also integrate effectively with QG-QA evaluators. The use of LLMs for evaluation is straightforward and convenient, offering flexibility in designing evaluation criteria based on specific tasks and enabling multi-faceted assessments. However, despite increasing confidence in LLMs as their size and capabilities expand, ensuring their stability and reliability in evaluation tasks remains an open challenge. Enhancing LLMs' capabilities in judgment, retrieval, and self-improvement represents a critical direction for future research.

5 Discussion and Future Directions

While existing AHE methods have demonstrated substantial progress, critical gaps persist in hallucination detection and evaluation. Particularly in cutting-edge task domains, certain hallucinations remain complex and difficult to detect and evaluate, which deserve further investigation.

5.1 Discussion Questions

Hallucination vs. Text Error It can be challenging to distinguish between hallucinations and other text errors (Guerreiro et al., 2023), such as less severe entity mistranslations. According to the traditional definition of hallucination (smooth but incorrect), any response from a large model that differs from the ground truth can be considered as hallucination, which is obviously unreasonable and can mislead researchers. Some works in NMT have already made progress in this area (Dale et al., 2023a), and future evaluation methods should aim to accurately identify real hallucinations.

Fine vs. Coarse Fact Granularity The studies surveyed in this work attempt to evaluate hallucinations at various granularities, ranging from fine-grained units such as tokens and entities to more coarse-grained units like phrase spans, claims, sentences, and document chunks. Which fact granularity is the best? Some studies have explored different levels of fact granularity (Hu et al., 2024a), or sought to integrate multiple granularities (Xie et al., 2021; Zhao et al., 2024). Determining the optimal granularity is challenging, as it is highly context-dependent and task-specific.

Hallucination vs. Imagination Is hallucination always bad? Not necessarily. In certain contexts, such as discussions about a sci-fi novel, imaginative content is expected, and the dialogue should be creative. In such cases, the line between hallucination and imagination becomes subtle. Differentiating between these two phenomena can help models more effectively evaluate diverse types of text (Zhou et al., 2024).

5.2 Future Directions

Supporting Theories Previous evaluations and detection of hallucinations have primarily focused on examining the final output of the model, specifically the hallucinations manifested in the generated text. Some preliminary studies have explored the feasibility of using internal states for hallucination evaluation (Chuang et al., 2024; Hu et al., 2024b). However, the underlying mechanisms remain under investigation.

Interpretability Identifying fact granularity and analyzing the reasons behind hallucination can provide significant assistance in solving hallucination problems. Some reasoning methods (Akbar et al., 2024) have the potential to analyze the underlying causes of hallucinations and offer better evaluation. Other observing aspects lie in the internal state of model generation (Su et al., 2024), which provide more analytical perspectives.

Complex Context It is crucial to address hallucinations caused by the model's difficulty in understanding complex inputs, including the long or multi-form context. Hallucinations caused by contradictions between the beginning and end of long outputs are also worth further exploration (Wei et al., 2024), such as detecting inconsistencies in character behavior within model-generated narratives. Furthermore, investigating multi-evidence verification during hallucination evaluation also presents a promising direction for future research (Wang et al., 2024).

Other Applications Moreover, the latest research focuses on expanding LLMs to areas such as multilingual, multimodality, autonomous agents, and real-world applications, which bring about new types of hallucinations, such as code hallucination (Qian et al., 2023), tool hallucination (Zhang et al., 2024), visual hallucination, cross-lingual halluciantion (Dale et al., 2023b), multimodal hallucination (Huang et al., 2024a), and so on. Evaluating such hallucinations is a very interesting and worthwhile direction to explore.

6 Conclusion

Evaluating hallucination in NLG is essential, as it influences the direction and future trends in developing more robust models. In this survey, we present the works of AHE by organizing it according to the steps of the evaluation pipeline, covering both SF and WF fields. Traditionally, most evaluation metrics have been task-specific, given the relative ease of defining criteria for task performance. However, with the growing focus on LLMs, new demands and challenges have emerged, prompting researchers to reconsider evaluation frameworks.

7 Limitation

In this paper, we collect a broad range of related papers and reports, categorize and compare various methods, and provide insights into discussion and potential future directions. However, this paper does have several limitations.

First of all, we did not do comprehensive experiments to revisit the above evaluators, because the evaluators usually focus on different types of hallucinations for various tasks, and it wouldn't be fair to compare across the categories. For example, evaluators for LLMs intend to build their own datasets with human annotation, which vary in categories and schemes. Secondly, content related to fact-checking and human evaluation is provided in Appendix C and Appendix D. Meanwhile, this survey focuses exclusively on text-to-text hallucinations. Due to space limitations, a comprehensive discussion of these topics is not included, as such details may divert attention from the primary focus of this paper. Last but not least, the case study we provide in Appendix B only includes a few representative cases on selective models for reference.

Acknowledgments

References

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica Salinas, Victor Alvarez, and Erwin Cornejo. 2024. Hallumeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15020–15037.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023a. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. pages 6325–6341, Singapore.

Shiqi Chen, Siyang Gao, and Junxian He. 2023b. Evaluating factual consistency of summaries with large language models. *arXiv preprint arXiv:2305.14069*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Kun Gai, and Ji-Rong Wen. 2024. Small agent can also rock! empowering small language models as hallucination detector. *arXiv preprint arXiv:2406.11277*.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R Costa-jussà. 2023b. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. *arXiv preprint arXiv:2305.11746*.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*.

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 Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar R Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022b. Evaluating attribution in dialogue systems: The begin benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander Richard Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *arXiv preprint arXiv:2310.18344*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. pages 5693–5709, Seattle, United States.

Mingqi Gao, Xiaojun Wan, Jia Su, Zhefeng Wang, and Baoxing Huai. 2023. Reference matters: Benchmarking factual error correction for dialogue summarization with fine-grained evaluation framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13932–13959, Toronto, Canada. Association for Computational Linguistics.

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 & data mining*, pages 166–175.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Anas Himmi, Guillaume Staerman, Marine Picot, Pierre Colombo, and Nuno M Guerreiro. 2024. Enhanced hallucination detection in neural machine translation through simple detector aggregation. *arXiv preprint arXiv:2402.13331*.

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 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

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*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024a. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6953–6975.

Xiaomeng Hu, Yiming Zhang, Ru Peng, Haozhe Zhang, Chenwei Wu, Gang Chen, and Junbo Zhao. 2024b. Embedding and gradient say wrong: A white-box method for hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1959.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024a. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683*.

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*.

Zhaoheng Huang, Zhicheng Dou, Yutao Zhu, and Jirong Wen. 2024b. Ufo: a unified and flexible framework for evaluating factuality of large language models. *arXiv preprint arXiv:2402.14690*.

Andrew Jesson, Nicolas Beltran-Velez, Quentin Chu, Sweta Karlekar, Jannik Kossen, Yarin Gal, John P Cunningham, and David Blei. 2024. Estimating the hallucination rate of generative ai. *arXiv preprint arXiv:2406.07457*.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical annotation of hallucinations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.

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 Computing Surveys*, 55(12):1–38.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. pages 7561–7583, Singapore.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2024. Realtime qa: What's the answer right now? *Advances in Neural Information Processing Systems*, 36.

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 (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics.

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*, 10:163–177.

Barrett Lattimer, Patrick CHen, Xinyuan Zhang, and Yi Yang. 2023. Fast and accurate factual inconsistency detection over long documents. pages 1691–1703, Singapore.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Zhaohui Wy, Dawei He, Peng Cheng, Zhonghao Wang, et al. 2023. Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. *arXiv preprint arXiv:2311.15296*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yu Lu Liu, Rachel Bawden, Thomas Scialom, Benoît Sagot, and Jackie Chi Kit Cheung. 2022. Maskeval: Weighted mlm-based evaluation for text summarization and simplification. *arXiv preprint arXiv:2205.12394*.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023a. MQAG: Multiple-choice question answering and generation for assessing information consistency in summarization. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 39–53, Nusa Dua, Bali. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023b. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. pages 9004–9017, Singapore.

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*, pages 1906–1919, Online. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. pages 12076–12100, Singapore.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Jio Oh, Soyeon Kim, Junseok Seo, Jindong Wang, Ruochen Xu, Xing Xie, and Steven Euijong Whang. 2024. Erbench: An entity-relationship based automatically verifiable hallucination benchmark for large language models. *arXiv preprint arXiv:2403.05266*.

OpenAI. 2022. Chatgpt.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2023. Fine-tuning or retrieval? comparing knowledge injection in llms. *arXiv preprint arXiv:2312.05934*.

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*, pages 4812–4829, Online. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. pages 314–334, Singapore.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 2021. Don't be contradicted with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system. pages 2357–2367, Online and Punta Cana, Dominican Republic.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. pages 624–643, Online.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. Questeval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604.

Thomas Scialom and Felix Hill. 2021. Beametrics: A benchmark for language generation evaluation evaluation. *arXiv preprint arXiv:2110.09147*.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Julius Steen, Juri Opitz, Anette Frank, and Katja Markert. 2023. With a little push, nli models can robustly and efficiently predict faithfulness. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 914–924.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. *arXiv preprint arXiv:2403.06448*.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. pages 11626–11644, Toronto, Canada.

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, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

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*, pages 5008–5020, Online. Association for Computational Linguistics.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022. Analyzing and evaluating faithfulness in dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4897–4908, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Binjie Wang, Ethan Chern, and Pengfei Liu. 2023a. Chinesefacteval: A factuality benchmark for chinese llms.

Binjie Wang, Steffi Chern, Ethan Chern, and Pengfei Liu. 2024. Halu-j: Critique-based hallucination judge. *arXiv preprint arXiv:2407.12943*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023b. Is ChatGPT a good NLG evaluator? a preliminary study. pages 1–11, Singapore.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. 2023c. Factcheck-gpt: End-to-end fine-grained document-level fact-checking and correction of llm output. *arXiv preprint arXiv:2311.09000*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2018. Dialogue natural language inference. *arXiv preprint arXiv:1811.00671*.

Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Sujian Li, and Yajuan Lyu. 2023. WeCheck: Strong factual consistency checker via weakly supervised learning. pages 307–321, Toronto, Canada.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. pages 100–110, Punta Cana, Dominican Republic.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3898–3908.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9333–9347, Bangkok, Thailand. Association for Computational Linguistics.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023a. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. pages 15445–15458, Singapore.

Shiyue Zhang, David Wan, and Mohit Bansal. 2023b. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2153–2174, Toronto, Canada. Association for Computational Linguistics.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023c. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen Wan, Yujiu Yang, et al. 2024. Toolbehonest: A multi-level hallucination diagnostic benchmark for tool-augmented large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11388–11422.

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2024. Felm: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 36.

Yilun Zhou, Caiming Xiong, Silvio Savarese, and Chien-Sheng Wu. 2024. Shared imagination: Llms hallucinate alike. *arXiv preprint arXiv:2407.16604*.

A Evaluator Taxonomy and Meta-Info

We present Figure 2 to clearly display the taxonomy for AHE methods according to the pipeline we proposed. We also provide a table of meta information for the evaluators here, as in Table 1 and Table 2. For the *Based-model* column, it means the models that evaluators use to perform evaluation or generate synthetic data. *Metric* column means the calculating method to get the final score. ✓and ✗ in *SF* and *WF* columns mean the aspects that the evaluators focus on.

B Case Study

Among the SF and WF errors discussed in this paper, we present a four-quadrant diagram in Figure 3 to more effectively illustrate these errors.

Here we present some results of selected evaluators on different kinds for SF or WF errors on summarization data in Table 3. The data we used are from XEnt dataset (Cao et al., 2022) and FactCollect (Ribeiro et al., 2022). We selected evaluators that use the GPT series and those that do not, covering both models that evaluate SF and WF facets. For the models utilizing LLMs, we specifically employed GPT-3.5-turbo. Although FacTool is not directly applicable for evaluating summarization tasks, we conducted experiments under its KBQA (Knowledge-Based Question Answering) setup to see its transfer ability.

The results of different models on these cases show considerable variation. In the SF-WF case, only FacTool made an incorrect judgment, which might be attributed to its insufficient transfer ability. SelfCheckGPT uses a zero-shot approach in its prompt to assess the consistency, whereas HaluEval's prompt provides examples for judgment. However, the SFE cases indicate that the results of these two evaluators remain unstable. For the WFE cases, FacTool provides the correct answers, and surprisingly, WeCheck also made correct judgments. Currently, to our best knowledge, no such labeled data is available for full evaluation. More accurate data is needed for further experiments to validate the preferences of different evaluators.

C Fact-checking

Fact-checking or fact-verification task is another line of work that has been paid much attention. The fact-checking framework can be divided into three components: claim detection, evidence retrieval, and claim verification (Guo et al., 2022), which is a relatively mature pipeline. Distinct from the evaluation methods discussed before in this paper, it typically involves assessing the factual accuracy of individual claims, mostly focusing on their WF. Wikipedia is a commonly used source for world knowledge (Thorne et al., 2018; Schuster et al., 2021; Kamoi et al., 2023; Gupta et al., 2022; Schuster et al., 2021), not only for fact-checking, but also for factuality evaluation. Especially when extracting evidence from a specific source, the WF turns into SF, which also demonstrates the dialectical unity of WF and SF. Benefiting from LLMs, fact-checking can process longer and more complex texts with more confidence and efficiency (Wang et al., 2023c). Due to the nature of the fact-checking task, it can be seen as a WF evaluator for text generation with a binary (true/false) checker.

D Human Evaluation

For hallucination evaluation, human perspectives can play a pivotal role, providing datasets and establishing benchmarks for the development of automatic models. To build a human annotation framework, there are three aspects requiring consideration: 1) How to design error categories and unify guidelines for annotators; 2) How to ensure the reliability of human annotation; And 3) how to digitally present annotated results. Human evaluation can be time-consuming and is particularly inefficient for large-scale evaluations, but still is the most trustworthy way of model evaluation.

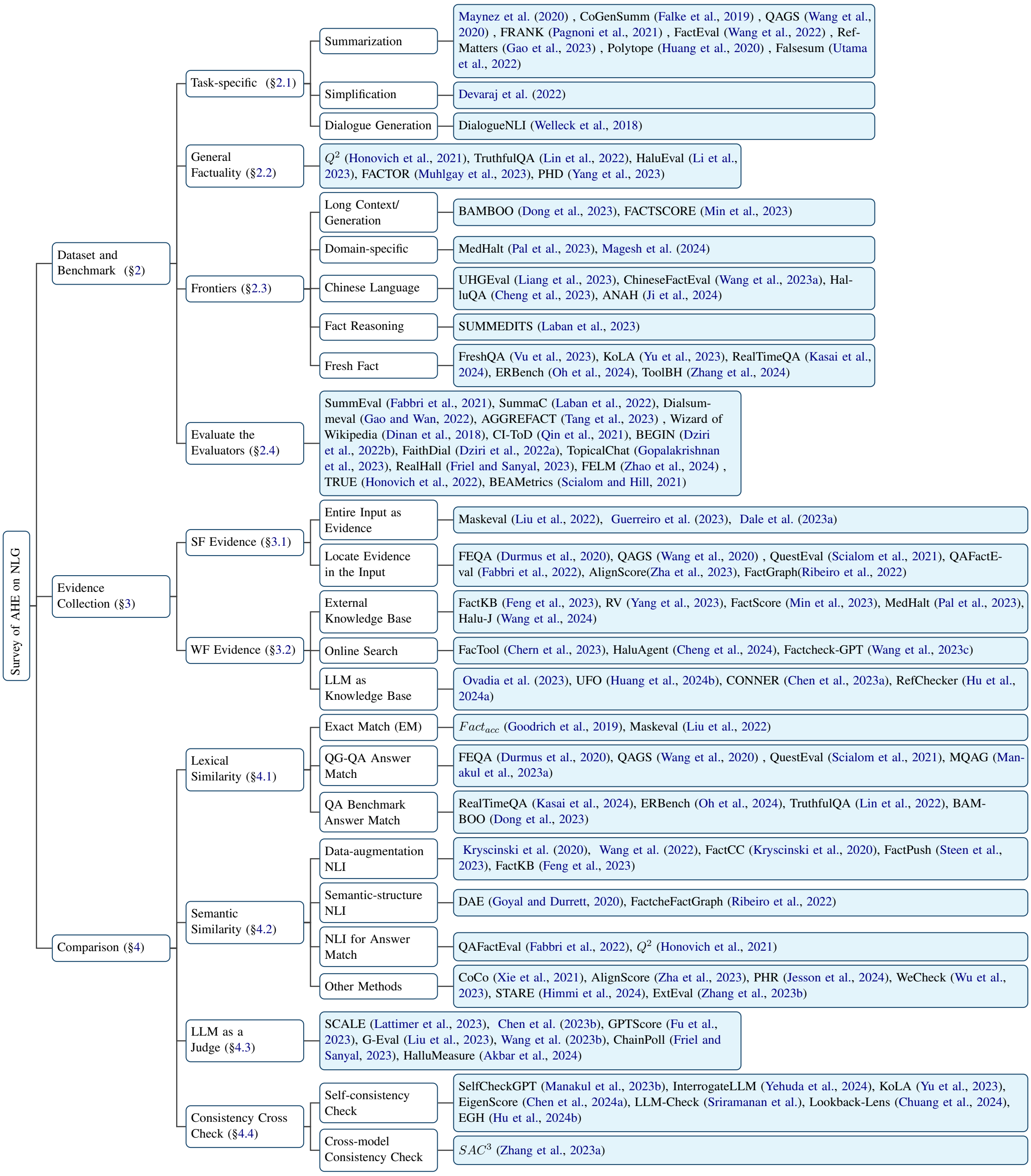

Figure 2: Taxonomy of AHE methods based on the distinct techniques employed at each stage of the pipeline.

Era	Name	New Dataset	Data Source	Fact Definition	Task	Based-model	Method	Metric	SF	WF
Before LLM Era	$Fact_{acc}$	WikiFact	Wikipedia, Wikidata KB	Triplet	Summ	Transformer	Triplet Extraction	P, R, F1	✓	✗
	FactCC	FactCC	CNN/DM, XSumFaith	Sent	Summ	BERT	NLI (2-class)	Likelihood	✓	✗
	DAE	DAE	PARANMT50M	Dependency	Summ	ELECTRA	NLI (2-class)	Likelihood	✓	✗
	Maskeval	/	CNN/DM, WikiLarge, ASSET	Word	Summ, Simp	T5	Word Weighting	Weighted Match Score	✓	✗
	Guerreiro et al. (2023)	Haystack	WMT2018, DE-EN	Text Span	NMT	Transformer	Uncertainty Measure	Avg. Similarity	✓	✗
	Dale et al. (2023a)	/	Haystack	Text Span	NMT	Transformer	Source Contribution	Percentage	✓	✗
	FEQA	FEQA	CNN/DM, XSum	Sent Span	Summ	BART (QG), BERT (QA)	QG-QA	Avg. F1	✓	✗
	QAGS	QAGS	CNN/DM, XSum	Ent, Noun Phrase	Summ	BART (QG), BERT (QA)	QG-QA	Avg. Similarity	✓	✗
	QuestEval	/	CNN/DM, Xsum	Ent, Noun	Summ	T5 (QG, QA)	QG-QA	P, R, F1	✓	✗
	QAFactEval	/	SummaC	NP Chunk	Summ	BART (QG), ELECTRA (QA)	QG-QA, NLI	LERC	✓	✗
	MQAG	/	QAGS, XSumFaith, Podcast, Assessment, SummEval	Sent Span	Summ	T5 (QG), Longformer (QA)	Multi-Choice QA	Choice Statistical Distance	✓	✗
	CoCo	/	QAGS, SummEval	Token, Span, Sent, Doc	Summ	BART	Counterfactual Estimation	Avg. Likelihood Diff	✓	✗
	FactGraph	FactCollect	CNN/DM, XSum	Dependency	Summ	ELECTRA	Classification	BACC, F1	✓	✗
	FactKB	FactKB	CNN/DM, XSum	Triplet	Summ	RoBERTa	Classification	BACC, F1	✓	✗
	ExtEval	ExtEval	CNN/DM	Discourse, Coreference, Sentiment	Summ	SpanBERT, RoBERTa	Direct Prediction, Statistic	Summation of Sub-scores	✓	✗
	Q^2	Q^2	WOW	Sent Span	Diag	T5 (QG), Albert-Xlarge (QA), RoBERTa (NLI)	QG-QA, NLI	Likelihood	✗	✓
	FactPush	/	TRUE	Span	Diag, Summ, Paraphrase	DeBERTa	NLI	AUC	✓	✗
	AlignScore	/	22 datasets from 7 tasks	Sent	NLI, QA, Paraphrase, Fact Verification, IR, Semantic Similarity, Summ	RobERTa	3-way Classification	Likelihood	✓	✗
	WeCheck	/	TRUE	Response	Summ, Diag, Para, Fact Check	DeBERTaV3	Weakly Supervised NLI	Likelihood	✓	✗

Table 1: AHE Meta-Info Table before LLM era, which means the methods do not rely on the ability of LLMs such as ChatGPT.

Era	Name	New Dataset	Data Source	Fact Definition	Task	Based-model	Method	Metric	SF	WF
After LLM Era	SCALE	ScreenEval	LLM, Human	Sentence	Long Diag	Flan-T5	NLI	Likelihood	✓	✗
	Chen et al. (2023b)	/	SummEval, XSumFaith, Goyal21, CLIFF	Response	Summ	Flan-T5, code-davinci-002, text-davinci-003, ChatGPT, GPT-4	Vanilla/COT/ Sent-by-Sent Prompt	Balanced Acc	✓	✗
	GPTScore	/	37 datasets from 4 tasks	Various	Summ, Diag, NMT, D2T	GPT-2, OPT, FLAN, GPT-3	Direct Assessment	Direct Score	✓	✗
	G-Eval	/	SummEval, Topical-Chat, QAGS	Response	Summ, Diag	GPT-4	COT, Form-filling	Weighted Scores	✓	✗
	Wang et al. (2023b)	/	5 datasets from 3 tasks	Response	Summ, D2T, Story Gen	ChatGPT	Direct Assessment, Rating	Direct score	✓	✗
	ChainPoll	RealHall-closed, RealHall-open	COVID-QA, DROP, Open Ass prompts, TriviaQA	Response	Hallu Detect	gpt-3.5-turbo	Direct Assessment (2-class)	Acc	✓	✗
	EigenScore	/	CoQA, SQuAD, TriviaQA Natural Questions	Inner State	Open-book QA Closed-book QA	LLaMA, OPT	Semantic Consistency/ Diversity in Dense Embedding Space	AUROC, PCC	✓	✗
	TruthfulQA	TruthfulQA	LLM, Human	Response	Multi-Choice QA, Generation	GPT-3-175B	Answer Match	Percentage, Likelihood	✗	✓
	HaluEval	Task-specific, General	Alpaca, Task datasets ChatGPT	Response	QA, Summ, Knowledge-grounded Diag, Generation	ChatGPT	Direct Assessment	Acc	✓	✓
	FACTOR	Wiki-/News-/ Expert-FACTOR	Wikipedia, Refin-edWeb, ExpertQA	Sent Span	Generation	/	FRANK Error Classification	likelihood	✗	✓
	FELM	FELM	TruthfulQA, Quora, MMLU, GSM8K, ChatGPT, Human	Text Span, Claim	World Knowledge, Sci and Tech, Math, Writing and Recom-mendation, Reasoning	Vicuna, ChatGPT, GPT4	Direct Assessment	F1, Balanced Acc	✓	✓
	FreshQA	Never/Slow Fast-changing, false-premise	Human	Response	Generation	/	Answer Match	Acc	✗	✓
	RealTimeQA	RealTimeQA	CNN, THE WEEK, USA Today	Response	Multi-Choice QA, Generation	GPT-3, T5	Answer Match	Acc, EM, F1	✗	✓
	ERBench	ERBench Database	5 datasets from Kaggle	Ent-Rel	Binary/ Multiple -choice QA	/	Direct Assessment, String Matching	Ans/Rat/ Ans-Rat Acc, Hallu Rate	✗	✓
	FactScore	/	Biographies in Wikipedia	Atomic Fact	Generation	InstructGPT, ChatGPT, PerplexityAI	Binary Classification	P	✗	✓
	BAMBOO	SenHallu, AbsHallu	10 datasets from 5 tasks	Response	Multi-choice tasks, Select tasks	ChatGPT	Answer Match	P, R, F1	✓	✗
	MedHalt	MedHalt	MedMCQA, Medqa USMILE, Medqa (Taiwan), Headqa, PubMed	Response	Reasoning Hallu Test, Memory Hallu Test	ChatGPT	Answer Match	Pointwise Score, Acc	✗	✓
	ChineseFactEval	ChineseFactEval	/	Response	Generation	/	FacTool, Human annotator	Direct Score	✗	✓
	UHGEval	UHGEval	Chinese News Websites	Keywords	Generative/ Discriminative/ Selective Evaluator	GPT-4	Answer Match, Similarity	Acc, Similarity Score	✗	✓
	HalluQA	HalluQA	Human	Response	Generation	GLM-130B, ChatGPT, GPT-4	Direct Assessment	Non-hallu Rate	✗	✓
	FacTool	/	RoSE, FactPrompts, HumanEval, GSM-Hard, Self-instruct	Claim, Response	Knowledge-based QA, Code Generation, Math Reasoning, Sci-literature Review	ChatGPT	Claim Extraction, Query Generation, Tool Querying, Evidence Collection, Agreement Verification	P, R, F1	✓	✓
	UFO	/	NQ, HotpotQA, TruthfulQA, CNN/DM, Multi-News, MS MARCO	Ent	Open-domain/ Web Retrieval-based/ Expert-validated/ Retrieval-Augmented QA, News Fact Generation	ChatGPT (gpt-3.5-turbo-1106)	Fact Unit Extraction, Fact Source Verification, Fact Consistency Discrimination	Avg. Sub-scores	✓	✓
	CONNER	/	NQ, WoW	Sentence	Open-domain QA, Knowledge-grounded Dialogue	NLI-RoBERTa -large, ColBERTv2	3-way NLI	Acc	✗	✓
	SelfCheckGPT	SelfCheckGPT	WikiBio	Response	Hallu Detect	GPT-3	NLI, Ngram, QA, BERTScore, Prompt	AUC-PR	✓	✗
	InterrogateLLM	/	The Movies Dataset, GCI The Book Dataset (Kaggle)	Response	Hallu Detect	GPT-3, LLaMA-2	Query Consistency	AUC, Balanced Acc	✗	✓
	SAC^3	/	HotpotQA, NQ-open	Response	QA Generation	gpt-3.5-turbo, Falcon-7b-instruct, Guanaco-33b	Cross-checking, QA Pair Consistency	AUROC	✓	✓
	KoLA	KoLA	Wikipedia, Updated News and Novels	Response	Knowledge Memorization /Understanding/Applying /Creating	/	Self-contrast Answer Match	Similarity	✗	✓
	RV	PHD	Human Annotator	Ent	Generation	ChatGPT	Construct Query, Access Databases, Entity-Answer Match	P, R, F1	✓	✗
	SummEdits	SummEdits	9 datasets from Summ task	Span	Summ, Reasoning	gpt-3.5-turbo	Seed summary verify, Summary edits, Annotation	Balanced Acc	✓	✗
	LLM-Check	/	FAVA-Annotation, RAGTruth, SelfcheckGPT	Response	Fact-checking	Llama-2, Llama-3, GPT4. Mistral-7b	Analyze internal attention kernel maps, hidden activations and output prediction probabilities	AUROC, FPR, Acc	✗	✓
	PHR	synthetic	/	Response	ICL	Llama-2, Gemma-2	Posterior Hallucination Rate (Baysian)	Hallu Rate	✓	✗
	HalluMeasure	TechNewsSumm	CNN/DM, SummEval	claim	Summ	Claude	COT, Reasoning	P, R, F1	✓	✗
	EGH	/	HADES, HaluEval, SelfcheckGPT	Response	QA, Diag Summ	LLaMa2, OPT, GPT-based	Taylor expansion on embedding difference	Acc, P, R, F1, AUC, G-Mean, BSS	✓	✓
	STARE	/	LfaN-Hall, HalOmi	Sentence	NMT	COMET-QE, LASER, XNLI and LaBSE	Aggregate hallucination scores	AUROC, FPR	✓	✗
	HaluAgent	/	HaluEval-QA, WebQA, Ape210K, HumanEval, WordCnt,	Response, Sent	knowledge-based QA, math, code generation, and conditional text generation.	Baichuan2-Chat, GPT-4	Sentence Segmentation, Tool Selection and Verification, Reflection	Acc, P, R, F1	✓	✓
	RefChecker	KnowHalBench	Natural Questions, MS MARCO, databricks -dolly15k	claim-triplet	Closed-Book QA, RAG, Summ, Closed QA Information Extraction	Mistral-7B, GPT-4, NLI	Extractor and Checker	Acc, P, R, F1	✓	✓
	Lookback Lens	/	CNN/DM, XSum, Natural Questions, MT-Bench	Response	Summ, QA, Multi-turn conversation	LLaMA-2-7B-Chat, GPT-based	Attention Map	AUROC, EM	✓	✓
	Halu-J	ME-FEVER	FEVER	Claim	Fact-checking	GPT-4, Mistral-7B-Instruct	Reasoning	Acc	✗	✓

Table 2: AHE Meta-Info Table after LLM era, which means the methods utilize the ability of LLMs such as ChatGPT.

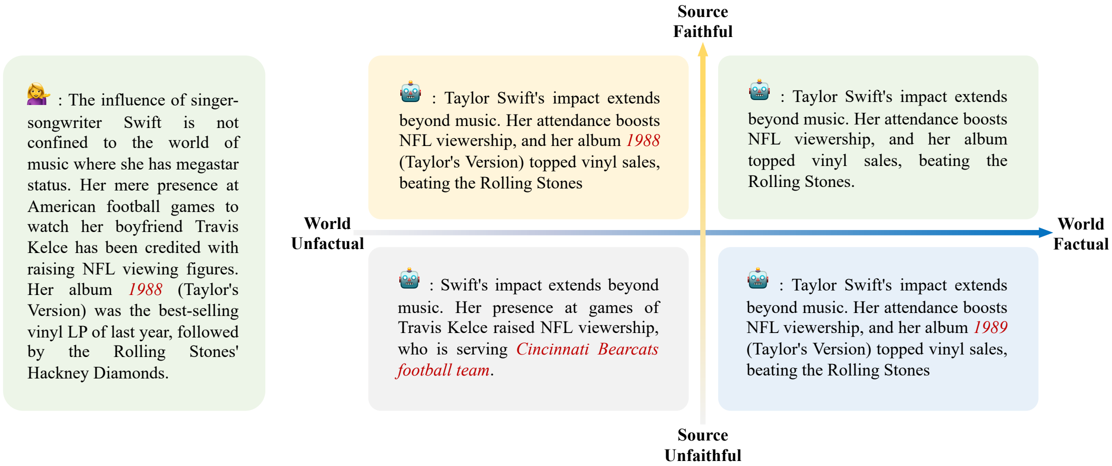

Figure 3: Source Faithful Error (SFE) and World Factual Error (WFE) examples. The correct album is *"1989"*, but the source document contains incorrect information. If the generated text says *"1988"*, it is SF but has WFE. If it corrects to *"1989"*, it is WF but has SFE. When the text exhibits both SFE and WFE, it often includes non-factual content not from the source, e.g. the incorrect statements about *Travis Kelce not serving the Cincinnati Bearcats football team*. Otherwise, if no such errors are present, the text should be both SF and WF.

	Document	**Summary**	**Note**	**WeCheck**	**SelfCheckGPT**	**HaluEval**	**FacTool**
SF-WF	... Harry Kane has been given the nod by Youssouf Mulumbu for this season's players' Player of the Year award The West Brom midfielder has picked Chelsea wideman Eden Hazard for the young player of the year prize Congo international Mulumbu posted his votes for this year's PFA awards to Twitter on Wednesday Mulumbu challenges QPR defender Yun Suk-Young during West Brom's 4-1 defeat at The Hawthorns Goalkeepe ...	The DR Congo international has picked Chelsea wideman Eden Hazard for the young player of the year prize .	The summary is correct.	TRUE	TRUE	TRUE	FALSE
SF-WFE	... Since the end of March, the Vikings' only wins have been in the Challenge Cup against lower-league sides. "We've got the personnel and we've got the people to spark us back into life," **Chris Betts** told BBC Radio Merseyside. "When we get rolling again I'm sure, or I'm positive, that we can really turn this year around for ourselves." ... "The players are hurting and we've got to win," added England assistant coach Betts. ...	Widnes Vikings can turn their poor start to the Super League season around if they can find a winning streak, says assistant coach **Chris Betts**.	"Chris Betts" is in the document but is incorrect essentially.	FALSE	TRUE	TRUE	FALSE
SFE-WF	The panther chameleon was found on Monday by a dog walker in the wooded area at **Marl Park**. It had to be put down after X-rays showed all of its legs were broken and it had a deformed spine. RSPCA Cymru said it was an "extremely sad example of an abandoned and neglected exotic pet".	A chameleon has been put down by RSPCA Cymru after it was found injured and abandoned in a **Cardiff park**.	The Marl Park is in Cardiff but not mentioned in the document.	TRUE	FALSE	TRUE	TRUE
SFE-WFE	A number of men, two of them believed to have been carrying guns, forced their way into the property at Oakfield Drive shortly after 20:00 GMT on Saturday. They demanded money before assaulting a man aged in his 50s. ... Alliance East Antrim MLA Stewart Dickson has condemned the attack. ...	A man has been assaulted by a gang of armed men during a robbery at a house in Ballymena, County Antrim.	"**Ballymena**" is neither in the document nor correct according to external knowledge.	FALSE	TRUE	TRUE	FALSE

Table 3: Examples of the results from selected evaluators on the SFE and WFE. "TRUE" means the evaluator labeled it as correct while "FALSE" means incorrect.