A Survey of Automatic Hallucination Evaluation on Natural Language Generation

Anonymous authors
Paper under double-blind review

Abstract

The proliferation of Large Language Models (LLMs) has introduced a critical challenge: accurate hallucination evaluation that ensures model reliability. While Automatic Hallucination Evaluation (AHE) has emerged as essential, the field suffers from methodological fragmentation, hindering both theoretical understanding and practical advancement. This survey addresses this critical gap through a comprehensive analysis of 74 evaluation methods, revealing that 74% specifically target LLMs, a paradigm shift that demands new evaluation frameworks. We formulate a unified evaluation pipeline encompassing datasets and benchmarks, evidence collection strategies, and comparison mechanisms, systematically documenting the evolution from pre-LLM to post-LLM methodologies. Beyond taxonomical organization, we identify fundamental limitations in current approaches and their implications for real-world deployment. To guide future research, we delineate key challenges and propose strategic directions, including enhanced interpretability mechanisms and integration of application-specific evaluation criteria, ultimately providing a roadmap for developing more robust and practical hallucination evaluation systems.

1 Introduction

Hallucination in Natural Language Generation (NLG) typically refers to situations where generated text contradicts or lacks support from source input or external knowledge. Like an elephant in the room, this phenomenon has persisted since NLG's inception but was largely overlooked in early developments (van Deemter, 2024; Ji et al., 2023; Gatt & Krahmer, 2018). As text generation models evolved, technologies like Large Language Models (LLMs) achieved grammatical correctness and fluency nearly indistinguishable from human writing (Dou et al., 2022; Brown et al., 2020). Consequently, hallucination has emerged as a prominent concern demanding urgent attention. Automatic hallucination evaluation proves crucial for advancing LLMs toward greater reliability and safety. This paper presents a comprehensive survey of Automatic Hallucination Evaluation (AHE) methods, documenting current advances in hallucination detection while identifying future research directions.

The concept of hallucination initially described grammatically correct but semantically inaccurate content relative to source input (Lee et al., 2018). This phenomenon appeared commonly in tasks like Summarization (Maynez et al., 2020) and Neural Machine Translation (NMT) (Raunak et al., 2021), where source information remained well-defined. The paradigm shifted dramatically with LLMs like ChatGPT (OpenAI, 2022). Many NLG tasks became achievable through prompting LLMs with designed instructions (Ouyang et al., 2022). However, their responses occasionally contain hallucinations deviating from input or established world knowledge (Jesson et al., 2024), presenting significant evaluation challenges.

Faithfulness and factuality represent two closely related yet distinct concepts for describing hallucinations. Faithfulness measures output consistency with given source input, while factuality assesses alignment with established real-world knowledge. Despite frequent usage, these terms often become conflated (Huang et al., 2023a; Dong et al., 2022; Xie et al., 2021), creating evaluation ambiguity. This paper provides clearer distinctions by introducing precise terminology: Source Faithfulness (SF) and World Factuality (WF). SF measures how accurately generated output reflects source input consistency. SF operates within limited

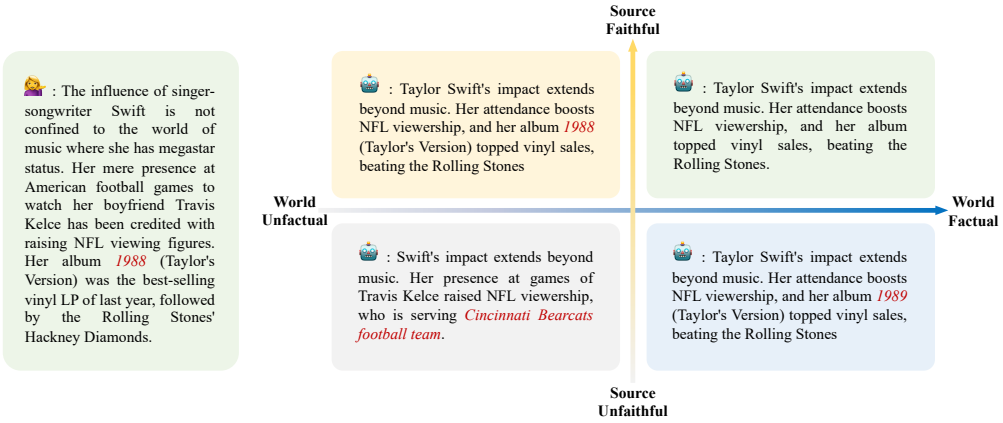

Figure 1: Source Faithful Error (SFE) and World Factual Error (WFE) examples. The correct album is *"1989"*, but the source document contains incorrect information. If the generated text says *"1988"*, it is SF but has WFE. If it corrects to *"1989"*, it is WF but has SFE. When the text exhibits both 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.

scope, as specific sources can substantiate generated text. WF assesses whether the generated output aligns with general world knowledge and facts. WF presents more expansive challenges, extending beyond specific sources to consider broader common sense and established knowledge, which proves difficult to collect and encode comprehensively (Gupta et al., 2024; Garrido et al., 2024). Recent studies increasingly recognize the critical importance of measuring SF and WF in generated text.

Evaluating SF versus WF aspects requires different source information, closely tied to specific tasks. In NMT, translations detached from source text are deemed unfaithful (Dale et al., 2023a). In summarization, summaries should maintain source document faithfulness, though some hallucinations may align with external facts (Dong et al., 2022). In tasks involving LLMs, hallucinations exhibit greater diversity, often encompassing both SF and WF issues simultaneously. LLMs face unique challenges, including outdated world information and false-premise questions (Kasai et al., 2023; Yuan et al., 2024). Figure 1 illustrates these error types through a four-quadrant framework. In light of these task-specific differences and evolving error types, we define the scope of this survey to clearly delimit the boundaries of our analysis and maintain conceptual clarity.

1.1 Scope of the Survey

This survey systematically organizes AHE methods across three core dimensions: dataset construction, evidence collection, and comparison mechanisms. Our goal is to provide a comprehensive account of how hallucination has been assessed across different eras of models, specifically contrasting the pre-LLM era, marked by smaller, task-specific systems, with the post-LLM era, characterized by powerful, instruction-tuned models with broader generative capabilities and increased unpredictability. We analyze and compare methods from both periods through a central framework that distinguishes between the SF and WF perspectives, which shape the definition, detection, and measurement of hallucination across evaluation techniques. We exclude hallucination mitigation techniques, purely human evaluation methods, and multimodal systems to maintain focused scope on text-only automated evaluation.

1.2 Compare with Existing Surveys

Several surveys have touched upon methods for evaluating hallucinations in large language models (LLMs), though often only briefly or without detailed analysis (Huang et al., 2023b; Zhang et al., 2023c; Ji et al., 2023; Huang et al., 2021). These surveys primarily focus on either pre-LLM or early-stage LLM techniques

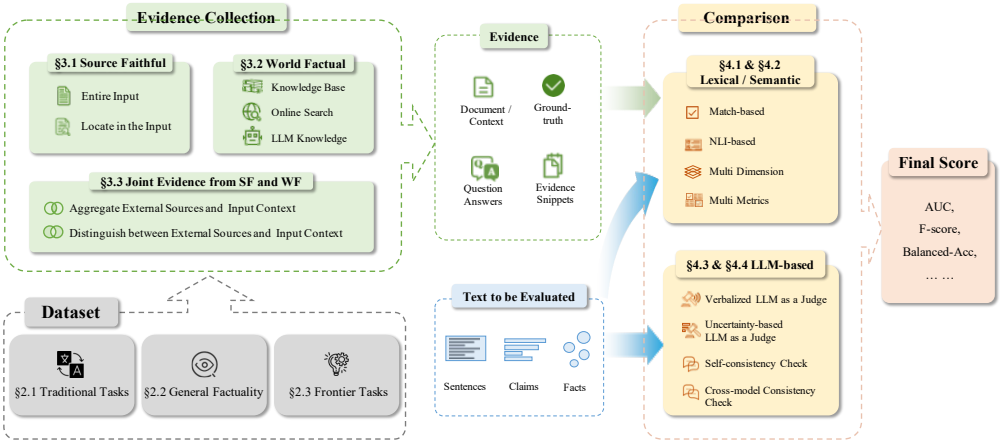

Figure 2: Automatic Hallucination Evaluation (AHE) methods typically follow a pipeline that includes dataset construction, evidence collection, and comparison between the generated output and reference evidence, resulting in a final score that reflects the level of hallucination.

and do not cover more recent developments in the field. Consequently, they do not provide a comprehensive categorization of benchmarks, nor do they systematically summarized evaluator processes. Furthermore, they lack a comparative analysis of methods across different stages, leading to an absence of in-depth analysis regarding their details, strengths, and weaknesses. In contrast, our survey presents a unified and up-to-date overview of AHE methodologies, structured around a standardized evaluation pipeline, as illustrated in Figure 2 and Figure 3.

1.3 Structure of the Survey

This survey examines three sequential components of AHE research. We begin with datasets and benchmarks as the foundational basis (§2), focusing on enhancing data availability and diversity across tasks and evaluation perspectives. Next, we discuss evidence collection for identifying SF and WF evidence (§3). While not mandatory, this step proves crucial for approaches comparing outputs against input-based and external evidence. Some methods bypass explicit evidence collection, evaluating hallucinations directly from internal states or output logits. Subsequently, we examine comparison and judgment mechanisms that utilize collected evidence or analyze implicit model representations to produce quantitative results (§4). This structured framework provides a coherent understanding of diverse approaches and their evolution from pre-LLM to post-LLM eras, though not all methods incorporate every pipeline stage. We also present Table 3, Table 4, and Table 5 for all the methods surveyed in this paper, including key aspects discussed in the following sections. Finally, following the pipeline, this survey 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, 51.4% 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

Although common task-specific datasets are not originally curated with hallucination detection in mind, they often contain instances of hallucinated content as a byproduct of the task, making them valuable resources for hallucination evaluation. In particular, the summarization task has seen substantial efforts in this regard,

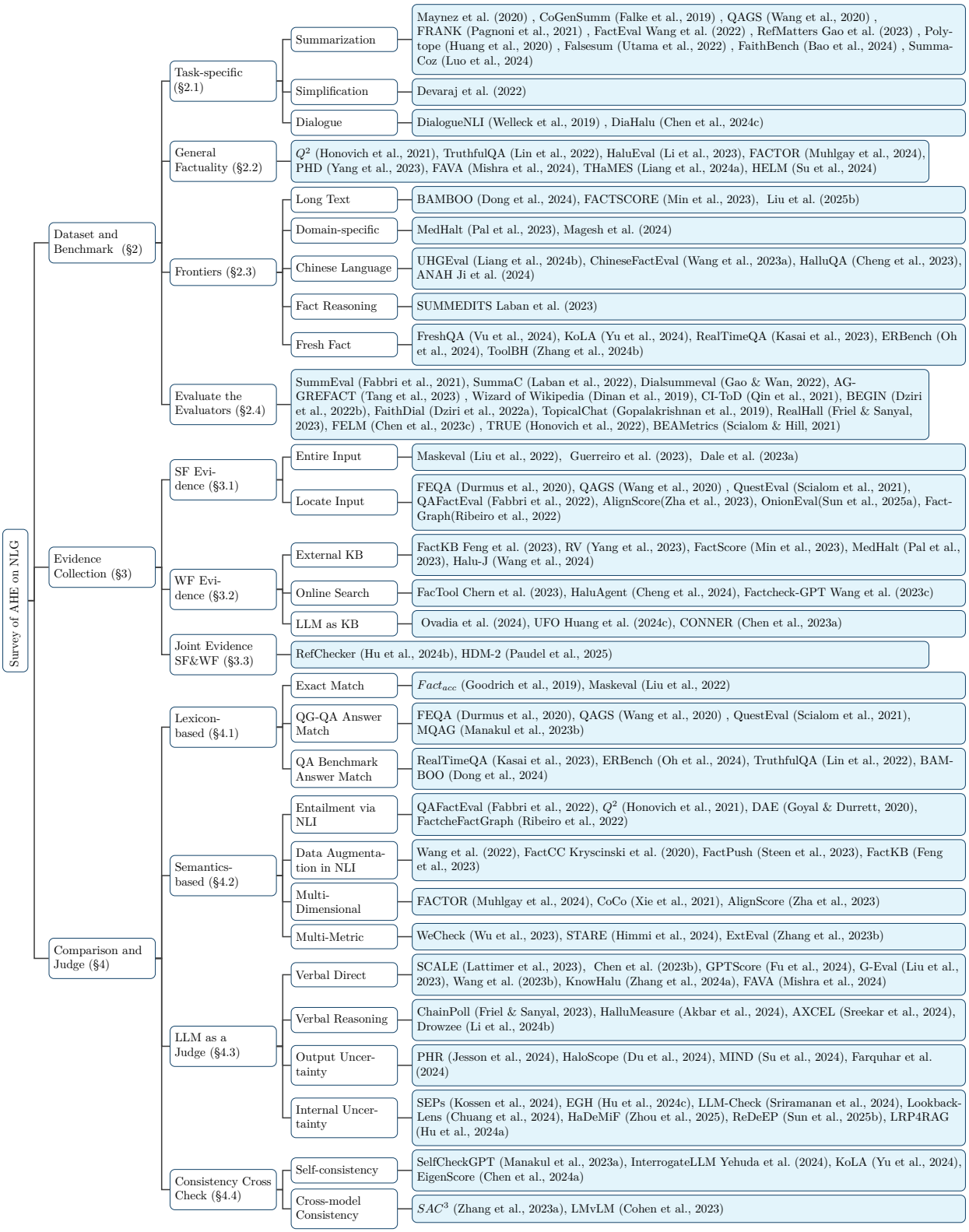

Figure 3: Taxonomy of AHE methods (highlighted in blue nodes) based on the distinct techniques employed at each stage of the pipeline.

where many studies have manually assessed model-generated summaries and released annotated datasets to facilitate research. For example, on widely-used news summarization datasets such as XSum and CNN/DM, Maynez et al. (2020) introduce XSumFaith, which includes fine-grained span-level annotations of hallucination types, distinguishing between intrinsic and extrinsic hallucinations. Similarly, CoGenSumm (Falke et al., 2019) provides human annotations on the CNN/DM dataset and shows out-of-the-box Natural Language Inference (NLI) models do not perform well on correctness evaluation. Across both datasets, QAGS (Wang et al., 2020) annotates each sentence with a binary label of SF. Additionally, Polytope (Huang et al., 2020) contributes annotations for both SF and WF, enabling evaluation across both extractive and abstractive summarization approaches.

However, binary classification of text as whether hallucinated often lacks granularity and fails to capture the nuanced nature of hallucinations. To address this, FRANK (Pagnoni et al., 2021) introduces a more fine-grained typology of factual errors and collects annotations accordingly. In the domain of dialogue summarization, FactEval (Wang et al., 2022) explicitly incorporates hallucination errors during the annotation process, while RefMatters (Gao et al., 2023) further refines error categorization by integrating both content-based and form-based factual inconsistencies, and TofuEval (Tang et al., 2024b) analyzes on multiple LLMs on different dialogue topics. In addition, FaithBench (Bao et al., 2024) isolates challenging summaries identified by state-of-the-art hallucination detection models. These summaries are drawn from 10 modern LLMs spanning 8 different model families, and are annotated with detailed hallucination types. SummaCoz (Luo et al., 2024) takes this a step further by providing annotations not only on hallucination categories but also on the underlying explanations. These explanations are derived through a combination of LLM-generated and human-curated insights, enabling a deeper understanding of the mechanisms behind hallucinations.

Beyond summarization, Devaraj et al. (2022) propose a taxonomy of factual errors, namely, information insertion, deletion, and substitution, in the context of the text simplification task, using data from the Newsela (Xu et al., 2015) and Wikilarge (Zhang & Lapata, 2017) datasets. In the domain of dialogue generation, factual consistency has also received growing attention. DialogueNLI (Welleck et al., 2019) provides sentence-level entailment labels to assess the logical consistency between utterances. Going beyond sentence-level evaluation, DiaHalu (Chen et al., 2024c) introduces a comprehensive benchmark at the dialogue level, incorporating both SF and WF annotations. Expanding to other generation settings, RAGTruth (Niu et al., 2024) addresses hallucination in Retrieval-Augmented Generation (RAG) systems. It offers fine-grained annotations that distinguish between evident and subtle hallucinations, thereby supporting more robust and nuanced evaluation in retrieval-based contexts.

In addition to annotating existing model-generated outputs, data augmentation serves as a complementary strategy for enriching datasets. A growing body of work focuses on automatically generating diverse, controllable hallucinated data aligned with specific hallucination typologies, in order to support model training and benchmarking. For instance, Falsesum (Utama et al., 2022) introduces an automated augmentation pipeline capable of controlling the insertion of intrinsic and extrinsic errors in summaries. Similarly, MFMA (Lee et al., 2022) generates hallucinations by masking key information in the reference, while NonFactS (Soleimani et al., 2023) produces non-factual summaries through random word seeding.

Task-specific annotation and augmentation methods are progressively evolving toward detailing granularity, automation, and scalability. As LLMs continue to advance, the boundaries between tasks are becoming increasingly blurred, indicating that future data development efforts should aim to support more general and cross-domain applications.

2.2 General Factuality

Moving forward from task-specific datasets, recent studies have been increasingly toward more generalized evaluation protocols designed to assess LLMs overall capacity to avoid hallucinations across a broader range of scenarios. These evaluations are often implemented through multi-turn Question-Answer (QA) setups, which enable more dynamic and flexible probing of factual consistency, reasoning fidelity, and the model's ability to maintain coherence over extended interactions.

Within knowledge-grounded dialogue, Q^2 (Honovich et al., 2021) gives an annotated dataset of consistency with respect to a given knowledge. FACTOR (Muhlgay et al., 2024) follows the fine-grained 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, HaluEval (Li et al., 2023) verifies hallucinations in ChatGPT, PHD (Yang et al., 2023) focuses on passage-level entity-centric knowledge, and FAVA (Mishra et al., 2024) offers more fine-grained annotations through tagged elements in the model-generated text. THaMES (Liang et al., 2024a), on the other hand, pairs hallucinated answers with correct ones using NLI models as well as hallucination evaluation models to jointly assess SF and WF across a range of texts, including political news articles, academic papers, and Wikipedia entries. In addition to textual outputs, snapshots of each model's internal states are also valuable for analyzing model behavior, where HELM (Su et al., 2024) provides such data to facilitate deeper investigation. 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. Building on this perspective, HalluLens (Bang et al., 2025) further underscores the importance of appropriately refusing to answer when confronted with non-existent instances.

The evaluation of hallucinations in LLMs often emphasizes WF, leading to the widespread use of large-scale common knowledge corpora, such as Wikipedia, as reference sources for constructing evaluation datasets. Broadly speaking, research on the general factuality of LLMs adopts two main data structuring approaches. The first involves human annotation of model-generated outputs, with fine-grained labeling of hallucination types. The second assesses whether LLMs possess specific knowledge based on their accuracy in answering multiple-choice questions. While the former captures more complex and nuanced forms of hallucination, the latter typically focuses on knowledge-centric hallucinations and examines the model's behavior in cases where it chooses not to answer. Overall, efforts to evaluate general factuality in LLMs aim to probe hallucination tendencies in broader, more open-ended, and everyday scenarios, making such evaluations more universally relevant and reflective of real-world use cases.

2.3 Frontiers

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

Long Context/Generation Despite recent advancements that have improved the ability of LLMs to process and generate long-form text, evaluating hallucinations in extended contexts remains a significant challenge. This difficulty arises in part because long-form outputs often contain a complex mixture of factual and hallucinated information, making accurate assessment more nuanced (Liu et al., 2025b). Benchmarks in this area primarily focus on handling complex topics or decomposing texts into fine-grained factual units for more precise evaluation. For instance, BAMBOO (Dong et al., 2024) incorporates hallucination detection as part of its multi-task benchmark for long-context scenarios. Similarly, FactScore (Min et al., 2023) offers long-form biographies sampled from Wikipedia, breaking down the generated content into fine-grained atomic facts, each annotated with a binary factuality label.

Domain-specific Hallucinations in specialized domains such as medicine and law can have serious real-world consequences, making the development of domain-specific evaluation datasets especially critical. To address this need in the medical domain, MedHalt (Pal et al., 2023) introduces a structured evaluation pipeline comprising three fact-based tests: the False Confidence Test (FCT), the None of the Above Test (Nota), and the Fake Questions Test (FQT). These tests are designed to systematically detect and filter hallucinated content generated by models in clinical contexts. Similarly, MedHallu (Pandit et al., 2025) provides introduces synthetic hallucination QA pairs, where hallucinated answers are generated through a controlled pipeline built upon PubMedQA. In the legal domain, Magesh et al. (2024) compile a reference-based QA dataset that covers legal questions across five dimensions: general legal knowledge, jurisdiction-specific text, time-sensitive scenarios, false-premise contexts, and fact-recall tasks. This dataset enables more rigorous evaluation of hallucination in legal response generation.

Non-English Languages Alongside the global trend in LLM development, numerous Chinese LLMs have emerged, with hallucination remaining a critical concern. To address this, several benchmarks have been proposed for evaluating hallucinations in Chinese-language contexts. UHGEval (Liang et al., 2024b) focuses on hallucinations generated by Chinese LLMs in the news domain, while ChineseFactEval (Wang

et al., 2023a) offers a comprehensive benchmark spanning seven real-world application scenarios—including a dedicated section on modern Chinese history, to assess factual consistency in everyday use cases. Drawing inspiration from TruthfulQA (Lin et al., 2022), HalluQA (Cheng et al., 2023) systematically summarizes question types in Chinese, incorporating cultural context to classify hallucinations into imitative falsehoods and factual errors. Additionally, the ANAH benchmark (Ji et al., 2024), which supports both Chinese and English, prompts models to annotate hallucinations at the sentence level, encompassing tasks such as reference retrieval, type classification, and correction. Beyond Chinese, multilingual resources like HalOmi (Dale et al., 2023b) facilitate hallucination evaluation across languages and are designed to disentangle hallucinations from mere translation errors. In summary, while hallucination evaluation shares common challenges across languages, such as distinguishing of knowledge, there are also language-specific aspects influenced by cultural context and the training corpus (Cheng et al., 2023).

Fact Reasoning Reasoning with LLMs in hallucination evaluation presents significant challenges due to the inherently multi-step nature of the reasoning process. To address this, Laban et al. (2023) propose SUMMEDITS, a benchmark that introduces a structured three-step protocol for constructing inconsistency detection datasets. The framework is implemented across ten domains and is designed to assess LLMs' capabilities in factual reasoning and error identification. However, while this approach provides a valuable lens for evaluating factual reasoning, it does not fully capture hallucinations that emerge within the reasoning process itself, hallucinations that can propagate and ultimately affect the final output. Such issues remain underexplored and merit further attention in future research.

Fresh Fact As the world is constantly changing, a critical question arises: how can we assess whether LLMs possess up-to-date, dynamic knowledge? To address this, several benchmarks have been developed that focus on constructing time-sensitive datasets, enabling systematic evaluation of LLMs' ability to reflect and reason over recent information. Some efforts approach this by explicitly defining categories of fresh events, incorporating both current and recently updated content (Vu et al., 2024). These benchmarks are often paired with carefully designed submission workflows (Kasai et al., 2023) that continuously collect emerging events or news (Yu et al., 2024), thereby offering a sustainable mechanism for tracking and evaluating the evolving world knowledge of LLMs. Other approaches integrate external resources, such as structured databases (Oh et al., 2024) or search engines (Zhang et al., 2024b), to support real-time applications. By leveraging these tools, researchers can evaluate LLMs' hallucination tendencies in a more realistic context, examining both the depth and breadth of factual accuracy across a range of domains and task settings. Findings from these investigations indicate that larger model sizes do not necessarily improve factuality. Instead, factors such as the quality of training data and the design of response strategies play critical roles in determining a model's ability to minimize hallucinations.

2.4 Evaluate the Evaluators

Building on the large amount of automatic evaluation metrics, a number of specialized "meta-benchmarks" have emerged to systematically reassess and compare these metrics' abilities to detect hallucinations across different NLG tasks. SummEval (Fabbri et al., 2021), SummaC (Laban et al., 2022), DialSumMeval (Gao & Wan, 2022), and AGGREFACT (Tang et al., 2023) each assemble human-annotated summaries, from newswire to conversational transcripts, and then measure how well a variety of metrics correlate with human judgments of factual consistency and coherence. In parallel, dialogue-oriented datasets like Wizard of Wikipedia (Dinan et al., 2019), CI-ToD (Qin et al., 2021), BEGIN (Dziri et al., 2022b), FaithDial (Dziri et al., 2022a), and TopicalChat (Gopalakrishnan et al., 2019) provide turn-level and multi-turn annotations of whether model responses remain grounded in provided knowledge, enabling direct evaluation of dialogue metrics. Extending beyond single domains, RealHall (Friel & Sanyal, 2023) bridges closed- and open-domain scenarios to benchmark both SF and WF, while FELM (Chen et al., 2023c) further diversifies the evaluation landscape by incorporating scientific explanations, mathematical problem solving, recommendation dialogues, and complex reasoning tasks into its coverage. Finally, generalist frameworks such as TRUE (Honovich et al., 2022) and BEAMetrics (Scialom & Hill, 2021) evaluate metric performance across NLG tasks, ranging from summarization and translation to style transfer and code generation, thereby illuminating each metric's cross-task robustness and highlighting both universal strengths and domain-specific weaknesses.

Category	Dataset	Task	Size	Label Type	Links
Traditional Task	DialogueNLI (Welleck et al., 2019)	Dialogue	343k pairs	Entailment/contradiction/neutral	GitHub
	CoGenSumm (Falke et al., 2019)	Summarization	100 articles	Sentence correct/incorrect	Dataset Link
	XSumFaith (Maynez et al., 2020)	Summarization	500 articles	Span intrinsic/extrinsic hallucination	GitHub
	QAGS (Wang et al., 2020)	Summarization	474 articles	Consistent/inconsistent	GitHub
	Polytope (Huang et al., 2020)	Summarization	1.5k summaries	Intrinsic/extrinsic hallucination	GitHub
	FRANK (Pagnoni et al., 2021)	Summarization	2.25k summaries	Relation/entity/circumstance/coreference/ discourse/out-of-article/gramma errors	GitHub
	Falsesum (Utama et al., 2022)	Summarization	2.97k articles	Consistent/inconsistent	GitHub
	FactEval (Wang et al., 2022)	Dialogue summarization	150 dialogues	Consistent/inconsistent	GitHub
	Devaraj et al. (2022)	Text simplification	1.56k pairs	Insertion/deletion/substitution	GitHub
	NonFactS (Soleimani et al., 2023)	Augmented summarization	400k samples	Non-factual summaries	GitHub
	RefMatters (Gao et al., 2023)	Dialogue summarization	4k pairs	FRANK errors	GitHub
	DiaHalu (Chen et al., 2024c)	Dialogue generation	1.0k samples	Dialogue-level factuality/faithfulness	GitHub
	TofuEval (Tang et al., 2024b)	Dialogue summarization	1.5k pairs	Consistent/inconsistent	GitHub
	RAGTruth (Niu et al., 2024)	RAG systems	2.97k samples	Evident/subtle conflict/baseless	GitHub
	SummaCoz (Luo et al., 2024)	Summarization	6.07k summaries	Explanation	HF Dataset
	FaithBench (Bao et al., 2024)	Summarization	750 samples	Questionable/benign/unwanted	GitHub
General Factuality	Q2 (Honovich et al., 2021)	Knowledge-based dialogue QA	750 samples	Consistent/inconsistent	GitHub
	TruthfulQA (Lin et al., 2022)	Truthfulness QA	817 pairs	QA truthfulness	GitHub
	FACTOR (Muhlgay et al., 2024)	Multi-choice	4.27k samples	FRANK errors	GitHub
	HaluEval (Li et al., 2023)	QA/Summarization/ dialog/general	35K samples	Hallucinations yes/no	GitHub
	PHD (Yang et al., 2023)	Passage-level QA	300 entities	factual/non-factual/unverifiable	GitHub
	FAVA (Mishra et al., 2024)	General queries	200 queries	Entity/relation/contradictory/invented/ subjective errors/unverifiable	Project Page
	THaMES (Liang et al., 2024a)	General QA	2.1k samples	Correct/hallucinated	GitHub
	HELM (Su et al., 2024)	LLM continue generation	1.2k passages	Hallucination/non-hallucination	GitHub
	HalluLens (Bang et al., 2025)	LLM generation	130k instances	Intrinsic/extrinsic hallucination / factuality	GitHub
Frontiers	FactScore (Min et al., 2023)	Long-form biography	6.5k samples	Support/unsupport	GitHub
	BAMBOO (Dong et al., 2024)	Long-context	1.5k samples	SenHallu, AbsHallu	GitHub
	ChineseFactEval (Wang et al., 2023a)	Chinese multi-domain	125 prompts	Factual/non-factual	Project Page
	HalluQA (Cheng et al., 2023)	Chinese QA	450 questions	Misleading/misleading-hard/knowledge	GitHub
	UHGEval (Liang et al., 2024b)	Chinese news	5k samples	Hallucination/non-halluciantion	GitHub
	ANAH (Ji et al., 2024)	Chinese/English LLM generation	4.3k generation	Contradictory/unverifiable/no fact	GitHub
	SUMMEDITS (Laban et al., 2023)	Multi-domain	6.35k samples	Consistency/inconsistency	HF Dataset
	MedHalt (Pal et al., 2023)	Medical tests	25.64k samples	Groundedness/hallucination	Project Page
	MedHallu (Pandit et al., 2025)	Medical QA	10k samples	Hard/medium/easy hallucination	Project Page
	LegalHallu (Magesh et al., 2024)	Legal QA	745k samples	Correctness/groundedness	HF Dataset
	HalOmi (Dale et al., 2023b)	Multilingual translation	18 language (144-197 pairs each)	Hallucination, omission	GitHub
	FreshLLMs (Vu et al., 2024)	Time-sensitive QA	599(June,2025) pairs	Fast/slow/never changing/ false premise	GitHub
	RealtimeQA (Kasai et al., 2023)	Real-time knowledge	4.3k(June,2023) pairs	Correct/retrieval/ reading comprehension error	GitHub
	ERBench (Oh et al., 2024)	Knowledge-based LLM QA	Not specified	Binary/multi-choice	GitHub
	KOLA (Yu et al., 2024)	Knowledge-based LLM generation	2.15k samples	Correct/incorrect	GitHub
	ToolBeHonest (Zhang et al., 2024b)	Tool-augmented LLM	700 samples	missing necessary tools/potential tools/limited functionality tools	GitHub
Evaluate the Evaluators	Wizard of Wikipedia (Dinan et al., 2019)	Knowledge-based dialogue eval	22.3k dialogues	Knowledge selection, response generation	Project Page
	TopicalChat (Gopalakrishnan et al., 2019)	Knowledge-based dialogue eval	10.79k dialogues	Knowledge source	GitHub
	SummEval (Fabbri et al., 2021)	Summarization metric eval	1.6k summaries	Consistent/inconsistent	GitHub
	BEAMetrics (Scialom & Hill, 2021)	Multi-task metric eval	Not specified	Coherence	GitHub
	CI-ToD (Qin et al., 2021)	Task-oriented dialogue	3.19k dialogues	Consistent/inconsistent	GitHub
	SummaC (Laban et al., 2022)	Summarization metric eval	Not specified	Consistent/inconsistent	GitHub
	BEGIN (Dziri et al., 2022b)	Knowledge-based dialogue	12k turns	Fully/not attributable/generic	GitHub
	FaithDial (Dziri et al., 2022a)	Dialogue eval	5.65k ddialogues	BEGIN, VRM	HF Dataset
	DialSumMeval (Gao & Wan, 2022)	Dialogue summarization metric eval	1.5k summaries	Consistent/inconsistent	GitHub
	TRUE (Honovich et al., 2022)	Cross-task metric eval	~200k samples	Consistent/inconsistent	GitHub
	AGGREFACT (Tang et al., 2023)	Summarization metric eval	59.7k samples	Consistent/inconsistent	HF Dataset
	FELM (Chen et al., 2023c)	Multi-task metric eval	847 samples	Factuality positive/negative	GitHub

Table 1: Overview of AHE datasets/benchmarks by time and task category. "HF" indicates "HuggingFace".

2.5 Summary

We provide a comprehensive overview of the datasets in Table 1, including metadata and access links for reference and reproducibility. As a summary, a wide array of datasets and benchmarks across diverse domains has been developed to facilitate more robust evaluation of hallucinations in language models. However, despite the growing volume, many of these datasets are constrained by small sample sizes and a narrow alignment between specific datasets and evaluation approaches, limiting their generalizability and reuse. To address these limitations, future efforts in dataset construction should prioritize the integration of variable sources, the adoption of standardized annotation protocols, and the balancing of both data quality and scale. Additionally, the aforementioned datasets either support the evaluation of generation models, or function as resources for assessing the performance of evaluation metrics themselves. In fact, several datasets have emerged as by-products of evaluation method development. Building on this foundation, the next sections introduce the categorization of AHE methods, which we organize into two core stages based on their typical pipeline: evidence collection and comparison.

3 Evidence Collection

Datasets and benchmarks provide the foundation for AHE. The next step involves comparing generated text with relevant evidence to quantify the degree of hallucinations. This step is grounded in the premise that hallucinations emerge from inconsistencies in SF or WF, depending on whether the discrepancy lies with internal context or external knowledge. While ground-truth references, typically curated or authored by domain experts, serve as gold-standard evidence for detecting hallucinations, their construction is often resource-intensive and difficult to scale. As a result, a more practical alternative is the automated collection of relevant evidence to support the evaluation process. Large-scale automation in evidence gathering is thus pivotal to advancing AHE in real-world applications.

In this section, we focus on evidence collection strategies that do not rely on human-annotated ground truths. For SF evaluation, evidence is typically extracted directly from the input or surrounding context. In contrast, WF evaluation usually draws upon external resources or the model's own latent knowledge to verify the factual consistency of generated content.

3.1 SF Evidence

To assess the faithfulness of the generated content and detect potential hallucinations, the source input can be utilized in two principal ways: either by treating it as a complete reference for evaluation, or by extracting specific pieces of information, referred to as SF Evidence, that support or contradict the generated output to enable more fine-grained verification.

Entire Input as Evidence Utilizing the entire input as evidence implies that the evaluation process does not involve extracting specific sentences or spans. 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). 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. 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, where answers are often generated from the document using extractive QA models, such as BERT (Devlin et al., 2019) fine-tuned on QA datasets. 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 abstractive QA approaches. This suggests that QA capability is not the primary bottleneck in the task. For answer selection granularity, Fabbri et al. (2022) demonstrate that selecting noun phrase chunks as answers yields better performance than entities. While evidence is often represented as individual words or short spans in the above methods, more comprehensive approaches have been proposed. These include segmenting the context into discrete segments (Zha et al., 2023), constructing hierarchical layers of context with progressively richer information (Sun et al., 2025a), and representing the core content of the source input using semantic graphs (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, effectively treating the KBs as extended source inputs, making this approach closely resemble SF evaluation. 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). When multiple pieces of evidence are available, identifying the most relevant ones becomes a critical step prior to making factuality judgments (Wang et al., 2024). Hallucination evaluation in specific domains can benefit from domain-specific KBs. For instance, domain-wide KBs such as PubMed play a crucial role in biomedical information retrieval (Pal et al., 2023). In the legal domain, Magesh et al. (2024) provide valuable evaluation references for constructing law-related benchmarks.

LLM as KB LLMs have massive learned knowledge while training, and powerful LLMs can serve as KBs. In a closed-book setting, the model generates answers solely based on its parametric knowledge, without accessing any external resources. In contrast, in an open-book setting, LLMs can be further enhanced through fine-tuning or by incorporating retrieved information at inference time (Ovadia et al., 2024; Chen et al., 2024b). Several approaches leverage the parametric knowledge of LLMs directly for hallucination evaluation. UFO (Huang et al., 2024c) introduces a fact verification framework that integrates multiple sources of evidence, including knowledge internal to LLMs. Similarly, CONNER (Chen et al., 2023a) uses LLMs to generate relevant knowledge as supplementary evidence for evaluation purposes. These methods are particularly well-suited for knowledge-intensive tasks, such as open-domain question answering and knowledge-grounded dialogue, where both the breadth and depth of hallucination understanding are critical.

Online Search While static KBs and LLMs parametric knowledge offer a solid foundation of common facts, they cannot keep pace with rapidly changing events or breaking news. To bridge this gap, many recent approaches integrate online search engines as dynamic, real-time sources of evidence. FacTool (Chern et al., 2023) first decomposes the input text into independent atomic claims, then issues targeted web search queries to retrieve evidence that either confirms or refutes each claim. Building on this idea, HaluAgent (Cheng et al., 2024) combines smaller language models with search tool plugins. However, issuing separate searches for every possible claim can become costly and unwieldy. To avoid unnecessary queries, Factcheck-GPT (Wang et al., 2023c) introduces a check-worthiness pre-filtering module. Before any search is sent, this module scores each claim on its potential impact—only high-priority claims are forwarded to the search engine. Together, these methods address three core challenges in dynamic verification: (1) how to transform complex text into precise, searchable queries; (2) how to efficiently prioritize which queries warrant external lookup; and (3) how to integrate and interpret retrieved web evidence back into a coherent judgment. By combining claim decomposition, strategic filtering, and search-model interaction, they ensure that facts are checked against the most current information available.

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

3.3 Joint Evidence for SF and WF

The evidence extraction methods discussed above typically focus on a single aspect, either SF or WF. However, text generated by models is often complex, potentially containing both unfaithful and unfactual content simultaneously. To address this, several approaches have been developed to enable joint evaluation of both SF and WF. For instance, RefChecker (Hu et al., 2024b) distills a smaller model from GPT-4 to perform evidence extraction within input context, external sources, and the LLM's internal memory independently. However, this approach evaluates with a unified framework in different scenarios, rather than truly performing a joint evaluation. In contrast, HDM-2 (Paudel et al., 2025) explicitly distinguishes between context-based and common-knowledge hallucinations within a single text. This categorization more accurately reflects the nature of hallucinations in real-world scenarios and offers a more comprehensive evaluation.

3.4 Summary

The effectiveness of evidence derived from fixed sources, such as SF evidence and those based on static KBs, is largely determined on the accuracy of the extraction process. When relying on LLMs for evidence retrieval, there is a risk of circular verification, where the model may "lie to verify a lie," given that LLMs themselves are prone to hallucinations. Online search offers broader coverage and access to up-to-date information, but the multi-step retrieval pipeline can introduce information loss, and the overall effectiveness often hinges on the quality of search queries, retrieval results, and subsequent interpretation. When both SF and WF evidence are considered, challenges arise not only in ensuring sufficient coverage for both levels but also in resolving potential conflicts between the two. Ultimately, how the retrieved evidence is leveraged, particularly in how it is aligned and compared with the generated text, plays a critical role in determining the evaluation outcome.

4 Comparison and Judge

The final and critical step is to compare the generated text against the corresponding ground truths or the collected evidence. To achieve more precise and reliable evaluation, a wide array of techniques has been proposed, addressing the problem from different angles. Some approaches directly employ external evidence to compute similarity or entailment scores. Others operate without explicit evidence, instead drawing on the model's internal knowledge and applying self-supervised or learned scoring functions. In this section, we organize these comparison strategies into well-defined categories and provide an overview of the principal scoring metrics along with representative methods that utilize them.

4.1 Lexicon-based Metrics

Lexicon-based metrics usually refer to the measurement of the closeness or similarity between two pieces of text based on their word usage. Traditional n-gram-based methods, such as ROUGE (Lin, 2004), assess the overlap of n-grams between the texts. However, these approaches have demonstrated weak correlation with human evaluation (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 triplets (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 comparison 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., 2023b) 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 § 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., 2023; Oh et al., 2024). Another line of research involves using multiple-choice tasks (Lin et al., 2022; Kasai et al., 2023; Oh et al., 2024; Dong et al., 2024), where accuracy or F-score is computed as the final performance metrics.

4.2 Semantics-based Metrics

The approaches presented in this section diverge from the lexicon-based metrics, 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 models to classify upon the degree of semantic similarity.

Entailment Evaluation via NLI A common strategy for semantics-based metrics involves evaluating the degree of entailment using a NLI model, wherein the predicted likelihood is utilized as a measure of the entailment score. 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 highlight that QA-based and NLI-based metrics can provide complementary insights. With more focus on the encoding processes, some studies leverage sentence or document structure to construct semantic representations. For example, DAE (Goyal & 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. In this framework, an entailment indicates that the relationship between a dependency arc's head and tail is supported by the ground-truth sentence, and vice versa. Expanding on this, FactGraph (Ribeiro et al., 2022) improves discourse understanding by encoding semantic structures as graphs for both the input and output. These graph representations are concatenated with the corresponding textual embeddings and fed into a classifier. This graph-based approach facilitates a more nuanced analysis of semantic relationships, aiding in capturing SF consistency.

Data Augmentation in NLI To train an NLI model that is well-suited for a specific task, data augmentation has become a widely adopted strategy to enhance model performance. Recent studies have focused on constructing both positive and negative examples to improve the model's ability to discriminate between entailment and non-entailment. Positive instances are often generated by paraphrasing or back-translation, thereby preserving meaning while varying surface form (Kryscinski et al., 2020; Wang et al., 2022). Negative samples, by contrast, may be created through word swapping and noise injection as in FactCC and FactCCX (Kryscinski et al., 2020), by appending random or misleading phrases as in FactPush (Steen et al., 2023), or by integrating external KB triples to enrich factual context as in FactKB (Feng et al., 2023). With these balanced datasets, supervised classification or contrastive learning objectives can be employed to train the NLI model to reliably distinguish between entailed and non-entailed text pairs.

Multi-Dimensional Evaluation The aforementioned NLI methods focus on evaluating entailment within a binary classification framework. However, hallucinations can be assessed from a broader range of perspectives, allowing for more nuanced evaluation. FACTOR (Muhlgay et al., 2024) follows the error types from FRANK (Pagnoni et al., 2021) and performs the multi-choice factual 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 an LM and 3 individual linear layers as the 3-way classification (aligned, contradict, neutral), binary classification (aligned, not-aligned), and regression heads.

Multi-Metric Evaluation 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., 2024a). Such methods can be further categorized into verbalized judge and judge with uncertainty, depending on whether the judgment is based on verbalized generation outputs or derived from internal model states, such as logits, attention maps, or layer-wise representations.

Verbalized Direct Judge The evaluation process usually 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., 2024) and G-Eval (Liu et al., 2023) both provide multi-facet evaluation frameworks in which consistency serves as a core metric. KnowHalu (Zhang et al., 2024a) further enables comparison by integrating both structured and unstructured forms of knowledge. Different from methods that directly predict a hallucination category, FAVA (Mishra et al., 2024) trains an LLM to explicitly tag hallucinated segments within its generated output.

Verbalized Judge with Reasoning Chain-of-thought (CoT) prompting also can enable the reasoning capabilities of LLMs (Liu et al., 2023; Friel & Sanyal, 2023; Akbar et al., 2024), as it improves the handling of complex and nuanced judgments by explicitly outlining the intermediate reasoning steps, thereby increasing transparency. Reasoning not only facilitates more interpretable evaluation processes but also provides justifications for the model's decisions. For instance, AXCEL (Sreekar et al., 2024) offers explanations for consistency scores by presenting detailed reasoning traces and highlighting the specific text spans that exhibit inconsistency. Similarly, Drowzee (Li et al., 2024b) detects fact-conflicting hallucinations by applying logic-reasoning-based data mutation through five custom-designed rules and deploying two semantic-aware oracles to automatically assess the reasoning consistency of LLM-generated answers.

Judge with Output Uncertainty In addition to examining the generated text, the semantic information embedded in output representations can provide valuable signals for hallucination detection. PHR (Jesson et al., 2024) estimates hallucination rates by assessing the log probabilities of responses from conditional generative models. Leveraging unlabeled data, HaloScope (Du et al., 2024) clusters outputs based on their representations and flags outliers as potential hallucinations. Similarly, MIND (Su et al., 2024) employs an unsupervised strategy to train a classifier using the embeddings of generated outputs. Furthermore, the semantic entropy of these representations can serve as a measure of generative uncertainty. For instance, Farquhar et al. (2024) quantify uncertainty at the semantic level, focusing on meaning rather than surface-level lexical variation.

Judge with Internal Uncertainty Beyond output uncertainty, the latent representations within LLMs deserve deeper investigation. SEPs (Kossen et al., 2024) introduces linear probes trained on hidden states to capture semantic entropy. EGH (Hu et al., 2024c) models the distributional distance between embeddings and gradients of conditional versus unconditional outputs, using a Taylor expansion framework with contextual input as the generation condition. LLM-Check (Sriramanan et al., 2024) utilizes internal attention kernel maps, hidden activations, and output prediction probabilities to identify hallucinations, while Lookback-Lens (Chuang et al., 2024) focuses on attention maps to detect contextual inconsistencies. To further improve the interpretability of LLMs, HaDeMiF (Zhou et al., 2025) calibrating model predictions using a deep dynamic

decision tree and multilayer perception, whereas BTProp (Hou et al., 2024) implements a hidden Markov tree to model the uncertainty in the LLM generation process. In RAG settings, ReDeEP (Sun et al., 2025b) examines the model mechanisms on both external and parametric knowledge by analyzing feed-forward layers and attention weights. LRP4RAG (Hu et al., 2024a) leverages layer-wise relevance propagation to compute input-output relevance in RAG generators, followed by resampling and classification to detect hallucinations.

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., 2023a) 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., 2024) develops a self-contrast metric by contrasting two completions generated by the same model and gets the similarity score. Based on multiple generations, EigenScore (Chen et al., 2024a) leverages eigenvalues of responses' covariance matrix to measure self-consistency.

Cross-model Consistency Check Although self-inconsistency in LLMs is often linked to hallucinations, achieving self-consistency does not necessarily guarantee the correctness of generated content. This limitation arises because self-consistency checking typically focuses solely on model outputs, overlooking the critical role of input information, particularly in tasks like QA, where inputs can be highly informative and even decisive. To address this, SAC^3 (Zhang et al., 2023a) incorporates verifier LMs to perform cross-checking, taking into account both the input questions and the output answers when evaluating semantic consistency. Similarly, LMvLM (Cohen et al., 2023) enables multi-turn interactions between the claim-generating LM and a separate examiner LM, which poses follow-up questions to uncover potential inconsistencies.

4.5 Summary

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

While many challenges have been addressed or mitigated by existing AHE methods, there are still some questions that need to be investigated. In this section, we provide a discussion and analysis of several existing questions, study the relationship between SF and WF evaluation, and offer introductions, comparisons and complementary insights related to fact-checking and human evaluation.

5.1 Questions Concerning AHE

How to distinguish hallucination and text error? According to the conventional definition, hallucinations are typically characterized as fluent yet incorrect outputs. However, under this definition, any model-generated response that diverges from the reference could be labeled a hallucination, an overly broad

interpretation that risks misleading researchers and conflating distinct error types. To address this ambiguity, recent studies in NMT have proposed more refined criteria (Dale et al., 2023a), such as evaluating whether the generated content is detached from the input. Other lines of research focus on knowledge-based hallucinations (Hu et al., 2024b), which involve inconsistencies between the model's internal representations and external factual sources, aiming to identify errors rooted in knowledge misalignment rather than surface-level divergence.

Which fact granularity is the best? The studies reviewed in this work evaluate hallucinations across various granularities, ranging from fine-grained units, such as individual tokens and entities, to more coarse-grained elements, including phrase spans, claims, sentences, and even document-level segments. This naturally raises a fundamental question: What is optimal for factual granularity for evaluation? Prior work has explored this issue by examining various granularity levels (Hu et al., 2024b), or by proposing multi-level approaches that integrate multiple granularities for a more comprehensive assessment (Xie et al., 2021; Chen et al., 2023c). Broadly speaking, different stages in the development of AHE have demonstrated varying preferences for factual granularity. Early methods tended to represent knowledge using entities or entity-relation triplets. With the advent of LLMs, direct judgments of model responses became increasingly common. More recently, finer-grained representations, such as claims or span-level atomic facts, have been adopted to enable more precise fact comparisons. However, identifying a universally optimal level of granularity remains an open challenge, as the most appropriate choice often depends on the specific task and application context.

Is hallucination always bad? Not necessarily. In certain domains, such as legal summarization, hallucinated content can be beneficial when it involves the integration of relevant external knowledge to enhance the informativeness or coherence of summaries (Bendahman et al., 2025). In such scenarios, factual hallucinations are not only tolerated but may be desirable, provided they are controllable and contextually appropriate. Conversely, in more imaginative settings, such as discussions surrounding science fiction novels, creative and speculative content is expected. In these cases, the boundary between hallucination and imagination becomes increasingly ambiguous. Recognizing and distinguishing between these phenomena is crucial for enabling models to appropriately evaluate and generate text across a wide range of use cases (Zhou et al., 2024).

5.2 Comparing SF and WF Evaluation

SF and WF evaluations are two subcategories of AHE. While they assess hallucination from different perspectives, they also share certain points of convergence. For example, some evaluation methods in both SF and WF rely on reference texts for comparison in order to produce a final judgment. Moreover, both types of evaluation can be conducted by analyzing the model's internal states. In this section, we discuss how the evaluation of SF and WF may influence each other when using the same evaluator, by presenting the evaluation results of four SF and WF evaluators across the four quadrants shown in Figure 1.

The cases presented here are based on the summarization data shown in Table 2, which are drawn from the XEnt dataset (Cao et al., 2022) and FactCollect (Ribeiro et al., 2022). We selected four evaluators representing different perspectives, including both GPT-based (SelfCheckGPT, HaluEval, FacTool) and non-GPT-based models (WeCheck), and covering evaluators designed for assessing both SF and WF aspects. SelfCheckGPT uses a zero-shot approach in its prompt to assess the consistency, HaluEval's prompt provides examples for judgment, and FacTool aggregates online search to judge the factuality. For GPT-based models, we specifically used GPT-3.5-turbo. Although FacTool is not originally designed for summarization evaluation, we adapted it to the KBQA (Knowledge-Based Question Answering) setting in order to explore its transferability to this task. All the evaluators only provide binary classification results.

The results of different models on these cases show considerable variation. The SFE cases indicate that the results of SelfCheckGPT and HaluEval remain unstable. For the WFE cases, FacTool provides the correct answers, and surprisingly, WeCheck also made correct judgments. This result aligns with Qi et al. (2025), which suggests that the model's ability in one aspect may subconsciously influence its evaluation in the other. In other words, SF and WF evaluations can affect each other, primarily due to the presence of misaligned information within the model.

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

5.3 Fact-checking and AHE

Fact-checking or fact-verification is a related line of research that has previously received considerable attention. Compared to the more complex nature of AHE, fact-checking addresses relatively simpler problems, primarily focusing on the comparison of facts, which corresponds to the comparison component in section 4 discussed in this paper. It typically involves assessing the factual accuracy of individual claims, with an emphasis 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) to check the correctness of a fact. Benefiting from the capabilities of LLMs, fact-checking systems are now able to handle longer and more complex texts with greater confidence and efficiency (Xie et al., 2025; Wang et al., 2023c). Given the nature of the task, fact-checking can be viewed as a WF evaluator for text generation, typically functioning as a binary (true/false) checker. Moreover, when evidence is extracted from a specific source for verification, the focus shifts from WF to SF, further illustrating the dialectical relationship between the two dimensions (Tang et al., 2024a).

5.4 Human Evaluation and AHE

For hallucination evaluation, human perspectives can play a pivotal role, providing datasets and establishing benchmarks for the development of automatic models. To develop a robust human annotation framework, three key aspects must be carefully considered. First, it is essential to design a clear and comprehensive evaluation criteria, along with unified annotation guidelines to ensure consistency across annotators. Second, effective evaluation requires annotators with relevant domain expertise and strong linguistic skills. Researchers should carefully select annotators, provide essential training on evaluation criteria, offer extensive practice examples, and implement robust quality control measures. Third, effective digital frameworks must be established for representing annotated results that enable systematic analysis and integration with downstream applications. Standardized Although manual evaluation is time-consuming and inefficient for large-scale assessments, it remains the most reliable method for evaluating model outputs, particularly in domains with a low tolerance for hallucination, such as the medical field (Asgari et al., 2025).

6 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.

Interpretability Previous hallucination evaluation efforts primarily focused on model outputs rather than underlying mechanisms. However, analyzing factual granularity and underlying causes can substantially enhance our understanding of these phenomena. Future research directions show significant promise across multiple fronts. Reasoning-based approaches (Liu et al., 2025a; Akbar et al., 2024) demonstrate potential for uncovering hallucination origins and providing more informative evaluations. Emerging studies investigate leveraging internal model states for assessment (Chuang et al., 2024; Hu et al., 2024c; Su et al., 2024), examining how the origin, distribution, and layer-wise dynamics of neural representations relate to hallucination phenomena. Advanced interpretability methods, including sparse autoencoder(SAE)-based approaches, attempt to project neurons into more interpretable spaces for systematic analysis. These internal mechanism investigations represent a critical frontier, as the fundamental drivers of hallucinations remain poorly understood and offer substantial opportunities for breakthrough insights into model reliability.

Complex Context Effectively addressing hallucinations arising from a model's difficulty in processing complex inputs, such as long or multi-format contexts, is of critical importance. Current research on LLMs in long-context scenarios primarily focuses on handling extended input sequences, for instance, incorporating entire contexts or multi-turn dialogue histories. However, hallucinations from inconsistencies within long outputs, particularly contradictions between the beginning and the end of a generated text, remain underexplored (Wei et al., 2024), such as detecting inconsistencies in character behavior within model-generated narratives. In addition, the incorporation of multi-evidence verification into hallucination

evaluation offers a promising research direction (Wang et al., 2024), as it can enhance the robustness of factuality assessments by grounding model outputs against multiple corroborating sources.

Efficiency Looking ahead, improving the efficiency of hallucination evaluation will be critical for enabling large-scale, real-time assessment of generated text. First, developing lightweight meta-evaluators that can rapidly approximate detailed SF/WF judgments without invoking full LLM inference will reduce both latency and cost. For example, distilled classifiers or probing models trained on intermediate representations could flag likely hallucinations. Second, the integration of multi-granular caching and incremental evaluation pipelines will allow reasonable allocation of compute resources. Third, we should explore hybrid annotation strategies, whereby human annotators are only queried for the most uncertain or high-impact examples, thereby minimizing annotation overhead while maximizing the information gain of each label. Finally, exploring hybrid human–machine evaluation workflows, where automated scorers pre-screen outputs and human experts validate only edge cases, will create scalable yet trustworthy systems (Schiller, 2024). Together, these directions will move hallucination evaluation from an expensive research, only procedure toward a practical component of everyday NLG pipelines.

Emerging Hallucination Types Recent research has expanded LLMs into diverse domains including multilingual communication, multimodal understanding, and autonomous systems, introducing novel hallucination types distinct from traditional text generation. These include code hallucination, syntactically valid but semantically incorrect code (Qian et al., 2023); tool hallucination from false assumptions about external tool behavior (Zhang et al., 2024b); visual hallucination involving inaccurate content descriptions (Huang et al., 2024a); cross-lingual hallucination where meaning distorts across languages (ul Islam et al., 2025; Kang et al., 2024); and multimodal hallucination featuring cross-modal inconsistencies (Huang et al., 2024b). These domain-specific challenges require specialized evaluation frameworks that can transfer knowledge across contexts while addressing unique characteristics of each application domain. Developing robust evaluation methods for these emerging hallucination forms proves both intellectually compelling and essential for ensuring LLM system reliability and safety in diverse application contexts.

7 Conclusion

Evaluating hallucination in NLG remains a critical challenge, as it directly affects the reliability, safety, and overall applicability of language models across diverse tasks and domains. Accurate and systematic hallucination evaluation not only informs the design of more robust models but also shapes the research trajectory and future development trends within the broader NLG community. In this survey, we have systematically reviewed recent advances in the field of automatic hallucination evaluation, structuring our discussion along the key stages of the evaluation pipeline. This includes both the source faithfulness and world factuality, which differ in their grounding requirements and thus present distinct evaluation challenges.

Historically, the majority of hallucination evaluation methods have been designed in a task-specific manner, as defining clear performance criteria is often more straightforward within narrowly scoped applications. However, the rise of LLMs has brought new demands and exposed limitations in existing approaches. These models are typically deployed in open-ended, multi-domain contexts, where traditional task-based metrics fall short in capturing nuanced hallucinations. Consequently, the community has been driven to reconsider and refine existing evaluation paradigms, aiming to develop more generalizable and scalable evaluation frameworks.

Going forward, addressing hallucination in LLMs will require continued efforts in both benchmark construction and metric development, particularly those that are sensitive to domain-specific knowledge, real-world reasoning, and the dynamic nature of factual correctness. Collaboration across different fields, using ideas from linguistics, knowledge representation, cognitive science, and human evaluation, will be key to advancing the field. As such, hallucination evaluation is not merely a downstream concern but a foundational issue that will define the next generation of trustworthy NLG systems.

References

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica Salinas, Victor Alvarez, and Erwin Cornejo. HalluMeasure: Fine-grained Hallucination Measurement Using Chain-of-thought Reasoning. In *EMNLP*, pp. 15020–15037, 2024.

Elham Asgari, Nina Montaña Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digit. Medicine*, 2025.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM Hallucination Benchmark. *CoRR*, 2025.

Forrest Sheng Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, Mike Qi, Ruixuan Tu, Chenyu Xu, Matthew Gonzales, Ofer Mendelevitch, and Amin Ahmad. FaithBench: A Diverse Hallucination Benchmark for Summarization by Modern LLMs. *CoRR*, 2024. doi: 10.48550/ARXIV.2410.13210.

Nihed Bendahman, Karen Pinel-Sauvagnat, Gilles Hubert, and Mokhtar Boumedyen Billami. Not all Hallucinations are Good to Throw Away When it Comes to Legal Abstractive Summarization. In *ACL*, pp. 5331–5344, 2025.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-shot Learners. In *NeurIPS*, 2020.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization. In *ACL*, pp. 3340–3354, 2022. doi: 10.18653/V1/2022.ACL-LONG.236.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. In *ICLR*, 2024a.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking Large Language Models in Retrieval-augmented Generation. In *AAAI*, pp. 17754–17762, 2024b. doi: 10.1609/AAAI.V38I16.29728.

Kedi Chen, Qin Chen, Jie Zhou, Yishen He, and Liang He. DiaHalu: A Dialogue-level Hallucination Evaluation Benchmark for Large Language Models. In *EMNLP Findings*, pp. 9057–9079, 2024c.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. Beyond Factuality: A Comprehensive Evaluation of Large Language Models as Knowledge Generators. In *EMNLP*, pp. 6325–6341, 2023a. doi: 10.18653/V1/2023.EMNLP-MAIN.390.

Shiqi Chen, Siyang Gao, and Junxian He. Evaluating Factual Consistency of Summaries with Large Language Models. *CoRR*, 2023b. doi: 10.48550/ARXIV.2305.14069.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. FELM: Benchmarking Factuality Evaluation of Large Language Models. In *NeurIPS*, 2023c.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. Evaluating Hallucinations in Chinese Large Language Models. *CoRR*, 2023. doi: 10.48550/ARXIV.2310.03368.

Xiaoxue Cheng, Junyi Li, Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Kun Gai, and Ji-Rong Wen. Small Agent Can Also Rock! Empowering Small Language Models as Hallucination Detector. In *EMNLP*, pp. 14600–14615, 2024.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. FacTool: Factuality Detection in Generative AI - A Tool Augmented Framework for Multi-task and Multi-domain Scenarios. *CoRR*, 2023. doi: 10.48550/ARXIV.2307.13528.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. In *EMNLP*, pp. 1419–1436, 2024.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. LM vs LM: Detecting Factual Errors via Cross Examination. In *EMNLP*, pp. 12621–12640, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.778.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. Detecting and Mitigating Hallucinations in Machine Translation: Model Internal Workings Alone Do Well, Sentence Similarity Even Better. In *ACL*, pp. 36–50, 2023a. doi: 10.18653/v1/2023.acl-long.3.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R. Costa-jussà. HalOmi: A Manually Annotated Benchmark for Multilingual Hallucination and Omission Detection in Machine Translation. In *EMNLP*, pp. 638–653, 2023b. doi: 10.18653/V1/2023.EMNLP-MAIN.42.

Ashwin Devaraj, William Sheffield, Byron C. Wallace, and Junyi Jessy Li. Evaluating Factuality in Text Simplification. In *ACL*, pp. 7331–7345, 2022. doi: 10.18653/V1/2022.ACL-LONG.506.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pp. 4171–4186, 2019. doi: 10.18653/V1/N19-1423.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of Wikipedia: Knowledge-powered Conversational Agents. In *ICLR*, 2019.

Yue Dong, John Wieting, and Pat Verga. Faithful to the Document or to the World? Mitigating Hallucinations via Entity-linked Knowledge in Abstractive Summarization. In *EMNLP Findings*, pp. 1067–1082, 2022. doi: 10.18653/V1/2022.FINDINGS-EMNLP.76.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. BAMBOO: A Comprehensive Benchmark for Evaluating Long Text Modeling Capacities of Large Language Models. In *COLING*, pp. 2086–2099, 2024.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *ACL*, pp. 7250–7274, 2022. doi: 10.18653/V1/2022.ACL-LONG.501.

Xuefeng Du, Chaowei Xiao, and Sharon Li. HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection. In *NeurIPS*, 2024.

Esin Durmus, He He, and Mona T. Diab. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In *ACL*, pp. 5055–5070, 2020. doi: 10.18653/V1/2020.ACL-MAIN.454.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar R. Zaïane, Mo Yu, Edoardo Maria Ponti, and Siva Reddy. FaithDial: A Faithful Benchmark for Information-seeking Dialogue. *Trans. Assoc. Comput. Linguistics*, pp. 1473–1490, 2022a. doi: 10.1162/TACL_A_00529.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark. *Trans. Assoc. Comput. Linguistics*, pp. 1066–1083, 2022b. doi: 10.1162/TACL_A_00506.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. QAFactEval: Improved QA-based Factual Consistency Evaluation for Summarization. In *NAACL*, pp. 2587–2601, 2022. doi: 10.18653/v1/2022.naacl-main.187.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. SummEval: Re-evaluating Summarization Evaluation. *Trans. Assoc. Comput. Linguistics*, pp. 391–409, 2021. doi: 10.1162/TACL_A_00373.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In *ACL*, pp. 2214–2220, 2019. doi: 10.18653/V1/P19-1213.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nat.*, pp. 625–630, 2024. doi: 10.1038/S41586-024-07421-0.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. FactKB: Generalizable Factuality Evaluation using Language Models Enhanced with Factual Knowledge. In *EMNLP*, pp. 933–952, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.59.

Robert Friel and Atindriyo Sanyal. Chainpoll: A high efficacy method for LLM hallucination detection. *CoRR*, 2023. doi: 10.48550/ARXIV.2310.18344.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. GPTScore: Evaluate as You Desire. In *NAACL*, pp. 6556–6576, 2024. doi: 10.18653/V1/2024.NAACL-LONG.365.

Mingqi Gao and Xiaojun Wan. DialSummEval: Revisiting Summarization Evaluation for Dialogues. In *AACL*, pp. 5693–5709, 2022. doi: 10.18653/v1/2022.naacl-main.418.

Mingqi Gao, Xiaojun Wan, Jia Su, Zhefeng Wang, and Baoxing Huai. Reference Matters: Benchmarking Factual Error Correction for Dialogue Summarization with Fine-grained Evaluation Framework. In *ACL*, pp. 13932–13959, 2023. doi: 10.18653/V1/2023.ACL-LONG.779.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and Leveraging World Models in Visual Representation Learning. *CoRR*, 2024. doi: 10.48550/ARXIV.2403.00504.

Albert Gatt and Emiel Krahmer. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, pp. 65–170, 2018. doi: 10.1613/JAIR.5477.

Ben Goodrich, Vinay Rao, Peter J. Liu, and Mohammad Saleh. Assessing The Factual Accuracy of Generated Text. In *KDD*, pp. 166–175, 2019. doi: 10.1145/3292500.3330955.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-grounded Open-domain Conversations. In *INTERSPEECH*, pp. 1891–1895, 2019. doi: 10.21437/INTERSPEECH.2019-3079.

Tanya Goyal and Greg Durrett. Evaluating Factuality in Generation with Dependency-level Entailment. In *EMNLP Findings*, pp. 3592–3603, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.322.

Nuno Miguel Guerreiro, Elena Voita, and André F. T. Martins. Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation. In *EACL*, pp. 1059–1075, 2023. doi: 10.18653/V1/2023.EACL-MAIN.75.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. DialFact: A Benchmark for Fact-checking in Dialogue. In *ACL*, pp. 3785–3801, 2022. doi: 10.18653/v1/2022.acl-long.263.

Sharut Gupta, Chenyu Wang, Yifei Wang, Tommi S. Jaakkola, and Stefanie Jegelka. In-context Symmetries: Self-supervised Learning through Contextual World Models. In *NeurIPS*, 2024.

Anas Himmi, Guillaume Staerman, Marine Picot, Pierre Colombo, and Nuno Guerreiro. Enhanced Hallucination Detection in Neural Machine Translation through Simple Detector Aggregation. In *EMNLP*, pp. 18573–18583, 2024.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2: Evaluating Factual Consistency in Knowledge-grounded Dialogues via Question Generation and Question Answering. In *EMNLP*, pp. 7856–7870, 2021. doi: 10.18653/v1/2021.emnlp-main.619.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansky, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating Factual Consistency Evaluation. In *NAACL*, pp. 3905–3920, 2022. doi: 10.18653/V1/2022.NAACL-MAIN.287.

Bairu Hou, Yang Zhang, Jacob Andreas, and Shiyu Chang. A Probabilistic Framework for LLM Hallucination Detection via Belief Tree Propagation. *CoRR*, 2024. doi: 10.48550/ARXIV.2406.06950.

Haichuan Hu, Yuhan Sun, and Quanjun Zhang. LRP4RAG: Detecting Hallucinations in Retrieval-augmented Generation via Layer-wise Relevance Propagation. *CoRR*, 2024a. doi: 10.48550/ARXIV.2408.15533.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-centric Hallucination Detection. In *EMNLP*, pp. 6953–6975, 2024b.

Xiaomeng Hu, Yiming Zhang, Ru Peng, Haozhe Zhang, Chenwei Wu, Gang Chen, and Junbo Zhao. Embedding and Gradient Say Wrong: A White-box Method for Hallucination Detection. In *EMNLP*, pp. 1950–1959, 2024c.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. What Have We Achieved on Text Summarization? In *EMNLP*, pp. 446–469, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.33.

Kung-Hsiang Huang, Hou Pong Chan, and Heng Ji. Zero-shot Faithful Factual Error Correction. In *ACL*, pp. 5660–5676, 2023a. doi: 10.18653/V1/2023.ACL-LONG.311.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, pp. 42:1–42:55, 2023b. doi: 10.1145/3703155.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. Visual Hallucinations of Multi-modal Large Language Models. In *ACL Findings*, pp. 9614–9631, 2024a. doi: 10.18653/V1/2024.FINDINGS-ACL.573.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. Visual Hallucinations of Multi-modal Large Language Models. In *ACL Findings*, pp. 9614–9631, 2024b. doi: 10.18653/V1/2024.FINDINGS-ACL.573.

Yi-Chong Huang, Xia-Chong Feng, Xiao-Cheng Feng, and Bing Qin. The Factual Inconsistency Problem in Abstractive Text Summarization: A Survey. *CoRR*, 2021.

Zhaoheng Huang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. UFO: a Unified and Flexible Framework for Evaluating Factuality of Large Language Models. *CoRR*, 2024c. doi: 10.48550/ARXIV.2402.14690.

Andrew Jesson, Nicolas Beltran-Velez, Quentin Chu, Sweta Karlekar, Jannik Kossen, Yarin Gal, John P. Cunningham, and David M. Blei. Estimating the Hallucination Rate of Generative AI. In *NeurIPS*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, pp. 248:1–248:38, 2023. doi: 10.1145/3571730.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. ANAH: Analytical Annotation of Hallucinations in Large Language Models. In *ACL*, pp. 8135–8158, 2024. doi: 10.18653/V1/2024.ACL-LONG.442.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world Entailment for Claims in Wikipedia. In *EMNLP:2023:main*, pp. 7561–7583, 2023. doi: 10.18653/v1/2023.emnlp-main.470.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. Comparing Hallucination Detection Metrics for Multilingual Generation. *CoRR*, 2024. doi: 10.48550/ARXIV.2402.10496.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. RealTime QA: What's the Answer Right Now? In *NeurIPS*, 2023.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal. Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs. *CoRR*, 2024. doi: 10.48550/ARXIV.2406.15927.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the Factual Consistency of Abstractive Text Summarization. In *EMNLP*, pp. 9332–9346, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.750.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based Models for Inconsistency Detection in Summarization. *TACL*, pp. 163–177, 2022. doi: 10.1162/tacl_a_00453.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. SummEdits: Measuring LLM Ability at Factual Reasoning Through The Lens of Summarization. In *EMNLP*, pp. 9662–9676, 2023. doi: 10.18653/v1/2023.emnlp-main.600.

Barrett Martin Lattimer, Patrick Chen, Xinyuan Zhang, and Yi Yang. Fast and Accurate Factual Inconsistency Detection Over Long Documents. In *EMNLP*, pp. 1691–1703, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.105.

Hwanhee Lee, Kang Min Yoo, Joonsuk Park, Hwaran Lee, and Kyomin Jung. Masked Summarization to Generate Factually Inconsistent Summaries for Improved Factual Consistency Checking. In *ACL Findings*, pp. 1019–1030, 2022. doi: 10.18653/V1/2022.FINDINGS-NAACL.76.

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

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. *CoRR*, 2024a. doi: 10.48550/ARXIV.2411.16594.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A Large-scale Hallucination Evaluation Benchmark for Large Language Models. In *EMNLP*, pp. 6449–6464, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.397.

Ningke Li, Yuekang Li, Yi Liu, Ling Shi, Kailong Wang, and Haoyu Wang. Drowzee: Metamorphic Testing for Fact-conflicting Hallucination Detection in Large Language Models. *Proc. ACM Program. Lang.*, pp. 1843–1872, 2024b. doi: 10.1145/3689776.

Mengfei Liang, Archish Arun, Zekun Wu, Cristian Munoz, Jonathan Lutch, Emre Kazim, Adriano Koshiyama, and Philip Treleaven. THaMES: An End-to-End Tool for Hallucination Mitigation and Evaluation in Large Language Models, 2024a.

Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Yezhaohui Wang, Dawei He, Cheng Peng, Zhonghao Wang, and Haiying Deng. UHGEval: Benchmarking the Hallucination of Chinese Large Language Models via Unconstrained Generation. In *ACL*, pp. 5266–5293, 2024b. doi: 10.18653/V1/2024.ACL-LONG.288.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *ACL*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

MingShan Liu, Shi Bo, and Jialing Fang. Enhancing Mathematical Reasoning in Large Language Models with Self-Consistency-based Hallucination Detection. *CoRR*, 2025a.

Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. Towards Long Context Hallucination Detection. In *ACL Findings*, pp. 7827–7835, 2025b.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *EMNLP*, pp. 2511–2522, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.153.

Yu Lu Liu, Rachel Bawden, Thomas Scaliom, Benoît Sagot, and Jackie Chi Kit Cheung. MaskEval: Weighted MLM-based Evaluation for Text Summarization and Simplification. *CoRR*, 2022. doi: 10.48550/ARXIV.2205.12394.

Ge Luo, Weisi Fan, Miaoran Li, Guoruizhe Sun, Runlong Zhang, Chenyu Xu, and Forrest Sheng Bao. SummaCoz: A Dataset for Improving the Interpretability of Factual Consistency Detection for Summarization. In *EMNLP Findings*, pp. 3689–3702, 2024. doi: 10.18653/v1/2024.findings-emnlp.210.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *CoRR*, 2024. doi: 10.48550/ARXIV.2405.20362.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource Black-box Hallucination Detection for Generative Large Language Models. In *EMNLP:2023:main*, pp. 9004–9017, 2023a. doi: 10.18653/v1/2023.emnlp-main.557.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. MQAG: Multiple-choice Question Answering and Generation for Assessing Information Consistency in Summarization. In *AACL*, pp. 39–53, 2023b. doi: 10.18653/V1/2023.IJCNLP-MAIN.4.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On Faithfulness and Factuality in Abstractive Summarization. In *ACL*, pp. 1906–1919, 2020. doi: 10.18653/V1/2020.ACL-MAIN.173.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *EMNLP:2023:main*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. Fine-grained Hallucination Detection and Editing for Language Models. *CoRR*, 2024. doi: 10.48550/ARXIV.2401.06855.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. Generating Benchmarks for Factuality Evaluation of Language Models. In *EACL*, pp. 49–66, 2024.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-augmented Language Models. In *ACL*, pp. 10862–10878, 2024. doi: 10.18653/v1/2024.acl-long.585.

Jio Oh, Soyeon Kim, Junseok Seo, Jindong Wang, Ruochen Xu, Xing Xie, and Steven Whang. ERBench: An Entity-relationship based Automatically Verifiable Hallucination Benchmark for Large Language Models. In *NeurIPS*, 2024.

OpenAI. Chatgpt, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or Retrieval? Comparing Knowledge Injection in LLMs. In *EMNLP*, pp. 237–250, 2024.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In *NAACL*, pp. 4812–4829, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.383.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-HALT: Medical Domain Hallucination Test for Large Language Models. In *CoNLL*, pp. 314–334, 2023. doi: 10.18653/V1/2023.CONLL-1.21.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. MedHallu: A Comprehensive Benchmark for Detecting Medical Hallucinations in Large Language Models. *CoRR*, 2025.

Bibek Paudel, Alexander Lyzhov, Preetam Joshi, and Puneet Anand. HalluciNot: Hallucination Detection Through Context and Common Knowledge Verification, 2025.

Siya Qi, Rui Cao, Yulan He, and Zheng Yuan. Evaluating LLMs' Assessment of Mixed-context Hallucination Through the Lens of Summarization. *CoRR*, 2025. doi: 10.48550/ARXIV.2503.01670.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative Agents for Software Development. *CoRR*, 2023. doi: 10.48550/ARXIV.2307.07924.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. Don't be Contradicted with Anything! CI-ToD: Towards Benchmarking Consistency for Task-oriented Dialogue System. In *EMNLP:2021:main*, pp. 2357–2367, 2021. doi: 10.18653/v1/2021.emnlp-main.182.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. The Curious Case of Hallucinations in Neural Machine Translation. In *NAACL*, pp. 1172–1183, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.92.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. FactGraph: Evaluating Factuality in Summarization with Semantic Graph Representations. In *NAACL*, pp. 3238–3253, 2022. doi: 10.18653/v1/2022.naacl-main.236.

Christian A. Schiller. The Human Factor in Detecting Errors of Large Language Models: A Systematic Literature Review and Future Research Directions. *CoRR*, 2024. doi: 10.48550/ARXIV.2403.09743.

Tal Schuster, Adam Fisch, and Regina Barzilay. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *NAACL*, pp. 624–643, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.52.

Thomas Scialom and Felix Hill. BEAMetrics: A Benchmark for Language Generation Evaluation Evaluation. *CoRR*, 2021.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. QuestEval: Summarization Asks for Fact-based Evaluation. In *EMNLP*, pp. 6594–6604, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.529.

Amir Soleimani, Christof Monz, and Marcel Worring. NonFactS: NonFactual Summary Generation for Factuality Evaluation in Document Summarization. In *ACL Findings*, pp. 6405–6419, 2023. doi: 10.18653/v1/2023.findings-acl.400.

P. Aditya Sreekar, Sahil Verma, Suransh Chopra, Abhishek Persad, Sarik Ghazarian, and Narayanan Sadagopan. AXCEL: Automated eXplainable Consistency Evaluation using LLMs. In *EMNLP Findings*, pp. 14943–14957, 2024.

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 *NeurIPS*, 2024.

Julius Steen, Juri Opitz, Anette Frank, and Katja Markert. With a Little Push, NLI Models can Robustly and Efficiently Predict Faithfulness. In *ACL*, pp. 914–924, 2023. doi: 10.18653/V1/2023.ACL-SHORT.79.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. Unsupervised Real-time Hallucination Detection based on the Internal States of Large Language Models. In *ACL Findings*, pp. 14379–14391, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.854.

Chongren Sun, Yuran Li, Di Wu, and Benoit Boulet. OnionEval: An Unified Evaluation of Fact-conflicting Hallucination for Small-large Language Models. *CoRR*, 2025a.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. ReDeEP: Detecting Hallucination in Retrieval-augmented Generation via Mechanistic Interpretability. In *ICLR*, 2025b.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin F. Rousseau, and Greg Durrett. Understanding Factual Errors in Summarization: Errors, Summarizers, Datasets, Error Detectors. In *ACL*, pp. 11626–11644, 2023. doi: 10.18653/V1/2023.ACL-LONG.650.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient Fact-checking of LLMs on Grounding Documents. In *EMNLP*, pp. 8818–8847, 2024a. doi: 10.18653/v1/2024.emnlp-main.499.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. TofuEval: Evaluating Hallucinations of LLMs on Topic-focused Dialogue Summarization. In *NAACL*, pp. 4455–4480, 2024b. doi: 10.18653/v1/2024.naacl-long.251.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *NAACL*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074.

Saad Obaid ul Islam, Anne Lauscher, and Goran Glavas. How Much Do LLMs Hallucinate across Languages? On Multilingual Estimation of LLM Hallucination in the Wild. *CoRR*, 2025.

Prasetya Ajie Utama, Joshua Bambrick, Nafise Sadat Moosavi, and Iryna Gurevych. Falsesum: Generating Document-level NLI Examples for Recognizing Factual Inconsistency in Summarization. In *NAACL*, pp. 2763–2776, 2022. doi: 10.18653/V1/2022.NAACL-MAIN.199.

Kees van Deemter. The Pitfalls of Defining Hallucination. *Comput. Linguistics*, pp. 807–816, 2024. doi: 10.1162/COLI_A_00509.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry W. Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc V. Le, and Thang Luong. FreshLLMs: Refreshing Large Language Models with Search Engine Augmentation. In *ACL Findings*, pp. 13697–13720, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.813.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. In *ACL*, pp. 5008–5020, 2020. doi: 10.18653/V1/2020.ACL-MAIN.450.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. Analyzing and Evaluating Faithfulness in Dialogue Summarization. In *EMNLP*, pp. 4897–4908, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.325.

Binjie Wang, Ethan Chern, and Pengfei Liu. ChineseFactEval: A Factuality Benchmark for Chinese LLMs, 2023a.

Binjie Wang, Steffi Chern, Ethan Chern, and Pengfei Liu. Halu-J: Critique-based Hallucination Judge. *CoRR*, 2024. doi: 10.48550/ARXIV.2407.12943.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In *NEWSUM:2023:1*, pp. 1–11, 2023b. doi: 10.18653/v1/2023.newsum-1.1.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. Factcheck-GPT: End-to-end Fine-grained Document-level Fact-checking and Correction of LLM Output. *CoRR*, 2023c. doi: 10.48550/ARXIV.2311.09000.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. Long-form factuality in large language models. In *NeurIPS*, 2024.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue Natural Language Inference. In *ACL*, pp. 3731–3741, 2019. doi: 10.18653/V1/P19-1363.

Wenhao Wu, Wei Li, Xinyan Xiao, Jiachen Liu, Sujian Li, and Yajuan Lyu. WeCheck: Strong Factual Consistency Checker via Weakly Supervised Learning. In *ACL:2023:long*, pp. 307–321, 2023. doi: 10.18653/v1/2023.acl-long.18.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. Factual Consistency Evaluation for Text Summarization via Counterfactual Estimation. In *EMNLP Findings*, pp. 100–110, 2021. doi: 10.18653/V1/2021.FINDINGS-EMNLP.10.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. FIRE: Fact-checking with Iterative Retrieval and Verification. In *Proc. of ACL Findings*, pp. 2901–2914, 2025.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. Problems in Current Text Simplification Research: New Data Can Help. *Trans. Assoc. Comput. Linguistics*, pp. 283–297, 2015. doi: 10.1162/TACL_A_00139.

Shiping Yang, Renliang Sun, and Xiaojun Wan. A New Benchmark and Reverse Validation Method for Passage-level Hallucination Detection. In *EMNLP Findings*, pp. 3898–3908, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.256.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. InterrogateLLM: Zero-resource Hallucination Detection in LLM-generated Answers. In *ACL*, pp. 9333–9347, 2024. doi: 10.18653/V1/2024.ACL-LONG.506.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng Yun, Linlu Gong, Nianyi Lin, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. KoLA: Carefully Benchmarking World Knowledge of Large Language Models. In *ICLR*, 2024.

Hongbang Yuan, Pengfei Cao, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. Whispers that Shake Foundations: Analyzing and Mitigating False Premise Hallucinations in Large Language Models. In *Proc. of EMNLP*, pp. 2670–2683, 2024.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating Factual Consistency with A Unified Alignment Function. In *ACL*, pp. 11328–11348, 2023. doi: 10.18653/V1/2023.ACL-LONG.634.

Jiawei Zhang, Chejian Xu, Yu Gai, Freddy Lécué, Dawn Song, and Bo Li. KnowHalu: Hallucination Detection via Multi-form Knowledge Based Factual Checking. *CoRR*, 2024a. doi: 10.48550/ARXIV.2404.02935.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. SAC^3: Reliable Hallucination Detection in Black-box Language Models via Semantic-aware Cross-check Consistency. In *FINDINGS:2023:emnlp*, pp. 15445–15458, 2023a. doi: 10.18653/v1/2023.findings-emnlp.1032.

Shiyue Zhang, David Wan, and Mohit Bansal. Extractive is not Faithful: An Investigation of Broad Unfaithfulness Problems in Extractive Summarization. In *ACL*, pp. 2153–2174, 2023b. doi: 10.18653/V1/2023.ACL-LONG.120.

Xingxing Zhang and Mirella Lapata. Sentence Simplification with Deep Reinforcement Learning. In *EMNLP*, pp. 584–594, 2017. doi: 10.18653/V1/D17-1062.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *CoRR*, 2023c. doi: 10.48550/ARXIV.2309.01219.

Yuxiang Zhang, Jing Chen, Junjie Wang, Yaxin Liu, Cheng Yang, Chufan Shi, Xinyu Zhu, Zihao Lin, Hanwen Wan, Yujiu Yang, Tetsuya Sakai, Tian Feng, and Hayato Yamana. ToolBeHonest: A Multi-level Hallucination Diagnostic Benchmark for Tool-augmented Large Language Models. In *EMNLP*, pp. 11388–11422, 2024b.

Xiaoling Zhou, Mingjie Zhang, Zhemg Lee, Wei Ye, and Shikun Zhang. HaDeMiF: Hallucination Detection and Mitigation in Large Language Models. In *ICLR*, 2025.

Yilun Zhou, Caiming Xiong, Silvio Savarese, and Chien-Sheng Wu. Shared Imagination: LLMs Hallucinate Alike. *CoRR*, 2024. doi: 10.48550/ARXIV.2407.16604.

A Evaluator Meta Information

We present a set of tables summarizing the meta-information of the surveyed evaluators, as shown in Table 3, Table 4, and Table 5. In the *New Dataset* column, if the dataset name is identical to the evaluator's name, it indicates that the authors did not explicitly name the dataset; instead, we assign the evaluator's name for clarity and reference. The *Based-model* column refers to the underlying models used by each evaluator either for performing evaluation or for generating synthetic data. The *Method* column describes the evaluation pipeline, methodological framework, or the primary novel contribution introduced by the evaluator. The *Metric* column specifies the scoring strategy or computational approach used to produce the final evaluation score. Lastly, the *SF* (Source Faithfulness) and *WF* (World Factuality) columns use ✓ and ✗ to indicate whether an evaluator explicitly addresses each respective aspect.

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	Text 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 3: 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, RefinedWeb, 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 Recommendation, 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	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, 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	✓	✓
	HDM-2	HDMBENCH	RAGTruth, enterprise support tickets, MS Marco, SQuAD, Red Pajama v2.	Word, Response	Generation	Qwen-2.5-3B-Instruct	Classification	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	✓	✓

Table 4: AHE Meta-Info Table after LLM era (Part 1), which means the methods utilize 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	KnowHalu	/	HaluEval, HotpotQA, CNN/DM	Response	QA, Summ	Starling-7B, GPT-3.5	Identify non-fabrication, multi-form fact-checking	TPR, TNR, Avg Acc	✓	✓
	AXCEL	/	SummEval, QAGS	claim	Summ, Generation, Data2text	Llama-3-8B,Claude-Haiku, Claude-Sonnet	Direct Assessment	P, R, F1, Auc	✓	✓
	Drowzee	Drowzee	/	Response	QA	GPT-3.5-turbo, GPT-4, Llama2-7B,70B, Mistral-7B-v0.2,8x7B	Direct Assessment	FCH Ratio	✗	✓
	MIND	HELM	/	Span	Continual writing	MLP	Embedding MLP classification	AUC, Pearson corr	✗	✓
	BTProp	/	Wikibio-GPT3, FELM-Science, FactCheckGPT	Response	Generation	gpt-3.5 -turbo, Llama3-8B Instruct	hidden Markov tree	AUROC, AUC-PR, F1, Acc	✗	✓
	FAVA	FAVABENCH	Open prompts	Span	Information retrieving	Llama2-Chat 7B	Hallucination tags generation	F1	✗	✓
	Semantic Entropy	/	BioASQ, TriviaQA, NQ Open, SQuAD	Response	QA	LLaMA 2 Chat-7B,13B,70B, Falcon Instruct-7B,40B, Mistral Instruct-7B	Semantic Entropy	AUROC, AURAC	✗	✓
	SEPs	/	BioASQ, TriviaQA, NQ Open, SQuAD	Response	QA	Llama-2-7B,70B, Mistral-7B, Phi-3-3.8B	Semantic Entropy Probes	AUROC	✗	✓
	HaloScope	/	TruthfulQA, TriviaQA, CoQA, TydiQA-GP	Response	QA	LLaMA-2-chat-7B,13B, OPT6.7B,13B	Unsupervised learning	AUROC, BLUERT, ROUGE	✗	✓
	LRP4RAG	/	RAGTruth	Response	QA	Llama-2-7B/13B-chat	Internal state classification	Acc, P, R, F1	✓	✓
	Halu-J	ME-FEVER	FEVER	Claim	Fact-checking	GPT-4, Mistral-7B-Instruct	Reasoning	Acc	✗	✓
	NonFactS	NonFactS	CNN/DM	Word	Summa	BART-base, ROBERTa ALBERT	NLI	Balanced Acc	✓	✗
	MFMA	/	CNN/DM, XSum	Span, Ent	Summ	BART-base, T5-small, Electra-base-discriminator	Classification	F1, Balanced Acc	✓	✗
	HADEMIF	/	Response	QA	Llama2-7B	hidden state calibration	Expected Calibration Error, Brier Score	acc@q, cov@p	✗	✓
	REDEEP	/	RAGTruth, Dolly (AC)	Span	RAG	Llama2-7B/13B/70B, Llama3-8B	External Context Score, Parametric Knowledge Score	AUC, PCC, Acc, R, F1	✗	✓
	LMvLM	/	LAMA, TriviaQA, NQ, PopQA	Response	QA	ChatGPT, text-davinci-003, Llama-7B	LMs multi-turn judge	P, R, F1	✗	✓
	OnionEval	OnionEval	/	Ent, Atomic fact	QA	SLLMs (Llama, Qwen, Gemma)	Layered Evaluation	Acc, Context-influence Score	✓	✗

Table 5: AHE Meta-Info Table after LLM era (Part 2), which means the methods rely on the ability of LLMs such as ChatGPT.