Trust but Verify: A Comprehensive Survey of Faithfulness Evaluation Methods in Abstractive Text Summarization

Salima Lamsiyah[1]* **Aria Nourbakhsh**[1]* **Christoph Schommer**[1]
Department of Computer Science, Faculty of Science, Technology and Medicine,
University of Luxembourg
{salima.lamsiyah,aria.nourbakhsh,christoph.schommer}@uni.lu

Abstract

Abstractive text summarization systems have advanced significantly with the rise of neural language models. However, they frequently suffer from issues of *unfaithfulness* or *factual inconsistency*, generating content that is not verifiably supported by the source text. This survey provides a comprehensive review of over 40 studies published between 2020 and 2025 on methods for evaluating faithfulness in abstractive summarization. We present a unified taxonomy that covers human evaluation techniques and a variety of automatic metrics, including question answering (QA)-based methods, natural language inference (NLI)-based methods, graph-based approaches, and large language model (LLM)-based evaluation. We also discuss meta-evaluation protocols that assess the quality of these metrics. In addition, we analyze a wide range of benchmark datasets, highlighting their design, scope, and relevance to emerging challenges such as long-document and domain-specific summarization. In addition, we identify critical limitations in current evaluation practices, including poor alignment with human judgment, limited robustness, and inefficiencies in handling complex summaries. We conclude by outlining future directions to support the development of more reliable, interpretable, and scalable evaluation methods. This work helps researchers navigate the evolving landscape of faithfulness evaluation in summarization. Related papers and updates are available at [1].

1 Introduction

Automatic Text Summarization (ATS) is a fundamental task in Natural Language Processing (NLP) that aims to condense a single document or a collection of documents into concise, coherent, and informative summaries (Nenkova et al., 2011). This process significantly reduces the time and effort required for document comprehension (Zhang et al., 2024a). Generally, ATS methods were predominantly extractive, selecting important sentences or phrases directly from the original text. While effective in preserving factual accuracy by design, extractive methods often yield less fluent or coherent summaries (Zhang et al., 2024a).

With the advent of deep learning, particularly sequence-to-sequence (seq2seq) architectures such as recurrent neural networks (Medsker et al., 2001) and the Transformer (Vaswani et al., 2017), the field has witnessed a transformative shift. These models enable the generation of entirely new sentences that paraphrase the core ideas of the source text (Zhang et al., 2023). More recently, the rise of large language models (LLMs) has further enhanced these capabilities, producing summaries that demonstrated remarkable fluency, coherence, and human-like quality (Li et al., 2022).

Despite significant advancements in fluency and coherence, neural abstractive summarization models continue to face a major challenge: they often generate content that is either *unfaithful* to the input document or factually incorrect. *Faithfulness* refers to the extent to which a summary is directly supported by its source, without adding, omitting, or contradicting information (Maynez et al., 2020; Dong et al., 2022; Li et al., 2022). In contrast, *factuality* concerns whether the content is true with respect to real-world knowledge, even if it is not explicitly stated in the input (Maynez et al., 2020). Errors in generation are commonly described as *hallucinations*, which can be classified as *intrinsic*, when the summary contradicts the source, or *extrinsic* when it introduces unverifiable or fabricated content (Zhao et al., 2020; Cao and Wang, 2021; Huang et al., 2021; Ji et al., 2023).

Since 2020, research on faithfulness and factual consistency in abstractive summarization has

*Equal contribution

[1] https://github.com/lamsiyah/faithfulness-eval-summarization

Proceedings of Recent Advances in Natural Language Processing,pages 633–643
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-074

grown rapidly. Studies have introduced a range of model-based evaluation metrics, such as those using natural language inference and question answering, and released new annotated datasets that support more systematic benchmarking. However, recent evaluations (e.g., Fabbri et al. (2021)) show that no single metric consistently aligns with human judgments across models that highlight the complexity and evolving nature of the field. To clarify this landscape, this survey provides a comprehensive overview of research from 2020 to 2025, with a primary focus on faithfulness evaluation, while also addressing related concepts such as factuality and hallucination.

Contributions: (1) We present a comprehensive survey of 40+ works (2020–2025) on faithfulness evaluation in abstractive summarization. (2) We propose a unified taxonomy that encompasses human evaluations, automatic metrics (QA-, NLI-, graph-, LLM-based), and meta-evaluations. (3) We provide a comparative analysis of datasets and metrics, highlighting trends, strengths, and limitations. (4) We outline key gaps and open problems to guide future research. (5) We release all resources to support the community.

2 Related Work

Several surveys have addressed the issue of faithfulness in text generation, both generally and with a specific focus on abstractive text summarization. For instance, Li et al. (2022) provide a broad overview of faithfulness across diverse Natural Language Generation (NLG) tasks, highlighting its systemic nature and the potential for cross-task solutions. Similarly, Ji et al. (2023) offer a wide-ranging survey of hallucination in natural language generation, including summarization, and establish key definitions for intrinsic versus extrinsic hallucinations. Shifting to a narrower scope, Pagnoni et al. (2021) focus on abstractive text summarization, categorizing evaluation methods designed to address factual inconsistency. In a related vein, Huang et al. (2021) review recent advances in neural abstractive summarization, highlighting the challenge of factual inconsistencies and focusing on fact-aware evaluation metrics and models designed to improve factual consistency. Furthermore, early work by Cao et al. (2018) marks the beginning of explicit research into faithfulness in summarization, identifying the problem of "fake facts" and proposing initial mitigation strategies. At the level of summarization evaluation as a whole, Fabbri et al. (2021) address the limitations of current summarization evaluation by comprehensively re-evaluating metrics and models, releasing large-scale benchmark datasets and human annotations, and providing a unified toolkit to promote more consistent and reliable evaluation protocols aligned with human judgment.

In contrast to prior surveys, our work provides a focused and up-to-date (2020–2025) review of evaluation methods for faithfulness in abstractive summarization. We offer a fine-grained taxonomy of human, automatic, and meta-evaluation approaches, integrate recent advances in LLM-based methods, and analyze benchmarks with attention to long-document and domain-specific evaluation. This survey aims to equip readers with a detailed understanding of current evaluation practices, trade-offs between methods, and emerging research challenges.

3 A Taxonomy of Faithfulness Evaluation Methods

Faithfulness evaluation in abstractive summarization can be grouped into three categories: (1) ***human evaluations***, offering nuanced judgments but limited by subjectivity and scalability; (2) ***automatic metrics***, efficient yet often misaligned with human judgment; and (3) **meta-evaluation frameworks**, which assess the reliability of the metrics themselves. This section reviews each category, and Figure 1 illustrates the taxonomy of methods, benchmarks, challenges, and future directions.

3.1 Human Evaluation Methods

Human evaluation remains the gold standard for assessing summarization quality, including faithfulness, due to its ability to capture details that automatic metrics often miss (Kryscinski et al., 2020; Wang et al., 2020; Krishna et al., 2023). However, human evaluation is essentially slow, costly, and challenging to scale. A persistent challenge in human evaluation is achieving high inter-annotator agreement (IAA). Studies consistently report low IAA, especially for highly abstractive summaries, indicating the subjective and difficult nature of consistent judgment (Pagnoni et al., 2021). For example, SummEval found "no correlation" between expert and crowd-sourced judgments for summarization quality (Fabbri et al., 2021). This consistent finding of low IAA and the need for new guidelines

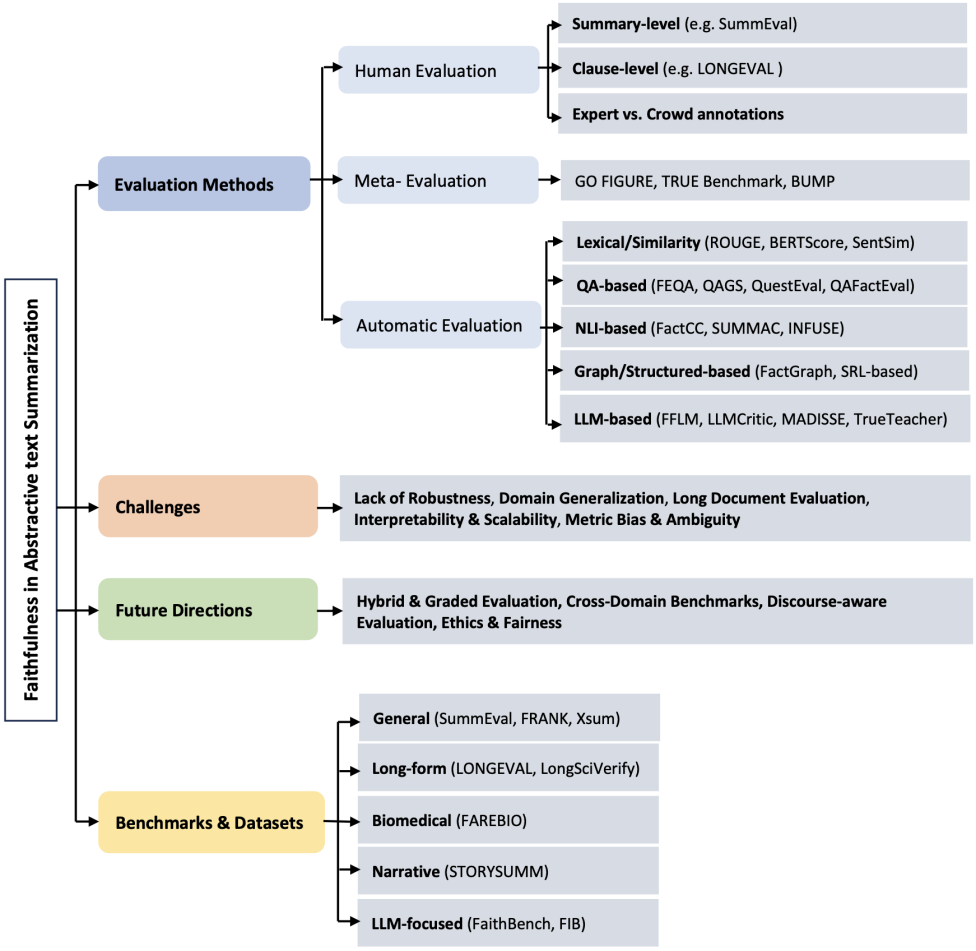

Figure 1: A Taxonomy of Faithfulness Evaluation in Abstractive Summarization: Methods, Benchmarks, Challenges, and Future Directions.

and nuanced labels underscore that faithfulness is not a universally objective concept. If human evaluators themselves struggle to consistently agree on what constitutes a "faithful" summary, then the "gold standard" for evaluating automatic metrics is noisy. This implies that achieving perfect correlation with human judgment might be an unrealistic goal. Instead, automatic metrics should aim for robustness against human variability, perhaps by focusing on clearly verifiable factual statements or by providing scores that reflect the "graded" nature of faithfulness. This situation highlights a fundamental limitation in the evaluation pipeline, where the target (human judgment) itself is a moving target, pushing research towards more objective, machine-verifiable definitions of faithfulness or more sophisticated aggregation of human input (Durmus et al., 2020; Pagnoni et al., 2021; Fabbri et al., 2021).

To improve human evaluation, especially for long-form summarization, several guidelines and benchmarks have been introduced. **LONGEVAL Guidelines** (Krishna et al., 2023) recommend clause-level and partial annotation, which improve inter-annotator agreement (IAA) and reduce annotator workload. Yet low IAA persists for abstractive and narrative summaries, reflecting the inherent subjectivity of human judgment and variability in acceptable paraphrasing or omission. Complementary resources include **FIB** (Tam et al., 2023), which tests whether LLMs prefer factually consistent continuations, and **FaithBench** (Bao et al., 2025), which introduces graded labels ("unwanted," "questionable," "benign") to capture degrees of inconsistency. FaithBench further shows that even strong protocols miss subtle errors in free-form summaries, underscoring the need for more flexible, multi-layered annotation strategies.

3.2 Automatic Evaluation Methods

Given the limitations of human evaluation, automatic metrics have been developed. Early met-

rics like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) mainly capture n-gram overlap and fail to ensure factual correctness (Dong et al., 2020; Fabbri et al., 2021). Recent work has shifted to model-based approaches, which we classify into five categories: Lexical/Semantic Similarity, QA-based, NLI/Entailment-based, Graph-based, and LLM-based metrics.

Lexical Overlap and Semantic Similarity Metrics. Some studies have used lexical overlap and semantic similarity as proxies for assessing factual consistency. Traditional metrics such as ROUGE and BLEU, which rely on n-gram overlap, have been shown to correlate poorly with human judgments of faithfulness (Goodrich et al., 2019; Maynez et al., 2020; Wang et al., 2020). Their insensitivity to semantic changes means they can assign high scores even to factually incorrect summaries. To address the limitations of n-gram-based methods, later work adopted semantic similarity approaches using contextual embedding models. Notably, **BERTScore** (Zhang et al., 2019) computes similarity scores between tokens in candidate and reference summaries using contextual embeddings from BERT. While BERTScore generally correlates better with human judgments than ROUGE, its effectiveness is still limited, particularly for highly abstractive summaries that require deeper semantic understanding (Durmus et al., 2020; Fischer et al., 2022). Moreover, BERTScore can be sensitive to contextual variations and may overpredict faithfulness in some cases (Gabriel et al., 2021). Similarly, the **SentSim** metric (Fischer et al., 2022), which measures sentence-level semantic similarity, has demonstrated promising alignment with human judgments on datasets like SummEval, offering a better alternative to surface-based metrics.

Question Answering (QA)-based Metrics. QA-based metrics evaluate faithfulness by generating questions from the summary and verifying answers in the source. The intuition, introduced by Wang et al. (2020), is that faithful summaries should contain only source-verifiable content. Durmus et al. (2020) proposed **FEQA**, which compares QA-derived answers with expected ones using F1, showing strong correlation with human judgments on abstractive datasets like XSum (Narayan et al., 2018), though limited by QA model quality and rigid scoring. **QAGS** (Wang et al., 2020) similarly uses QG/QA but improves interpretability by identifying inconsistent tokens, demonstrating robustness across domains. **QuestEval** (Scialom et al., 2021) introduced a reference-less QA framework with question weighting, improving correlations across factual consistency, coherence, fluency, and relevance. **QAFactEval** (Fabbri et al., 2022) further optimized QG (BART-large), answer selection (NP chunking), and scoring (LERC), yielding a 14% gain on SummaC.

Overall, QA-based metrics align closely with human fact-checking and provide interpretable error traces, but remain computationally intensive and sensitive to QG/QA quality. Despite these limitations, they represent the most effective tools for assessing factual consistency.

Natural Language Inference (NLI)/Entailment-based Metrics. NLI-based metrics assess factual consistency by checking whether the summary is entailed by the source, or at least neutral, but never contradictory (Falke et al., 2019). Early work by Falke et al. (2019) used pretrained NLI models to re-rank summaries, but these struggled to distinguish factual from non-factual content. **FactCC** (Kryscinski et al., 2020) addressed this with a weakly supervised BERT model trained on synthetic perturbations (e.g., entity swaps, paraphrases, negation), outperforming general NLI models like MNLI or FEVER in less abstractive domains. Dependency Arc Entailment **(DAE)** (Goyal and Durrett, 2020, 2021) advanced this idea by scoring factuality at the dependency arc level, while **SUMMAC** (Laban et al., 2022) computed entailment probabilities between each summary and document sentence, aggregating scores via max, average (SCZS), or convolutional methods (SCConv).

TRUE (Honovich et al., 2022) benchmarked 12 metrics across 11 datasets, finding NLI- and QA-based approaches strongest, though less reliable for longer inputs. To mitigate this, **INFUSE** (Zhang et al., 2024b) introduced fine-grained reasoning that incrementally selects minimal context per summary sentence and applies sub-sentence decomposition, improving evaluation of complex summaries. More recently, Zhong and Litman (2025) incorporated discourse analysis (e.g., RST parsing) into NLI, showing the value of document structure and sentence importance for long-document summarization.

Entailment-based metrics directly model the "supported by source" notion of factual consistency, typically outputting probabilities or scores. How-

ever, they depend on NLI model quality and suitable training data. Limitations include: (a) fragility to domain shifts (e.g., news vs. dialogue); (b) penalizing paraphrasing or aggregation unseen in training; and (c) binary entailment labels that obscure error severity or location. Despite these challenges, entailment-based methods consistently outperform surface-level approaches like n-gram overlap in correlating with human judgments (Honovich et al., 2022).

Graph-based and Structured Representation Metrics. These methods aim to move beyond surface-level textual analysis by converting summaries and source texts into semantically rich structures such as Abstract Meaning Representations (AMRs), semantic role labels (SRLs), named entities, or relational triples to enable more precise factual consistency comparisons.

FactGraph (Ribeiro et al., 2022) decomposes document and summary sentences into AMR graphs and uses a dual-encoder architecture to jointly encode textual and graph modalities. This abstraction away from surface form improves the detection of semantic errors and unverifiable content, though AMR parsing remains computationally demanding. Extending this direction, **AMRFACT** (Qiu et al., 2024) generates high-quality negative training samples by injecting factual inconsistencies into AMR-parsed reference summaries. These are converted back into text and filtered using a dedicated module (NEGFILTER), resulting in more coherent and diverse training data for factual consistency detectors.

Complementary to AMR-based techniques, SRL-based metrics compare semantic frames or argument structures derived from summaries and sources (Fischer et al., 2022). While less semantically detailed than AMRs, SRL approaches introduce useful linguistic structure and have shown moderate correlation with human judgments. Similarly, named entity recognition (NER)-based methods target entity hallucination by aligning named entities across source and summary, though their effectiveness diminishes when errors extend beyond named entities (Fischer et al., 2022).

Finally, triple-based methods extract relational triples (e.g., <subject, relation, object>) from both source and summary and assess their factual alignment. For example, Goodrich et al. (2019) trained a Transformer-based fact extractor on Wikidata-derived data. However, open-schema extraction faces challenges in aligning semantically equivalent but structurally different triples, whereas fixed-schema approaches improve alignment consistency at the cost of reduced coverage.

Large Language Model-based Evaluation. The emergence of LLMs has introduced new paradigms for evaluating summary faithfulness, either by leveraging LLMs as direct evaluators or by using them to generate training data for smaller models. Jia et al. (2023) proposed **FFLM**, a zero-shot metric using moderately-sized LLMs (e.g., LLaMa) to assess faithfulness by comparing the generation probability of a summary under consistent and inconsistent document contexts, without task-specific fine-tuning. Adlakha et al. (2024) introduced **LLMCritic**, where LLMs like GPT-4 are prompted to assess faithfulness in a QA setup. This L*LM-as-a-judge* approach has gained traction for its strong correlation with human evaluations. Similarly, Koupaee et al. (2025) proposed **MADISSE**, a multi-agent debate framework where LLMs with opposing stances argue about a summary's faithfulness, often revealing ambiguous cases. Bao et al. (2025) used GPT-4o in FaithBench to detect challenging hallucination cases where faithfulness detectors disagree, and also benchmarked GPT-4o as an evaluator itself. In the same context, Gekhman et al. (2023) introduced **TrueTeacher**, an approach where an LLM annotates factual errors in model-generated summaries to train a smaller student model, eliminating the need for manually perturbed reference summaries.

Despite their promise, LLM-based evaluation approaches face notable challenges: they tend to be costly, often operate as black-boxes with opaque reasoning, and can yield inconsistent judgments across different prompts or runs, especially when relying on closed-source models (e.g., see Stureborg et al. (2024), Li et al. (2024)). Hybrid methods, by contrast, show promise: for instance, using LLMs to generate explanations or rationale while entrusting scoring to simpler models, or distilling LLM behaviors into smaller, task-specific models (e.g. Hsieh et al. (2023)).

3.3 Meta-Evaluation of Faithfulness Metrics

Meta-evaluation refers to the process of evaluating the evaluation metrics themselves. This is crucial for understanding the reliability, strengths, and weaknesses of proposed faithfulness metrics.

Several benchmarks and protocols have been

proposed to rigorously evaluate these metrics. For instance, Gabriel et al. (2021) introduced **GO FIGURE**, a framework that outlines five essential criteria for faithfulness metrics: boundedness, sensitivity to factuality levels, robustness across error types, generality across domains, and human correlation. It employs diagnostic datasets with both synthetic and human-annotated errors in news and dialogue summaries. Honovich et al. (2022) proposed the **TRUE** benchmark, which uses example-level meta-evaluation with ROC AUC to quantify how well a metric distinguishes consistent from inconsistent summaries. Ma et al. (2023) developed **BUMP**, a benchmark of minimally perturbed summary pairs designed to test a metric's sensitivity to subtle factual errors, enabling fine-grained assessment of consistency and discrimination ability. In addition to dedicated benchmarks, many studies perform correlation analysis between proposed metrics and human ratings on established datasets such as **SummEval** (Fabbri et al., 2021) and **FRANK** (Pagnoni et al., 2021) that serve as a common form of informal meta-evaluation.

Meta-evaluation provides an objective basis for comparing metrics and helps identify which metrics truly reflect human judgment. It improves reproducibility and metric robustness across domains. However, it depends on high-quality human annotations and curated benchmarks, which may introduce biases or fail to represent all real-world summarization challenges. Moreover, correlation with human judgment does not guarantee interpretability or error localization, which highlights the need for complementary analysis tools.

4 Benchmarks for Faithfulness Evaluation

A wide range of benchmarks and datasets has been created for faithfulness evaluation in abstractive summarization. **General-purpose datasets**, such as SummEval and XSUM-based resources (e.g., XSF, QAGS-XSUM) (Maynez et al., 2020), provide human judgments of consistency, with XSUM widely used for its abstractiveness. **Consolidated benchmarks** like the TRUE Benchmark (Honovich et al., 2022) and AggreFact (Tang et al., 2023) unify multiple datasets (e.g., FRANK, SummEval, QAGS, MNBM) for standardized meta-evaluation; FRANK offers fine-grained error labels, while BUMP tests metric sensitivity with minimally perturbed pairs. **Domain- and task-specific datasets** address specialized needs, including DiverSumm (Zhang et al., 2024b), LongSciVerify (Bishop et al., 2023), LONGEVAL (Krishna et al., 2023), FAREBIO (Fang et al., 2024), FIB (Tam et al., 2023), and STORYSUMM (Subbiah et al., 2024). Weakly supervised training is supported by synthetic resources such as FactCC (Kryscinski et al., 2020), LLM-labeled datasets like TrueTeacher (Gekhman et al., 2023), and **structurally perturbed datasets** such as AMRFACT (Qiu et al., 2024). Together, these resources play complementary roles: general-purpose datasets enable broad comparisons, consolidated suites support standardized meta-evaluation, domain-specific collections address long-form and LLM-generated summaries, and synthetic/perturbed sets facilitate stress-testing and training. Table 1 summarizes these datasets.

5 Challenges and Limitations of the Existing Evaluation Methods

Despite progress, evaluating faithfulness in abstractive summarization remains challenging, as both human and automatic methods have notable limitations. A comparative analysis of these challenges is available at [2]. Persistent challenges include:

Correlation with Human Judgment and Interpretability. A key challenge for automatic faithfulness metrics is their often low to moderate correlation with human judgments, indicating they don't fully capture human nuances (Huang et al., 2021). Furthermore, many advanced model-based metrics, particularly those using deep learning or LLMs, lack interpretability, making it difficult to understand their scoring logic and derive actionable insights. While some metrics like QAGS (Wang et al., 2020) (through its question-answer pairs) and MADISSE (through debate arguments) (Koupaee et al., 2025) offer better interpretability, this remains a significant challenge for the field.

Robustness to Adversarial Inputs and Domain Shifts Faithfulness metrics frequently lack robustness, which shows inconsistent performance across different domains or out-of-distribution data (Fischer et al., 2022). They can also be overly sensitive to minor paraphrasing or lexical variations that do not impact factual accuracy. For example, NLI-based models can be misled by superficial

[2] https://github.com/lamsiyah/faithfulness-eval-summarization

Benchmark	Focus	Annotation Granularity	Size / Scope	Includes LLM-Generated Summaries?
SummEval (Fabbri et al., 2021)	General News (CNN/DM)	Summary-level (consistency, coherence, etc.)	100 articles, 16 models, ∼1,600 summaries	Yes (various neural models)
FRANK (Pagnoni et al., 2021)	General News (CNN/DM, XSum)	Sentence-level error typology	2,250 summaries	Yes (various neural models)
TRUE (Honovich et al., 2022)	Diverse (Summarization, Dialogue, Fact Verification)	Binary summary/text-level consistency	11 datasets consolidated	Yes (from constituent datasets)
BUMP (Ma et al., 2023)	General News (CNN/DM)	Minimal pairs (faithful vs. unfaithful sentence)	Human-written minimal pairs	No (focus is on human-written variations)
AggreFact (Tang et al., 2023)	General News (CNN/DM, XSum)	Binary summary-level	9 datasets consolidated, FTSOTA split focuses	Yes (BART, PEGASUS, etc.)
DiverSumm (Zhang et al., 2024b)	Diverse (Long-form, Multi-doc, Meeting, XSum)	Sentence-level faithfulness, error types	563 instances, tasks like MultiNews, QMSUM, ArXiv, GovReport	Yes (includes GPT-3.5)
LongSciVerify (Bishop et al., 2023)	Long Scientific Docs (PubMed, ArXiv)	Fine-grained summary-level factual consistency	270 annotated summaries	Yes (LongT5, BigBird-PEGASUS)
LONGEVAL (Krishna et al., 2023)	Long-form (Literary, Scientific)	Fine-grained (clause-level) binary faithfulness	120 SQuALITY & PubMed summaries	Yes (model outputs used in study)
FIB (Tam et al., 2023)	News Summarization (LLM Evaluation)	Binary (LLM preference for consistent summary)	CNN/DM, XSum; evaluates 23 LLMs	Yes (evaluates LLMs' scoring, uses model-generated inconsistent summaries)
FaithBench (Bao et al., 2025)	LLM Hallucinations (Challenging Samples)	Span-level hallucination type (questionable, benign, unwanted)	750 summaries from 10 modern LLMs	Yes (core focus)
STORYSUM (Subbiah et al., 2024)	Narrative Summarization (LLM Evaluation)	Sentence-level binary faithfulness, error explanations	96 short stories, summaries from GPT/Claude series	Yes (core focus)
FAREBIO (Fang et al., 2024)	Plain Biomedical Summaries (LLM Evaluation)	Sentence-level faithfulness, supporting evidence	175 summaries from 7 LLMs, expert MD annotations	Yes (core focus)

Table 1: Overview of key benchmarks and datasets for faithfulness evaluation.

lexical overlap, resulting in false assessments of entailment or contradiction (Huang et al., 2021).

Coverage of Diverse Faithfulness Error Types. Existing metrics often struggle to comprehensively cover the diverse spectrum of faithfulness errors, such as subtle distortions, omissions, or fabricated information (Pagnoni et al., 2021). Some may excel at detecting direct contradictions but fail on other error types, and metrics trained on synthetic data may not generalize to errors from advanced models (Fischer et al., 2022). A particular challenge is identifying "Out-of-Article" errors, which are factually correct externally but unsupported by the source document (Fang et al., 2024; Qiu et al., 2024).

Scalability, Cost, and Efficiency. The practical utility of evaluation methods is significantly impacted by their scalability, cost, and efficiency. While human evaluation is often considered the gold standard, it is resource-intensive (Lee et al., 2024). Automatic metrics present a trade-off: simpler lexical methods are efficient but less accurate, whereas sophisticated model-based approaches, especially those involving multiple LLM inferences or complex graph parsing, can be computationally expensive and slow, which limits their widespread adoption.

Issues in Long-Form and Multi-Document Summarization Evaluation. Evaluating faithfulness in long-form and multi-document summarization presents unique complexities not typically addressed by methods designed for shorter, single-document texts. Challenges include handling extended context within models that often have in-

put length limitations and the difficulty of identifying relevant supporting evidence across vast amounts of text. Several methods are emerging to specifically address these challenges, but this area requires further research (Bishop et al., 2023; Krishna et al., 2023).

Limitations of Current Benchmarking Practices. Current benchmarking practices for faithfulness evaluation face several limitations, including historical lack of standardization (partially addressed by initiatives like TRUE (Honovich et al., 2022)), potential issues with annotation quality and consistency, and the relevance of older benchmarks to errors produced by state-of-the-art models (prompting newer benchmarks like FaithBench). Furthermore, the size and diversity of existing benchmarks can restrict the generalizability of findings (Subbiah et al., 2024).

Challenges with LLM-based Evaluation. While LLMs offer significant potential for faithfulness evaluation, their application introduces specific challenges such as biases (Fang et al., 2024), opacity in their decision-making, high operational costs for powerful models, inconsistent reliability influenced by prompting, and struggles with particularly *challenging* hallucination cases (Adlakha et al., 2024; Bao et al., 2025). Addressing these issues is crucial for effectively leveraging LLMs in this evolving field where human evaluation also remains indispensable.

6 Future Directions

Future Directions in Faithfulness Evaluation. The future of faithfulness evaluation is moving toward more robust solutions that combine multiple metric types, nuanced benchmarks, and human-in-the-loop assessments. Rather than relying on a single universal metric, research is trending toward hybrid approaches that integrate QA, NLI, and semantic similarity methods to capture multiple facets of faithfulness, including factual accuracy, completeness, and distortion detection. There is a growing emphasis on developing fine-grained, interpretable metrics capable of localizing specific errors, providing actionable explanations, and classifying error types using established taxonomies. Expanding faithfulness evaluation to low-resource languages and domains such as legal or scientific texts remains a key frontier, with few-shot and zero-shot methods offering promising directions.

LLMs, Ethical Concerns, and Graded Evaluation. LLMs are reshaping evaluation strategies, acting as both summary evaluators and refiners. Research is focusing on improving LLM-based evaluators through better prompting and fine-tuning, while also building robust meta-evaluation protocols to assess their reliability and biases. Ethical considerations such as fairness across demographics, detecting evaluation biases, and preventing metric exploitation are increasingly important. Additionally, moving beyond binary classification toward graded faithfulness scales and ambiguity modeling is essential for reflecting the nuanced nature of summarization quality. These advances aim to ensure evaluation methods evolve in tandem with the growing capabilities and complexity of summarization models.

Limitations

While this survey aims to provide an exhaustive overview of methods for evaluating faithfulness in abstractive text summarization based on the collected references, certain limitations should be acknowledged. The field is evolving rapidly, particularly with the emergence of LLMs, and some of the most recent or pre-print research may not be comprehensively covered. Due to the broad scope, this work emphasizes breadth over depth and does not offer deeply technical analyses of every algorithm or benchmark. Furthermore, our evaluation of the strengths and weaknesses of various methods is primarily based on the findings and claims reported in the original publications; we did not undertake an independent empirical re-evaluation of all discussed metrics. Additionally, the primary focus is on faithfulness evaluation in general abstractive text summarization. While some concepts are broadly applicable, dedicated explorations of faithfulness in highly specialized domains or other NLG tasks (e.g., dialogue, story generation) are only addressed when they introduce significantly novel evaluation paradigms relevant to abstractive summarization. Finally, this survey concentrates on evaluation methods rather than methods aimed at improving faithfulness through novel generation architectures or mitigation strategies.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2024. Evaluating correctness and faithfulness of instruction-

following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699.

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. 2025. FaithBench: A diverse hallucination benchmark for summarization by Modern LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 448–461, Albuquerque, New Mexico. Association for Computational Linguistics.

Jennifer A Bishop, Qianqian Xie, and Sophia Ananiadou. 2023. Longdocfactscore: Evaluating the factuality of long document abstractive summarisation. *arXiv preprint arXiv:2309.12455*.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. Association for Computational Linguistics.

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

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

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

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

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

Biaoyan Fang, Xiang Dai, and Sarvnaz Karimi. 2024. Understanding faithfulness and reasoning of large language models on plain biomedical summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9890–9911, Miami, Florida, USA. Association for Computational Linguistics.

Tim Fischer, Steffen Remus, and Chris Biemann. 2022. Measuring faithfulness of abstractive summaries. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 63–73, Potsdam, Germany. KONVENS 2022 Organizers.

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

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Ben Goodrich, Vinay Rao, Peter J Liu, and Mohammad Saleh. 2019. Assessing the factual accuracy of generated text. In *proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 166–175.

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

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

pages 1449–1462, Online. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Qi Jia, Siyu Ren, Yizhu Liu, and Kenny Zhu. 2023. Zero-shot faithfulness evaluation for text summarization with foundation language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11017–11031, Singapore. Association for Computational Linguistics.

Mahnaz Koupaee, Jake W. Vincent, Saab Mansour, Igor Shalyminov, Han He, Hwanjun Song, Raphael Shu, Jianfeng He, Yi Nian, Amy Wing-mei Wong, Kyu J. Han, and Hang Su. 2025. Faithful, unfaithful or ambiguous? multi-agent debate with initial stance for summary evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12209–12246, Albuquerque, New Mexico. Association for Computational Linguistics.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Yukyung Lee, Joonghoon Kim, Jinhyuk Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2024. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. *arXiv preprint arXiv:2403.18771*.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227*.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024. Split and merge: Aligning position biases in llm-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liang Ma, Shuyang Cao, Robert L Logan IV, Di Lu, Shihao Ran, Ke Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. BUMP: A benchmark of unfaithful minimal pairs for meta-evaluation of faithfulness metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12788–12812, Toronto, Canada. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Larry R Medsker, Lakhmi Jain, et al. 2001. Recurrent neural networks. *Design and Applications*, 5(64-67):2.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Ani Nenkova, Kathleen McKeown, et al. 2011. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Haoyi Qiu, Kung-Hsiang Huang, Jingnong Qu, and Nanyun Peng. 2024. AMRFact: Enhancing summarization factuality evaluation with AMR-driven negative samples generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 594–608, Mexico City, Mexico. Association for Computational Linguistics.

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

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

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Thomas Adams, Lydia Chilton, and Kathleen McKeown. 2024. STORYSUMM: Evaluating faithfulness in story summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9988–10005, Miami, Florida, USA. Association for Computational Linguistics.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2024a. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*.

Huajian Zhang, Yumo Xu, and Laura Perez-Beltrachini. 2024b. Fine-grained natural language inference based faithfulness evaluation for diverse summarisation tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1701–1722, St. Julian's, Malta. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.

Yang Zhong and Diane Litman. 2025. Discourse-driven evaluation: Unveiling factual inconsistency in long document summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2050–2073, Albuquerque, New Mexico. Association for Computational Linguistics.