Attribution, Citation, and Quotation: A Survey of Evidence-based Text Generation with Large Language Models

Tobias Schreieder[1] **and Tim Schopf**[1,2] **and Michael Färber**[1]
[1]TU Dresden & ScaDS.AI Dresden/Leipzig, Dresden, Germany
[2]National Institute of Informatics, Tokyo, Japan
{tobias.schreieder, tim.schopf, michael.faerber}@tu-dresden.de

Abstract

The increasing adoption of large language models (LLMs) has raised serious concerns about their reliability and trustworthiness. As a result, a growing body of research focuses on *evidence-based text generation with LLMs*, aiming to link model outputs to supporting evidence to ensure traceability and verifiability. However, the field is fragmented due to inconsistent terminology, isolated evaluation practices, and a lack of unified benchmarks. To bridge this gap, we systematically analyze 134 papers, introduce a unified taxonomy of evidence-based text generation with LLMs, and investigate 300 evaluation metrics across seven key dimensions. Thereby, we focus on approaches that use citations, attribution, or quotations for evidence-based text generation. Building on this, we examine the distinctive characteristics and representative methods in the field. Finally, we highlight open challenges and outline promising directions for future work.

1 Introduction

Recent LLMs have demonstrated remarkable capabilities in language understanding and generation (Brown et al., 2020; Ouyang et al., 2022). Despite these advances, LLMs continue to face challenges such as the tendency to generate hallucinations (Ji et al., 2023; Huang et al., 2025) and their knowledge being limited to training data (Zhang et al., 2023c; Li and Goyal, 2025).

To address these limitations and increase trust, an emerging line of research focuses on generating text that is traceable to supporting evidence, allowing verification of LLM-generated content (Huang and Chang, 2024). However, despite growing interest, there is no shared understanding of the field, as prior works have used varied terminology to describe similar research efforts. For instance, while retrieval-augmented generation (RAG) has gained prominence in recent years, our survey identifies it as just one of seven closely related approaches. Consequently, research efforts in this area are highly fragmented, with studies typically aiming to cite (used in 75% of papers), attribute (62%), or quote (13%) evidence.[1] Generating text with citations involves the insertion of citation markers that explicitly reference supporting evidence sources (Gao et al., 2023b). Attributed text generation constitutes a broader notion that generally focuses on linking generated content back to the underlying sources used for grounding (Slobodkin et al., 2024). Further, generating text with quotes focuses on incorporating excerpts from evidence sources into the generated text (Menick et al., 2022). While these research efforts differ in focus, they share the common paradigm of "*evidence-based text generation with LLMs*", where LLMs generate texts, accompanied by explicit references that make the outputs traceable to supporting evidence. Figure 1 illustrates this task across different scenarios.

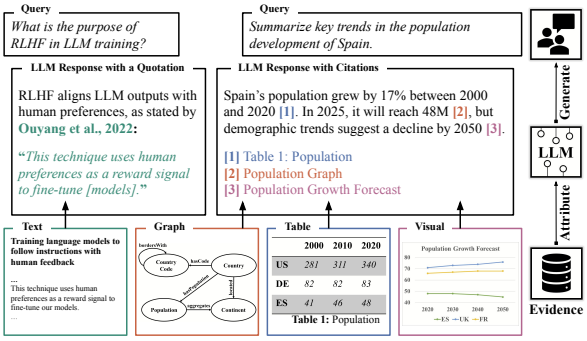

Figure 1: Illustration of evidence-based text generation with LLMs with various citation modalities and styles.

Given the rapid advancement of LLMs and diverse research on evidence-based text generation, a wide range of approaches has emerged, calling for consolidation. However, no existing study offers a comprehensive and systematic review of the full research landscape in this area. To address this gap,

[1]Some studies use multiple terminologies and approaches.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30956–31000
July 2-7, 2026

we conducted an extensive survey, categorizing key concepts, identifying trends, and outlining promising directions for future work. To the best of our knowledge, this is the first survey dedicated to this paradigm. Our analysis of 134 papers, 300 evaluation metrics, 19 frameworks, 231 datasets, and 11 benchmarks serves as a valuable resource for understanding and navigating the field. We make our annotated dataset available in a public repository.[2]

The main contributions of this study are:

1. We provide the first taxonomy of evidence-based text generation with LLMs.

2. We review 300 evaluation metrics, classify them by seven dimensions and six methods, and identify common benchmarks.

3. We outline emerging research trends, key limitations, and promising future directions.

2 Related Work

Most existing surveys address related but distinct topics, such as LLM hallucinations (Ji et al., 2023; Zhang et al., 2023b; Sahoo et al., 2024; Huang et al., 2025), knowledge-enhanced text generation (Yu et al., 2022), grounding capabilities of LLMs (Lee et al., 2024b; Qiu et al., 2024; Jokinen, 2024), generative information retrieval (Li et al., 2025), and RAG (Fan et al., 2024; Arslan et al., 2024; Gao et al., 2024; Chen et al., 2024). While these works study important components of evidence-based text generation, they focus on isolated aspects such as factuality, retrieval, or grounding, without providing a unified perspective.

Research on AI-generated plagiarism detection is related, as both address attribution for semantic reuse (Pudasaini et al., 2025; Wu et al., 2025b). However, plagiarism detection focuses on identifying uncredited content, often from a legal or ethical perspective, emphasizing post-hoc analysis. In contrast, evidence-based text generation with LLMs aims to generate text supported by verifiable sources through explicit citation.

Prior literature on evidence-based text generation with LLMs remains limited in scope. In their position paper, Huang and Chang (2024) emphasize the importance of citation mechanisms but do not systematically review prior work. Li et al. (2023a) discuss early research on LLM attribution but do not cover the broader paradigm of evidence-based text generation, lack a systematic literature search, and are already outdated, as over 75% of studies in our dataset were published after 2023. In contrast, our survey systematically reviews the full landscape of evidence-based text generation with LLMs, covering attribution approaches, citation characteristics, tasks, and evaluation resources.

[2]Dataset: https://github.com/faerber-lab/AttributeCiteQuote

3 Evidence-based Text Generation

This section presents the findings of our literature review. We conducted a systematic mapping study following the PRISMA protocol, yielding 805 deduplicated papers, of which 134 were identified as relevant through manual screening. Figure 3 presents the multidimensional taxonomy we developed to characterize evidence-based text generation with LLMs, using a facetted classification approach (Crowston and Kwasnik, 2004). For each taxonomy dimension, we provide a descriptive overview along with key findings and future directions, while methodological details and extended analyses are reported in Appendices B, C, and D.

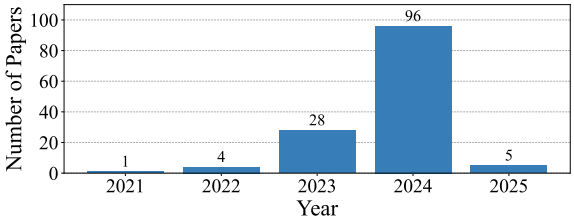

Figure 2: Number of studies per year.

We analyze the number of published papers per year as a proxy for research interest. Figure 2 shows that after a few studies in 2021–2022, papers increased to 28 in 2023 and surged to 96 in 2024, a 3.4-fold rise. Over 75% of studies were published after 2023, highlighting limitations of earlier surveys such as Li et al. (2023a), which omit much of the recent literature. As our search was conducted in February 2025, the dataset includes only five papers from that year. Given the trend, we anticipate continued growth, underscoring sustained interest in evidence-based text generation with LLMs.

3.1 Attribution Approach

Huang and Chang (2024) categorize attribution into parametric and non-parametric approaches, distinguished by the nature of knowledge used during text generation. *Parametric approaches* rely on

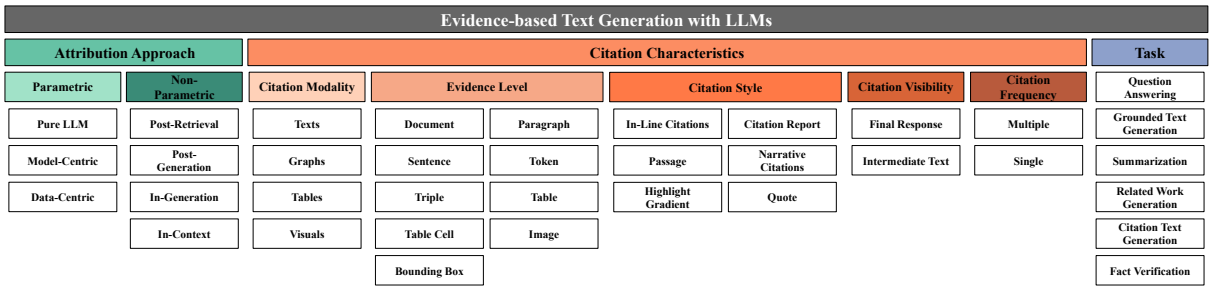

Figure 3: Multidimensional taxonomy of evidence-based text generation with LLMs. The taxonomy categorizes papers along three independent dimensions: attribution approach, citation characteristics, and task, which together capture the core design choices of evidence-based text generation. Table 7 in Appendix lists all annotated papers.

knowledge encoded within model parameters during training and are suited for analyzing knowledge within LLMs by enabling attribution without reliance on external sources, supporting explainability. In contrast, *non-parametric approaches* incorporate external sources at inference time and are the predominant choice in current evidence-based text generation systems, as they enable the use of explicit and up-to-date evidence. We extend this categorization to include approaches outside the original framework. In total, we identify 126 non-parametric and 25 parametric attribution approaches, with post-retrieval being most common. A detailed distribution across tasks is shown in Figure 7 in Appendix.

3.1.1 Parametric Attribution

Previous work treats parametric attribution as a single category (Huang and Chang, 2024; Li et al., 2023a), while our review identifies three types. **Pure LLMs** rely on the inherent attribution capabilities of existing LLMs without altering their architecture, training procedures, or underlying data (Byun et al., 2024; Zuccon et al., 2023). **Model-Centric** approaches aim to improve attribution by modifying the LLM architecture or training objectives (Chu et al., 2025; Khalifa et al., 2024). Finally, **Data-Centric** attribution curates, augments, or synthetically generates data without changing LLM architectures or internal behavior (Li et al., 2024b; Huang et al., 2024b).

Takeaways. Parametric attribution remains largely underexplored, with only 25 papers in our study, most of which (72%) focus on evaluating attribution behavior of pure LLMs. Model- and data-centric approaches receive limited attention, with few methods targeting even widely studied tasks and no clear growth trend (see Figure 7), reflecting the technical difficulty and limited scalability of current approaches. While evaluation of pure LLMs remains important, future work could increasingly focus on model- and data-centric approaches to strengthen the intrinsic ability of LLMs to generate attributable text. This is essential for understanding model-internal knowledge and data provenance, enabling more effective analysis of hallucinations, and supporting assessments of privacy and copyright.

3.1.2 Non-Parametric Attribution

Non-parametric attribution relies on the integration of external evidence. Li et al. (2023a) distinguish between *post-retrieval* and *post-generation* approaches. We extend this categorization with the two classes *in-generation* and *in-context*.

Post-Retrieval attribution, also known as *pre-hoc* (Huang and Chang, 2024), refers to a class of methods which retrieve relevant external information before text generation (Li et al., 2023a). The retrieved content is incorporated as additional context to condition the LLM's response. A prominent architecture in this category is RAG, wherein the LLM is instructed to base its output solely on the retrieved documents. However, standard RAG does not inherently support attribution, and must be extended with mechanisms for providing explicit references to the retrieved source documents.

Post-Generation attribution, also known as *post-hoc* (Huang and Chang, 2024), first generates a response and then retrieves relevant evidence based on the LLM output (Li et al., 2023a). These approaches resemble classical citation recommendation (Färber and Jatowt, 2020) by retrieving sources for each generated claim, leveraging evidence retrieval on both the user input (e.g., a question) and the LLM output to attribute each generated claim.

In-Generation attribution is a recent paradigm

in which LLMs dynamically determine the need for additional evidence and retrieve it during generation (Asai et al., 2024b; Li et al., 2024c). This contrasts with traditional post-retrieval or post-generation methods by tightly integrating retrieval decisions into the generation process.

In-Context attribution differs from previous approaches, as it does not require retrieval. Instead, the user explicitly provides evidence as part of the prompt, embedding it directly into the context (Cohen-Wang et al., 2024; Zhang et al., 2025a).

Takeaways. Non-parametric attribution dominates the literature, comprising 126 works in our study. Post-retrieval attribution is the most prevalent paradigm, covering 58% of non-parametric approaches. This prevalence reflects the practical effectiveness of RAG pipelines. Post-generation attribution remains comparatively limited (18%) and is concentrated primarily in question answering (see Figure 7 in Appendix). In-context attribution (20%) is task-dependent, mainly appearing in settings where evidence is provided as input, such as summarization and citation text generation. In contrast, in-generation attribution remains underexamined (4%) but represents a promising direction for reducing the limitations of rigid evidence retrieval for non-parametric attribution.

3.2 Citation Characteristics

The five citation characteristics focus on the nature and presentation of evidence.

3.2.1 Citation Modality

Although evidence-based text generation requires LLMs to generate text, the modality of the underlying evidence can differ. We identify four citation modalities, with the most prevalent being **texts**, which covers all forms of unstructured textual data (Malaviya et al., 2024a). Studies also cite **graphs** as structured representations of entities and relations (He et al., 2025), **tables** as data organized in rows and columns (Mathur et al., 2024), and **visuals** such as images (Ma et al., 2025).

Takeaways. Text overwhelmingly dominates citation modalities in current work (96% of studies). Citing non-textual evidence remains largely unexplored, pointing to multimodality as an important direction for future research.

3.2.2 Evidence Level

Each citation modality can be cited at different levels of granularity, which we refer to as the evidence level. For *text*, cited evidence can correspond to a full **document**, such as a scientific article (Anand et al., 2023b), a **paragraph**, for example a retrieved chunk (Gao et al., 2023b), a single **sentence** (Xu et al., 2025), or even individual **tokens** (Phukan et al., 2024). For *graphs*, evidence is cited at the level of a **triple** (Li et al., 2024f), while no approach cites entire knowledge graphs. The *tables* modality includes evidence at the level of a **table** (Suri et al., 2025) or individual **table cells** (Mathur et al., 2024). Similarly, *visual* evidence is cited either at the level of an **image** (Suri et al., 2025) or a **bounding box** (Ma et al., 2025).

Takeaways. The evidence level is dominated by coarse-grained texts, with document- and paragraph-level evidence accounting for the majority of studies (43% and 40%). As shown in Figure 8 in Appendix, these levels emerged earlier and continue to dominate the literature, while finer-grained evidence such as sentences and tokens remain less prevalent (12% and 2%), but exhibit stronger growth. Evidence levels for non-textual modalities remain underexplored, limiting conclusions about their granularity.

3.2.3 Citation Style

As illustrated in Figure 10 in Appendix, we identify six citation styles that differ in how evidence is presented to users. **In-line citations** place references directly after citation-worthy claims (Huang et al., 2024a). **Citation reports** provide a separate list of references alongside the LLM output (Bohnet et al., 2023). Some approaches display only the **passage** retrieved or used during generation, particularly in evaluation settings (Muller et al., 2023). **Narrative citations** integrate references into the natural flow of the generated text to improve contextual clarity (Shaier et al., 2024). **Highlight gradients** visually indicate source influence by coloring relevant tokens or sentences in the output and the supporting evidence (Do et al., 2024). Finally, **quotes** embed verbatim excerpts from the evidence directly into the generated response (Xiao et al., 2025).

Takeaways. In-line citations dominate current practice, appearing in 62% of studies (see Figure 9 in Appendix). For user-facing applications, citation style strongly affects verifiability: in-line citations enable claim-level verification of LLM-generated text, while styles such as highlight gradients or quotes additionally allow users to directly identify the supporting evidence spans.

3.2.4 Citation Visibility

Citation visibility determines whether citations are shown to users. Most approaches include citations in the **final response**, enabling users to trace claims to their sources (Gao et al., 2023b). In contrast, some approaches generate citations only in an **intermediate text**, where citations are used internally by the LLM rather than exposed for direct user traceability (Fang et al., 2024).

Takeaways. Citation visibility is predominantly user-facing, with 91% of studies providing citations in the final response. Intermediate citation generation is rare and appears only in question answering and grounded text generation.

3.2.5 Citation Frequency

Citation frequency captures the number of citations assigned to an LLM-generated claim. Existing approaches differ in whether they provide a **single** citation (Shaier et al., 2024) or **multiple** citations per claim (Khalifa et al., 2024).

Takeaways. Most studies support multiple citations per claim (64%), particularly in non-parametric attribution settings where retrieval naturally enables citing several sources. In contrast, parametric approaches are often constrained by model architecture and may support only single citations. This suggests an open design space around when multiple citations are beneficial versus when single citation strategies suffice.

3.3 Task

We identify six frequent tasks in evidence-based text generation with LLMs. **Question answering (QA)** (Gao et al., 2023b) evaluates an LLM's ability to answer questions by generating answers grounded in evidence. **Grounded text generation** (Cheng et al., 2025) captures more general generation settings, such as open-ended text or dialogue. **Summarization** (Deng et al., 2024) assesses whether generated summaries attribute source documents, while **fact verification** (Buchmann et al., 2024) focuses on determining the correctness of claims with respect to supporting evidence. Finally, **citation text generation** (Anand et al., 2023b) focuses on generating citation contexts in scientific writing where references are inserted into an existing manuscript. In contrast, **related work generation** (Byun et al., 2024) targets broader document-level synthesis, generating literature overviews that organize and relate prior works.

Takeaways. The task landscape of evidence-based text generation with LLMs reflects a gradual expansion from early tasks such as question answering and grounded text generation toward more specialized settings, including citation text generation and fact verification. However, since most approaches and evaluation practices were developed around these dominant tasks, they may inadequately capture the requirements of newer tasks that involve reasoning over multiple sources or operate at different evidence levels. Future research could examine how existing approaches and evaluation practices transfer across tasks, and where task-specific adaptations are required.

4 LLM Integration

This section provides a complementary analysis of how LLMs are operationalized in evidence-based text generation. In practice, LLMs can be integrated at multiple stages of a system beyond text generation, for example to support attribution, citation generation, or task-specific requirements. Because these integration mechanisms span multiple taxonomy dimensions, we analyze LLM integration separately to preserve conceptual clarity. Across the reviewed literature, we identify two complementary integration strategies: *training* and *prompting*. Training-based approaches modify model behavior through pretraining or fine-tuning, whereas prompting-based approaches guide the model at inference time through structured inputs.

4.1 Training

We annotate whether and how studies modify LLMs through training, distinguishing between pretraining and fine-tuning. Few approaches employ **pretraining** either to incorporate attribution-specific objectives or to adapt models to broader settings. Several works treat pretraining as a necessary component for parametric attribution (Khalifa et al., 2024; Lu et al., 2025), while others leverage it to support multilingual or multimodal scenarios (Abbas et al., 2025; Patel et al., 2024). More commonly, studies apply fine-tuning to adapt LLMs to evidence-based text generation tasks. Fine-tuning is used to improve generation quality and attribution behavior, most often through **supervised fine-tuning** (Li et al., 2024b; Ye et al., 2024). In contrast, **self-supervised fine-tuning** (Huang et al., 2024c; Chen et al., 2023) and **reinforcement learning** (Huang et al., 2024a,b) appear less frequently.

Takeaways. Training-based integration is used selectively in evidence-based text generation (45% of studies). It is most common in model- and data-centric attribution, where attribution behavior is embedded in the LLM through fine-tuning or pre-training. Overall, fine-tuning primarily improves attribution quality but is also used for task adaptation when prompting alone is insufficient, indicating that training serves targeted methodological needs rather than a universal solution. This selective use highlights open questions about the limits of prompting, specifically which attribution behaviors can be achieved through prompting alone and which require parameter-level adaptation.

4.2 Prompting

In practice, the majority of studies rely on prompting to steer LLMs toward evidence-based text generation without modifying LLM parameters. Common prompting techniques include **zero-shot**, **few-shot**, and **chain-of-thought** prompting, which are used to provide task instructions, in-context examples, or explicit reasoning steps (Tahaei et al., 2024; Ateia and Kruschwitz, 2025; Li et al., 2024d). These techniques are frequently combined, for example by pairing few-shot demonstrations with explicit reasoning steps (Shaier et al., 2024). Beyond general-purpose prompting, several studies introduce strategies designed to improve citation behavior. These include **chain-of-citation** and **chain-of-quote** prompting (Li et al., 2024g), which encourage explicit alignment between reasoning steps and cited evidence, as well as **conflict-aware** prompting (Patel and Anand, 2024). Additional prompting strategies are summarized in Table 7 in Appendix.

Takeaways. Prompting is the predominant approach for operationalizing LLMs in evidence-based text generation (78% of studies). Most approaches rely on standard prompting strategies to guide the model behavior without parameter updates, reflecting the flexibility and low overhead of inference-time control. More specialized prompting strategies explicitly target attribution quality. Overall, prompting remains the state of the art for non-parametric attribution. The diversity of prompting strategies highlights the need for standardized templates in this domain.

5 Evaluation

This section provides an overview of evaluation approaches for evidence-based text generation with LLMs. In total, we identified 300 distinct metrics, each targeting different aspects of evaluation. Figure 4 offers a structured overview of frequently reused metrics categorized by *evaluation method* and *evaluation dimension*. We define *reused* resources as those employed at least twice among surveyed studies. We observe that only two frameworks are reused across studies, namely ALCE (Gao et al., 2023b) and G-Eval (Liu et al., 2023b). Overall, our survey comprises 19 frameworks (Table 10), 11 benchmarks (Table 11), and 231 datasets (Table 12), detailed in Appendix E.

5.1 Evaluation Methods

The evaluation methods characterize the underlying strategies used to compute evaluation metric scores. Our initial categorization was derived from Dziri et al. (2022) and iteratively refined and extended during data annotation. **Human evaluation** relies on human judges who rate LLM-generated texts along predefined criteria. **Inference-based** metrics use natural language inference (NLI) models to assess whether an LLM-generated text is entailed by a reference, while **lexical overlap** metrics focus on surface-level word matching. **LLM-as-a-judge** metrics automatically assess the quality of texts generated by other LLMs. **Retrieval-based** metrics evaluate how effectively relevant evidence is incorporated by comparing retrieved evidence with ground truth. **Semantic similarity-based** metrics assess the similarity between generated and reference texts using dense vector embeddings. More details are described in Table 8 in Appendix E.

5.2 Evaluation Dimensions

Evaluation dimensions specify which aspects of evidence-based text generation are evaluated.

Attribution. These metrics evaluate whether an LLM-generated output can be attributed to sources, without requiring citations or assessing factual correctness. Attribution is among the most critical and widely studied evaluation dimensions and is predominantly assessed via inference-based methods. The *Citation NLI* metric cluster (Gao et al., 2023b) uses NLI models to determine whether LLM outputs are supported by sources and reports *precision*, *recall*, and F_1 scores. Related inference-based metrics include *Auto-AIS-sentence* (Gao et al., 2023a), which extends AIS (Rashkin et al., 2023) to estimate the proportion of attributable sentences, and *FActScore* (Min et al., 2023), which decomposes outputs into atomic facts and verifies their

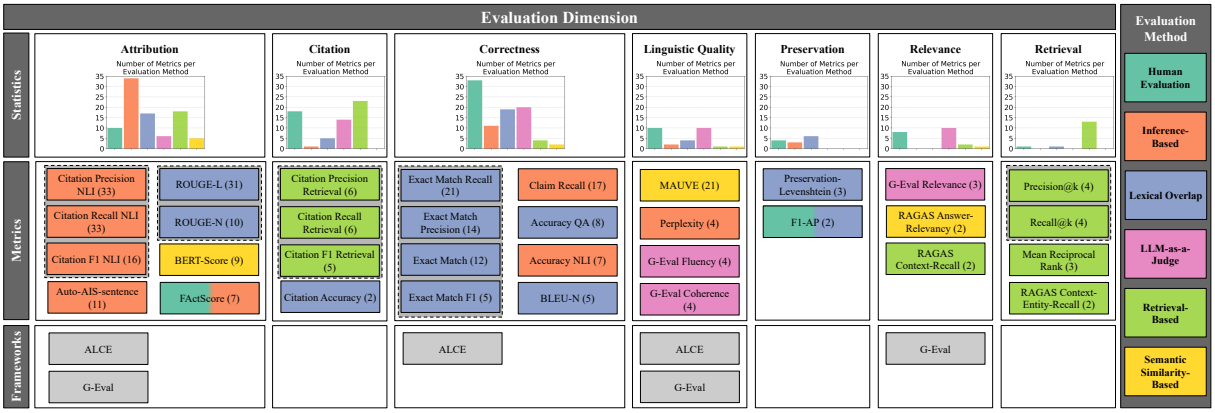

Figure 4: Frequently reused evaluation metrics and frameworks for evidence-based text generation. Numbers in parentheses indicate how many studies used each metric. Metrics grouped by dashed lines represent complementary metrics, recommended being used together. Additional, less reused metrics are listed in Appendix E.3.

support via NLI or human annotation. Lexical overlap metrics such as the *ROUGE* cluster (Lin, 2004) assess attribution by measuring surface-level overlap between generated text and evidence, while semantic similarity–based metrics, including *BERTScore* (Zhang et al., 2020), compare dense embeddings of LLM texts and references.

Citation. Metrics in this dimension assess whether an LLM-generated output cites appropriate evidence. Unlike attribution metrics, which focus on whether content can be traced to evidence, citation metrics emphasize the correctness and completeness of citations and are predominantly evaluated using retrieval-based methods. The most frequently adopted approach is the *Citation Retrieval* metric cluster (Deng et al., 2024). These metrics compare cited sources in LLMs-generated texts against a gold standard to assess whether LLM-generated citations support the content and are cited appropriately, reporting *precision*, *recall*, and F_1 scores. In addition, *Citation Accuracy* (Shaier et al., 2024) evaluates whether citation strings in LLM outputs correspond to correct sources by matching them against a gold standard.

Correctness. This dimension evaluates whether an LLM-generated text is semantically accurate, such as generating answers without hallucinations or summaries without omitting key content. Correctness is predominantly assessed through human evaluation, reflecting limitations of automated metrics in capturing semantic accuracy. However, due to their heterogeneous implementation, human evaluation metrics are not frequently reported in Figure 4. Among automated metrics, lexical overlap is widely used. The *Exact Match* cluster, including *recall*, *precision*, F_1, and *accuracy*, is used in factoid-style QA (Gao et al., 2023b), while *BLEU-N* measures n-gram overlap with ground-truth texts (Papineni et al., 2002). In multiple-choice QA settings, correctness is evaluated using *Accuracy QA* (Chu et al., 2025). For long-form generation, inference-based metrics and LLM-as-a-judge approaches are increasingly employed, including *Claim Recall*, which uses NLI models to verify whether generated outputs entail gold claims (Gao et al., 2023b), and *Accuracy NLI*, which assesses factual consistency between texts (Malaviya et al., 2024a).

Linguistic Quality. These metrics capture aspects such as fluency, referring to the naturalness and grammatical correctness of text, and coherence, which reflects the logical flow and consistency of ideas. Most metrics rely on human evaluation or the LLM-as-a-judge paradigm, including *G-Eval Fluency* and *G-Eval Coherence* (Liu et al., 2023b). In addition, *MAUVE* (Pillutla et al., 2021) is a widely used semantic similarity–based metric that assesses fluency and coherence by comparing distributions of generated and reference texts. The inference-based metric *Perplexity* (Liu et al., 2021) measures a model's uncertainty in next-token prediction, with lower values typically indicating more fluent text.

Preservation. Introduced by Gao et al. (2023a) in the context of post-retrieval attribution, preservation evaluates how much a revised output y retains from the original LLM-generated text x after incorporating retrieved evidence. The evaluation focuses on limiting unnecessary changes during revision. We identify two frequently used preservation metrics. *Preservation Levenshtein* (Gao et al., 2023a) is a lexical overlap metric that penalizes

Evaluation Dimension	When to Use	Explanation
Core		
Attribution	When no annotated evidence is available	Attribution is used when gold evidence is not provided (e.g., question answering with only ground-truth questions and answers). LLM-generated texts are compared against retrieved sources to determine whether claims can be supported by external evidence.
Citation	When annotated ground-truth evidence is available	Applicable when evidence is explicitly annotated (e.g., question–answer–evidence triples), enabling direct evaluation of whether the approach cites correct sources. In such settings, citation can be evaluated instead of, or in addition to, attribution.
Correctness	Always	Assesses factual accuracy of generated content. Correctness is fundamental to evidence-based text generation with LLMs and should be evaluated regardless of task, system design, or evidence availability.
Contextual		
Linguistic Quality	When the LLM is modified, or when linguistic quality is critical for the use case	Linguistic quality is relevant when the LLM generation process is altered (e.g., through pretraining or fine-tuning) or when high-quality language is essential for the application.
Preservation	When a post-generation attribution approach is applied	Preservation measures how much revised outputs deviate from the original generation. This dimension is particularly important for post-generation attribution approaches, which use external evidence to revise text generated by an LLM.
Relevance	When the approach is deployed in user-centric settings	Evaluates how well the LLM-generated output aligns with the user query or task requirements. In user-centric settings, relevance directly affects perceived utility and helpfulness of an LLM output.
Retrieval	When a non-parametric attribution approach is used and its quality depends on retrieval performance	Retrieval assesses the performance of the retrieval system, which strongly influences downstream attribution quality.

Table 1: Evaluation guidelines for evidence-based text generation with LLMs. Not all approaches need to be assessed across all evaluation dimensions. We distinguish between core evaluation dimensions, which should be evaluated, and contextual evaluation dimensions, whose applicability depends on the task and system design.

character-level edits between x and y. *F1-AP* (Gao et al., 2023a) combines attribution and preservation by computing a harmonic mean that incorporates AIS-sentence with preservation signals.

Relevance. This dimension evaluates how well an LLM-generated text aligns with the user query or task, commonly associated with utility or helpfulness. Relevance is frequently assessed using LLM-as-a-judge approaches, such as *G-Eval Relevance* (Liu et al., 2023b). In addition, the RAGAS framework (Es et al., 2024) introduces relevance-focused metrics. *RAGAS Answer-Relevancy* computes the semantic similarity between the input question and questions automatically generated from the answer, while *RAGAS Context-Recall* measures how well the retrieved context reflects the ground-truth answer, focusing on whether key information is retrieved and incorporated.

Retrieval. In non-parametric attribution settings, retrieval assesses whether relevant evidence is provided to the LLM. As expected, evaluation relies on retrieval-based metrics. Frequent metrics include *Precision@k* and *Recall@k*, which assess relevance and coverage of retrieved documents (Ramu et al., 2024), as well as *Mean Reciprocal Rank*, which captures the rank position of the first relevant item (Xiao et al., 2025). In addition, *RAGAS Context-Entity-Recall* evaluates entity-level recall by computing the fraction of ground-truth entities present in the retrieved context (Es et al., 2024).

5.3 Evaluation Guidelines

In Table 1, we outline guidelines for evaluating evidence-based text generation with LLMs. Although we identify seven evaluation dimensions, their relevance varies across approaches. We define *core* evaluation dimensions, namely attribution or citation, which depend on evidence availability, and correctness, which should always be evaluated. In addition, *contextual* dimensions include linguistic quality, preservation, relevance, and retrieval, whose applicability depends on the task, system design, and application setting. This distinction ensures that core aspects of evidence-based text generation with LLMs are consistently assessed, promoting more standardized evaluation practices.

For the core evaluation dimensions, we provide a comparative overview of frequently reused evaluation metrics in Table 9 in the appendix, which summarizes their measured aspects, applicability, and limitations to support informed metric selection. Additional evaluation metrics are discussed in Appendix E.3, while task-specific analyses and trends of evaluation dimensions and methods are presented in Appendix E.1 and Appendix E.2.

Takeaways. Evaluation of evidence-based text generation with LLMs is inherently multidimensional, yet current practices often assess dimensions in isolation. While not every approach requires evaluation along all dimensions, their interactions are critical for meaningful assessment, and evaluation choices implicitly prioritize certain LLM behaviors. For correctness in long-form text generation, there is a fundamental trade-off between scalability and semantic coverage. Human evaluation remains dominant due to the difficulty of capturing nuanced factual errors. Automated metrics are more scalable but capture only partial signals: lexical and inference-based methods rely on surface or entailment cues, while LLM-as-a-judge approaches extend coverage but require careful alignment with human judgments. Consequently, current automated correctness metrics provide only indicative signals, leaving substantial room for future research.

6 Discussion and Future Directions

Evidence-based text generation with LLMs has emerged as a rapidly growing research area, with 75% of surveyed works published after 2023. Despite this progress, key limitations identified by Huang and Chang (2024) remain largely unresolved. Building on these insights, we synthesize our findings to highlight four central limitations that define important directions for future research.

Parametric and Hybrid Attribution. Parametric attribution remains an open and underexplored challenge, with existing model- and data-centric approaches facing significant limitations in scalability and generalizability. Hybrid attribution, which combines parametric and non-parametric signals, offers a pragmatic pathway to improve attribution even when parametric methods remain imperfect. By allowing partial or coarse parametric signals to complement retrieved evidence, hybrid approaches can help surface the boundaries of model-internal knowledge, mitigate retrieval limitations, and provide richer attribution signals than either approach alone. At the same time, advances in parametric attribution remain critical for strengthening hybrid methods, suggesting a co-evolution rather than a strict sequential dependency between the two.

Evaluation Standards. Despite 300 identified evaluation metrics, only two frameworks and two benchmarks are frequently reused among studies. This emphasizes the urgent need for standardized evaluation frameworks to enable fair and consistent comparison across methods. Future work should ensure these frameworks are adaptable to the diverse tasks within evidence-based text generation with LLMs and comprehensively cover all seven evaluation dimensions outlined in this survey. Additionally, the wide variation in human evaluation, with many studies introducing unique and non-standardized metrics, underscores the need for automated and scalable evaluation approaches.

Explainable Citations. The citation behavior of LLMs remains underexplored. These models may exhibit citation-related biases similar to those seen in human authorship. To improve explainability of LLM-generated texts, users should understand why LLMs select a source from multiple candidates. Transparent citation reasoning is a prerequisite to identify biases and increase trust. Future work should systematically analyze the citation behavior of current approaches and improve the explainability of citation reasoning.

Multimodal Evidence. Evidence-based text generation with LLMs remains text-centric, with text constituting the citation modality in 96% of studies, indicating that multimodal evidence is largely underexplored. A key open challenge is combining evidence across modalities and evidence levels, such as integrating visual evidence, tables, and graphs within a single response. This requires determining how heterogeneous evidence supports shared claims, and how it should be selected, weighted, and cited in LLMs outputs.

7 Conclusion

We surveyed 134 papers on evidence-based text generation with LLMs, introduced a novel taxonomy, categorized 300 evaluation metrics into seven dimensions and six methods and extracted 231 unique datasets. The field is still fragmented with varied terminology and evaluation standards. Our synthesis offers a clear reference point for advancing the reliability and verifiability of LLMs.

Limitations

Even though we followed a systematic study design, including a thorough keyword-based literature search, careful inclusion classification, and iterative refinement of paper annotation, certain limitations apply to our study. The restriction to a single primary search string may have led to the omission of some relevant studies. To assess this risk, we conducted a sensitivity analysis using an expanded set of adjacent search terms, which identified only a small additional proportion of relevant studies (4%). This indicates that the risk from using a single primary search string is small, and that the original search strategy achieved strong coverage within the major sources considered. Maintaining a manageable number of retrieved studies remains indispensable for manual inclusion screening, particularly since current LLM-based approaches remain limited compared to human expertise. To further mitigate incomplete literature coverage, we selected nine widely used databases that collectively cover a substantial portion of the examined field. We also tested multiple search strings against an initial literature sample to optimize retrieval completeness while minimizing overall retrieval volume. For more details, see Appendix B.

The inclusion screening and categorization of relevant studies inherently involve some degree of subjectivity, as both rely on the judgments of the researchers. While each decision was informed by the authors' domain expertise, this process may still introduce bias. To mitigate these limitations, we established explicit inclusion criteria, resolved disagreements through discussion until consensus was reached, and measured inter-annotator agreement to ensure consistent application of the annotation schema. Additionally, the annotation schema was regularly reviewed and refined to maintain a shared and consistent understanding of each category.

Finally, certain coverage biases may remain. Our restriction to English-language studies may underrepresent non-English research and regional venues, although English is the dominant language in this field. Moreover, focusing on publicly accessible sources may underrepresent non-public or industry-internal work. Including arXiv helped capture recent developments and reflects common dissemination practices in rapidly evolving research areas, but does not address the limited visibility of proprietary research. These factors should be considered when interpreting our findings.

Acknowledgments

We thank AFM Mohimenul Joaa and Kyuri Im for their efforts during our data collection.

The authors acknowledge the financial support by the Federal Ministry of Research, Technology and Space of Germany (BMFTR) and by Sächsische Staatsministerium für Wissenschaft, Kultur und Tourismus in the programme Center of Excellence for AI-research „Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig", project identification number: ScaDS.AI.

Tobias Schreieder is supported by the BMFTR through a Software Campus project, project identification number: 16lS23070.

Tim Schopf is supported by a scholarship of the German Academic Exchange Service (DAAD).

We used AI-based assistance tools to support language editing, minor formatting, and coding tasks. These tools did not contribute to the intellectual content or scientific conclusions. All content was reviewed by the authors, who assume full responsibility for the publication.

References

Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, and 22 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Amin Abolghasemi, Leif Azzopardi, Seyyed Hadi Hashemi, Maarten de Rijke, and Suzan Verberne. 2025. Evaluation of attribution bias in generator-aware retrieval-augmented large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21105–21124, Vienna, Austria. Association for Computational Linguistics.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy Dj Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025. LitLLMs, LLMs for literature review: Are we there yet? *Transactions on Machine Learning Research*.

Mia Allen, Usman Naeem, and Sukhpal Singh Gill. 2024. Q-module-bot: A generative ai-based question and answer bot for module teaching support. *IEEE Transactions on Education*, 67(5):793–802.

Rami Aly, Zhiqiang Tang, Samson Tan, and George Karypis. 2024. Learning to generate answers with

citations via factual consistency models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11876–11896, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. 2023. QAMPARI: A benchmark for open-domain questions with many answers. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 97–110, Singapore. Association for Computational Linguistics.

Avinash Anand, Mohit Gupta, Kritarth Prasad, Ujjwal Goel, Naman Lal, Astha Verma, and Rajiv Ratn Shah. 2023a. Kg-ctg: Citation generation through knowledge graph-guided large language models. In *Big Data and Artificial Intelligence: 11th International Conference, BDA 2023, Delhi, India, December 7–9, 2023, Proceedings*, page 37–49, Berlin, Heidelberg. Springer-Verlag.

Avinash Anand, Kritarth Prasad, Ujjwal Goel, Mohit Gupta, Naman Lal, Astha Verma, and Rajiv Ratn Shah. 2023b. Context-enhanced language models for generating multi-paper citations. In *Big Data and Artificial Intelligence: 11th International Conference, BDA 2023, Delhi, India, December 7–9, 2023, Proceedings*, page 80–94, Berlin, Heidelberg. Springer-Verlag.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024a. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *Preprint*, arXiv:2411.14199.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024b. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Samy Ateia and Udo Kruschwitz. 2025. Bioragent: A retrieval-augmented generation system for showcasing generative query expansion and domain-specific search for scientific q&a. In *Advances in Information Retrieval*, pages 1–5, Cham. Springer Nature Switzerland.

Moshe Berchansky, Daniel Fleischer, Moshe Wasserblat, and Peter Izsak. 2024. CoTAR: Chain-of-thought attribution reasoning with multi-level granularity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 236–246, Miami, Florida, USA. Association for Computational Linguistics.

Nitzan Bitton-Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szpektor, Amir Globerson, and Yuval Elovici. 2024. Visual riddles: a commonsense and world knowledge challenge for large vision and language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, and 3 others. 2023. Attributed question answering: Evaluation and modeling for attributed large language models. *Preprint*, arXiv:2212.08037.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jan Buchmann, Xiao Liu, and Iryna Gurevych. 2024. Attribute or abstain: Large language models as long document assistants. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8113–8140, Miami, Florida, USA. Association for Computational Linguistics.

Courtni Byun, Piper Vasicek, and Kevin Seppi. 2024. This reference does not exist: An exploration of LLM citation accuracy and relevance. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 28–39, Mexico City, Mexico. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2024. Verifiable generation with subsentence-level fine-grained citations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15584–15596, Bangkok, Thailand. Association for Computational Linguistics.

Arie Cattan, Alon Jacovi, Alex Fabrikant, Jonathan Herzig, Roee Aharoni, Hannah Rashkin, Dror Marcus, Avinatan Hassidim, Yossi Matias, Idan Szpektor, and Avi Caciularu. 2025. Doubledipper: Improving long-context llms via context recycling. *Preprint*, arXiv:2406.13632.

Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. 2025. Scalable influence and fact tracing for large language model pre-

training. In *The Thirteenth International Conference on Learning Representations*.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *Preprint*, arXiv:2305.14908.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya Sakai, Ji-Rong Wen, and Zhicheng Dou. 2025. CORAL: Benchmarking multi-turn conversational retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1308–1330, Albuquerque, New Mexico. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2024. Merging facts, crafting fallacies: Evaluating the contradictory nature of aggregated factual claims in long-form generations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2734–2751, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Chu, Jingchang Chen, Zhongjie Wang, Guo Tang, Qianglong Chen, Ming Liu, and Bing Qin. 2025. Towards faithful multi-step reasoning through fine-grained causal-aware attribution reasoning distillation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2291–2315, Abu Dhabi, UAE. Association for Computational Linguistics.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Douglas B. Craig and Sorin Drăghici. 2024. What's the data say? an llm-based system for interrogating experimental data. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1457–1462.

Kevin Crowston and Barbara H. Kwasnik. 2004. A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 4 - Volume 4*, HICSS '04, page 40100.1, USA. IEEE Computer Society.

Meghal Dani, Muthu Jeyanthi Prakash, Zeynep Akata, and Stefanie Liebe. 2024. SemioLLM: Assessing large language models for semiological analysis in epilepsy research. In *ICML 2024 AI for Science Workshop*.

Mohammad Dehghan, Mohammad Alomrani, Sunyam Bagga, David Alfonso-Hermelo, Khalil Bibi, Abbas Ghaddar, Yingxue Zhang, Xiaoguang Li, Jianye Hao, Qun Liu, Jimmy Lin, Boxing Chen, Prasanna Parthasarathi, Mahdi Biparva, and Mehdi Rezagholizadeh. 2024. EWEK-QA : Enhanced web and efficient knowledge graph retrieval for citation-based question answering systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14169–14187, Bangkok, Thailand. Association for Computational Linguistics.

Haolin Deng, Chang Wang, Li Xin, Dezhang Yuan, Junlang Zhan, Tian Zhou, Jin Ma, Jun Gao, and Ruifeng Xu. 2024. WebCiteS: Attributed query-focused summarization on Chinese web search results with citations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15095–15114, Bangkok, Thailand. Association for Computational Linguistics.

Peter Devine. 2025. Aloftrag: Automatic local fine tuning for retrieval augmented generation. *Preprint*, arXiv:2501.11929.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Yifan Ding, Matthew Facciani, Ellen Joyce, Amrit Poudel, Sanmitra Bhattacharya, Balaji Veeramani, Sal Aguinaga, and Tim Weninger. 2025. Citations and trust in llm generated responses. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Hanane Djeddal, Pierre Erbacher, Raouf Toukal, Laure Soulier, Karen Pinel-Sauvagnat, Sophia Katrenko, and Lynda Tamine. 2024. An evaluation framework for attributed information retrieval using large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 5354–5359, New York, NY, USA. Association for Computing Machinery.

Hyo Jin Do, Rachel Ostrand, Justin D. Weisz, Casey Dugan, Prasanna Sattigeri, Dennis Wei, Keerthiram Murugesan, and Werner Geyer. 2024. Facilitating human-llm collaboration through factuality scores and source attributions. *Preprint*, arXiv:2405.20434.

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

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. 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.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Yihao Fang, Stephen Thomas, and Xiaodan Zhu. 2024. HGOT: Hierarchical graph of thoughts for retrieval-augmented in-context learning in factuality evaluation. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 118–144, Mexico City, Mexico. Association for Computational Linguistics.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11397–11417, Bangkok, Thailand. Association for Computational Linguistics.

Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2024. Qlarify: Recursively expandable abstracts for dynamic information retrieval over scientific papers. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA. Association for Computing Machinery.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Aidan Gilson, Xuguang Ai, Thilaka Arunachalam, Ziyou Chen, Ki Xiong Cheong, Amisha Dave, Cameron Duic, Mercy Kibe, Annette Kaminaka, Minali Prasad, Fares Siddig, Maxwell Singer, Wendy Wong, Qiao Jin, Tiarnan D. L. Keenan, Xia Hu, Emily Y. Chew, Zhiyong Lu, Hua Xu, and 3 others. 2024. Enhancing large language models with domain-specific retrieval augment generation: A case study on long-form consumer health question answering in ophthalmology. *Preprint*, arXiv:2409.13902.

Lotem Golany, Filippo Galgani, Maya Mamo, Nimrod Parasol, Omer Vandsburger, Nadav Bar, and Ido Dagan. 2024. Efficient data generation for source-grounded information-seeking dialogs: A use case for meeting transcripts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1908–1925, Miami, Florida, USA. Association for Computational Linguistics.

K. Gopalakrishnan, B. Hedayatnia, Q. Chen, A. Gottardi, S. Kwatra, A. Venkatesh, R. Gabriel, and D. Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Nianlong Gu and Richard Hahnloser. 2024. Controllable citation sentence generation with language models. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 22–37, Bangkok, Thailand. Association for Computational Linguistics.

Deepak Gupta, Dina Demner-Fushman, William R. Hersh, Steven Bedrick, and Kirk Roberts. 2024. Overview of trec 2024 biomedical generative retrieval (biogen) track. *CoRR*, abs/2411.18069.

George Hannah, Rita T. Sousa, Ioannis Dasoulas, and Claudia d'Amato. 2025. On the legal implications of large language model answers: A prompt engineering approach and a view beyond by exploiting knowledge graphs. *Journal of Web Semantics*, 84:100843.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand. Association for Computational Linguistics.

Jie He, Yijun Yang, Wanqiu Long, Deyi Xiong, Victor Gutierrez Basulto, and Jeff Z. Pan. 2025. Evaluating and improving graph to text generation with large language models. 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 10219–10244, Albuquerque, New Mexico. Association for Computational Linguistics.

Lucas Torroba Hennigen, Zejiang Shen, Aniruddha Nrusimha, Bernhard Gapp, David Sontag, and Yoon Kim. 2024. Towards verifiable text generation with symbolic references. In *First Conference on Language Modeling*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Aliyah R. Hsu, James Zhu, Zhichao Wang, Bin Bi, Shubham Mehrotra, Shiva K. Pentyala, Katherine Tan, Xiang-Bo Mao, Roshanak Omrani, Sougata Chaudhuri, Regunathan Radhakrishnan, Sitaram Asur, Claire Na Cheng, and Bin Yu. 2025. Rate, explain and cite (rec): Enhanced explanation and attribution in automatic evaluation by large language models. *Preprint*, arXiv:2411.02448.

I-Hung Hsu, Zifeng Wang, Long Le, Lesly Miculicich, Nanyun Peng, Chen-Yu Lee, and Tomas Pfister. 2024. CaLM: Contrasting large and small language models to verify grounded generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12782–12803, Bangkok, Thailand. Association for Computational Linguistics.

Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, and Dexter Pratt. 2025a. Evaluation of large language models for discovery of gene set function. *Nature Methods*, 22(1):82–91.

Nan Hu, Jiaoyan Chen, Yike Wu, Guilin Qi, Hongru Wang, Sheng Bi, Yongrui Chen, Tongtong Wu, and Jeff Z. Pan. 2025b. Can LLMs evaluate complex attribution in QA? automatic benchmarking using knowledge graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17096–17118, Vienna, Austria. Association for Computational Linguistics.

Chengyu Huang, Zeqiu Wu, Yushi Hu, and Wenya Wang. 2024a. Training language models to generate text with citations via fine-grained rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2926–2949, Bangkok, Thailand. Association for Computational Linguistics.

Jie Huang and Kevin Chang. 2024. Citation: A key to building responsible and accountable large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 464–473, Mexico City, Mexico. Association for Computational Linguistics.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Xiachong Feng, Weijiang Yu, Weihua Peng, Duyu Tang, Dandan Tu, and Bing Qin. 2024b. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14095–14113, Bangkok, Thailand. Association for Computational Linguistics.

Lei Huang, Xiaocheng Feng, Weitao Ma, Liang Zhao, Yuchun Fan, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024c. Advancing large language model attribution through self-improving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3822–3836, Miami, Florida, USA. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

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

Zhuoran Jin, Hongbang Yuan, Tianyi Men, Pengfei Cao, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2025. RAG-RewardBench: Benchmarking reward models in retrieval augmented generation for preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17061–17090, Vienna, Austria. Association for Computational Linguistics.

Kristiina Jokinen. 2024. The need for grounding in LLM-based dialogue systems. In *Proceedings of the*

Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024, pages 45–52, Torino, Italia. ELRA and ICCL.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Daniel Jung, Alex Butler, Joongheum Park, and Yair Saperstein. 2024. Evaluating the impact of a specialized llm on physician experience in clinical decision support: A comparison of ask avo and chatgpt-4. *Preprint*, arXiv:2409.15326.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *Preprint*, arXiv:2307.16883.

Hyeonsu B Kang, Tongshuang Wu, Joseph Chee Chang, and Aniket Kittur. 2023. Synergi: A mixed-initiative system for scholarly synthesis and sensemaking. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. 2024. Source-aware training enables knowledge attribution in language models. In *First Conference on Language Modeling*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Philippe Laban, Alexander Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context LLMs and RAG systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9885–9903, Miami, Florida, USA. Association for Computational Linguistics.

Dongyub Lee, Eunhwan Park, Hodong Lee, and Heuiseok Lim. 2024a. Ask, assess, and refine: Rectifying factual consistency and hallucination in LLMs with metric-guided feedback learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2422–2433, St. Julian's, Malta. Association for Computational Linguistics.

Dongyub Lee, Taesun Whang, Chanhee Lee, and Heuiseok Lim. 2023. Towards reliable and fluent large language models: Incorporating feedback learning loops in qa systems. *Preprint*, arXiv:2309.06384.

Hyunji Lee, Se June Joo, Chaeeun Kim, Joel Jang, Doyoung Kim, Kyoung-Woon On, and Minjoon Seo. 2024b. How well do large language models truly ground? 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 2437–2465, Mexico City, Mexico. Association for Computational Linguistics.

Seongmin Lee, Zijie J. Wang, Aishwarya Chakravarthy, Alec Helbling, ShengYun Peng, Mansi Phute, Duen Horng (Polo) Chau, and Minsuk Kahng. 2025. Llm attributor: Interactive visual attribution for llm generation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Aochong Oliver Li and Tanya Goyal. 2025. Memorization vs. reasoning: Updating llms with new knowledge. *Preprint*, arXiv:2504.12523.

Dongfang Li, Xinshuo Hu, Zetian Sun, Baotian Hu, Shaolin Ye, Zifei Shan, Qian Chen, and Min Zhang. 2024a. TruthReader: Towards trustworthy document assistant chatbot with reliable attribution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 89–100, Miami, Florida, USA. Association for Computational Linguistics.

Dongfang Li, Zetian Sun, Baotian Hu, Zhenyu Liu, Xinshuo Hu, Xuebo Liu, and Min Zhang. 2024b. Improving attributed text generation of large language models via preference learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5079–5101, Bangkok, Thailand. Association for Computational Linguistics.

Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. 2023a. A survey of large language models attribution. *Preprint*, arXiv:2311.03731.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on*

Empirical Methods in Natural Language Processing, pages 6449–6464, Singapore. Association for Computational Linguistics.

Junyi Li and Hwee Tou Ng. 2025. Think&cite: Improving attributed text generation with self-guided tree search and progress reward modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9928–9942, Vienna, Austria. Association for Computational Linguistics.

Minghan Li, Xilun Chen, Ari Holtzman, Beidi Chen, Jimmy Lin, Wen tau Yih, and Xi Victoria Lin. 2024c. Nearest neighbor speculative decoding for LLM generation and attribution. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. 2024d. Citation-enhanced generation for LLM-based chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1451–1466, Bangkok, Thailand. Association for Computational Linguistics.

Xiangci Li and Jessica Ouyang. 2025. Explaining relationships among research papers. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1080–1105, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2024e. LLatrieval: LLM-verified retrieval for verifiable 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 5453–5471, Mexico City, Mexico. Association for Computational Linguistics.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From matching to generation: A survey on generative information retrieval. *ACM Trans. Inf. Syst.*, 43(3).

Xinze Li, Yixin Cao, Liangming Pan, Yubo Ma, and Aixin Sun. 2024f. Towards verifiable generation: A benchmark for knowledge-aware language model attribution. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 493–516, Bangkok, Thailand. Association for Computational Linguistics.

Yanyang Li, Shuo Liang, Michael Lyu, and Liwei Wang. 2024g. Making long-context language models better multi-hop reasoners. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2462–2475, Bangkok, Thailand. Association for Computational Linguistics.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024h. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14919–14935, Bangkok, Thailand. Association for Computational Linguistics.

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

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

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

Yanting Liu, Tao Ji, Changzhi Sun, Yuanbin Wu, and Xiaoling Wang. 2024. Generation with dynamic vocabulary. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18931–18948, Miami, Florida, USA. Association for Computational Linguistics.

Yixin Liu, Graham Neubig, and John Wieting. 2021. On learning text style transfer with direct rewards. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4262–4273, Online. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Xinyang Lu, Jingtan Wang, Zitong Zhao, Zhongxiang Dai, Chuan-Sheng Foo, See-Kiong Ng, and Bryan Kian Hsiang Low. 2025. WASA: WAtermark-based source attribution for large language model-generated data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23791–23824, Vienna, Austria. Association for Computational Linguistics.

Xueguang Ma, Shengyao Zhuang, Bevan Koopman, Guido Zuccon, Wenhu Chen, and Jimmy Lin. 2025. VISA: Retrieval augmented generation with visual source attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30154–30169, Vienna, Austria. Association for Computational Linguistics.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2025. Hallucination-free? assessing the reliability of

leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242.

Himanshu Maheshwari, Sambaran Bandyopadhyay, Aparna Garimella, and Anandhavelu Natarajan. 2024. Presentations are not always linear! GNN meets LLM for text document-to-presentation transformation with attribution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15948–15962, Miami, Florida, USA. Association for Computational Linguistics.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024a. ExpertQA: Expert-curated questions and attributed answers. 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 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Chaitanya Malaviya, Subin Lee, Dan Roth, and Mark Yatskar. 2024b. What if you said that differently?: How explanation formats affect human feedback efficacy and user perception. 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 3046–3065, Mexico City, Mexico. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Puneet Mathur, Alexa Siu, Nedim Lipka, and Tong Sun. 2024. MATSA: Multi-agent table structure attribution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 250–258, Miami, Florida, USA. Association for Computational Linguistics.

James Mayfield, Eugene Yang, Dawn Lawrie, Sean MacAvaney, Paul McNamee, Douglas W. Oard, Luca Soldaini, Ian Soboroff, Orion Weller, Efsun Kayi, Kate Sanders, Marc Mason, and Noah Hibbler. 2024. On the evaluation of machine-generated reports. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1904–1915, New York, NY, USA. Association for Computing Machinery.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes. *Preprint*, arXiv:2203.11147.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1211–1228, New York, NY, USA. Association for Computing Machinery.

Benjamin Muller, John Wieting, Jonathan Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. Evaluating and modeling attribution for cross-lingual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 144–157, Singapore. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Yixin Mao, and Chien-Sheng Wu. 2025. Search engines in the ai era: A qualitative understanding to the false promise of factual and verifiable source-cited responses in llm-based search. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 1325–1340, New York, NY, USA. Association for Computing Machinery.

Kazuya Nishimura, Kuniaki Saito, Tosho Hirasawa, and Yoshitaka Ushiku. 2024. Toward structured related work generation with novelty statements. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 38–57, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and

262 others. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, and 7 others. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maya Patel and Aditi Anand. 2024. Factuality or fiction? benchmarking modern llms on ambiguous qa with citations. *Preprint*, arXiv:2412.18051.

Nilay Patel, Shivashankar Subramanian, Siddhant Garg, Pratyay Banerjee, and Amita Misra. 2024. Towards improved multi-source attribution for long-form answer 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 3906–3919, Mexico City, Mexico. Association for Computational Linguistics.

Vinzent Penzkofer and Timo Baumann. 2024. Evaluating and fine-tuning retrieval-augmented language models to generate text with accurate citations. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 57–64, Vienna, Austria. Association for Computational Linguistics.

Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. 2008. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*, pages 1–10.

Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. 2024. Peering into the mind of language models: An approach for attribution in contextual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11481–11495, Bangkok, Thailand. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems*.

David Pride, Matteo Cancellieri, and Petr Knoth. 2023. Core-gpt: Combining open access research and large language models for credible, trustworthy question answering. In *Linking Theory and Practice of Digital Libraries*, pages 146–159, Cham. Springer Nature Switzerland.

Shushanta Pudasaini, Luis Miralles-Pechuán, David Lillis, and Marisa Llorens Salvador. 2025. Survey on AI-Generated plagiarism detection: The impact of large language models on academic integrity. *Journal of Academic Ethics*, 23(3):1137–1170.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6037–6053, Miami, Florida, USA. Association for Computational Linguistics.

Haosheng Qian, Yixing Fan, Ruqing Zhang, and Jiafeng Guo. 2025. On the capacity of citation generation by large language models. In *Information Retrieval*, pages 109–123, Singapore. Springer Nature Singapore.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2024. Are large language model temporally grounded? 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 7064–7083, Mexico City, Mexico. Association for Computational Linguistics.

Sampath Rajapaksha, Ruby Rani, and Erisa Karafili. 2025. A rag-based question-answering solution for cyber-attack investigation and attribution. In *Computer Security. ESORICS 2024 International Workshops*, pages 238–256, Cham. Springer Nature Switzerland.

Pritika Ramu, Koustava Goswami, Apoorv Saxena, and Balaji Vasan Srinivasan. 2024. Enhancing post-hoc attributions in long document comprehension via coarse grained answer decomposition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17790–17806, Miami, Florida, USA. Association for Computational Linguistics.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David

Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Felicia Redelaar, Romy Van Drie, Suzan Verberne, and Maaike De Boer. 2024. Attributed question answering for preconditions in the Dutch law. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 154–165, Miami, FL, USA. Association for Computational Linguistics.

Joel Rorseth, Parke Godfrey, Lukasz Golab, Divesh Srivastava, and Jaroslaw Szlichta. 2024. Towards explainability in retrieval-augmented llms. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 5669–5670.

Rishiraj Saha Roy, Joel Schlotthauer, Chris Hinze, Andreas Foltyn, Luzian Hahn, and Fabian Kuech. 2025. Evidence contextualization and counterfactual attribution for conversational qa over heterogeneous data with rag systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25, page 1040–1043, New York, NY, USA. Association for Computing Machinery.

Furkan Şahinuç, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2024. Systematic task exploration with LLMs: A study in citation text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4832–4855, Bangkok, Thailand. Association for Computational Linguistics.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA. Association for Computational Linguistics.

Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. Towards faithful and robust LLM specialists for evidence-based question-answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1931, Bangkok, Thailand. Association for Computational Linguistics.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.

Tal Schuster, Adam Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William Cohen, and Donald Metzler. 2024. SEMQA: Semi-extractive multi-source question answering. 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 1363–1381, Mexico City, Mexico. Association for Computational Linguistics.

Sagi Shaier, Ari Kobren, and Philip V. Ogren. 2024. Adaptive question answering: Enhancing language model proficiency for addressing knowledge conflicts with source citations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17226–17239, Miami, Florida, USA. Association for Computational Linguistics.

Kartik Sharma, Peeyush Kumar, and Yunqing Li. 2025. OG-RAG: Ontology-grounded retrieval-augmented generation for large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32950–32969, Suzhou, China. Association for Computational Linguistics.

Jiajun Shen, Tong Zhou, Yubo Chen, and Kang Liu. 2024. Citekit: A modular toolkit for large language model citation generation. *Preprint*, arXiv:2408.04662.

Kaushal Shetty, Santosh Kumar Bojanki, and Adwait Ratnaparkhi. 2024. Sovereign risk summarization. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 779–786, New York, NY, USA. Association for Computing Machinery.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.

Karthik Soman, Andrew Langdon, Catalina Villouta, Chinmay Agrawal, Lashaw Salta, Braian Peetoom, Gianmarco Bellucci, and Orion J Buske. 2024. Zebra-llama: A context-aware large language model for democratizing rare disease knowledge. *Preprint*, arXiv:2411.02657.

Maojia Song, Shang Hong Sim, Rishabh Bhardwaj, Hai Leong Chieu, Navonil Majumder, and Soujanya Poria. 2025. Measuring and enhancing trustworthiness of LLMs in RAG through grounded attributions and learning to refuse. In *The Thirteenth International Conference on Learning Representations*.

Dirk H. R. Spennemann. 2025. The origins and veracity of references 'cited' by generative artificial intelligence applications: Implications for the quality of responses. *Publications*, 13(1).

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yize Sui, Jing Ren, Huibin Tan, Huan Chen, Zhaoye Li, and Ji Wang. 2024. Enhancing llm's reliability by iterative verification attributions with keyword fronting. In *Machine Learning and Knowledge Discovery in Databases. Research Track*, pages 251–268, Cham. Springer Nature Switzerland.

Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy X. R. Wang. 2021. D2S: Document-to-slide generation via query-based text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1418, Online. Association for Computational Linguistics.

Hao Sun, Hengyi Cai, Bo Wang, Yingyan Hou, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2024. Towards verifiable text generation with evolving memory and self-reflection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8211–8227, Miami, Florida, USA. Association for Computational Linguistics.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2025. VisDoM: Multi-document QA with visually rich elements using multimodal retrieval-augmented generation. 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 6088–6109, Albuquerque, New Mexico. Association for Computational Linguistics.

Marzieh Tahaei, Aref Jafari, Ahmad Rashid, David Alfonso-Hermelo, Khalil Bibi, Yimeng Wu, Ali Ghodsi, Boxing Chen, and Mehdi Rezagholizadeh. 2024. Efficient citer: Tuning large language models for enhanced answer quality and verification. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4443–4450, Mexico City, Mexico. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Deepa Tilwani, Revathy Venkataramanan, Amit P. Sheth, and Amit Sheth. 2024. Neurosymbolic ai approach to attribution in large language models. *IEEE Intelligent Systems*, 39(6):10–17.

Stavros Vassos, Stratos Goudelis, Dimi Balaouras, Giannis Vitalis, Vasilis Nakos, Glykeria Pigka, Loukia Tsagkli, Menia Hatzikou, Zachos Tsionas, Alexandros Chasanis, Stan van de Burgt, Mark Pors, Stratos Papadoudis, and Lefteris Loukas. 2024. Now i know! empowering voters with rag-enabled llms to eliminate political uncertainty. In *Proceedings of the 13th Hellenic Conference on Artificial Intelligence*, SETN '24, New York, NY, USA. Association for Computing Machinery.

Juraj Vladika, Luca Mülln, and Florian Matthes. 2024. Enhancing answer attribution for faithful text generation with large language models. *Preprint*, arXiv:2410.17112.

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 22–32, New York, NY, USA. Association for Computing Machinery.

Theodora Worledge, Tatsunori Hashimoto, and Carlos Guestrin. 2024a. The extractive-abstractive spectrum: Uncovering verifiability trade-offs in llm generations. *Preprint*, arXiv:2411.17375.

Theodora Worledge, Judy Hanwen Shen, Nicole Meister, Caleb Winston, and Carlos Guestrin. 2024b. Unifying corroborative and contributive attributions in large language models. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 665–683.

Jiayi Wu, Hengyi Cai, Lingyong Yan, Hao Sun, Xiang Li, Shuaiqiang Wang, Dawei Yin, and Ming Gao. 2025a. PA-RAG: RAG alignment via multi-perspective preference optimization. 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 9091–9112, Albuquerque, New Mexico. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025b. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.

Sirui Xia, Xintao Wang, Jiaqing Liang, Yifei Zhang, Weikang Zhou, Jiaji Deng, Fei Yu, and Yanghua Xiao. 2025a. Ground every sentence: Improving retrieval-augmented LLMs with interleaved reference-claim generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 969–988, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Jun Chen, and Haifeng Huang. 2025b. Improving retrieval augmented language model with self-reasoning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Jin Xiao, Bowei Zhang, Qianyu He, Jiaqing Liang, Feng Wei, Jinglei Chen, Zujie Liang, Deqing Yang, and Yanghua Xiao. 2025. Quill: Quotation generation enhancement of large language models. *Preprint*, arXiv:2411.03675.

Rui Xing, Timothy Baldwin, and Jey Han Lau. 2025. Evaluating evidence attribution in generated fact checking explanations. 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 5475–5496, Albuquerque, New Mexico. Association for Computational Linguistics.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2025. ALiiCE: Evaluating positional fine-grained citation generation. 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 545–561, Albuquerque, New Mexico. Association for Computational Linguistics.

Siqiao Xue, Fan Zhou, Yi Xu, Ming Jin, Qingsong Wen, Hongyan Hao, Qingyang Dai, Caigao Jiang, Hongyu Zhao, Shuo Xie, Jianshan He, James Zhang, and Hongyuan Mei. 2024. Weaverbird: Empowering financial decision-making with large language model, knowledge base, and search engine. *Preprint*, arXiv:2308.05361.

Zhichao Yan, Jiapu Wang, Jiaoyan Chen, Xiaoli Li, Jiye Liang, Ru Li, and Jeff Z. Pan. 2025. Atomic fact decomposition helps attributed question answering. *IEEE Transactions on Knowledge & Data Engineering*, 37(12):6959–6972.

Dongjie Yang, Ruifeng Yuan, Yuantao Fan, Yifei Yang, Zili Wang, Shusen Wang, and Hai Zhao. 2023. RefGPT: Dialogue generation of GPT, by GPT, and for GPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2511–2535, Singapore. Association for Computational Linguistics.

Xikang Yang, Xuehai Tang, Jizhong Han, and Songlin Hu. 2024. The dark side of trust: Authority citation-driven jailbreak attacks on large language models. *Preprint*, arXiv:2411.11407.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation 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 6237–6251, Mexico City, Mexico. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Comput. Surv.*, 54(11s).

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635, Singapore. Association for Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. 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 5628–5643, Mexico City, Mexico. Association for Computational Linguistics.

Jiajie Zhang, Yushi Bai, Xin Lv, Wanjun Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025a. LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5098–5122, Vienna, Austria. Association for Computational Linguistics.

Jiasheng Zhang, Ali Maatouk, Jialin Chen, Ngoc Bui, Qianqian Xie, Leandros Tassiulas, Hua Xu, Jie Shao, and Rex Ying. 2025b. Litfm: A retrieval augmented structure-aware foundation model for citation graphs. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 3728–3739, New York, NY, USA. Association for Computing Machinery.

Jiebin Zhang, Eugene J. Yu, Qinyu Chen, Chenhao Xiong, Dawei Zhu, Han Qian, Mingbo Song, Weimin Xiong, Xiaoguang Li, Qun Liu, and Sujian Li. 2025c. WIKIGENBENCH:exploring full-length Wikipedia generation under real-world scenario. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5191–5210, Abu Dhabi, UAE. Association for Computational Linguistics.

Kepu Zhang, Weijie Yu, Sunhao Dai, and Jun Xu. 2025d. CitaLaw: Enhancing LLM with citations in legal domain. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11183–11196, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Weijia Zhang, Mohammad Aliannejadi, Yifei Yuan, Jiahuan Pei, Jia-hong Huang, and Evangelos Kanoulas. 2024. Towards fine-grained citation evaluation in generated text: A comparative analysis of faithfulness metrics. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 427–439, Tokyo, Japan. Association for Computational Linguistics.

Yang Zhang, Yufei Wang, Kai Wang, Quan Z. Sheng, Lina Yao, Adnan Mahmood, Wei Emma Zhang, and Rongying Zhao. 2023a. When large language models meet citation: A survey. *Preprint*, arXiv:2309.09727.

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. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023c. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. Chatgpt hallucinates when attributing answers. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP '23, page 46–51, New York, NY, USA. Association for Computing Machinery.

A Data Availability

Our dataset, which includes all 134 annotated publications, 300 extracted evaluation metrics, and 231 datasets, is publicly available under the BSD 3-Clause license at `https://github.com/faerber-lab/AttributeCiteQuote`.

This dataset facilitates the reproduction of the experiments and analyses described in this study and serves as a foundation for future research.

B Methodology

To provide a comprehensive overview of the research landscape, we conducted a systematic mapping study following the guidelines outlined by Petersen et al. (2008). To ensure transparency and reproducibility in the literature review process, we report the identification, screening, and inclusion of studies using a PRISMA 2020 flow diagram (Page et al., 2021), shown in Figure 5. The main steps are detailed in the following subsections.

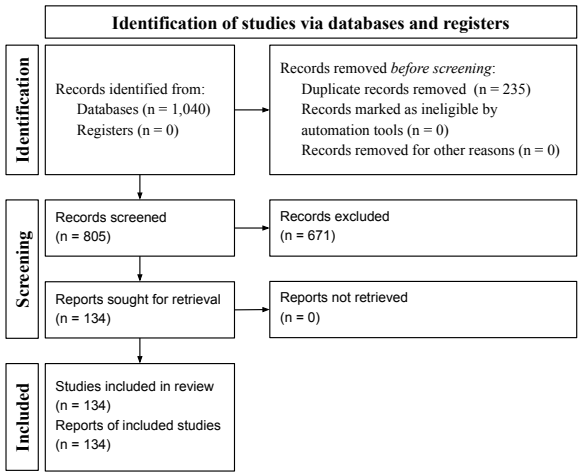

Figure 5: Systematic literature review process following the PRISMA protocol.

B.1 Literature Search and Inclusion Criteria

We identified targeted keywords to guide the literature search on evidence-based text generation with LLMs. To stay focused on the LLM-based paradigm while covering related concepts, we applied the following search string:

("large language model" OR "llm") AND ("citation" OR "attribution" OR "quote")

To identify the most appropriate keywords, we compiled an initial set of literature, including, but not limited to, the position paper by Huang and Chang (2024) and the survey papers by Li et al. (2023a) and Zhang et al. (2023a). Using this initial literature, we systematically evaluated 15 search strings to identify those offering both high recall (i.e., the proportion of papers from the initial list successfully retrieved) and a limited result set size (i.e., the number of retrieved papers). Controlling the result set size was critical to ensure that man-

ual screening of each candidate paper remained feasible. We deliberately decided against using LLM-based screening, which would have been necessary with a very high result set size, to ensure the quality and reliability of our annotations and the overall survey. Even minor changes to the search terms, such as applying stemming to the keywords "citation", "attribution" and "quote", led to a 1,207% increase in retrieved papers (exceeding 10,000 publications) with negligible recall improvement, and were therefore deemed unsuitable. Similarly, adding terms such as "reference", "source", or "evidence" did not yield further recall gains. Consequently, we adopted the keyword configuration above that achieved the optimal balance between recall and result set size.

The final search string was matched against the title and abstract of each publication. To ensure comprehensive coverage, we queried nine literature databases, following Schneider et al. (2022). The ACL Anthology is a primary repository for leading conferences and journals in natural language processing (NLP). The ACM Digital Library, IEEE Xplore, ScienceDirect, and Springer Nature cover key venues across the broader computer science domain. We also included arXiv to capture recent publications not yet in the queried databases, as well as relevant proceedings from major machine learning conferences.

In February 2025, we ran the search and initially identified 1,040 publications across all queried databases. After removing duplicate records, 805 unique publications remained and were used for screening. Subsequently, we used the following inclusion criteria to identify studies relevant to our research focus: (1) studies that aim to generate natural language text with LLMs, (2) studies that deliberately incorporate references to sources of evidence during text generation, and (3) studies that are written in English with the full texts electronically accessible. The accessibility requirement served to facilitate full-text screening and did not lead to the exclusion of any otherwise eligible publications. Publications not meeting all criteria were excluded from the final dataset. The first two authors screened all titles and abstracts, consulting full texts as needed. The screening workload was divided between the annotators, each reviewing a subset of publications. Both annotators are domain experts in computer science with relevant research experience. Each annotator flagged uncertain cases, which were subsequently discussed jointly until consensus was reached. This process resulted in 134 included publications, as shown in Table 2.

Literature Database	No. of Papers
ACL Anthology	54
ACM Digital Library	7
arXiv	59
ICML Proceedings	0
ICLR Proceedings	3
IEEE Xplore	4
NeurIPS Proceedings	3
ScienceDirect	0
Springer Nature	4
Total	**134**

Table 2: Overview of the queried literature databases and the number of studies included.

Inter-Annotator Agreement. To assess the reliability of the inclusion screening, both annotators independently labeled a random sample of 50 studies prior to discussion. Inter-annotator agreement was measured using Krippendorff's α and yielded a value of 1.0, indicating perfect agreement.

Sensitivity Analysis. To assess the robustness of our literature search, we conducted a sensitivity analysis using an expanded set of adjacent terms commonly used in the field:

("llm" OR "language model") AND (evidence-based" OR evidence linking" OR grounded generation" OR source-grounded" OR provenance tracing" OR source linking" OR verifiable generation")

The query was applied to the ACL Anthology and arXiv, which together cover 84% of the studies included in our review. The expanded search retrieved 126 studies. After deduplication with the initial literature corpus, 103 unique papers remained and were screened according to our protocol. This process identified five additional relevant studies. Overall, the results indicate that the original search strategy already captured the vast majority of relevant literature within these major sources, with the sensitivity search yielding only a small additional fraction of relevant studies (4%). The additional studies did not affect the structure of the proposed taxonomy. The five relevant but non-critical studies identified through this process are listed here for transparency (Schimanski et al., 2024; Hsu et al., 2024; Li et al., 2024e; Yue et al., 2024; Hennigen et al., 2024). These were not incorporated into the main corpus, as their inclusion did not affect the structure of the proposed taxonomy.

Contribution Type	Description
Approach	An approach consists of a set of novel methods, techniques, and procedures that need to be systematically executed to achieve a concrete goal.
Application	An application is a documented implementation of an existing approach, technique, or method in the form of a software library, prototype, or full application system.
Resource	A resource is a published dataset that supports approaches, techniques, methods, or applications, e.g., text corpora or benchmarks.
Evaluation	Evaluations of existing approaches, techniques, or methods as well as the introduction of new evaluation approaches including, e.g., new metrics or frameworks.
Survey	A survey analyses and synthesizes findings from multiple studies to systematically review a research field or gather evidence on a topic.
Position	A position paper presents a personal perspective on the suitability or direction of a specific research aspect, without presenting new empirical evidence.

Table 3: Categorization scheme for contribution types adapted from Schneider et al. (2022).

B.2 Categorization Scheme and Method

We categorized each publication by contribution type: approach, application, resource, evaluation, survey, and position. The detailed categorization scheme, adapted from Schneider et al. (2022), is presented in Table 3. Other dimensions are based on Huang and Chang (2024) and Li et al. (2023a), and were iteratively refined following the methodology of Petersen et al. (2008). All 134 studies were categorized along the defined dimensions, with annotation split between the first two authors. Regular calibration rounds refined the scheme and resolved labeling ambiguities. Each paper could be assigned multiple values per dimension if needed. Based on this categorization scheme, we derived a multidimensional taxonomy for evidence-based text generation with LLMs, which organizes the annotated studies along conceptually distinct facets. In addition, we annotated the publication date of each study, using the earliest available version such as preprints. This ensures that emerging trends can be analyzed as they appear, without the delay introduced by formal publication timelines.

B.3 Design of the Taxonomy

We adopt a multidimensional taxonomy based on a facetted classification approach, following the principles outlined by Crowston and Kwasnik (2004). In a facetted design, independent conceptual dimensions describe different aspects of the study, enabling flexible representation of complex approaches without enforcing artificial mutual exclusivity. This approach is well suited for evidence-based text generation with LLMs, where methods often combine multiple mechanisms that cannot be captured by a single hierarchical structure.

Our taxonomy consists of three dimensions that together characterize papers on evidence-based text generation with LLMs: (1) attribution approach, (2) citation characteristics, and (3) task. These dimensions represent distinct analytical perspectives on approach design choices. Since real-world approaches frequently combine mechanisms, the taxonomy allows multi-label assignments within a dimension, while each dimension captures a separate facet of the approach.

The **attribution approach** dimension describes how generated text is linked to supporting evidence. Within this dimension, parametric and non-parametric attribution form subdimensions, and approaches are further organized into finer-grained classes that capture different ways of tracing evidence back to LLM-generated text. The **citation characteristics** dimension describes what evidence is made available to users and how it is presented. This includes the modality, evidence level, style, visibility, and frequency of citations. The **task** dimension specifies the functional goal of the approach, such as QA, grounded text generation, summarization, related work generation, citation text generation, and fact verification.

By combining these dimensions, the multidimensional taxonomy supports fine-grained analysis of heterogeneous approaches and enables consistent comparison of approaches while acknowledging the multifaceted nature of evidence-based text generation with LLMs. The taxonomy in Figure 3 and the annotated dataset of 134 papers in Table 7 can be read from left to right, where each paper is categorized across all dimensions, with multi-label annotations when a paper spans multiple classes.

Inter-Annotator Agreement. To assess the reliability of the taxonomy annotation, both annotators independently labeled a random subset of 30 publications. Inter-annotator agreement was measured using Krippendorff's α. Given the multidi-

Taxonomy Dimension	Krippendorff's α
Contribution Type	0.73
Non-Parametric	0.82
Citation Modality	0.79
Evidence Level	0.83
Citation Style	0.77
Citation Visibility	1.00
Citation Frequency	0.92
Task	0.88

Table 4: Inter-annotator agreement for literature annotation measured using Krippendorff's α. The scores are macro-averaged across labels within each dimension.

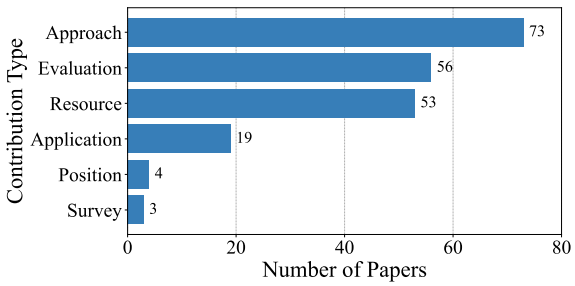

Figure 6: Number of studies by contribution type. Each paper can be assigned to multiple contribution types, so totals exceed 134.

mensional and multi-label nature of the taxonomy, agreement was computed separately for each taxonomy dimension. Within each dimension, α was calculated for individual labels that appeared at least five times in the annotated subset to ensure stable estimates. For each dimension, we report the macro-average across its labels.

The agreement scores are shown in Table 4. No studies employing parametric attribution approaches were present in the annotated sample, so agreement could not be computed for this dimension. The observed values indicate acceptable to strong agreement across all taxonomy dimensions, according to commonly used interpretation guidelines for Krippendorff's α. The comparatively lower agreement for contribution type is consistent with its multi-label nature.

C Research Landscape

Figure 6 shows the distribution of publications by contribution type. We observe that most studies propose novel approaches for evidence-based text generation with LLMs. A substantial number also introduce new resources and focus on evaluation, underscoring the growing attention these aspects receive within the community. Further, this highlights the necessity of not only reviewing methodological contributions but also systematically mapping existing evaluation approaches and resources, a gap this survey addresses in Section 5 and Appendix E. Application, position, and survey studies are relatively underrepresented.

D Extended Analysis and Trends of Evidence-based Text Generation

This section presents an extended analysis of evidence-based text generation with LLMs, complementing the analyses in Section 3. In particular, we highlight representative works associated with each dimension of the proposed taxonomy, offer task-specific analyses, and discuss trends that have emerged over time.

D.1 Attribution Approach

Attribution refers to the process of tracing LLM-generated text back to supporting evidence. In our taxonomy, we distinguish between parametric and non-parametric attribution approaches. Non-parametric attribution is applied substantially more often, with 126 studies, compared to 25 studies on parametric attribution. Within parametric attribution, 18 studies evaluate the ability of pure LLMs to correctly attribute a generated text to a source of evidence (e.g., by generating citations), while four studies address data-centric attribution and three studies focus on model-centric attribution. For non-parametric attribution, where external evidence sources are incorporated into the LLM, the majority of approaches adopt post-retrieval attribution, with 73 studies. In-context attribution is used in 25 studies. Post-generation attribution is explored in 22 studies, whereas in-generation attribution is comparatively rare, with only five studies.

D.1.1 Parametric Attribution

Pure LLMs rely solely on the model's inherent attribution behavior. Studies in this category focus on evaluating the attribution behavior of LLMs, for example, by assessing the presence and quality of citations generated by the model (Byun et al., 2024; Malaviya et al., 2024a). Early work highlights substantial limitations. Pride et al. (2023) report that citations generated by GPT-3.5 are factually correct only 22% of the time, and that GPT-4 does not improve this rate (20%). Similarly, Zuccon et al. (2023) find that while GPT-3.5 generates correct or partially correct answers in 51% of cases, its accompanying references correspond to real sources

only 14% of the time. While pure LLMs are typically prompted to provide citations, Moayeri et al. (2024) observed that GPT models occasionally generate citations spontaneously.

Model-Centric attribution adjust the model architecture or training objectives. In our review, we found that two papers with the contribution type "approach" exemplify this strategy. First, Chu et al. (2025) propose FARD, which trains a student model to imitate citation-grounded rationales distilled from a teacher model, replacing chain-of-thought reasoning to improve causal-aware attribution. Second, Khalifa et al. (2024) introduce source-aware training, where the LLM learns to associate knowledge with document identifiers during pretraining, followed by instruction tuning that encourages explicit citations.

Data-Centric attribution curates, augments, or synthetically generates data. We identified three papers with the contribution type "approach" that fall under this category. Li et al. (2024b) introduce an automatic preference optimization framework that models attribution as a preference learning task using both curated and automatically synthesized citation preference data. Huang et al. (2024b) propose FRONT, a training approach that supervises LLMs with fine-grained supporting quotes to guide citation generation. Lu et al. (2025) frame attribution as a watermarking task, enabling LLMs to embed source-identifying signals into the text.

Task-specific Analysis. Figure 7 shows how attribution approaches distribute across tasks. Parametric attribution appears only in QA, grounded text generation, citation text generation, and related work generation (Zuccon et al., 2023; Lu et al., 2025; Huang and Chang, 2024; Byun et al., 2024). As observed from our annotated dataset, summarization always relies on non-parametric attribution because the task requires one or more input documents to serve as evidence for the summary. Fact verification, in principle, could be performed using only parametric model knowledge, but the lack of parametric attribution in current studies indicates that LLMs are not yet considered sufficiently reliable for this purpose (Buchmann et al., 2024; Asai et al., 2024b). As a result, existing work predominantly relies on non-parametric attribution to ensure factuality. 72% of parametric attribution approaches evaluate pure LLMs without modifying the model architecture or training data. Model-centric and data-centric attribution is explored only in tasks that are already widely studied and that align closely with the capabilities of state-of-the-art LLM chatbots (OpenAI et al., 2023). This makes them natural targets for developing parametric attribution approaches, as these tasks benefit most directly from models that can generate text attributed to parametric knowledge. The observed patterns indicate that parametric attribution in pure LLMs remains unreliable for tasks with strong factuality requirements (Pride et al., 2023), while data-centric and model-centric approaches exist only for a small subset of tasks and still face significant limitations (Huang and Chang, 2024). As a result, tasks that demand robust evidence grounding continue to rely predominantly on non-parametric methods, which highlights key limitations of current LLM architectures.

Trends. Non-parametric attribution approaches appeared substantially earlier than parametric ones, with the first non-parametric studies published in Q1 2022 and the first parametric attribution approach emerging only in Q3 2023. Since then, parametric attribution has shown little momentum: new approaches appear only occasionally and do not form a clear upward trend. This stagnation reflects the technical difficulty of adapting or training LLMs to emit reliable attribution signals, as well as the limited scalability and generalizability of current data-centric and model-centric approaches (Lu et al., 2025; Khalifa et al., 2024). In contrast, non-parametric attribution continues to grow steadily, indicating that parametric attribution has not yet matured into a widely adopted methodology.

D.1.2 Non-Parametric Attribution

Post-Retrieval attribution retrieves external information before text generation and uses it to ground the model's output in evidence. The first post-retrieval approach included in our survey is GopherCite, which is trained to generate answers together with in-line citations (Menick et al., 2022). Attribution is achieved by training the model to generate verbatim quotes from retrieved documents within a fixed output format, with supervised fine-tuning and reinforcement learning from human feedback encouraging answers that are both correct and explicitly supported by the cited evidence. In contrast to GopherCite's joint generation of answers and citations, PURR adopts a revision-based perspective: it first generates search queries from an ungrounded given statement to retrieve relevant evidence, which is then used by a small editor model to revise the original statement (Chen

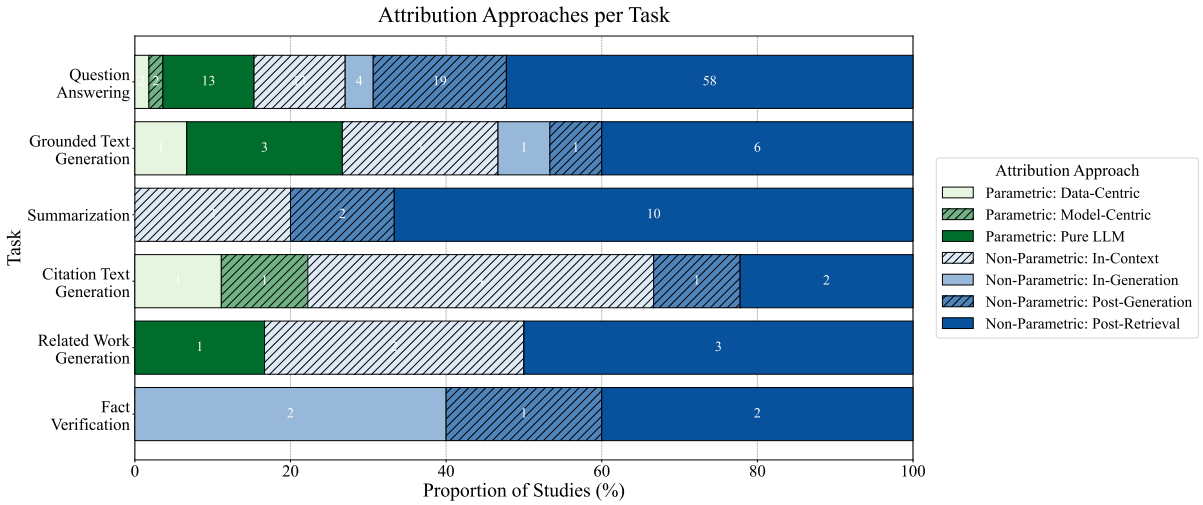

Figure 7: Distribution of attribution approaches across tasks in evidence-based text generation with LLMs. The bars represent the relative percentages of parametric attribution approaches (green) and non-parametric attribution approaches (blue) for each task. The absolute number of papers for each attribution approach is reported alongside the bars. Tasks include QA, grounded text generation, summarization, citation text generation, related work generation, and fact verification. Each paper can be assigned to multiple tasks, so totals exceed 134.

et al., 2023). PURR can be interpreted as either post-retrieval or post-generation. From a post-generation perspective, it revises an already generated statement, while from a post-retrieval perspective, the attributed output depends on evidence retrieved beforehand. Since we define attribution as the process of tracing generated text back to supporting evidence, we classify PURR as post-retrieval, as the grounding evidence is explicitly retrieved before text generation. Huang et al. (2024a) focus on improving post-retrieval attribution through fine-grained reward signals applied during training. Within their method, retrieved passages serve as supervision signals, and fine-grained rewards evaluate how well segments of the generated text align with the evidence. In contrast to previous approaches, MIRAGE derives attribution from model internals by estimating how strongly each retrieved context token contributes to each generated answer token (Qi et al., 2024). This enables token-level attribution without modifying the underlying model, as saliency scores are used to trace the generated output back to specific parts of the retrieved evidence.

Post-Generation attribution reverses the order of post-retrieval attribution by first generating an answer using parametric model knowledge and then revising it using retrieved non-parametric evidence. Gao et al. (2023a) introduce a revision step in their system, RARR, which performs posterior verification by assessing the correctness of the LLM response with reference to the retrieved documents and making adjustments when necessary. Yan et al. (2025) extend this idea with ARE, which decomposes a generated long-form answer into atomic facts and retrieves evidence for each fact individually, enabling more fine-grained verification and edits than the single-step, answer-level revision used in RARR. In contrast, CEG follows a regeneration-based strategy in which evidence is retrieved after an initial answer is generated, and the response is regenerated until all statements are supported by citations (Li et al., 2024d). This differs from RARR, which performs a single targeted revision rather than iteratively regenerating the full answer. Huang et al. (2024c) extend post-generation attribution with START, a self-improving framework that begins with synthetic attribution data and then iteratively samples candidate responses, retrieves evidence, and applies fine-grained preference optimization to improve attribution.

In-Generation attribution dynamically determines when additional evidence is needed and activates the retrieval system during the generation process. A prominent example is Self-RAG (Asai et al., 2024b), which introduces self-reflection tokens enabling the LLM to adaptively retrieve relevant evidence and iteratively critique and revise its outputs during inference. Other notable approaches include AGREE (Ye et al., 2024), which trains LLMs to retrieve and cite evidence dynamically during generation, and Think&Cite (Li and

Ng, 2025), which employs self-guided tree search combined with reward modeling to progressively enhance attribution through iterative evidence gathering. Additionally, NEST (Li et al., 2024c) performs token-level retrieval at each generation step, incorporating relevant spans into the output.

In-Context attribution does not require retrieval models, as the evidence is directly provided within the prompt. This setting has been adopted in a variety of scenarios. For instance, several studies use in-context attribution to evaluate their attribution approaches by supplying a curated set of source documents, rather than relying on retrieval systems (Zhang et al., 2025a; Cohen-Wang et al., 2024; Slobodkin et al., 2024). Furthermore, some tasks inherently require in-context attribution. A typical example is document-level summarization, where multiple documents are given to the LLM for attributed summarization. Another example is citation text generation, where the LLM is prompted to generate citation-worthy text that integrates the content of a given scientific paper into a surrounding context, such as a related work section (Anand et al., 2023a; Gu and Hahnloser, 2024).

Task-specific Analysis. Figure 7 shows how non-parametric attribution paradigms distribute across the six tasks. Post-retrieval attribution is by far the dominant approach, appearing in every task and accounting for the largest proportion within non-parametric approaches in QA (62%), grounded text generation (55%), summarization (67%), and related work generation (60%). Its prevalence reflects the simplicity and effectiveness of current RAG pipelines. In contrast, post-generation attribution is adopted less frequently, representing only 20% of non-parametric approaches and appearing mainly in QA. Only a limited set of tasks benefits from workflows that combine parametric and non-parametric knowledge (Gao et al., 2023a; Ramu et al., 2024), while most settings rely effectively on post-retrieval attribution with outputs grounded directly in retrieved evidence. Post-generation attribution remains valuable, however, in scenarios where reduced reliance on retrieval is advantageous. For example, in QA tasks that require models to provide an output even under incomplete or imperfect evidence, models can first draw on parametric knowledge and then verify or refine the output using external evidence, whereas post-retrieval approaches may return no answer when retrieval fails. In-context attribution exhibits clear task specificity. It is widely used in summarization, citation text generation, and related work generation, tasks where relevant evidence is naturally provided as input by the user (Slobodkin et al., 2024; Gu and Hahnloser, 2024; Nishimura et al., 2024). In many evaluation settings, evidence is provided directly rather than retrieved so that the attribution behavior of LLMs can be assessed independently of retrieval performance (Cao and Wang, 2024; Lee et al., 2024a). This applies in particular to QA and grounded text generation, where in-context attribution is used when retrieval components are removed and the model's reasoning is evaluated in isolation. In-generation attribution remains the least explored non-parametric paradigm. To date, it has only been applied to QA and fact verification (Asai et al., 2024b; Li and Ng, 2025), and has been proposed for grounded text generation (Tilwani et al., 2024). Overall, non-parametric attribution shows clear strengths and constraints. In particular, its reliability depends on the availability of accurate evidence and the quality of the underlying retrieval model. It also does not provide insight into the model's reasoning process. Nevertheless, it remains the prevailing approach for evidence-based text generation with LLMs, with post-retrieval attribution dominating the literature.

Trends. Non-parametric attribution has exhibited a clear and steady increase since its first appearance in Q1 2022. Post-retrieval attribution shows the strongest growth trajectory and has rapidly become the dominant paradigm. Although only 14 post-retrieval attribution approaches were published in 2023, this number increased notably to 56 in 2024. This indicates a significant trend toward adopting RAG as the state-of-the-art architecture for evidence-based text generation with LLMs. Post-generation and in-context attribution follow similar patterns, with initial work emerging in Q4 2022 and the number of studies increasing from 3 in 2023 to 15 in 2024 for each paradigm. This demonstrates consistent growth, although at a smaller scale compared to post-retrieval attribution. In-generation attribution, by contrast, is the most recent paradigm, appearing first in Q4 2023 with only two approaches in 2023 and three in 2024. These early and sparse instances do not yet form a noticeable trend. Overall, the field continues to grow primarily along non-parametric attribution, driven by rapid growth in post-retrieval attribution approaches, while post-generation and in-context attribution show gradual growth and in-generation attribution remains exploratory.

D.2 Citation Characteristics

Following the attribution approaches, this section provides a detailed analysis of the five citation characteristics for evidence-based text generation with LLMs: citation modality, evidence level, citation style, citation visibility, and citation frequency.

D.2.1 Citation Modality

The citation modality specifies the type of evidence cited by an approach. We distinguish four modalities: **texts**, **graphs**, **tables**, and **visuals**. The vast majority of studies rely on textual evidence, with 96% of the surveyed papers citing unstructured text as their primary source of evidence. Within text-based approaches, the underlying evidence predominantly comes from a small set of well-established domains (Gao et al., 2023b; Malaviya et al., 2024a). Table 12 shows the most frequently reused datasets per task. As shown in the table, encyclopedic sources, most notably Wikipedia, are the most frequently reused, followed by scientific literature and general web search content. News articles and synthetic datasets appear less often, while domains such as social media, government documents, and health-related texts are only sparsely represented. The predominance of these general-purpose textual domains suggests that current approaches are primarily evaluated on easily accessible evidence sources, leaving the robustness of evidence-based text generation underrepresented for more specialized or regulated domains. In contrast, non-textual modalities are used less frequently. Graph-based evidence appears in only a small subset of studies, with four papers using this modality, and is typically used to represent structured relationships between entities (He et al., 2025; Dehghan et al., 2024). Tabular evidence is used even more sparsely, appearing in three studies, and is generally applied when factual grounding requires access to structured numerical or categorical data (Saha Roy et al., 2025; Mathur et al., 2024). Visual evidence, such as images, is explored in only two approaches, reflecting early efforts toward multimodal evidence (Ma et al., 2025; Suri et al., 2025).

Task-specific Analysis. The distribution of citation modalities varies across tasks but remains strongly dominated by textual evidence. Summarization, citation text generation, related work generation, and fact verification rely exclusively on text-based evidence. QA is the only task that spans all identified citation modalities, although non-textual evidence remains clearly underrepresented. Grounded text generation also exhibits limited modality diversity, with only a single approach incorporating tabular evidence. This task-level distribution highlights that modality diversity in evidence-based text generation with LLMs is uneven and largely concentrated in QA.

Trends. Textual evidence has been present since the earliest work on evidence-based text generation, with the first paper appearing in Q2 2021. The number of text-based approaches increased substantially from 26 publications in 2023 to 93 in 2024, reinforcing text as the dominant citation modality over time. Non-textual modalities emerge considerably later and remain rare. Graph-based evidence first appears in Q4 2023 and grows modestly from one study in 2023 to four in 2024. Tabular and visual evidence are the most recent modalities, both first introduced in Q4 2024, with three and two studies published in 2024, respectively. Overall, modality-level trends indicate that non-textual citation modalities remain underrepresented but are beginning to attract increasing research interest.

D.2.2 Evidence Level

The evidence level describes the granularity at which cited evidence is linked to LLM-generated text and is closely related to the underlying citation modality. Rather than specifying the type of evidence, this dimension captures how fine-grained evidence is referenced within a given citation modality. For textual evidence, multiple evidence levels are observed. Evidence can be cited at the level of a full **document**, such as an entire scientific article (Anand et al., 2023b), or at a finer granularity such as a **paragraph**, often corresponding to a retrieved text chunk (Gao et al., 2023b). Some approaches operate at the **sentence** level (Xu et al., 2025), while a small number attempt attribution at the level of individual **tokens** (Phukan et al., 2024). Across all reviewed studies, document-level evidence is used in 43% of papers and paragraph-level evidence in 40%, making these the most common choices. Sentence-level attribution appears in 12% of studies, whereas token-level attribution remains rare at 2%. For non-textual modalities, the evidence level is largely determined by the structure of the citation modality itself. Graph-based approaches typically cite evidence at the level of individual **triples** (Li et al., 2024f), while no surveyed study cites an entire graph as a single evidence unit. Tabular evidence is referenced either at the level of a full **table** or at the level of individual **table**

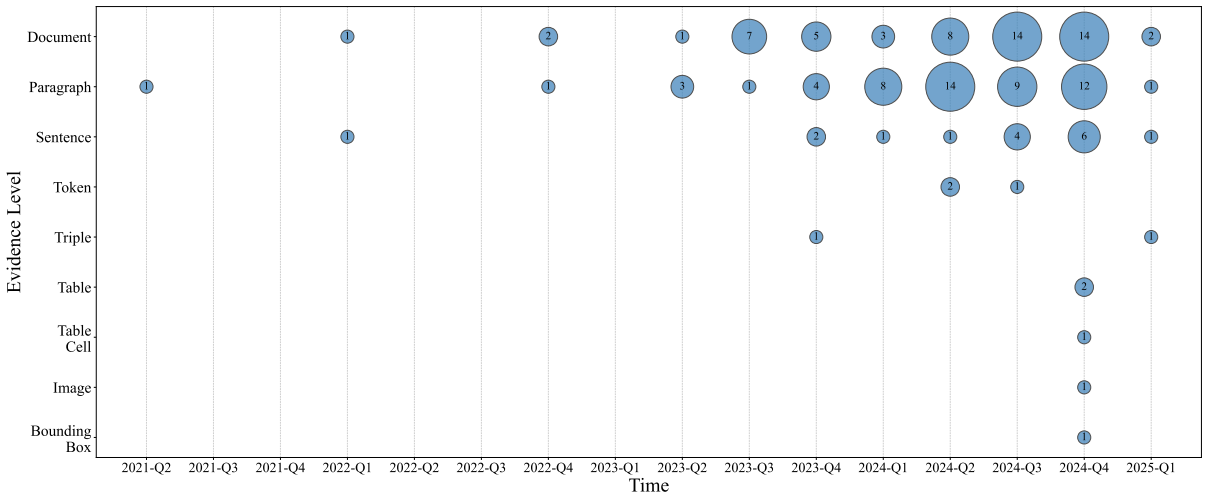

Figure 8: Temporal trends of evidence levels in evidence-based text generation with LLMs. For each evidence level, the figure reports the number of studies adopting that evidence level in each quarter from Q2 2022 to Q1 2025. The size of each bubble corresponds to the number of papers published in the respective quarter using the given evidence level. Q1 2025 includes relevant papers available up to the inclusion deadline in February. Each paper can be assigned to multiple evidence levels, so totals exceed 134.

cells (Mathur et al., 2024). Visual evidence is cited either as a complete **image** or as a specific region, such as a **bounding box** (Ma et al., 2025). Due to the limited number of studies employing non-textual modalities, differences in evidence level within these modalities are comparatively small.

Task-specific Analysis. Given the strong dominance of textual citation modalities, task-specific differences in evidence level are discussed primarily for text-based evidence. For non-textual citation modalities, the limited number of studies prevents meaningful task-specific differentiation. QA exhibits the widest range of textual evidence levels, spanning from coarse-grained document- and paragraph-level sources to fine-grained sentence- and token-level sources. This reflects the need to support both broad contextual grounding and precise factual claims. Grounded text generation also employs multiple textual evidence levels but predominantly relies on document- and paragraph-level evidence, with finer-grained attribution appearing less frequently. In contrast, citation text generation and related work generation rely on coarse-grained textual evidence, most often citing entire documents. This aligns with conventions in scientific writing, where citations reference complete papers rather than individual passages or sentences. Summarization occupies an intermediate position, with different approaches adopting different textual evidence levels. Most rely on document- or paragraph-level grounding, while a smaller number of studies explore sentence- or token-level evidence to support more fine-grained alignment between source content and generated summaries. Fact verification relies mainly on paragraph- and sentence-level textual evidence, reflecting its emphasis on localized factual support.

Trends. Evidence levels at coarse granularity appear earliest and dominate the literature over time. Paragraph-level evidence is the earliest to emerge, with the first studies appearing in Q2 2021, followed by document-level evidence in Q1 2022. Both show substantial growth, with document-level approaches increasing from 13 publications in 2023 to 39 in 2024, and paragraph-level approaches rising from 8 to 43 over the same period. Sentence-level evidence emerges later and remains less common, growing from two studies in 2023 to twelve in 2024. Token-level evidence remains rare. It appears only in Q2 2024 and is explored by a small number of studies. Evidence levels associated with non-textual modalities, such as triples, tables, table cells, images, and bounding boxes, appear only from late 2023 or 2024 onward and remain sparsely represented. Overall, trends indicate strong growth in coarse-grained evidence levels, with finer-grained and non-textual evidence levels emerging only recently and at a smaller scale.

D.2.3 Citation Style

The citation style determines how evidence for LLM-generated text is presented to the user. We

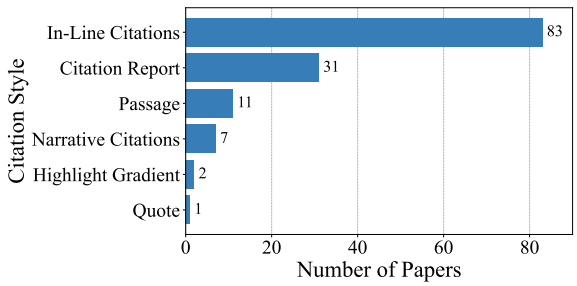

Figure 9: Number of studies using each citation style. Each paper can be assigned to multiple citation styles, so totals exceed 134.

identify six distinct citation styles: in-line citations, citation reports, passages, narrative citations, highlight gradients, and quotes. Figure 9 shows the distribution of citation styles across the reviewed studies, highlighting that 62% of studies rely on in-line citations. As illustrated in Figure 10, **in-line citations** are inserted directly after a citation-worthy claim (e.g., [1][2]) (Huang et al., 2024a), enabling users to trace individual statements back to their supporting evidence. Another frequently used style is the **citation report**, which provides a separate list of references alongside the LLM-generated output (Bohnet et al., 2023). While citations are still presented in textual form, users must manually associate reported references with specific generated claims, which makes citation reports less transparent. Several approaches display only the retrieved **passage** used by the LLM during generation, particularly in evaluation settings for attribution approaches (Muller et al., 2023). As shown in Figure 10, the passage from the evidence corpus is presented alongside the LLM-generated output without any textual or visual citation markers that explicitly link the evidence to specific claims. This requires users to perform the alignment implicitly and makes this citation style less suitable for transparent or user-facing evidence-based text generation with LLMs, while still being useful for methodological evaluation. **Narrative citations** integrate references into the natural flow of the generated text, for example by explicitly mentioning authors or sources (e.g., "Author et al. argue that ...") to improve contextual clarity (Shaier et al., 2024). A different approach is the **highlight gradient**, which visually encodes the influence of cited content by coloring relevant tokens or sentences in both the generated output and the source text, thereby providing a visual form of traceability (Do et al., 2024). Finally, we identify a single paper that employs direct **quotes**, embedding verbatim excerpts from the evidence source into the generated response (Xiao et al., 2025). Importantly, quotations are not limited to serving as a citation style. In our analysis, we find that 13% of the reviewed studies explicitly engage with quotations in some form. Moreover, quotations can function as a training or prompting strategy, as exemplified by approaches such as chain-of-quote (Li et al., 2024g). In their user study, Do et al. (2024) show that LLM responses with citations are perceived as significantly more trustworthy than uncited responses, while no significant difference in trust is observed between in-line citations and highlight gradient visualizations.

Task-specific Analysis. In-line citations are among the top-2 most frequently used citation style across all tasks, only for grounded text generation and fact verification citation reports appear slightly more often, making them the default style for presenting evidence in LLM-generated text. QA exhibits the greatest diversity of citation styles, including in-line citations, citation reports, passages, narrative citations, and highlight gradients. This diversity reflects the overall frequency of studies for QA, but also the wide range of different settings of this task. Grounded text generation shows a similar diversity of citation styles, but being currently the only task where quotes have been utilized. Summarization, citation text generation, and related work generation rely primarily on in-line citations, largely following conventions from document-centric and scientific writing, where citations are embedded directly in the text. While fact verification studies predominantly use citation reports, the limited sample size prevents drawing reliable conclusions about preferred citation mechanisms in this task. Overall, citation style choices show only limited task-specific variation. Across tasks, most studies employ fine-grained citation styles, such as in-line citations, that allow users to verify individual LLM-generated claims, while differences related to task requirements or user interaction remain marginal.

Trends. Citation styles exhibit a staggered emergence over time, with clear differences in adoption. Passages appear earliest, first emerging in Q2 2021, but grow only moderately. In-line citations are introduced later, first appearing in Q1 2022, and show the strongest growth trajectory, increasing substantially from 11 studies in 2023 to 67 in 2024. Citation reports also emerge in late 2022 and ex-

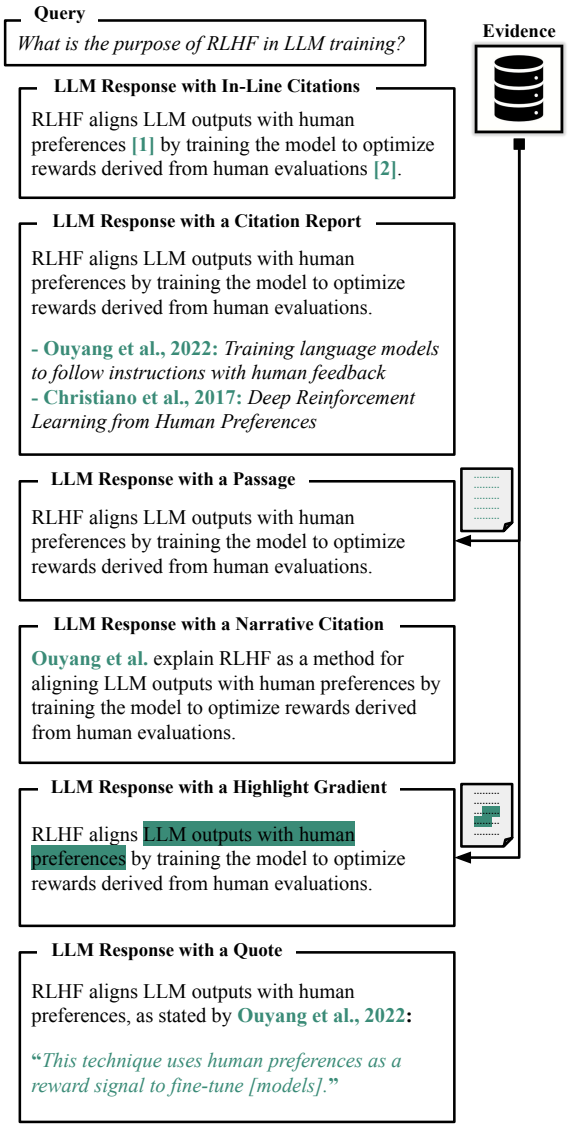

Figure 10: Visualization of different citation styles. For a given query, the LLM generates a response with citations presented in different styles, enabling users to verify LLM-generated text against supporting evidence.

pand steadily, rising from 10 publications in 2023 to 19 in 2024. Narrative citations appear only from Q3 2023 onward and remain comparatively rare, while highlight gradients and quotes emerge only in 2024 and are explored by very few studies. Overall, the literature shows a rapid consolidation around a small set of dominant citation styles, with more specialized citation styles appearing only recently and at limited scale.

D.2.4 Citation Visibility

Citation visibility describes whether citations are visible to users in the LLM-generated output, or are generated only internally in intermediate text. The majority of reviewed papers include citations directly in the **final response**, making supporting evidence visible to users and enabling explicit traceability for verification. Overall, 91% of the surveyed studies adopt this form of citation visibility. In contrast, 4% of studies generate citations exclusively in **intermediate text** that is not shown to users. In these cases, citations are used internally by the LLM as part of the generation or evaluation process. Across the reviewed studies, citations in intermediate text serve three main purposes: *internal verification* (Fang et al., 2024; Chiang and Lee, 2024), *guiding multi-step reasoning during generation* (Chu et al., 2025; Xia et al., 2025b), and *aligning evidence across modalities in multimodal settings* (Suri et al., 2025).

Task-specific Analysis. Citation visibility differs across tasks but is dominated by citations in the final response. Summarization, citation text generation, related work generation, and fact verification exclusively rely on citations that are visible to users. This aligns with the objective of these tasks, which is to explicitly present evidence to support generated content. QA and grounded text generation are the only tasks that also include approaches using citations in intermediate text. In these cases, citations are used internally to optimize aspects such as answer correctness, without necessarily exposing supporting evidence to users.

Trends. Citations shown in the final response have been present since the earliest work on evidence-based text generation, with the first studies appearing in Q2 2021. The number of approaches exposing citations to users increased substantially from 24 papers in 2023 to 79 in 2024. In contrast, citation visibility limited to intermediate text emerged only recently, first appearing in Q4 2023, and remains rare, with only a small number of studies in 2023 and 2024. Overall, trends indicate strong growth in citations in the final response, while intermediate text citations remain a sparsely utilized design choice.

D.2.5 Citation Frequency

Citation frequency describes the number of citations provided for each LLM-generated claim. We distinguish between approaches that provide a **single** citation per claim and those that support **multiple** citations. Overall, 64% of the reviewed studies support multiple citations per claim, indicating a preference for broader evidential coverage. This design choice reflects a trade-off between evidential coverage and simplicity. Multiple ci-

tations can strengthen support by drawing on diverse sources, whereas single citations (31% of studies) offer a more concise presentation and may reduce cognitive load for users. We observe that some parametric approaches are limited to generating a single citation per claim due to architectural constraints (Khalifa et al., 2024; Lu et al., 2025). In contrast, most non-parametric approaches naturally support multiple citations by aggregating evidence through retrieval mechanisms. Ding et al. (2025) found that while the presence of citations increases perceived trust in LLM-generated responses, adding more than one citation does not yield additional trust gains.

Task-specific Analysis. While all tasks include approaches that support multiple citations per LLM-generated claim, QA, grounded text generation, summarization, and fact verification predominantly rely on multiple citations, reflecting the need to aggregate evidence from several sources. In contrast, citation text generation and related work generation more often rely on a single citation per claim. This reflects differences in task formulation. In citation text generation, a single paper is typically provided to the LLM and incorporated into an existing text, naturally resulting in a single citation per claim. In related work generation, the goal is to discuss and compare individual papers, with claims commonly tied to specific works rather than to an aggregation of sources.

Trends. Studies supporting only single citations per claim appear earlier, with the first paper published in Q2 2021. Support for multiple citations emerges in Q4 2022, but shows substantially stronger growth. The number of approaches supporting multiple citations increased from 18 in 2023 to 64 in 2024, compared to an increase from 6 to 29 for single-citation approaches over the same period. Overall, while single citations remain preferred in certain task formulations, the ability to provide multiple citations has become widely supported and no longer represents a limiting factor for evidence-based text generation with LLMs.

D.3 Task

In this section, we summarize the distribution of tasks studied in evidence-based text generation with LLMs. Across the literature, we identified 16 distinct tasks, of which only six are addressed by more than one study. As shown in Table 5, QA dominates the field, while grounded text generation and summarization receive moderate attention.

Task	No. of Papers
Question Answering	95
Grounded Text Generation	15
Summarization	14
Citation Text Generation	6
Related Work Generation	5
Fact Verification	3

Table 5: Overview of the six most popular tasks in the literature on evidence-based text generation with LLMs. Each paper can be assigned to multiple tasks, so totals exceed 134.

Other tasks, including citation text generation, related work generation, and fact verification, are investigated less frequently.

Trends. Evidence-based text generation tasks exhibit distinct temporal adoption patterns. The earliest task in our annotated dataset is grounded text generation, which first appears in Q2 2021. After limited activity for several years, it gains momentum only recently, increasing from 2 publications in 2023 to 11 in 2024. QA emerges next, first appearing in Q1 2022 and subsequently becoming the dominant task. The number of papers in this category rise sharply from 18 in 2023 to 71 in 2024, reflecting its central role in evidence-based text generation with LLMs. Summarization enters the landscape in Q4 2022 and shows notable growth, increasing from 2 studies in 2023 to 11 in 2024. Citation text generation appears somewhat later, with three studies in 2023 and two in 2024, indicating steady but comparatively modest activity. Related work generation and fact verification are the most recent tasks, both first emerging in Q4 2023. These areas show early signs of uptake, with related work generation growing from one study in 2023 to four in 2024, and fact verification from one to two. Overall, the task-level trends reveal a broadening scope of evidence-based text generation with LLMs, with QA driving much of the recent growth and grounded text generation and summarization also showing increased adoption.

E Evaluation Resources

This section provides additional details on the evaluation of evidence-based text generation with LLMs, complementing Section 5. We define evaluation frameworks as standardized procedures that systematically apply a predefined set of metrics (at least two) over specific tasks to assess methodological performance along at least one evaluation dimension. In contrast, benchmarks are curated

combinations of datasets that facilitate reproducible and comparative evaluation, while individual metrics and datasets are not assigned to a framework or benchmark. Overall, we identify 19 evaluation frameworks and 11 benchmarks, of which only two from each group are reused across multiple studies, highlighting a persistent lack of consensus and standardization in current evaluation practices.

Section E.1 provides task-specific analyses and identifies trends across evaluation dimensions, while Section E.2 presents a complementary perspective of evaluation dimensions. In addition, Section E.3 discusses notable evaluation metrics that do not appear among the frequently reused metrics detailed in Section 5.2. Finally, Section E.4 and Section E.5 summarize the identified frameworks and benchmarks, respectively, and Section E.6 provides an overview of available evaluation datasets.

E.1 Evaluation Methods

We identified six evaluation methods, described in Section 5.1. Table 8 summarizes these evaluation methods along with supplementary details not included in the main text.

Task-specific Analysis. Figure 11 shows how evaluation methods distribute across tasks. QA, grounded text generation, and summarization exhibit highly similar evaluation patterns. Across these tasks, evaluation is dominated by inference-based and lexical overlap methods, which together account for 53% of evaluation usage in QA, 45% in grounded text generation, and 57% in summarization. This indicates a strong reliance on automatic metrics, while human evaluation represents the largest relative share of evaluation in summarization, highlighting the limitations of purely automatic methods. In contrast, LLM-as-a-judge evaluation remains relatively uncommon across tasks, reaching its highest share in grounded text generation (16%) while remaining below 10% in all other tasks, indicating that LLM-as-a-judge evaluation is still emerging and has not yet replaced established automatic approaches. Citation text generation and related work generation show a distinct evaluation pattern. In citation text generation, lexical overlap methods clearly dominate, accounting for 50% of evaluation usage. In related work generation, lexical overlap and retrieval-based methods are most prevalent, together accounting for 62% of evaluation usage. Fact verification exhibits the narrowest distribution of evaluation methods, with inference-based and lexical overlap approaches accounting for 55% of evaluation usage, while the absence of LLM-as-a-judge evaluation indicates continued reliance on deterministic and interpretable methods. Overall, the observed patterns indicate that traditional automated metrics, particularly lexical overlap and semantic similarity-based methods, continue to dominate the evaluation of shorter text segments, while long-form text is assessed mainly using inference-based methods, with LLM-as-a-judge approaches still used only to a limited extent.

Trends. Evaluation methods exhibit distinct temporal adoption patterns. The earliest methods in our annotated dataset are inference-based, lexical overlap, and semantic similarity-based metrics, all first appearing in Q2 2021. These traditional automated metrics have since experienced substantial growth. Inference-based evaluation shows the largest absolute increase, rising from 11 studies in 2023 to 51 in 2024, while lexical overlap grows from 14 to 50 over the same period. Semantic similarity-based evaluation also expands strongly, increasing from 5 to 29 studies during this interval. Human evaluation was applied slightly later, first appearing in Q1 2022, and shows steady growth, increasing from 8 to 23 studies over the same period. More recent evaluation paradigms appear later in the timeline. LLM-as-a-judge evaluation first emerges in Q4 2022 and grows moderately from 5 to 17 studies. Retrieval-based evaluation is the most recent approach in this domain, first appearing in Q3 2023, although the underlying methodology is well established. Nevertheless, it exhibits the fastest growth overall, increasing from 4 to 33 studies during this interval. Overall, these temporal patterns indicate that evaluation practices remain strongly anchored in traditional automated metrics, while newer paradigms that leverage retrieval signals or LLMs for evaluation are emerging but still less widely adopted.

E.2 Evaluation Dimensions

Task-specific Analysis. Evaluation dimensions show a high degree of consistency across tasks. The three core dimensions (attribution, citation, and correctness) clearly dominate evaluation practices in all tasks, together accounting for between 64% and 81% of evaluation instances. The evaluation of attribution and correctness is prominent across all tasks, while the evaluation dimension citation is applied whenever annotated ground-truth evidence is available. Notably, citation has not yet been evaluated in fact verification, where an-

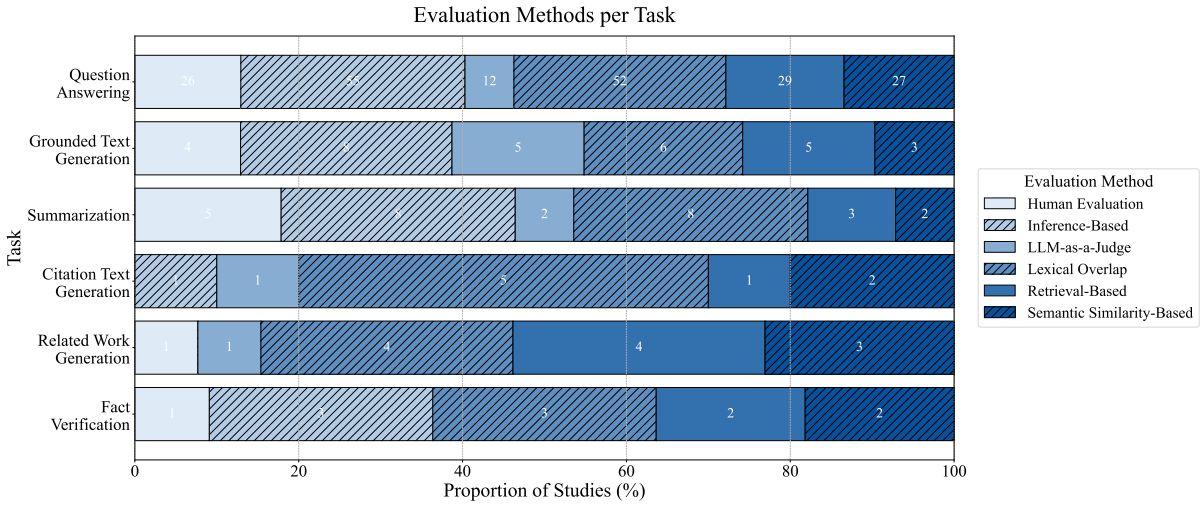

Figure 11: Distribution of evaluation methods across tasks in evidence-based text generation with LLMs. The bars represent the relative percentages of different evaluation methods for each task. The absolute number of papers for each evaluation method is reported alongside the bars. Tasks include question answering, grounded text generation, summarization, citation text generation, related work generation, and fact verification. Each paper can be assigned to multiple tasks, so totals exceed 134.

notated ground-truth evidence is often unavailable, and studies therefore rely primarily on attribution and correctness. In contrast, contextual evaluation dimensions exhibit greater variation across tasks. Linguistic quality is the most frequently applied contextual dimension and appears across most tasks, whereas preservation, relevance, and retrieval are used only selectively. These dimensions reflect task-specific evaluation needs rather than core requirements for verifying evidence-based text generation with LLMs. Overall, the observed patterns indicate strong adherence to the core evaluation dimensions defined in our guidelines. Attribution, citation, and correctness consistently form the foundation of evaluation across tasks, while contextual dimensions serve a complementary and task-dependent role.

Trends. Evaluation dimensions have been applied at different stages over time. The earliest dimension in our annotated dataset is attribution, first appearing in Q2 2021, followed by correctness in Q1 2022. These core evaluation dimensions have since experienced substantial growth. Correctness shows the largest absolute increase, rising from 13 studies in 2023 to 63 in 2024, while attribution grows from 16 to 60 over the same period. The evaluation of citation is applied later, first appearing in Q4 2022, but increases from 5 to 30 studies during this interval. Contextual evaluation dimensions are applied at later stages. Linguistic quality, first introduced in Q4 2022, shows steady growth from 7 to 31 studies over the same period. Preservation remains rarely applied, increasing only marginally from 1 to 3 studies. Relevance and retrieval are the most recent dimensions, both first appearing in Q3 2023, and show moderate growth from 2 to 13 studies. Overall, these temporal patterns indicate that evaluation practices remain strongly anchored in core dimensions, particularly attribution and correctness, while contextual dimensions have emerged more recently and continue to be applied more selectively depending on task requirements.

E.3 Additional Notable Evaluation Metrics

Given the total of 300 extracted metrics, a detailed description of each metric is beyond the scope of this survey. To ensure clarity and encourage evaluation standardization, Section 5.2 highlights the most frequently reused metrics. We acknowledge that some relevant, especially recent, evaluation metrics may not yet meet this criterion. Although not yet widely adopted, these additional evaluation metrics are still significant for the evolving evaluation landscape. Accordingly, this section presents further notable evaluation metrics that may shape future research. The full table of 300 categorized evaluation metrics is available in our repository, as described in Appendix A.

Attribution. Early work in multilingual QA attribution by Muller et al. (2023) introduced *AttrAcc*, an inference-based metric to evaluate the classification accuracy of attribution detection mod-

els. Building on this, Yue et al. (2023) proposed *AttrScore*, which extends binary attribution accuracy to a three-way classification, distinguishing whether a generated statement is fully supported, insufficiently supported, or contradicted by a cited reference. In the context of citation text generation, Şahinuç et al. (2024) revealed relationships between different metrics, including *ROUGE-L*, *BERT-Score*, *SciBERT-Score*, *BLEURT*, and *SummaC*. Extending attribution evaluation to a finer evidence level, Phukan et al. (2024) proposes the metrics *Token-level Recall*, *Precision*, and *F1*, which measure how accurately and completely the model identifies the correct source tokens, as well as *Span-level Accuracy*, which assesses the percentage of predicted spans that exactly match the ground truth. More recent work by Yan et al. (2025) proposes decomposing claims in LLM-generated answers into atomic facts and defines the inference-based metrics *Attribution Recall* and *Precision* to quantify the proportion of claims supported by evidence and the relevance of the evidence retrieved, respectively. Leveraging knowledge graphs to generate comprehensive attributions, *CAQA* enables Hu et al. (2025b) to evaluate 25 automatic and human evaluators, as well as LLMs fine-tuned on the task. Separately, Xing et al. (2025) introduce the metrics *Set Precision*, *Recall*, and *F1*, which assess how accurately an annotator, human or LLM, identifies citations within a masked explanation. For long-form Wikipedia generation, Zhang et al. (2025c) introduce the inference-based metric *Citation Rate*, defined as the word-length-weighted average of sentence-level Citation Recall NLI, which accounts for variation in sentence length. Qian et al. (2025) additionally extend the ALCE framework by using GPT models as NLI component, observing similar overall trends to classical NLI models but generally lower citation scores with the GPT-based approach. Wallat et al. (2025) report that up to 57% of citations generated by RAG systems are unfaithful, and suggest that attribution evaluation should also consider faithfulness, that is, whether sources were actually used during text generation, as attribution metrics primarily capture whether sources support a generated claim.

Correctness. Beyond the metrics presented in Section 5.2, several additional studies have proposed notable approaches. Chiang and Lee (2024) propose *D-FActScore*, an extension of *FActScore* (Min et al., 2023) designed to better handle entity ambiguity. Whereas FActScore evaluates each fact independently, D-FActScore groups related facts that a reader could reasonably interpret as referring to the same entity without prior knowledge.

Citation. To evaluate citations, Nishimura et al. (2024) complement existing metrics with *ARI'*, a modification of the Adjusted Rand Index that accounts for unmatched citations across multiple paragraphs, and *Citation Structure F1*, which measures how closely the citation patterns in generated paragraphs match those in the ground truth. In long-context summarization, Laban et al. (2024) proposed a *Citation Score*, measuring citation accuracy, and a *Coverage Score*, using LLM-as-a-judge to assess correctness. They also introduced a *Joint Score*, combining both metrics to evaluate whether LLMs and RAG systems preserve comprehensive coverage of cited content.

Preservation. Şahinuç et al. (2024) employed *N-Gram Overlap*, a lexical overlap metric, to measure n-gram overlap between the input and the model output in citation text generation, assessing the extent of lexical reuse from the prompt.

E.4 Frameworks

Figure 4 maps the two evaluation frameworks that have been reused for evidence-based text generation with LLMs to the dimensions they address. The first, *ALCE* (Gao et al., 2023b), is the most widely adopted, appearing in 12 studies and covering attribution, correctness, and linguistic quality. Beyond applying the entire framework, several additional studies use only individual ALCE metrics, such as Citation Precision NLI and Citation Recall NLI, without adopting the complete framework. While these partial uses are not counted among the 12 instances adopting the framework, they further underscore the influence and prominence of ALCE in current evaluation practices. The second framework, *G-Eval* (Liu et al., 2023b), appears in two studies and provides a broader evaluation of natural language generation through an LLM-as-a-judge approach, assessing attribution, linguistic quality, and relevance. Although only ALCE and G-Eval have seen reuse so far, several additional frameworks have been introduced since 2023. While not yet adopted in published research, they remain relevant to the evaluation landscape. Table 10 therefore presents a complete overview of all frameworks, both reused and not yet adopted, detailing their respective evaluation dimensions and methods.

In addition to these evaluation frameworks, Zhang et al. (2024) propose a meta-framework that

systematically evaluates the effectiveness of citation and faithfulness metrics themselves, highlighting that not only evaluation frameworks but also meta-frameworks have emerged in this area.

E.5 Benchmarks

Similar to the evaluation frameworks, we identified only two benchmarks that were used more than once to evaluate evidence-based text generation approaches. *ALCE* (Gao et al., 2023b) functions both as a framework and benchmark, and has been reused as a benchmark in nine studies. This benchmark supports multiple evaluation dimensions and is built on datasets focused on information-seeking tasks. *RAG-RewardBench* (Jin et al., 2025), designed to evaluate reward models for preference alignment in RAG, is used twice. It spans 18 datasets and targets evaluation scenarios such as multi-hop reasoning, fine-grained citation, appropriate abstention, and conflict robustness. Although only two benchmarks have been reused, several others have been introduced in the literature. While these have not yet been reused in published research, they remain relevant within the evaluation landscape. Therefore, Table 11 provides the full list of evaluation benchmarks, including their associated task and data domain.

E.6 Datasets

Overall, we extracted 231 highly task-specific datasets serving diverse purposes, including both training and evaluation. Due to this heterogeneity, we focus on the most frequently reused datasets that support standardized evaluation, so far primarily centered on QA with more than 64% of datasets, followed by grounded text generation (9%) and summarization (6%). Table 12 presents the 12 most frequent datasets for QA, the eight most frequent datasets for grounded text generation, and the four most frequent datasets for all other tasks used to evaluate evidence-based text generation with LLMs. Additionally, we identified the data domain for each frequent dataset and found that most datasets fall within the Wikipedia, scientific, and news domains, while others are less frequently considered. The full list of datasets is available in our Git repository, as described in Appendix A.

Contribution Type	**Parametric**	**Non-Parametric**	**Citation Modality**
APP: Approach	PLL: Pure LLM	PR: Post-Retrieval	TX: Texts
EVA: Evaluation	MC: Model-Centric	PG: Post-Generation	GR: Graphs
RES: Resource	DC: Data-Centric	IG: In-Generation	TB: Tables
APL: Application		IC: In-Context	VL: Visuals
POS: Position			
SUR: Survey			
Evidence Level	**Citation Style**	**Citation Visibility**	**Citation Frequency**
DOC: Document	ILC: In-Line Citations	FR: Final Response	SI: Single
PAR: Paragraph	CR: Citation Report	IT: Intermediate Text	MU: Multiple
SEN: Sentence	PAS: Passage		
TOK: Token	NC: Narrative Citations		
TRI: Triple	HG: Highlight Gradient		
TAB: Table	QU: Quote		
TC: Table Cell			
IMG: Image			
BB: Bounding Box			
Task	**Training**	**Prompting**	**Evaluation Dimension**
QA: Question Answering	SFT: Supervised Fine-Tuning	ZS: Zero-Shot	ATT: Attribution
GTG: Grounded Text Generation	SSFT: Self-Supervised Fine-Tuning	FS: Few-Shot	COR: Correctness
SUM: Summarization	RL: Reinforcement Learning	COT: Chain-Of-Thought	CIT: Citation
RWG: Related Work Generation	PT: Pretraining	SC: Self-Consistency	LQ: Linguistic Quality
CTG: Citation Text Generation		AO: Active-Oriented	PRE: Preservation
FV: Fact Verification		COC: Chain-Of-Citation	REL: Relevance
OTH: Other		COQ: Chain-Of-Quote	RET: Retrieval
		CA: Conflict-Aware	
		RP: Role-Play	
Evaluation Method			
HE: Human Evaluation			
IB: Inference-Based			
LO: Lexical Overlap			
LJ: LLM-as-a-Judge			
RB: Retrieval-Based			
SB: Semantic Similarity-Based			

Table 6: Legend of all abbreviations used for contribution types, attribution approaches, citation characteristics, task categories, LLM integration strategies, evaluation dimensions, and evaluation methods. These abbreviations are used throughout Table 7, Table 9, Table 10, Table 11, and Table 12.

Paper	Contribution Type	Attribution Approach		Citation Characteristics					Task	LLM Integration	
		Par.	Non-Par.	Mod.	Evi.	Style	Vis.	Freq.		Train.	Prompt.
Dziri et al. (2022)	EVA, RES	–	–	TX	PAR	PAS	FR	SI	GTG	SFT	–
Menick et al. (2022)	APP	–	PR	TX	DOC, SEN	ILC	FR	SI	QA	SFT, RL	FS
Gao et al. (2023a)	APP	–	PG	TX	PAR	CR	FR	MU	QA, SUM	SFT	FS
Gu and Hahnloser (2024)	APP, RES	–	IC	TX	DOC	ILC	FR	SI	CTG	SFT, RL	ZS
Bohnet et al. (2023)	EVA, RES	–	PR, PG	TX	DOC	CR	FR	SI	QA	SFT	FS
Yue et al. (2023)	EVA	–	–	TX	DOC	ILC	FR	SI	QA	SFT	ZS, FS
Gao et al. (2023b)	APP, EVA, RES	–	PR, PG	TX	PAR	ILC	FR	MU	QA	–	ZS
Muller et al. (2023)	APP, EVA, RES	–	PR	TX	PAR	PAS	FR	SI	QA	SFT	FS, COT
Chen et al. (2023)	APP, RES	–	PR	TX	PAR	CR	FR	MU	GTG	SSFT	ZS, FS, COT
Pride et al. (2023)	EVA, APL	PLL	PR	TX	DOC	CR	FR	MU	QA	–	ZS
Huang and Chang (2024)	POS	MC, DC	PR, PG	TX	–	ILC	FR	–	CTG	–	–
Kamalloo et al. (2023)	RES	–	–	TX	–	–	–	MU	QA	–	–
Kang et al. (2023)	APL	–	PR	TX	DOC	ILC, CR, NC	FR	MU	SUM	–	ZS
Xue et al. (2024)	APL	–	PR	TX	DOC	CR	FR	MU	QA	SFT	ZS
Spennemann (2025)	EVA	PLL	–	TX	DOC	CR	FR	MU	QA	–	–
Zuccon et al. (2023)	EVA	PLL	–	TX	DOC	CR	FR	MU	QA	–	ZS
Hu et al. (2025a)	EVA, APL	PLL	–	TX	DOC	CR	FR	MU	OTH	–	ZS
Malaviya et al. (2024a)	APP, EVA, RES	PLL	PR, PG	TX	DOC	ILC	FR	MU	QA	SFT	ZS
Lee et al. (2023)	APP, RES	–	PR	TX	PAR	ILC	FR	MU	QA	SFT	ZS
Zhang et al. (2023a)	SUR	–	–	TX	–	–	–	–	OTH	–	–
Fok et al. (2024)	APL	–	PR	TX	DOC, PAR	PAS	FR	SI	QA, SUM	–	ZS, FS
Asai et al. (2024b)	APP	–	IG	TX	PAR	CR	FR	MU	QA, FV	SFT	ZS
Lu et al. (2025)	APP	DC	–	TX	SEN	PAS	FR	SI	GTG	SFT, PT	–
Li et al. (2024f)	EVA, RES	–	PR	GR	TRI	ILC	FR	MU	QA	–	FS
Li et al. (2023a)	SUR	–	–	TX	–	–	–	–	OTH	–	–
Ye et al. (2024)	APP	–	IG	TX	PAR	ILC	FR	MU	QA	SFT	ZS
Schuster et al. (2024)	APP, RES	–	PR	TX	DOC	ILC	FR	MU	QA	SFT	ZS, FS
Worledge et al. (2024b)	APP	–	–	–	–	–	FR	–	QA	–	–
Malaviya et al. (2024b)	EVA	–	IC	TX	SEN	PAS	IT	SI	QA	–	FS
Anand et al. (2023b)	APP, RES	–	IC	TX	DOC	CR	FR	MU	CTG	SFT	ZS
Anand et al. (2023a)	APP	–	IC	TX	DOC	CR	FR	SI	CTG	SFT	ZS
Agarwal et al. (2025)	APP, RES	–	PR	TX	DOC	ILC	FR	MU	RWG	SFT	ZS
Sun et al. (2024)	APP	–	PR	TX	PAR	ILC	FR	MU	QA	–	ZS
Hu et al. (2025b)	EVA, RES	–	PR	TX	DOC	ILC	FR	MU	QA	SFT	ZS
Li et al. (2024h)	EVA, RES	–	–	TX	PAR	–	–	SI	OTH	–	–
Li et al. (2024d)	APP	–	PG	TX	PAR	ILC	FR	SI	QA	–	ZS, COT
Li and Ouyang (2025)	APP	–	PR, IC	TX	DOC	NC	FR	SI	RWG	–	ZS
Fang et al. (2024)	APP, EVA	–	PR	TX	PAR	ILC	IT	MU	QA	–	FS
Chiang and Lee (2024)	APP, EVA	–	PR	TX	PAR	ILC	IT	MU	GTG	–	FS
Huang et al. (2024a)	APP	–	PR	TX	PAR	ILC	FR	MU	QA	SFT, RL	ZS
Zhang et al. (2025c)	EVA, RES	–	PR	TX	DOC	ILC	FR	MU	SUM	SFT	ZS
Lee et al. (2024a)	APP, EVA	–	IC	TX	PAR	ILC	FR	MU	QA	–	ZS
Slobodkin et al. (2024)	APP	–	IC	TX	SEN	ILC	FR	MU	QA, SUM	SFT	FS, COT
Li et al. (2024b)	APP, RES	DC	–	TX	PAR	ILC	FR	MU	QA	SFT, RL	ZS
Deng et al. (2024)	APP, EVA, RES	–	PR	TX	PAR	ILC	FR	MU	QA, SUM	SFT	ZS, FS
Berchansky et al. (2024)	APP	–	PR	TX	PAR	ILC	FR	MU	QA	SFT	FS
Li et al. (2025)	SUR	–	–	TX	–	–	–	–	OTH	–	–
Lee et al. (2025)	APL	PLL	–	TX	DOC, PAR, TOK	CR, HG	FR	MU	QA	SFT	–
Fierro et al. (2024)	APP	–	PR	TX	PAR	ILC	FR	MU	QA, SUM	SFT	–
Khalifa et al. (2024)	APP	MC	–	TX	DOC	ILC	FR	SI	QA	SFT, PT	–
Golany et al. (2024)	RES, APL	–	IC	TX	PAR	CR	FR	MU	GTG	SFT	FS
Do et al. (2024)	EVA	PLL	–	TX	PAR	ILC, PAS, HG	FR	MU	OTH	–	–
Magesh et al. (2025)	EVA, RES	PLL	PR	TX	DOC	ILC	FR	MU	QA	–	–

Paper	Contribution Type	Attribution Approach		Citation Characteristics					Task	LLM Integration	
		Par.	Non-Par.	Mod.	Evi.	Style	Vis.	Freq.		Train.	Prompt.
Li et al. (2024c)	APP	–	PR, IG	TX	PAR	ILC	FR	SI	QA, FV, OTH	–	ZS
Mayfield et al. (2024)	POS	–	PR	TX	DOC	ILC	FR	MU	SUM	–	–
Phukan et al. (2024)	APP, EVA, RES	–	–	TX	TOK	PAS	FR	SI	QA	–	–
Maheshwari et al. (2024)	APP, EVA	–	IC	TX	PAR	CR	FR	MU	GTG	–	ZS
Rorseth et al. (2024)	POS	–	PR	TX	PAR	CR	FR	MU	QA	–	ZS
Xu et al. (2025)	EVA	–	PR	TX	SEN	ILC	FR	MU	QA	–	FS
Cattan et al. (2025)	APP, EVA	–	IC	TX	PAR	CR	FR	MU	QA	–	ZS, FS
Dehghan et al. (2024)	APP	–	PR	TX, GR	PAR	ILC	FR	MU	QA	–	FS, COT
Tahaei et al. (2024)	APP, RES	–	PR	TX	PAR	ILC	FR	MU	QA	SFT	ZS
Xing et al. (2025)	EVA, RES	–	PR, IC	TX	PAR	ILC	FR	SI	GTG	–	ZS
Aly et al. (2024)	APP	–	PR	TX	PAR	ILC	FR	MU	QA	SFT	–
Qi et al. (2024)	APP	–	PR	TX	DOC	ILC	FR	MU	QA	–	ZS
Byun et al. (2024)	APP, EVA, RES	PLL	–	TX	DOC	ILC, CR	FR	SI	GTG, RWG	–	ZS
Zhang et al. (2024)	EVA	–	–	TX	–	ILC	FR	MU	QA	–	–
Patel et al. (2024)	APP, RES	–	PR, PG	TX	DOC, PAR	ILC	FR	MU	QA	SFT, PT	–
Cao and Wang (2024)	EVA, RES	–	IC	TX	DOC	ILC	FR	SI	QA	SFT	ZS
Buchmann et al. (2024)	EVA, RES	–	PR, PG	TX	PAR, SEN	CR	FR	MU	QA, SUM, FV	SFT	ZS
Xia et al. (2025a)	APP, RES	–	PR	TX	PAR	CR, NC	FR	SI	QA	SFT	ZS, FS
Xia et al. (2025b)	APP	–	PR	TX	PAR	ILC	IT	MU	QA	SFT	ZS, FS
Dani et al. (2024)	EVA, RES, APL	PLL	–	TX	DOC	CR	FR	SI	GTG	–	ZS, FS, COT, SC
Laban et al. (2024)	EVA, RES	–	PR	TX	DOC	ILC	FR	MU	SUM	–	ZS
Şahinuç et al. (2024)	EVA, RES	–	IC	TX	PAR	ILC	FR	SI	CTG	–	FS, COT, RP
Bitton-Guetta et al. (2024)	APP, EVA, RES	–	IC	TX	DOC	ILC	FR	SI	QA	–	ZS
Rajapaksha et al. (2025)	RES, APL	PLL	PR	TX	DOC	CR	FR	SI	QA	–	ZS, FS
Shen et al. (2024)	EVA	PLL	PR, PG	TX	PAR	ILC	FR	MU	QA	–	ZS, FS, COT, SC
Sui et al. (2024)	APP	–	PR	TX	PAR	ILC	FR	MU	QA	–	FS
Hashemi et al. (2024)	EVA	–	–	TX	DOC	–	–	MU	GTG	–	–
Huang et al. (2024b)	APP	DC	PR	TX	PAR	ILC	FR	MU	QA	SFT, RL	ZS
Li et al. (2024g)	APP, RES	–	IC	TX	DOC	ILC	FR	MU	QA	SFT	COT, AO, COC, COQ
Allen et al. (2024)	APL	–	PR	TX	PAR	PAS	FR	SI	QA	–	ZS
Nishimura et al. (2024)	APP, EVA, RES	–	IC	TX	DOC	ILC	FR	SI	RWG	–	ZS
Djeddal et al. (2024)	EVA	–	PR, PG	TX	DOC	ILC	FR	MU	QA	–	ZS
Cohen-Wang et al. (2024)	APP	–	IC	TX	SEN, TOK	ILC	FR	SI	QA, SUM	–	ZS
Gilson et al. (2024)	APL	PLL	PR	TX	DOC	CR	FR	MU	QA	–	ZS
Ramu et al. (2024)	APP	–	PG	TX	SEN	ILC	FR	MU	QA	–	FS
Penzkofer and Baumann (2024)	APP, EVA, RES	–	PR	TX	DOC	ILC	FR	MU	QA	SFT, RL	ZS
Jung et al. (2024)	EVA	PLL	IC	TX	DOC	ILC	FR	–	QA	–	–
Zhang et al. (2025b)	APP, RES	–	PR	TX	DOC	ILC	FR	MU	SUM, RWG, CTG, OTH, OTH, OTH	SFT	ZS
Zhang et al. (2025a)	APP, EVA, RES	–	IC	TX	PAR, SEN	ILC	FR	MU	QA	SFT	ZS, FS
Song et al. (2025)	APP, EVA	–	PR	TX	DOC	ILC	FR	MU	QA	RL	–
Tilwani et al. (2024)	POS	PLL	IG	TX	DOC	CR	FR	–	GTG	–	–
Hannah et al. (2025)	APL	PLL	PG	TX	PAR	ILC	FR	SI	QA	–	FS
Shaier et al. (2024)	EVA, RES	–	IC	TX	DOC	NC	FR	MU	QA	SFT	ZS, FS, COT
Huang et al. (2024c)	APP	–	PG	TX	PAR	ILC	FR	MU	QA	SFT, SSFT	ZS

Paper	Contribution Type	Attribution Approach		Citation Characteristics					Task	LLM Integration	
		Par.	Non-Par.	Mod.	Evi.	Style	Vis.	Freq.		Train.	Prompt.
Yan et al. (2025)	APP, EVA	–	PG	TX	SEN	CR	FR	MU	QA	SFT	FS, COT
Cheng et al. (2025)	EVA, RES	–	PR	TX	PAR	ILC	FR	MU	GTG	SFT	ZS, FS
Vladika et al. (2024)	APP	–	PG	TX	DOC	CR	FR	SI	QA	–	ZS, FS
Abolghasemi et al. (2025)	EVA	–	PR	TX	PAR	ILC	FR	SI	QA	–	ZS
Liu et al. (2024)	APP	–	IC	TX	DOC	ILC	FR	SI	QA	SSFT	–
Qian et al. (2025)	APP, EVA	–	PR, PG	TX	PAR	–	FR	MU	QA	SFT	FS
Chang et al. (2025)	APP	–	PG	TX	PAR	PAS	FR	SI	OTH	–	–
Narayanan Venkit et al. (2025)	EVA, RES	–	PR	TX	DOC	ILC, CR	FR	MU	QA	–	–
Redelaar et al. (2024)	EVA, RES, APL	–	PR	TX	PAR	CR	FR	MU	QA	–	FS
Mathur et al. (2024)	APP, RES	–	PG	TB	TC	ILC	FR	MU	QA	–	FS
Asai et al. (2024a)	APP, RES	–	PR	TX	SEN	ILC	FR	SI	QA	SFT	ZS, COT
Gupta et al. (2024)	APP, RES	–	PR	TX	DOC	ILC	FR	MU	QA	–	ZS, FS
Xiao et al. (2025)	EVA, RES	–	PR	TX	SEN	NC, QU	FR	SI	GTG	–	ZS, FS, COT
Hsu et al. (2025)	APP, RES	–	PR	TX	SEN	ILC, CR	FR	MU	OTH	SFT	ZS
Shetty et al. (2024)	APL	–	IC	TX	DOC	ILC	FR	SI	SUM	–	ZS, FS, COT
Yang et al. (2024)	APP	–	IC	TX	DOC	NC	FR	SI	OTH	SFT, RL	ZS
Worledge et al. (2024a)	EVA	PLL	PR, PG	TX	SEN	ILC	FR	MU	QA	–	FS
Li et al. (2024a)	APL	–	PR	TX	DOC	ILC	FR	MU	QA, SUM	SFT	–
Soman et al. (2024)	APP	–	PR	TX	DOC	ILC	FR	SI	QA	SFT	ZS
Ateia and Kruschwitz (2025)	APL	–	PR	TX	DOC	ILC	FR	MU	QA	–	FS
Zhang et al. (2025d)	EVA, RES	–	PR, PG	TX	DOC	ILC	FR	MU	QA	SFT	ZS
Wallat et al. (2025)	EVA	–	IC	TX	DOC	ILC	FR	MU	QA	–	ZS
Saha Roy et al. (2025)	RES, APL	–	PR	TB	TAB	ILC	FR	MU	GTG	–	ZS
Patel and Anand (2024)	EVA	PLL	–	TX	DOC	NC	FR	MU	QA	–	ZS, CA
Vassos et al. (2024)	APL	–	PR	TX	DOC	PAS	FR	SI	QA	–	ZS
Sharma et al. (2025)	APP	–	PR	TX, GR	PAR	PAS	FR	MU	QA	–	ZS
Wu et al. (2025a)	APP	–	PR	TX	PAR	ILC	FR	MU	QA	SFT, RL	FS
Jin et al. (2025)	EVA, RES	–	PR	TX	SEN	ILC	FR	MU	QA	–	ZS
Li and Ng (2025)	APP	–	IG	TX	PAR	ILC	FR	MU	QA	–	FS
Ma et al. (2025)	APP, RES	–	PR	VL	BB	ILC	FR	SI	QA	SFT	ZS
Suri et al. (2025)	APP, RES	–	PR	TX, TB, VL	PAR, TAB, IMG	CR	IT	MU	QA	–	COT
Craig and Drăghici (2024)	APP	–	PR	TX	PAR	ILC, CR	FR	MU	QA	–	ZS
Devine (2025)	APP	–	PR	TX	DOC	–	FR	SI	QA	SSFT	ZS
Ding et al. (2025)	EVA	–	PG	TX	PAR	ILC	FR	MU	QA	–	ZS
He et al. (2025)	APP, RES	–	IC	GR	TRI	ILC	FR	SI	OTH	SFT	ZS, FS
Abbas et al. (2025)	APL	–	PG	TX	DOC	–	FR	MU	GTG	SFT, RL, PT	–
Chu et al. (2025)	APP	MC	–	TX, GR	SEN	ILC	IT	MU	QA	SFT	ZS, FS

Table 7: Overview of the 134 papers on evidence-based text generation with LLMs included in our survey. The annotation follows the multidimensional taxonomy described in Section 3. For each paper, we record the Contribution Type, Attribution Approach (Parametric: Par., Non-Parametric: Non-Par.), Citation Characteristics (Citation Modality: Mod., Evidence Level: Evi., Citation Style: Style, Citation Visibility: Vis., Citation Frequency: Freq.), Task, and LLM Integration (Training: Train., Prompting: Prompt.). Additional information such as evaluation metrics, frameworks, datasets, and benchmarks is available in our public repository (see Appendix A). Papers are ordered by the publication date of the earliest accessible version, including preprints. For compactness, we employ abbreviations, all of which are defined in Table 6.

Evaluation Method	Description
Human Evaluation	Human evaluation metrics use human judges to evaluate different aspects of LLM-generated texts, typically on numerical scales of 1–5 or 1–10. Most studies define custom evaluation criteria, which limits standardization, comparability, and reusability across approaches.
Inference-based	Inference-based metrics use NLI models to automatically classify whether an LLM-generated text is entailed by a reference text. This evaluation method is commonly used to determine whether a grounded LLM-generated text can be attributed to its source and to assess the factual correctness of an LLM-generated response based on entailment with a ground-truth reference.
Lexical Overlap	Lexical overlap metrics evaluate the degree of lexical overlap between an LLM-generated text and a reference (ground-truth).
LLM-as-a-Judge	LLM-as-a-judge metrics use LLMs to automatically evaluate the quality of texts generated by other LLMs, thereby eliminating the need for human judges. Typically, the LLM acting as a judge generates numerical scores, often normalized to a 0–1 scale.
Retrieval-based	Retrieval-based metrics quantitatively evaluate how effectively an approach incorporates relevant evidence by comparing the retrieved or integrated evidence with a ground-truth. This method is commonly used to assess whether the correct sources are accurately cited in LLM-generated texts, as well as to determine whether appropriate sources can be retrieved and attributed to LLM-generated texts post-hoc.
Semantic Similarity-based	Semantic similarity-based metrics assess the degree of semantic similarity between an LLM-generated text and a reference (ground-truth) text, typically by computing the cosine similarity between their dense vector embeddings.

Table 8: Explanation of the evaluation methods for evidence-based text generation with LLMs.

Dim.	Metric	Measured Aspect	Tasks	Evaluation Context	Limitations
ATT	Citation NLI	To what extent do the provided citations support the associated claims?	QA, GTG, SUM, FV	sentence-level attribution in long-form text generation	– requires pre-trained NLI model – assumes each citation sentence represents a single factual claim – cannot assess uncited claims
	Auto-AIS-sentence	To what extent are generated sentences attributable to supporting sources?	QA, GTG, SUM	sentence-level attribution in long-form text generation	– requires pre-trained NLI model – assumes each sentence represents a single factual claim – requires access to external evidence provided for generation – does not evaluate citation quality
	ROUGE	To what extent does the generated text overlap lexically with the source texts?	ALL	summarization quality, lexical overlap-based alignment with sources	– measures lexical similarity rather than attribution – requires access to external evidence provided for generation – does not evaluate citation quality – lacks claim-level granularity for evaluating long-form texts
	BERT-Score	To what extent is the generated text semantically similar to the source texts?	QA, GTG, SUM, CTG, RWG	summarization quality, semantic similarity-based alignment with sources	– measures semantic similarity rather than attribution – requires access to external evidence provided for generation – does not evaluate citation quality – lacks claim-level granularity for evaluating long-form texts
	FActScore	To what extent are generated atomic facts supported by available sources?	QA, GTG, FV	claim-level attribution in long-form text generation	– requires pre-trained NLI model – requires access to external evidence provided for generation – evaluates factual precision but not factual recall – typically requires human annotation for atomic fact decomposition
CIT	Citation Retrieval	To what extent do generated citations match the set of supporting sources?	QA, SUM	citation-level retrieval alignment with sources	– requires ground-truth citation annotations – depends on completeness of the oracle citation set – does not evaluate whether the cited sources support the generated claims
	Citation Accuracy	To what extent do generated citations align with the ground-truth citation texts?	QA	citation-level text alignment with sources	– requires ground-truth citation text annotations – evaluates citation text matching rather than source selection or support – sensitive to citation text wording differences
COR	Exact Match	To what extent do generated answers exactly match the ground-truth answer texts?	QA	exact match for short factoid answers	– evaluates exact text matching rather than semantic correctness – not applicable to long-form text generation
	Claim Recall	To what extent does the generated text entail the factual claims in the ground-truth answer?	QA, SUM	claim-level correctness in long-form text generation	– requires pre-trained NLI model – depends on claim decomposition accuracy – evaluates claim recall rather than precision
	Accuracy QA	To what extent is the generated answer correct with respect to the ground-truth answer?	QA	answer-level correctness for short answers	– may depend on answer normalization or extraction procedures – not applicable to long-form text generation
	Accuracy NLI	To what extent can the factuality of generated claims be correctly classified?	QA, SUM, CTG	claim-level factuality classification for long-form text generation	– requires pre-trained NLI model – requires human factuality annotations – depends on claim decomposition accuracy – may struggle to detect non-factual claims when counter-evidence is limited
	BLEU-N	To what extent does the generated text overlap lexically with the ground-truth text?	QA, GTG, FV	answer-level lexical overlap-based alignment	– evaluates lexical similarity rather than factual correctness – less reliable for long-form text generation

Table 9: Comparative overview of evaluation metrics for core dimensions. The table summarizes frequently reused evaluation metrics identified in Figure 4, organized by the three core evaluation dimensions (Dim.), namely attribution, citation, and correctness. For each metric, we describe the measured aspect, applicable task settings, evaluation context, and known limitations. While Figure 4 provides an overview of usage frequency, this table supports informed metric selection. All abbreviations are listed in Table 6.

Framework	No. of Papers	Evaluation Dimension	Evaluation Method	Source
ALCE	12	ATT, COR, LQ	IB, LO, SB	Gao et al. (2023b)
G-Eval	2	ATT, LQ, REL	LJ	Liu et al. (2023b)
AEE	1	ATT, COR, CIT, REL	LJ, RB	Narayanan Venkit et al. (2025)
ALiiCE	1	ATT, COR, CIT, LQ	IB, LO, SB	Xu et al. (2025)
Attributed Information Retrieval	1	ATT, COR	IB, LO, RB, SB	Djeddal et al. (2024)
Attribution Bias	1	COR, CIT	LO, RB	Abolghasemi et al. (2025)
BEGIN	1	ATT	IB, LO, SB	Dziri et al. (2022)
BioGen	1	COR, CIT, LQ, REL, RET	HE, RB, SB	Gupta et al. (2024)
CitaLaw	1	ATT, LQ	LO, SB	Zhang et al. (2025d)
CiteKit	1	ATT, COR, CIT, LQ	IB, LO, RB, SB	Shen et al. (2024)
LLM-Rubric	1	ATT, CIT	LJ	Hashemi et al. (2024)
QUILL	1	COR, CIT, LQ, REL, RET	LJ, RB	Xiao et al. (2025)
RAG-RewardBench	1	COR, CIT, REL	LJ	Jin et al. (2025)
RAGAS	1	ATT, COR, REL, RET	IB, LJ, RB, SB	Es et al. (2024)
RAGE	1	CIT	RB	Penzkofer and Baumann (2024)
REC	1	COR, CIT, LQ, REL	LJ	Hsu et al. (2025)
ScholarQABench	1	ATT, COR, LQ, REL	IB, LO, LJ	Asai et al. (2024a)
Trust-Score	1	ATT, COR	IB, RB	Song et al. (2025)
VisDoMBench	1	COR, RET	LO, RB	Suri et al. (2025)

Table 10: Evaluation frameworks ordered by the number of papers reusing them. For each framework, we annotate the evaluation dimensions covered: attribution (ATT), correctness (COR), citation (CIT), linguistic quality (LQ), preservation (PRE), relevance (REL), and retrieval (RET). We additionally annotate the evaluation methods used by each framework: human evaluation (HE), inference-based (IB), lexical overlap (LO), LLM-as-a-judge (LJ), retrieval-based (RB), and semantic similarity-based (SB). In addition, all abbreviations are listed in Table 6.

Benchmark	No. of Papers	Task	Domain	Source
ALCE	9	QA	Social Media, Wikipedia	Gao et al. (2023b)
RAG-RewardBench	2	QA	Multi-Domain	Jin et al. (2025)
AttributionBench	1	GTG, QA	Multi-Domain	Li et al. (2024h)
BEGIN	1	GTG	News, Social Media, Wikipedia	Dziri et al. (2022)
CitaLaw	1	QA	Legal	Zhang et al. (2025d)
LAB	1	FV, QA, SUM	Multi-Domain	Buchmann et al. (2024)
MultiAttr	1	QA	Wikipedia	Patel et al. (2024)
ScholarQABench	1	QA	Health, Scientific, Wikipedia	Asai et al. (2024a)
Trust-Align	1	QA	Social Media, Wikipedia	Song et al. (2025)
TabCite	1	QA	Finance, Wikipedia	Mathur et al. (2024)
VisDoMBench	1	QA	Scientific, Wikipedia	Suri et al. (2025)

Table 11: Evaluation benchmarks ordered by the number of papers applying them. For each benchmark, we annotate the tasks it has been applied to, including question answering (QA), grounded text generation (GTG), summarization (SUM), and fact verification (FV), as well as the data domains. In addition, all abbreviations are listed in Table 6.

Task	Dataset	No. of Papers	Domain	Source
Question Answering	ASQA	26	Wikipedia	Stelmakh et al. (2022)
	ELI5	22	Social Media	Fan et al. (2019)
	NaturalQuestions	19	Wikipedia	Kwiatkowski et al. (2019)
	QAMPARI	11	Wikipedia	Amouyal et al. (2023)
	HotpotQA	10	Wikipedia	Yang et al. (2018)
	ExpertQA	9	Web Search	Malaviya et al. (2024a)
	StrategyQA	7	Wikipedia	Geva et al. (2021)
	HAGRID	5	Wikipedia, Synthetic	Kamalloo et al. (2023)
	TriviaQA	5	Web Search, Wikipedia	Joshi et al. (2017)
	2WikiMultiHopQA	4	Wikipedia	Ho et al. (2020)
	GenSearch-Verifiability	4	Web Search	Liu et al. (2023a)
	PopQA	4	Wikipedia	Mallen et al. (2023)
Grounded Text Generation	Wizard of Wikipedia	4	Wikipedia	Dinan et al. (2019)
	CMU-DoG	2	Wikipedia	Zhou et al. (2018)
	TopicalChat	2	Wikipedia, News, Social Media	Gopalakrishnan et al. (2019)
	WikiBio GPT-3	1	Wikipedia, Synthetic	Manakul et al. (2023)
	SciDuet	1	Scientific	Sun et al. (2021)
	LLM-Rubric	1	Web Search, Synthetic	Hashemi et al. (2024)
	BioKALMA	1	Wikipedia	Li et al. (2024f)
	RefGPT-Fact	1	Wikipedia	Yang et al. (2023)
Summarization	GovReport	2	Government	Huang et al. (2021)
	SummEval	1	News	Fabbri et al. (2021)
	SummHay	1	News, Synthetic	Laban et al. (2024)
	CNN DailyMail	1	News	Nallapati et al. (2016)
Related Work Generation	Byun-CR	1	Scientific	Byun et al. (2024)
	Li-Citation-Graph	1	Web Search	Li and Ouyang (2025)
	RollingEval	1	Scientific	Agarwal et al. (2025)
	STRoGeNS	1	Scientific	Nishimura et al. (2024)
Citation Text Generation	S2ORC	2	Scientific	Lo et al. (2020)
	MCG-S2ORC	1	Scientific	Anand et al. (2023b)
	Gu-CCSG	1	Scientific	Gu and Hahnloser (2024)
	Sahinuç-ACLRelWork	1	Scientific	Şahinuç et al. (2024)
Fact Verification	FEVER	5	Wikipedia	Thorne et al. (2018)
	Min-Biography	2	Wikipedia	Min et al. (2023)
	PubHealth	1	Health, News	Kotonya and Toni (2020)
	HaluEval	1	Wikipedia, Synthetic	Li et al. (2023b)

Table 12: Frequent datasets per task in evidence-based text generation with LLMs. For each task in Section 3.3, we display the most frequent datasets. The domain specifies the scope of the data, characterized by its source. This table presents only a sample and does not show the complete list of 231 datasets. The complete dataset table is provided in our repository, as described in Appendix A. Note, both LLM-Rubric and STRoGeNS contain multiple datasets, which have been combined in this table for the sake of simplicity.