A Survey of Benchmarks and Evaluation Methods for Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) addresses the significant need to answer user queries truthfully, regarding external knowledge sources. There appear various RAG system implementations differing in its architecture, data sources and evaluation procedures. These RAG aspects are discussed in the recent RAG surveys, however no thorough RAG benchmark systematization is yet published. This survey aims to fill this gap for the benefit of RAG system development. We focus on methods for evaluating RAG systems, systematizing and analysing specialized benchmarks and metrics. The survey emphasizes practical user scenarios and provides recommendations for selecting an evaluation method.

1 Introduction

Retrieval-Augmented Generation (RAG) has been established as a strong approach to enhance the LLM's ability to provide factually accurate information to users, mitigating hallucinations. The complexity of RAG systems requires careful, multifaceted evaluations of retrieval, generation, and the system as a whole (Yu et al., 2025). Many recently published benchmarks approach the RAG evaluation problem from different perspectives, regarding RAG task statement, provided query types, domain specificity, hallucination reduction, dynamic nature of queries and retrieval. Current work aims at presenting a comprehensive survey of RAG benchmarks. The paper describes existing RAG benchmarks and evaluation methods along three directions: (i) specially prepared static datasets and tests across different domains, (ii) dynamic benchmarks oriented on time-sensitive knowledge, and (iii) automated evaluation frameworks, where the benchmark is formed from developer-provided data, enabling scalable assessment without manual labeling.

This paper makes the following contribution:

- we overview up-to-date RAG benchmarks and propose their multidimensional categorization;
- we compute relevance density for the analyzed benchmarks;
- we compare automated RAG evaluation frameworks in terms of their capabilities and implemented metrics;
- we aim at offering practical goal-based recommendations for selecting appropriate RAG benchmarks.

2 Related Works

Previous RAG surveys have examined RAG architecture and retrieval mechanisms (Arslan et al., 2024; Fan et al., 2024; Gao et al., 2024), RAG system construction pipeline (Gao et al., 2024; Wu et al., 2025), scenarios of RAG applications (Arslan et al., 2024; Fan et al., 2024; Wu et al., 2025). Despite being focus of several surveys (Gao et al., 2024; Gupta et al., 2024; Gu et al., 2025; Yu et al., 2025), RAG evaluation, needs further thorough research, as new benchmarks appear and evaluation approaches change through time.

Automated RAG evaluation frameworks were discussed in Gao et. al. (2024) and Gu et. al. (2025), highlighting the use of LLM-as-a-Judge methods during evaluation and end-to-end evaluation design. With the development of automated RAG evaluation tools, a recent survey has examined the reliability of LLM-based evaluation methods (Brehme et al., 2025). Emphasizing structured thorough evaluation approach, Yu et. al. (2025) proposes Auepora aimed at comprehensive RAG system evaluation from both data and metric perspectives. The current survey supports this approach and proposes benchmark systematization according to different development and evaluation

objectives of RAG systems. Overall, the work extends existing surveys by offering a comprehensive and benchmark-centered perspective on RAG evaluation, and by analyzing benchmarks through the lens of evaluation goals and user-oriented scenarios rather than architectural design alone.

3 Benchmark Profiles

The set of further observed RAG benchmarks was curated based on (i) status of the publishing conference, (ii) paper citation number, and (iii) specific unique qualities, e.g. narrow domain RAG or multilingual RAG evaluation. The set partially overlaps with the existing RAG surveys (Gao et al., 2024). This section comprises detailed overview of standard RAG benchmarks, including different qualitative and quantitative metrics, overview and comparison of 4 selected automated RAG evaluation framework, and lastly general overview of metrics in both benchmark types.

3.1 Benchmark Characteristics

Data collections used in RAG benchmarks varies in size dramatically (see Table 2 in Appendix A). Data is predominantly English and Chinese with the small exception for few multilingual benchmarks (Blandón et al., 2025; Thakur et al., 2023). The covered domains are usually broad, except for DelucionQA (2023), DomainRAG (2024), WeatherArchive-Bench (2026), and Legal RAG Bench (2026), covering very specific RAG application.

Data collections in the benchmark were derived either from established RAG datasets, or human-annotated QA, or synthetically generated data (see Table 2 in Appendix A), with these approaches sometimes being combined. However, synthetic QA generation is the most common among data collection approaches, varying from straightforward to sophisticated Q/QA generation. Specifically, RGB benchmark (Chen et al., 2023) demonstrates use of entity recognition and knowledge graphs for QA generation. We differentiate between various forms of generated content, namely generating queries (labeled as q), generating question-answer pairs (labeled as qa) and creating structured task (e.g. multiple document summary) alongside the question-answer pairs (labeled as d+qa).

As shown in Table 2 (see Appendix A), question-answer generation (qa) is the most widely adopted strategy across all language settings, making it

Figure 1: Heatmap of binary dataset attributes across languages. Each cell shows the proportion of datasets exhibiting the attribute, with color intensity representing the value.

the most general and broadly applicable approach. In contrast, question-only generation (q) is predominantly used in English benchmarks, whereas non-English benchmarks more frequently rely on document-grounded generation (d+qa).

Notably, human annotation in some cases was used to evaluate quality of generated data (manual verification of synthetics is specified in the corresponding column in Table 2 in Appendix A). According to Figure 1, Chinese benchmarks are largely based on collected data with limited manual verification of generated data, whereas English benchmarks more often reuse existing datasets and less frequently rely on data collection. Overall, this diversity in benchmark size, domain coverage, and data construction strategies reflects the heterogeneity of RAG evaluation settings and motivates the need for careful benchmark selection depending on the evaluation goals.

3.2 Benchmark Descriptors

Focusing on the retrieval sources provided by benchmarks, we characterize benchmarks by basic dataset descriptors (number of documents, number of queries, and domain) and metric *relevance density* (Thakur et al., 2021). These descriptors display the amount of relevant evidence available per query and the sparsity of queries relative to the corpus.

Let $|Q|$ the number of queries, and $|R|$ the number of relevance judgments (often referred to as qrels in IR evaluation). We define relevance density as RelDensity $= |R|/|Q|$ which corresponds to the average number of relevant documents per query (often reported as Avg. D/Q) (Thakur et al., 2021; Dai et al., 2024). This distinguishes single-evidence settings from multi-evidence or multi-hop benchmarks (Tang and Yang, 2024; Lyu et al., 2024). We compute RelDensity for all benchmarks summarized in Table 2 (see Appendix A), excluding CRAG, RECALL, and ReEval due to unavail-

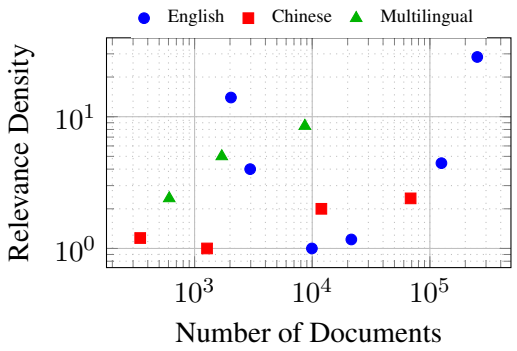

Figure 2: Scatter plot of RAG benchmarks in the space of corpus size and relevance density (log–log scale). Marker shapes denote language coverage.

able or non-comparable corpus-level statistics.

According to Figure 2, Chinese benchmarks tend to exhibit lower relevance density, indicating fewer relevant documents per query on average. Multilingual benchmarks, in turn, are generally characterized by smaller corpus sizes compared to English benchmarks. These trends suggest that evidence sparsity and corpus scale vary substantially across language settings, which may influence the difficulty and evaluation characteristics of RAG systems.

3.3 Benchmark Tasks

Main task. Question answering (QA) remains a basic task for LLMs, and thus, question answering is included in all overviewed benchmarks, excluding only a few that focus on hallucinations (Li et al., 2023; Friel et al., 2024). The QA task is affected by benchmark data collection and expected RAG application scenarios. Specifically, there appear RAG benchmarks investigating how multi-turn conversation influences RAG performance (Liu et al., 2024; Rosenthal et al., 2026). Solving routine user tasks may require *active retrieval* as both follow-up user queries and provided answers have effect on the target model answer (Katsis et al., 2025).

Multi-hop Question Answering. The real-world application scenarios of RAG encompass multi-hop queries that require retrieving and reasoning over evidence from multiple documents. The model capability to correctly answer multi-hop questions is crucial for it to be flexible in day-to-day conversations. MultiHop-RAG benchmark (Tang and Yang, 2024) is specifically designed to overcome this issue, presenting different types of multi-hop QA (inference queries, comparison queries, temporal queries). CRAG (Yang et al., 2024a) proposes more diverse multi-hop QA type range including also queries with false preposition or assumption, queries requiring aggregating the retrieval, queries requiring reasoning. Other benchmarks set the multi-hop task simply as synthesis of multiple paragraph information (Lyu et al., 2024; Yang et al., 2024a; Zhou et al., 2025), or include previously collected multi-hop datasets (Friel et al., 2024).

Specific application scenarios. Not limited by question answering, RAG may also be important in other application scenarios such as text generation. CRUD benchmark introduce such new RAG tasks as creative generation (continuing input text using external information sources), hallucination modification, and open-domain multi-document summarization (Lyu et al., 2024). RECALL benchmark also includes text generation scenario, drawing attention to robustness against counterfactual knowledge from external texts (Liu et al., 2023).

Detecting hallucinations. Hallucinations are a key issue in RAG, as LLM may guided both retrieved context and its internal knowledge, producing untruthful answers. The reviewed benchmarks either include special tasks to evaluate hallucinations in application scenario (Lyu et al., 2024), or are primarily focused on hallucinations appeared in the RAG pipeline (Li et al., 2023; Niu et al., 2024; Sadat et al., 2023). HaluEval benchmark (Li et al., 2023) differs from classical RAG benchmark focusing on recognizing hallucinations in the collected text generations. Including external knowledge in prompt or providing reasoning steps improves LLM hallucination recognition. RAGBench (Friel et al., 2024) addresses an issue of hallucinations in the existing RAG datasets.

Dynamic benchmarks. As both queries and textual information for RAG may become outdated quickly, dynamic RAG benchmarks are crucial for evaluating system stability and flexibility. Early RAG benchmarks typically used determined text collections, often based on Wikipedia articles (e.g. (Kwiatkowski et al., 2019)), whereas dynamic queries may concern facts that change over time.

CDQA benchmark (Xu et al., 2024) proposes a collection of human-annotated dynamic queries to evaluate model robustness towards changing facts and address training-data contamination by testing adaptation to up-to-date information. The CRAG benchmark also includes dynamic queries (Yang et al., 2024b) created from templates with entities from a knowledge graph, making query col-

lection more autonomous than in CDQA. Providing even more variability, DRAGON benchmark also updates the text corpus (Chernogorskii et al., 2025): the corpus is automatically updated with recent news texts, while queries are generated from a knowledge-graph subsample with further filtration. Additionally, DRAGON is released as a RAG evaluation framework similar to those discussed in Section 3.4. LiveSearchBench (Zhou et al., 2025) presents two separate retrieval snapshots for diachronic analysis, and authors aim to update continually benchmark.

Long-context reasoning benchmarks. BABILong (Kuratov et al., 2024) is a RAG-adjacent *reasoning-in-a-haystack* benchmark: it evaluates whether models can locate and reason over a few relevant facts buried inside extremely long texts. BABILong primarily targets long-context reasoning over the provided input. While retrieval can be added as a system component, the benchmark's scores mainly reflect how well models use very long contexts rather than end-to-end retrieval performance.

3.3.1 Critical analysis of benchmarks

Existing RAG-related benchmarks are valuable but not fully comparable, because they operationalize RAG differently. Some primarily evaluate grounded generation given provided evidence (e.g., RAGTruth, DelucionQA), others emphasize attribution judgments at finer granularity (AttributionBench's claim-level decomposition; RAGTruth's span-level annotation), while robustness suites such as RGB and NoMIRACL isolate specific failure modes (noise, non-relevance, rejection) rather than a single end-to-end corpus–query–qrels pipeline. As a result, reported scores frequently conflate retrieval quality, context construction, and generation behavior, limiting direct cross-benchmark ranking.

Several benchmarks also introduce reproducibility constraints. CDQA and other time-sensitive setups depend on external sources or search results that drift over time, while licensing and distribution choices can prevent releasing a stable corpus (e.g., DelucionQA's reconstructed contexts).

3.4 Automated Benchmarks

The set of observed automated RAG evaluation frameworks was selected based on (i) technical maturity and public availability, (ii) coverage of core RAG evaluation requirements, and (iii) applicability in real-world pipelines. These frameworks offer automated evaluation of RAG system given either annotated QA dataset or user document collection. The main advantages of automated evaluation frameworks are cross-lingual robustness, domain independence and automated test generation, scalability and cost-effectiveness, fine-grained component diagnostics, reproducibility and reduced subjectivity, confidentiality.

We also considered widely used industry-oriented frameworks, such as LangSmith RAG evaluation tools, LlamaIndex evaluation modules, and OpenAI Evals (LangChain; LlamaIndex; OpenAI). These systems are not encompassed by the detailed analysis, because the overview focuses on frameworks that match our selection criteria.

All frameworks share core features: LLM-based automated assessment of response quality, modular metric design, and a set of predefined metrics. Beyond answer-quality scoring, they probe retrieval quality and assess grounding or faithfulness of generated claims, including hallucination detection. Some frameworks also generate synthetic evaluation data from raw documents. The primary differences concern metric calculating strategies and the set of the available metrics. A summary comparison is presented in Table 1.

RAGAS. RAGAS (VibrantLabs, 2024) supports reference-free evaluation by generating an evaluation set from user-provided documents and scoring RAG outputs with an automatic metric suite. Test-set generation targets different query regimes along two axes: *single-hop* vs. *multi-hop* and *specific* vs. *abstract*, enabling a controllable mixture of question types. In practice, users can generate queries by building a document graph enriched with extracted properties, selecting evidence for single-hop or multi-hop settings, and prompting an LLM to produce queries and reference answers. *Scenarios* define the intent and constraints of the query template, while *Personas* condition the phrasing style. The resulting queries are run through the target RAG system to collect retrieved contexts and answers, after which RAGAS computes context- and answer-centric metrics. Most metrics rely on an LLM-as-a-judge, but the framework also supports classical metrics such as BLEU and ROUGE.

ARES. ARES (Saad-Falcon et al., 2023) evaluates RAG systems by training lightweight classifier-based judges on an in-domain synthetic dataset, then reporting system-level scores with Prediction-Powered Inference (PPI) confidence intervals cali-

brated on a small human-labeled validation set (Angelopoulos et al., 2023). In practice, users provide an in-domain passage set, a small human preference validation set, and a few-shot set of in-domain passage-to-query/answer examples to bootstrap synthetic data generation. A generator LLM produces synthetic question–answer pairs, low-quality queries are filtered out, and the remaining positives are augmented with weak and strong negatives. ARES then trains separate automated judges, calibrates them using the human-labeled set via PPI, and applies the judges to sampled outputs of the target RAG system to report scores with confidence intervals.

RAGChecker. RAGChecker(Ru et al., 2024) diagnoses RAG pipelines via claim-level entailment. Instead of a single response-level score, it decomposes outputs into atomic claims and checks whether each claim is supported by the retrieved context or entailed by a reference answer. The reference answer is used only for evaluation and is not shown to the model at inference time. This yields claim-level precision/recall/F1 for end-to-end comparison and component diagnostics that separate retrieval behavior from generation behavior. In practice, users provide instances as query–retrieved contexts–answer, optionally with a reference, run claim decomposition and entailment checks, and aggregate metrics across the evaluation set.

TruLens. TruLens (2025) differs from other tools by being primarily an observability layer for LLM/RAG applications: it wraps the application to record execution traces (inputs/outputs and intermediate steps), runs configurable feedback functions to score each run, and provides a dashboard for inspecting both aggregate metrics and per-example traces. In practice, users wrap their RAG pipeline with a TruLens integration, define feedback functions, run the system to log traces and compute feedback over recorded runs, and then inspect aggregate summaries and per-example traces in the dashboard. By default, TruLens provides a compact set of RAG-oriented feedback functions, and users can extend this set with custom metrics.

3.4.1 Meta-analysis of Auto-evaluation Frameworks

Automated assessment frameworks provide two capabilities: they scale and augment RAG analysis based on existing datasets and construct evaluation datasets from raw text. Among the reviewed frameworks, RAGAS implements one of the most advanced approaches to generating question–answer pairs by leveraging knowledge graphs, multi-hop questions and simulating user personas.

Framework	AutoQGen	REval	HEval	MHEeval
RAGAS	Yes	Yes	Yes	Yes
RAGChecker	No	Yes	Yes	No
ARES	Yes	Yes	Yes	No
TruLens	Yes	Yes	Yes	Limited

Table 1: Summary of capabilities across selected automated RAG evaluation frameworks. AutoQGen – automated question generation; REval – retrieval evaluation; HEval – hallucination evaluation; MHEval – multi-hop evaluation.

Many of metrics in observed frameworks are computed using an LLM-as-a-judge approach. Despite the existence of common metric types, frameworks generally differ in the specific supported metrics and evaluation protocols that vary from whole-response scoring to claim-level verification. As a result, there is an open question how to compare metrics across frameworks in a consistent way. Since many automated metrics rely on an LLM-as-a-judge approach, issues of reliability and reproducibility across different judge LLMs become important.

Comparative Reliability and Discriminative Power Comparative studies show substantial differences in evaluator reliability and discriminative power. ARES reports higher ranking correlation by using Prediction-Powered Inference and domain-specific judges, while RAGChecker shows that claim-level entailment better captures subtle hallucinations than response-level scoring. RAGBench further highlights the fragility of zero-shot LLM judges in industrial domains and shows stronger calibration and robustness for fine-tuned evaluation models. These findings suggest that reference-free evaluation with fixed prompts requires careful validation, especially in high-stakes settings (Saad-Falcon et al., 2023; Ru et al., 2024; Friel et al., 2024).

Thus, while automated evaluation frameworks offer several practical advantages, especially in cases where there is no ready-made benchmark for a specialized task, this area still requires further research and systematization.

3.5 RAG Evaluation Metrics

RAG evaluation usually starts with evaluating information retrieval with the use of standard metrics such as *MAP@K*, *MRR@K*, *Hit@K*, *NDCG@K* (Tang and Yang, 2024; Lyu et al., 2024; Trulens, 2025). Interestingly, the most of the overviewed benchmarks overlook those metrics, preferring to evaluate retrieval quality according to the given query using LLM-as-a-judge instead.

The overviewed automated frameworks apply these metric groups through different scoring protocols. RAGAS offers LLM-as-a-judge evaluation and supports classical metrics such as BLEU and ROUGE. ARES trains separate judges for *Context Relevance*, *Answer Faithfulness*, and *Answer Relevance*, and measures their estimates with PPI. RAGChecker employs claim-level analysis, including retriever-oriented and generator-oriented metrics. TruLens uses default feedback around the *RAG Triad*: context relevance, groundedness, and answer relevance.

Taking RAGAS metrics (VibrantLabs, 2024) as a starting point, *Context precision* (same in RAGChecker (2024)) aims to detect in both LLM and non-LLM variants relevant chunk share. More fine-grained metrics similar to Context precision would be *Context recall* (or *Claim recall* in RAGChecker (2024)) and *Context entity recall* intended to evaluate facts from the retrieved chunks separately. Similarly, there functions *Context relevance* (same in ARES (2023) and TruLens (2025)) aimed to assess the overall retrieval relevance based on multiple LLM judgements (specifically enhanced with PPI in ARES (2023)).

Following retrieval stage, generation is evaluated separately. *Response Groundedness* proposed in RAGAS framework (similar to *Context Utilization* (Ru et al., 2024; Friel et al., 2024)) assesses whether the generated LLM answer is grounded in the retrieved contexts. Taking into consideration correspondence between the generated response and the target, LLM answer may be evaluated using standard widely used metrics, namely *Exact match* (leveraged in all overviewed benchmarks), *token-level F1*, *BLEU*, *ROUGE* and *Cosine (also semantic) similarity*. Notably, CRUD (Lyu et al., 2024) includes *BERTScore* (Zhang et al., 2020) for generated answer and ground truth comparison, presumably because of the wide range of proposed RAG application tasks including summarization and text continuation.

Besides correspondence with the reference response, the LLM generated answer should be relevant to user query that is measured with *Response Relevancy* (same as *Answer Relevance* (Saad-Falcon et al., 2023; Trulens, 2025)). At the same time, the LLM is expected to incorporate in the answer only claims supported by the retrieved contexts that is measured with *Faithfulness* (VibrantLabs, 2024; Saad-Falcon et al., 2023; Ru et al., 2024) (same as *Groundedness* in Trulens (2025)), indicated among three main crucial LLM capability for RAG in ARES (2023). Similar to Faithfulness is *Completeness* metric proposed in RAGBench (Friel et al., 2024). RGB benchmark (Chen et al., 2023) also examines information integration, differing from Faithfulness in the narrow focus on incorporating information from multiple documents, but highlighting the multi-hop problem.

As LLMs are vulnerable to potential misleading claims in the retrieval, it is essential for LLMs to selectively extract significant facts, filtering noise and unreliable information (Chen et al., 2023). *Noise Sensitivity* (VibrantLabs, 2024; Ru et al., 2024) is aimed to evaluate fraction of incorrect retrieval claims in the LLM answer, and lower score indicates LLM noise robustness. *Factual Correctness* (similar to *Hallucination rate* in RAGChecker (2024), *Adherence* in RAGBench (2024) and *Counterfactual robustness* (Chen et al., 2023)) likewise measures response robustness. The key difference is the emphasis on claims in response being faithful and not hallucinated. First, the response is decomposed into claims, and then the claims are verified using NLI, namely assessing whether the generated answer is entailed in the retrieval or not, and finally the error rate or F1 are calculated (Yu et al., 2023; VibrantLabs, 2024; Ru et al., 2024).

Benchmark authors rarely address the issue of hallucinations caused by lack of information and LLM's need to respond except for a few papers. RGB (Chen et al., 2023) and ChatQA (Liu et al., 2024) benchmarks propose to decline answering queries that can not be supported by any chunk in retrieval. The *Negative Rejection* (Chen et al., 2023) may be measured as accuracy of model declining responding to unanswerable queries. Remaining metrics concern LLM alignment, e.g. evaluating harmfulness or toxicity, sentiment alignment, language match, fairness bias (Trulens, 2025).

3.5.1 Metric alignment and meta-analysis feasibility

We examined the feasibility of cross-benchmark meta-analysis for RAG evaluation, accounting for tasks, metrics, protocols, and model overlap. This analysis shows that, despite sharing metric names (e.g., BLEU, ROUGE-L, F1), most RAG benchmarks are not directly comparable. Identical metrics are frequently applied to different task formulations and evaluation targets. Consequently, metric name overlap does not imply semantic or methodological equivalence.

After aligning benchmarks by task type, metric definition, and evaluated models, we find that only a very limited subset of settings admits direct numerical comparison, namely token-level F1 for extractive QA evaluated on overlapping models. Even in this restricted case, the number of model–benchmark intersections is small, preventing robust quantitative conclusions or benchmark ranking.

More generally, this analysis highlights that RAG benchmarks differ not only in domain, language, and data construction, but also in their underlying evaluation objectives. As a result, benchmarks capture distinct aspects of RAG system behavior rather than forming a single comparable scale of difficulty. This limits the applicability of aggregate cross-benchmark performance analysis under current evaluation practices.

4 Practical Recommendations

Based on our analysis, we developed a decision tree for selecting benchmarks for RAG evaluation that depends on the evaluation goal, the target RAG component, label availability, the need for a dynamic knowledge base (KB), and the specific target evaluation aspects (Figure 3). The benchmark selection is determined by the evaluator's goal: research or production. When the required labels are missing or limited, auto-evaluation frameworks can be used to augment labels and provide additional metrics for in-depth analysis. They are also useful for domain-specific evaluation, either by enriching existing datasets with additional annotations or by generating evaluation datasets from a custom KB.

5 Open Challenges and Future Directions

Our analysis of RAG benchmarks and evaluation frameworks reveals several challenges that limit the reliability, comparability, and generalizability of current assessments.

Judge dependence and metric alignment Many automated RAG evaluation frameworks rely on LLMs to score factual correctness, context relevance, and claim verification. However, differences in judge models, prompt design, metric definitions, and evaluation protocols lead to inconsistent scores across frameworks. This makes cross-framework comparison difficult and highlights the need for calibrated judges, standardized protocols, and task-aware evaluation checklists.

Coverage gaps Most benchmarks focus on English or Chinese and widely studied domains. Although LLM-based automated benchmark generation can partially alleviate narrow-domain gaps, coverage remains limited and often lacks rigorous human validation. Community efforts to create shared multilingual RAG evaluation suites would help address these gaps.

Risks of synthetic data Many benchmarks rely on LLM-generated question-answer pairs or synthetic data with low rates of human verification. While this scales evaluation, it may introduce biases, hallucinations, and domain artifacts. Future work should quantify these risks and develop human-in-the-loop validation and bias-audit procedures. Combining synthetic data with an LLM-as-a-judge approach may increase the risk of circular evaluation, where data generation and scoring inherit similar biases and failure modes. Therefore, synthetic RAG evaluation pipelines should contain human-validated subsets, judge quality assessment, and advanced analysis of generator and LLM-as-a-judge models and prompts (Brehme et al., 2025; Saad-Falcon et al., 2023).

Lack of standardized evaluation protocols A major challenge is the lack of standardized evaluation protocols for RAG benchmarks. As demonstrated in Section 3.5.1, meaningful cross-benchmark comparison s only feasible in rare cases when models, metrics, and task definitions align. This fragmentation hinders reproducibility and limits general conclusions about RAG system performance.

Emerging evaluation directions Next-generation RAG benchmarks should better cover deployment and interaction settings that remain weakly represented in current evaluation

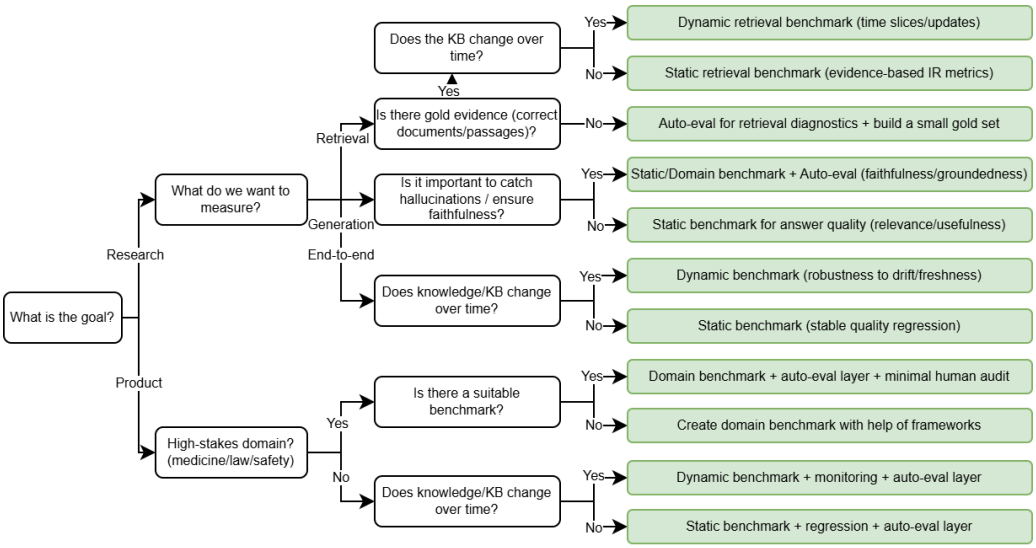

Figure 3: Decision tree for RAG evaluation resource selection by goal, component, labels, and KB dynamics.

suites. Alignment-aware evaluation should assess factual correctness, retrieval relevance, constraint following, evidence attribution, abstention under insufficient evidence, and robustness to conflicts between retrieved and parametric knowledge (Jin et al., 2024). Edge-side RAG evaluation is needed for local and resource-constrained pipelines, where retrieval quality interacts with memory footprint, indexing, caching, privacy, and latency (Seemakhupt et al., 2024). Cost- and energy-aware evaluation should complement accuracy metrics, since retrieval depth, reranking, context length, embedding size, and model choice affect latency, monetary cost, and energy consumption, often with accuracy trade-offs (Guo et al., 2026). Finally, human-centered RAG evaluation should measure not only answer quality, but also trust, verification effort, correction rate, citation usefulness, and decision-support utility (Mangold and Hoffmann, 2025). Together, these directions extend RAG evaluation toward reliability, deployability, efficiency, and human utility.

Addressing these challenges is essential to ensure reliable, reproducible, and practically meaningful assessment of RAG systems.

6 Conclusions

In this paper, we conducted a comprehensive survey of benchmarks for RAG evaluation. We categorized and systematized static and dynamic benchmarks and presented an overview of frameworks for automated evaluation of RAG systems. We provided an analysis of each benchmark, calculated its characteristics, and created meta-analyses of the observed research directions. Then, we reviewed metrics for RAG evaluation in standard benchmarks and frameworks for automatic assessment. Finally, we presented practical recommendations for RAG evaluation, and we highlighted open challenges and suggested future research directions.

7 Limitations

Current survey has several limitations. The scope of the reviewed benchmarks is restricted to text-based ones, excluding multimodal RAG benchmarks. Presented benchmark characteristics and benchmark tasks' summary involve qualitative judgments, to which different researchers might reasonably disagree. Same applies to practical recommendations, future work should incorporate also their reliability or validity tests. Despite described limitations, the surveys presents a comprehensive overview of the current RAG benchmarking landscape, highlighting practical benchmark uses, and future research directions.

Artifacts and licensing. We analyze publicly documented RAG resources for scholarly comparison and do not redistribute original artifacts. Users should consult the original sources for licenses and access conditions.

References

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic.

2023. Prediction-powered inference. *Preprint*, arXiv:2301.09633.

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).

María Andrea Cruz Blandón, Jayasimha Talur, Bruno Charron, Dong Liu, Saab Mansour, and Marcello Federico. 2025. Memerag: A multilingual end-to-end meta-evaluation benchmark for retrieval augmented generation. *Preprint*, arXiv:2502.17163.

Lorenz Brehme, Thomas Ströhle, and Ruth Breu. 2025. Can llms be trusted for evaluating rag systems? a survey of methods and datasets. In *2025 IEEE Swiss Conference on Data Science (SDS)*, pages 16–23. IEEE.

Abdur-Rahman Butler and Umar Butler. 2026. Legal rag bench: an end-to-end benchmark for legal rag. *Preprint*, arXiv:2603.01710.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation. *Preprint*, arXiv:2309.01431.

Fedor Chernogorskii, Sergei Averkiev, Liliya Kudraleeva, Zaven Martirosian, Maria Tikhonova, Valentin Malykh, and Alena Fenogenova. 2025. Dragon: Dynamic rag benchmark on news. *Preprint*, arXiv:2507.05713.

Sunhao Dai, Weihao Liu, Yuqi Zhou, Liang Pang, Rongju Ruan, Gang Wang, Zhenhua Dong, Jun Xu, and Ji-Rong Wen. 2024. Cocktail: A comprehensive information retrieval benchmark with mixed human and llm-generated content. In *Findings of ACL*.

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. *Preprint*, arXiv:2405.06211.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *Preprint*, arXiv:2407.11005.

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.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Zhinuan Guo, Chushu Gao, and Justus Bogner. 2026. On the effectiveness of proposed techniques to reduce energy consumption in RAG systems: A controlled experiment. *arXiv preprint arXiv:2601.02522*.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions. *Preprint*, arXiv:2410.12837.

Zhuoran Jin, Hongbang Yuan, Tianyi Men, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. RAG-RewardBench: Benchmarking reward models in retrieval augmented generation for preference alignment. *arXiv preprint arXiv:2412.13746*.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. Mtrag: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Preprint*, arXiv:2501.03468.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Preprint*, arXiv:2406.10149.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, and 1 others. 2019. Natural questions: A benchmark for question answering research. In *Transactions of the Association for Computational Linguistics*.

LangChain. Evaluate a RAG Application. `https://docs.langchain.com/langsmith/evaluate-rag-tutorial`. Accessed: 2026-05-25.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *Preprint*, arXiv:2305.11747.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024. Attributionbench: How hard is automatic attribution evaluation? *Preprint*, arXiv:2402.15089.

J Liu, C Lin, J Zhang, Z Zhu, Y Xu, W Xu, R Wang, Y Zhu, and P Zhang. 2024. Chatqa: Building gpt-4 level conversational qa benchmarks for retrieval-augmented generation. *Preprint*, arXiv:2401.10225.

Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Recall: A benchmark for llms robustness against external counterfactual knowledge. *Preprint*, arXiv:2311.08147.

LlamaIndex. Evaluating. `https://developers.llamaindex.ai/python/framework/module_guides/evaluating/`. Accessed: 2026-05-25.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2024. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *Preprint*, arXiv:2401.17043.

Aline Mangold and Kiran Hoffmann. 2025. Human-centered evaluation of RAG outputs: A framework and questionnaire for human-AI collaboration. *arXiv preprint arXiv:2509.26205*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Preprint*, arXiv:2401.00396.

OpenAI. Working with Evals. `https://developers.openai.com/api/docs/guides/evals`. Accessed: 2026-05-25.

Sara Rosenthal, Yannis Katsis, Vraj Shah, Lihong He, Lucian Popa, and Marina Danilevsky. 2026. Mtrag-un: A benchmark for open challenges in multi-turn rag conversations. *Preprint*, arXiv:2602.23184.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Preprint*, arXiv:2408.08067.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *Preprint*, arXiv:2311.09476.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh R Menon, Md Rizwan Parvez, and Zhe Feng. 2023. Delucionqa: Detecting hallucinations in domain-specific question answering. *Preprint*, arXiv:2312.05200.

Korakit Seemakhupt, Sihang Liu, and Samira Khan. 2024. EdgeRAG: Online-indexed RAG for edge devices. *arXiv preprint arXiv:2412.21023*.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *Preprint*, arXiv:2401.15391.

Nandan Thakur, Luiz Bonifacio, Xinyu Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. "Knowing When You Don't Know": A Multilingual Relevance Assessment Dataset for Robust Retrieval-Augmented Generation. *Preprint*, arXiv:2312.11361.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks Track*.

Trulens. 2025. Trulens. `https://github.com/truera/trulens`. [Online; accessed at 2026-01-12, release 2.5.2].

VibrantLabs. 2024. Ragas: Supercharge your llm application evaluations. `https://github.com/vibrantlabsai/ragas`.

Y Wang, Z Jiang, Y Huang, Y Li, S Hao, and L Liu. 2024. Domainrag: A chinese benchmark for evaluating domain-specific retrieval-augmented generation. *Preprint*, arXiv:2406.05654.

Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue Liu, Tei-Wei Kuo, Nan Guan, and Chun Jason Xue. 2025. Retrieval-augmented generation for natural language processing: A survey. *Preprint*, arXiv:2407.13193.

Zhikun Xu, Yinghui Li, Ruixue Ding, Xinyu Wang, Boli Chen, Yong Jiang, Hai-Tao Zheng, Wenlian Lu, Pengjun Xie, and Fei Huang. 2024. Let llms take on the latest challenges! a chinese dynamic question answering benchmark. *Preprint*, arXiv:2402.19248.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024a. Crag – comprehensive rag benchmark. *Preprint*, arXiv:2406.04744.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024b. Crag – comprehensive rag benchmark. *Preprint*, arXiv:2406.04744.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. *Evaluation of Retrieval-Augmented Generation: A Survey*, page 102–120. Springer Nature Singapore.

Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. 2023. Reeval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks. *Preprint*, arXiv:2310.12516.

Yongan Yu, Xianda Du, Qingchen Hu, Jiahao Liang, Jingwei Ni, Dan Qiang, Kaiyu Huang, Grant McKenzie, Renee Sieber, and Fengran Mo. 2026. Weatherarchive-bench: Benchmarking retrieval-augmented reasoning for historical weather archives. *Preprint*, arXiv:2510.05336.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Heng Zhou, Ao Yu, Yuchen Fan, Jianing Shi, Li Kang, Hejia Geng, Yongting Zhang, Yutao Fan, Yuhao Wu, Tiancheng He, Yiran Qin, Lei Bai, and Zhenfei Yin. 2025. Livesearchbench: An automatically constructed benchmark for retrieval and reasoning over dynamic knowledge. *Preprint*, arXiv:2511.01409.

A Complete Benchmark Overview Table

#	Benchmark	Language(s)	N Queries	N Docs	RelDensity	Domain(s)	Existing datasets	Data collection	Data generation	Manual verification
1	HaluEval (Li et al., 2023)	English	10K	10K	1	music; movies; books; sports; politics; health	+	–	q	–
2	RECALL (Liu et al., 2023)	English	–	–	–	commonsense; science	+	–	d+qa	–
3	ReEval (Yu et al., 2023)	English	84K	–	–	news articles; user questions	+	+	d+qa	–
4	DelucionQA (Sadat et al., 2023)	English	913	2K	13.98	car user manual	–	+	qa	+
5	RGB (Chen et al., 2023)	Chinese; English	978	8.7K	8.5	news	–	+	d+qa	+
6	NoMIRACL (Thakur et al., 2023)	multilingual	31K	69K	2.4	Wikipedia	–	+	–	–
7	AttributionBench (Li et al., 2024)	English	5.5K	21.5K	1.17	news; Wikipedia; science	+	–	–	–
8	RAGBench (Friel et al., 2024)	English	50.5K	125K	4.44	biomedical research; general knowledge; legal contracts; customer support; finance	+	–	q	–
9	RAGTruth (Niu et al., 2024)	English	989	3K	4	everyday QA; customer reviews; news summarization	+	–	q	+
10	ChatQA (ChatRAG Bench) (Liu et al., 2024)	English	34K	253K	28.4	Wikipedia; literature; cooking; news; school exams; travel; movies; finance	+	+	qa	+
11	CDQA (Xu et al., 2024)	Chinese	1.3K	1.3K	1	finance; daily life; politics; technology	–	+	qa	+
12	CRUD (Lyu et al., 2024)	Chinese	9.6K	12K	2	history	–	+	d+qa	–
13	DomainRAG (Wang et al., 2024)	Chinese	296	345	1.2	college admissions	–	+	qa	–
14	MultiHop-RAG (Tang and Yang, 2024)	English	2.2K	609	2.4	news	–	+	d+qa	–
15	CRAG (Yang et al., 2024b)	English	–	–	–	finance; sports; music; movies; open-domain	–	+	–	–
16	BABILong (Kuratov et al., 2024)	English	12K	12K	1	books	+	–	sampling	–
17	MEMERAG (Blandón et al., 2025)	multilingual	9.8K	1.7K	5	Wikipedia	+	–	qa	–
18	MTRAG (Katsis et al., 2025)	English	842	110	2.8	Wikipedia; finance; government; technical documentation	+	+	–	+
19	LiveSearchBench (Zhou et al., 2025)	English	600	–	–	Wikipedia	–	+	q	+
20	WeatherArchive-Bench (Yu et al., 2026)	English	335	334	1	historical weather news	–	+	q	+
21	MTRAG-UN (Rosenthal et al., 2026)	English	666	666	1	Wikipedia; finance; government; technical documentation; banking; telecommunications	+	+	–	+
22	Legal RAG Bench (Butler and Butler, 2026)	English	100	4.8K	1	legal textbook cases	–	+	–	+

Table 2: Summarized RAG benchmarks characteristics. Notation used in `Data generation` column: `qa` (question–answer pair generation), `d+qa` (generation of questions from collected data/events with corresponding answers), `sampling` (procedural construction of new instances by sampling from existing corpora and composing inputs via a fixed mixing/template rule).