Response Quality Assessment for Retrieval-Augmented Generation via Conditional Conformal Factuality

Naihe Feng
nfeng@dal.ca
Dalhousie University
Halifax, NS, Canada

Yi Sui
amy@layer6.ai
Layer 6 AI
Toronto, ON, Canada

Shiyi Hou
gloria@layer6.ai
Layer 6 AI
Toronto, ON, Canada

Jesse C. Cresswell
jesse@layer6.ai
Layer 6 AI
Toronto, ON, Canada

Ga Wu
ga.wu@dal.ca
Dalhousie University
Halifax, NS, Canada

Abstract

Existing research on Retrieval-Augmented Generation (RAG) primarily focuses on improving overall question-answering accuracy, often overlooking the quality of sub-claims within generated responses. Recent methods that attempt to improve RAG trustworthiness, such as through auto-evaluation metrics, lack probabilistic guarantees or require ground truth answers. To address these limitations, we propose Conformal-RAG, a novel framework inspired by recent applications of conformal prediction (CP) on large language models (LLMs). Conformal-RAG leverages CP and internal information from the RAG mechanism to offer statistical guarantees on response quality. It ensures group-conditional coverage spanning multiple sub-domains without requiring manual labelling of conformal sets, making it suitable for complex RAG applications. Compared to existing RAG auto-evaluation methods, Conformal-RAG offers statistical guarantees on the quality of refined sub-claims, ensuring response reliability without the need for ground truth answers. Additionally, our experiments demonstrate that by leveraging information from the RAG system, Conformal-RAG retains up to 60% more high-quality sub-claims from the response compared to direct applications of CP to LLMs, while maintaining the same reliability guarantee.

CCS Concepts

• **Information systems** → *Relevance assessment*.

Keywords

Retrieval Augmented Generation, Conformal Prediction

ACM Reference Format:
Naihe Feng, Yi Sui, Shiyi Hou, Jesse C. Cresswell, and Ga Wu. 2025. Response Quality Assessment for Retrieval-Augmented Generation via Conditional Conformal Factuality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730244

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730244

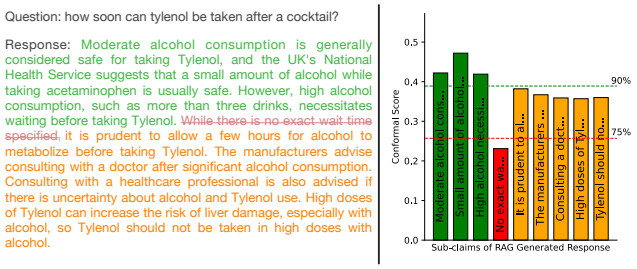

Figure 1: Conformal-RAG filters RAG's responses based on a calibrated factuality threshold. We show two example thresholds guaranteeing 75% and 90% factuality. Claims with scores below the threshold are removed from the final response.

1 Introduction

Existing research in Retrieval-Augmented Generation (RAG) [9, 17] mostly focuses on improving overall question-answering accuracy [35], but often overlooks the quality of sub-claims within generated responses, leading to partially incorrect outputs and hard-to-detect errors [20]. Human evaluations reveal that RAG-based question-answering systems sometimes misinterpret user queries [1, 31], struggle with reasoning in unseen scenarios [13, 21], and may generate claims that are irrelevant or even contradictory to the provided documents [23, 32].

Ensuring the trustworthiness of RAG systems remains a challenge, prompting research into various evaluation solutions. One straightforward way to quantify the trustworthiness of RAG systems is through auto-evaluation based on well-defined metrics. Unfortunately, popular auto-evaluation methods require ground truth answers at inference time, making them impractical in real applications [26, 28]. While some research has addressed this problem [7, 27], auto-evaluation methods still face criticism due to their lack of probabilistic guarantees. Compared to the evaluation techniques mentioned above, conformal prediction provides a stronger theoretical foundation for ensuring soundness of evaluations through statistical guarantees. In hallucination detection tasks, conformal factuality has provided remarkably robust guarantees on large language model (LLM) outputs, solely relying on the LLM's parametric knowledge [3, 22, 24]. Although recent work has integrated conformal prediction into RAG systems [15], it primarily focuses on analyzing generation risks based on adjustable parameters rather than verifying the factuality of sub-claims, leaving a critical research gap unfilled.

This paper presents Conformal-RAG, a conformal prediction [2, 5, 29] framework tailored for RAG systems. The proposed framework leverages contextual information (retrieved external knowledge) from a RAG system, and a high-quality conformal scoring

function, leading to substantially more retained response content compared to existing solutions when targeting the same factuality threshold. In particular, Conformal-RAG can ensure group-conditional factuality [8, 16, 30] spanning multiple sub-domains without requiring manual annotation of conformal set validity, making it highly adaptable for complex RAG applications. We empirically evaluate Conformal-RAG on four benchmark datasets from two domains, Wikipedia [18, 20, 34] and medicine [14]. The experimental results show that Conformal-RAG retains up to 60% more sub-claims from the output in question-answering tasks for the same factuality level compared to existing baselines.

2 Preliminaries and Related Work

Here, we briefly review conformal prediction and its role in ensuring the trustworthiness of question-answering (QA) tasks. Due to space constraints, we do not cover the broader literature on RAG system trustworthiness, as comprehensive surveys already provide an up-to-date literature review [36].

Conformal Prediction. Conformal Prediction (CP) [29] is a statistical framework that transforms heuristic uncertainty estimates into rigorous, calibrated confidence measures. It provides coverage guarantees over prediction sets, where larger sets indicate higher model uncertainty [2]. For a prediction task with possible outputs Y, given a conformity measure S and a tolerable error level α, the conformal prediction set for a new example x_{test} is

$$C_{\hat{q}}(x_{\text{test}}) = \{y \in Y \mid S(x_{\text{test}}, y) \leq \hat{q}\}, \tag{1}$$

where $\hat{q}$ is the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$-quantile of scores S over a calibration dataset containing n datapoints. When calibration and test data are drawn i.i.d. from a distribution $\mathbb{P}$, CP guarantees marginal coverage

$$\mathbb{P}(y^*_{\text{test}} \in C_{\hat{q}}(x_{\text{test}})) \geq 1 - \alpha. \tag{2}$$

Conformal Factuality for Open-ended QA. In classification tasks where Y is a finite label set, CP is straightforward to apply. However, in generative settings like open-ended QA, the output space is effectively infinite, with many semantically equivalent responses. One approach to constrain this space is to limit the output token count [15], however, explicit token limits are not well-suited for open-ended QA, where responses vary in length and structure.

A more principled approach to factuality assessment is to construct prediction sets implicitly as the set of all statements that entail the model's output [22]. An output y is factual if the ground truth y^* entails it (denoted by $y^* \Rightarrow y$), and CP enables calibration of the model's confidence about factuality. Inspired by FActScore [20], for long-form answers with multiple claims, one may estimate factuality per claim, filtering out low-confidence ones based on a threshold $\hat{q}$, while ensuring retained claims meet a factuality guarantee

$$\mathbb{P}(y^*_{\text{test}} \Rightarrow y_{\text{test}}(x_{\text{test}}; \hat{q})) \geq 1 - \alpha. \tag{3}$$

Despite the remarkable probabilistic guarantee offered by conformal factuality, LLMs relying solely on parametric knowledge often generate non-factual statements [19] and struggle with confidence calibration [33], which leads to high claim-rejection rates under strict factuality thresholds. While level-adaptive conformal prediction helps retain more claims, it comes with the cost of reducing overall factuality rates [3].

Algorithm 1 RAG Sub-claim Scoring

Require: Query x, retrieved documents $D = \{d_1, d_2, ...d_m\}$, generated answer $\hat{y} = \{c_1, c_2, ..., c_p\}$.

```
for c_k ∈ ŷ do
    for d_j ∈ D do
        s_kj = CosineSimilarity(x, d_j) · CosineSimilarity(c_k, d_j)
    end for
    r_k = max({s_kj}_{j=1}^m ∪ 0)        ▷ Sub-claim relevance scores
end for
return {r_k}_{k=1}^p
```

3 Methodology

We introduce Conformal-RAG, a framework leveraging CP and the RAG mechanism to offer statistical guarantees on response quality while remaining grounded in documents containing domain knowledge. Below we discuss the end-to-end application of the framework, followed by an in-depth examination of how concepts from CP are applied.

3.1 Conformal Factuality for RAG

Problem Formulation. Given a query $x \in X$, a RAG model retrieves a set of m relevant documents $D = \{d_1, d_2, ..., d_m\}$ from its knowledge corpus. The model then generates an answer $\hat{y}$ composed of p sub-claims $\hat{y} = \{c_1, c_2, ..., c_p\}$. The goal of Conformal-RAG is to modify $\hat{y}$ by filtering out sub-claims, producing y which satisfies eq. (3) where α is the predefined error tolerance level, and y consists of a subset of claims from $\hat{y}$, i.e. $y \subseteq \hat{y} = \{c_1, c_2, ..., c_p\}$.

Context Similarity-based Conformal Score. The first step of our method is to design and calibrate a function to score the relevance of claims. For each query x in the calibration set, we obtain the generated answer from RAG as $\hat{y} = \{c_1, c_2, ..., c_p\}$. Our scoring function $R(c \in \hat{y})$ assigns each claim c a relevance score as shown in algorithm 1. First we compute the cosine similarity between the claim and each of the m retrieved documents. These similarity scores are then multiplied by the cosine similarity between the corresponding document and the original query. Finally, the relevance score $R(c)$ takes the maximum of these values across all m documents (or zero if all scores are negative).

Automatic Calibration Set Annotation. The second step is to design an annotation function which takes advantage of the ground-truth answers from the calibration set to judge the factuality of claims. Specifically, we prompt an LLM [11] to annotate if a given sub-claim is factual by providing the query x, ground-truth answer y^*, as well as the retrieved documents D. The annotation function $A(c \in \hat{y}, x, y^*, D) = 1$ when the sub-claim c is factual and $A(c \in \hat{y}, x, y^*, D) = 0$ when it is non-factual.

Inference. Based on the relevance scores and annotations generated for each claim across queries in the calibration dataset, we apply CP to calibrate a threshold $\hat{q}$. Details on the marginal and conditional CP approaches are given below in section 3.2 and section 3.3. At inference, only queries and documents are available. Sub-claims and relevance scores are generated in the same way as during calibration. Then, claims are removed from the generated answer

if their relevance is below the calibrated threshold, creating the conformally factual output $y(x;\hat{q}) = \{c \in \hat{y} \mid R(c) \geq \hat{q}\}$.

Note that LLM-generated answers may not always be in the form of clearly separated sub-claims. Following previous work [22], we use an LLM to decompose the answer into sub-claims. Similarly, since removing sub-claims may affect the grammatical structure of the overall answer, the final set of claims is fed back into an LLM, which is prompted to merge them into a coherent response.

3.2 Marginal Conformal Factuality with RAG

Our marginal CP calibration builds off of work by Mohri and Hashimoto [22], but takes advantage of the RAG mechanism through our relevance scoring function. Our aim is to guarantee factuality of generated answers in the sense that the final generated output is entailed by the ground truth answer y^*_{test} with high probability, satisfying eq. (3).

We introduce a filtering function $F_q(\{c\})$ acting on a set of claims, and satisfying both $F_0(\{c\}) = \{c\}$ and $F_\infty(\{c\}) = \emptyset$. As the threshold q increases from 0, F_q progressively filters out more of the claims, and hence satisfies a nesting property: $F_q(\{c\}) \subseteq F_{q'}(\{c\})$ for $q \geq q'$ [12]. The filtering function is constructed using the relevance scores $R(c)$ described in algorithm 1 as

$$F_q(\hat{y}) = \{c \in \hat{y} \mid R(c) \geq q\}. \tag{4}$$

To determine the appropriate threshold q we use CP calibration over the conformal scores

$$S(x_i, y_i^*) := \inf\{q \in \mathbb{R}^+ \mid \forall q' \geq q, \forall c \in F_{q'}(\hat{y}_i), A(c, x_i, y_i^*, D) = 1\}. \tag{5}$$

That is, the score S is the smallest threshold q such that all retained claims are considered factual by the annotation function A from section 3.1. Then, the conformal threshold $\hat{q}$ is set as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the conformal scores over the calibration set.

On inference data we filter out claims with relevance score $R(c)$ less than $\hat{q}$, i.e. we return $y_{\text{test}} = F_{\hat{q}}(\hat{y})$. Under the assumption that the annotation function is correct on the calibration data, these sets of filtered claims will satisfy eq. (3) by Theorem 4.1 of Mohri and Hashimoto [22]. The core differences between Conformal-RAG and [22] are the relevance function $R(c)$ used for filtering which incorporates similarity information from the RAG mechanism, and the use of automatic annotation to provide ground truth on sub-claim factuality.

3.3 Conditional Conformal Factuality with RAG

Previous research [8, 10] shows that marginal CP can undercover some groups within the data, while overcovering others, leading to fairness concerns [4, 25]. To address this, one can aim to provide group-conditional coverage over a pre-specified grouping $g : X \to G = \{1 \dots n_g\}$:

$$\mathbb{P}(y^*_{\text{test}} \in C_{\hat{q}_a}(x_{\text{test}}) \mid g(x_{\text{test}}) = a) \geq 1 - \alpha \quad \forall\, a \in G. \tag{6}$$

Correspondingly, the conformal threshold $\hat{q}_a$ needs to depend on the group attribute a (e.g. topic category or difficulty of the query).

Cherian et al. [3] proposed to adapt the threshold per test datapoint. First, define the pinball loss $\ell_\alpha(r) := (1-\alpha)[r]_+ + \alpha[r]_-$. Then, the threshold specific to datapoint x_{test} is determined by the function $f_{\text{test}} : G \to \mathbb{R}$ defined as

$$f_{\text{test}} = \arg\min_{f \in \mathcal{F}} \frac{1}{n+1}\Big[\sum_{i=1}^{n} \ell_\alpha\big(S_i - f(g(x_i))\big) + \ell_\alpha\big(S_{\text{test}} - f(g(x_{\text{test}}))\big)\Big] \tag{7}$$

where $S_i = S(x_i, \hat{y}_i)$ (eq. (5)), S_{test} is imputed using quantile regression, and the optimization is over the family of linear functions $\mathcal{F} = \{f(a) = \beta^\top e_a\}$ for $\beta \in \mathbb{R}^{|G|}$ and e_a a basis vector of $\mathbb{R}^{|G|}$. The learned function f_{test} provides the adapted conformal quantile $\hat{q}_{\text{test}} = f_{\text{test}}(x_{\text{test}})$ which is used to filter out claims, i.e. the method returns $y_{\text{test}} = F_{\hat{q}_{\text{test}}}(\hat{y})$ as in eq. (4). This procedure satisfies group-conditional factuality [3],

$$\mathbb{P}(y^*_{\text{test}} \Rightarrow y_{\text{test}}(x_{\text{test}}; \hat{q}) \mid g(x_{\text{test}}) = a) \geq 1 - \alpha \quad \forall\, a \in G. \tag{8}$$

However, this method borders on impractical as it requires both a quantile regression to impute S_{test}, and an optimization over $\mathcal{F}$ for every inference datapoint. To simplify these procedures, Conformal-RAG follows the Mondrian CP paradigm [29, 30] which first partitions the calibration data by groups using g, then calibrates a distinct threshold $\hat{q}_a$ for each $a \in G$ using the procedure in section 3.2. At inference time, the threshold for group $a_{\text{test}} = g(x_{\text{test}})$ is used for filtering out claims, i.e. we return $y_{\text{test}} = F_{\hat{q}_{a_{\text{test}}}}(\hat{y})$. Since each group is calibrated independently, eq. (3) holds for each group, which implies eq. (8).

4 Experiments

Dataset. We evaluate Conformal-RAG on four benchmark datasets: FActScore [20], PopQA [18], HotpotQA [34], and MedLFQA [14]. The first three datasets use common knowledge from Wikipedia, whereas MedLFQA is a medical QA benchmark broken into five sub-datasets organized by topic and is considered more difficult than Wikipedia datasets for RAG. We follow the document curation process from previous work [3] for MedLFQA. For marginal Conformal-RAG, we evaluate the model on each of the four datasets individually. For conditional Conformal-RAG, we create a Wiki dataset by combining PopQA and HotpotQA, treating each as an individual group, while the MedLFQA dataset is divided into its underlying sub-datasets. In our experiment, the group labels are available during inference.

Experimental Setup. For our experiments, we use a RAG system with a FAISS retriever [6] and GPT-4o generator. For conformal calibration and inference, we adapted code from Mohri and Hashimoto [22]. In addition, we use a GPT-4o model for annotation, sub-claim decomposition, and sub-claim merging as described in section 3.1. For both marginal and conditional experiments, we test a range of error rates $\alpha \in [0.05, 0.40]$ and compare Conformal-RAG primarily to conformal factuality using confidence scoring directly from an LLM [22]. For clarity, we refer to our method as Conformal-RAG and the baseline as Conformal-LLM.

4.1 Results

Marginal Conformal Factuality. We plot the removal rate and empirical factuality achieved with different target factuality levels $1 - \alpha$ for both Conformal-RAG and Conformal-LLM in fig. 2 (a). For removal rate, Conformal-RAG consistently outperforms the baseline, which only uses an LLM's parametric knowledge, across all four

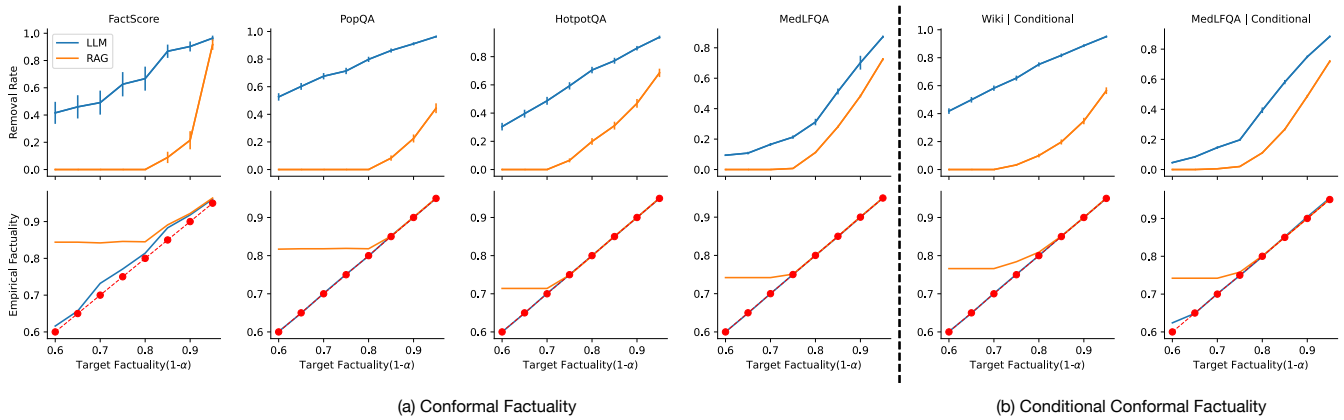

Figure 2: Sub-claim removal rates (top) and empirical factuality levels (bottom) for target factuality levels $1-\alpha$ using (a) marginal conformal prediction and (b) group-conditional conformal prediction, averaged over all test data. LLM is the baseline, while RAG is our method. The red dashed line shows the conformal factuality lower bound.

datasets. For example, Conformal-RAG's removal rate at target factuality level 85% for FActScore is only 8.9%, while Conformal-LLM removes 86.8% of sub-claims to guarantee the same factuality level. Hence, Conformal-RAG is able to return longer, more informative answers with the same guarantees on factuality. For empirical factuality, calculated as the average factuality using the ground-truth labels from the test data, we find that both Conformal-RAG and Conformal-LLM maintain a level at or above the target, as expected from the guarantee in eq. (3). Hence, Conformal-RAG does not sacrifice factuality even when retaining a much higher fraction of claims. Notably, in many cases Conformal-RAG reaches a plateau of empirical factuality when the target $1-\alpha$ is lowered enough. In these cases, essentially all claims can be retained because the RAG mechanism does not generate as many non-factual claims in the first place. This clearly demonstrates the advantages of grounding generation in domain knowledge.

The design of our relevance scoring function from section 3.1 also benefits the quality of retained claims. At the individual data point level, we observe that Conformal-RAG preferentially filters out claims that may be factually correct, but lack semantic or contextual relevance to the given query. For example, on the query "how soon can tylenol be taken after a cocktail?" from MedLFQA (fig. 1), one sub-claim states "there is no exact wait time specified [for alcohol metabolism]", which is factual but not relevant to the original question. This claim had low relevance $R(c) = 0.231$, leading to its removal at a relatively low target factuality of 75%, corresponding to a threshold of $\hat{q} = 0.257$. By comparison, claims with higher relevance like "The UK's National Health Service suggests that a small amount of alcohol while taking acetaminophen is usually safe" with higher score $R(c) = 0.472$, are both factual and more directly helpful for answering the query.

Conditional Conformal Factuality. In fig. 2 (b) we show results for conditional Conformal-RAG. We again observe that Conformal-RAG significantly reduces the removal rate while maintaining the (marginal) factuality guarantee. We further show the empirical factuality for each group on the MedLFQA dataset in fig. 3. Both Conformal-LLM and Conformal-RAG approximately achieve the

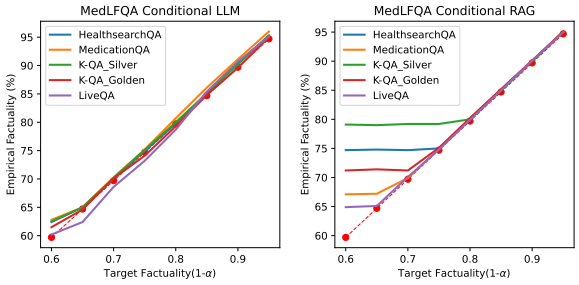

Figure 3: Empirical factuality by group for Conformal-LLM and Conformal-RAG on MedLFQA. The red dashed line shows the conformal factuality lower bound.

target factuality for every group, demonstrating their effectiveness across different subsets. However, Conformal-LLM shows slightly more variation, with some groups experiencing more deviation from the target factuality. For example, LiveQA shows a slight drop in factuality below the target when $1-\alpha < 0.8$. In contrast, Conformal-RAG exhibits less fluctuation in factuality across the groups, suggesting more stable performance. This stability can likely be attributed to the effective use of RAG's internal retrieval mechanism, which enhances the model's consistency in achieving the target factuality.

5 Conclusion

This paper introduced Conformal-RAG, a novel framework that applies conformal prediction (CP) to enhance RAG systems. An extension of Conformal-RAG to conditional CP ensures group-conditional coverage across multiple sub-domains without requiring manual annotation of conformal sets, making it well-suited for complex RAG applications. Experimental results showed that Conformal-RAG and its conditional extension retain up to 60% more high-quality sub-claims than direct applications of CP to LLMs, while maintaining the same factuality guarantees.

References

[1] Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Mindful-RAG: A Study of Points of Failure in Retrieval Augmented Generation. In *2024 2nd International Conference on Foundation and Large Language Models*. 607–611. doi:10.1109/FLLM63129.2024.10852457

[2] Anastasios N Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511* (2021).

[3] John Cherian, Isaac Gibbs, and Emmanuel Candes. 2024. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems*, Vol. 37.

[4] Jesse C. Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. 2025. Conformal Prediction Sets Can Cause Disparate Impact. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=fZK6AQXlUU

[5] Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. 2024. Conformal prediction sets improve human decision making. In *Proceedings of the 41th International Conference on Machine Learning*.

[6] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *arXiv:2401.08281* (2024).

[7] 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*, Nikolaos Aletras and Orphee De Clercq (Eds.). 150–158.

[8] Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. 2020. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA* 10, 2 (08 2020). doi:10.1093/imaiai/iaaa017

[9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv:2312.10997* (2023).

[10] Isaac Gibbs, John J Cherian, and Emmanuel J Candès. 2023. Conformal prediction with conditional guarantees. *arXiv:2305.12616* (2023).

[11] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A Survey on LLM-as-a-Judge. *arXiv:2411.15594* (2024).

[12] Chirag Gupta, Arun K. Kuchibhotla, and Aaditya Ramdas. 2022. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition* 127 (July 2022), 108496. doi:10.1016/j.patcog.2021.108496

[13] Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1049–1065. doi:10.18653/v1/2023.findings-acl.67

[14] Minbyul Jeong, Hyeon Hwang, Chanwoong Yoon, Taewhoo Lee, and Jaewoo Kang. 2024. OLAPH: Improving Factuality in Biomedical Long-form Question Answering. *arXiv:2405.12701* (2024).

[15] Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. 2024. C-RAG: Certified Generation Risks for Retrieval-Augmented Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235. 22963–23000.

[16] Jing Lei and Larry Wasserman. 2013. Distribution-free Prediction Bands for Non-parametric Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 76, 1 (07 2013), 71–96. doi:10.1111/rssb.12021

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. 9459–9474.

[18] 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)*. 9802–9822. doi:10.18653/v1/2023.acl-long.546

[19] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1906–1919. doi:10.18653/v1/2020.acl-main.173

[20] 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*. 12076–12100. doi:10.18653/v1/2023.emnlp-main.741

[21] Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=AjXkRZIvjB

[22] Christopher Mohri and Tatsunori Hashimoto. 2024. Language Models with Conformal Factuality Guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235. 36029–36047. https://proceedings.mlr.press/v235/mohri24a.html

[23] 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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10862–10878. doi:10.18653/v1/2024.acl-long.585

[24] Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal Language Modeling. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=pzUhfQ74c5

[25] Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. 2020. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review* 2, 2 (2020).

[26] 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. In *Advances in Neural Information Processing Systems*, Vol. 37. 21999–22027.

[27] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 338–354. doi:10.18653/v1/2024.naacl-long.20

[28] 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*. https://openreview.net/forum?id=Iyrtb9EJBp

[29] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer.

[30] Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. 2003. Mondrian confidence machine. *Technical Report* (2003).

[31] Di Wu, Jia-Chen Gu, Fan Yin, Nanyun Peng, and Kai-Wei Chang. 2024. Synchronous Faithfulness Monitoring for Trustworthy Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 9390–9406. doi:10.18653/v1/2024.emnlp-main.527

[32] Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E Ho, and James Zou. 2024. How well do LLMs cite relevant medical references? An evaluation framework and analyses. *arXiv:2402.02008* (2024).

[33] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=gjeQKFxFpZ

[34] 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*. 2369–2380. doi:10.18653/v1/D18-1259

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

[36] Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in retrieval-augmented generation systems: A survey. *arXiv preprint arXiv:2409.10102* (2024).