Retrieval-based Evaluation for LLMs: A Case Study in Korean Legal QA

Cheol Ryu **Seolhwa Lee** **Subeen Pang** **Chanyeol Choi**
Linq Labs
{cheol.ryu,seolhwa.lee,subeen.pang,jacob.choi}@getlinq.com

Hojun Choi **Myeonggee Min**
Law&Good
{hojun.choi,mgmin}@lawandgood.com

Jy-yong Sohn[*]
Yonsei University
jysohn1108@yonsei.ac.kr

Abstract

While large language models (LLMs) have demonstrated significant capabilities in text generation, their utilization in areas requiring domain-specific expertise, such as law, must be approached cautiously. This caution is warranted due to the inherent challenges associated with LLM-generated texts, including the potential presence of factual errors. Motivated by this issue, we propose Eval-RAG, a new evaluation method for LLM-generated texts. Unlike existing methods, Eval-RAG evaluates the validity of generated texts based on the related document that are collected by the retriever. In other words, Eval-RAG adopts the idea of retrieval augmented generation (RAG) for the purpose of evaluation. Our experimental results on Korean Legal Question-Answering (QA) tasks show that conventional LLM-based evaluation methods can be better aligned with Lawyers' evaluations, by combining with Eval-RAG. In addition, our qualitative analysis show that Eval-RAG successfully finds the factual errors in LLM-generated texts, while existing evaluation methods cannot.

1 Introduction

Recent advances of large language models (LLMs) are remarkable, in various natural language generative tasks such as translation or summarization. However, LLMs are prone to generating hallucinated text (Ji et al., 2023), i.e., the texts generated by LLMs are sometimes having factual errors, which implies the unreliability of the LLM-generated text. Consequently, it is necessary to develop a proper system of evaluating LLM-generated text, for using LLMs in real-world applications.

Various methods are proposed for evaluating natural language generation (NLG) (Papineni et al., 2002; Lin, 2004; Zhang et al., 2020). However, it is known that such classical evaluation methods have room for improvement compared with the human evaluation, especially for evaluating question answering (QA) tasks (Liu et al., 2023).

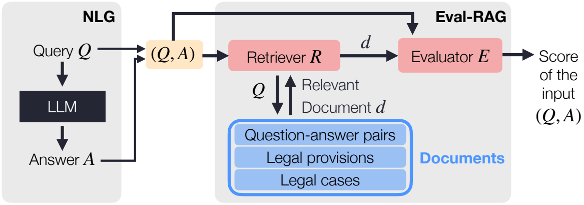

Figure 1: The proposed Eval-RAG method for Legal QA task. Suppose an LLM generates answer A for the query Q. Eval-RAG is a framework for evaluating the quality of the answer, using retriever R and evaluator E. First, R retrieves the document d relevant to the query Q. Second, E evaluates the score of (Q, A) based on the contents of the retrieved document d.

To address this issue of classical methods, recent approaches considered evaluating LLM-generated texts by another LLM (Chan et al., 2023; Liu et al., 2023; Wang et al., 2023b). Although the LLM-based evaluation methods outperforms classical methods, they are suffering from innate issues of LLMs, such as the lack of training data on specialized domains and the inability to reflect recent data.

Main Contributions In this paper, we propose Eval-RAG, a retrieval-based method evaluating LLM-generated texts. Eval-RAG first retrieves the document relevant to the generation task, and then evaluates the texts based on the relevant document. This method can be considered as applying retrieval-augmented generation (RAG) concept on the existing LLM-based evaluation systems. Our experimental results on Korean legal QA tasks show that Eval-RAG outperforms existing evaluation methods, in terms of higher correlation with the human evaluation. Our qualitative analysis also shows that Eval-RAG overcomes the limitation of existing LLM-based evaluation methods which cannot judge the hallucination of LLMs.

---

[*] Corresponding author

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 132–137
December 7, 2023 ©2023 Association for Computational Linguistics

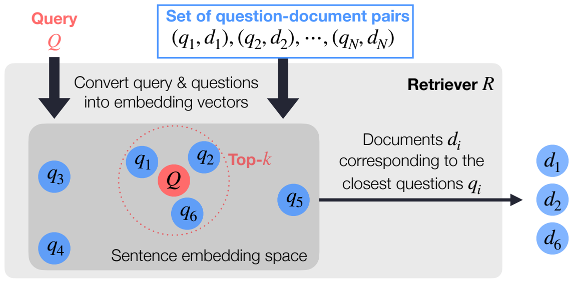
Query
Q
Set of question-document pairs
$(q_1, d_1), (q_2, d_2), \cdots, (q_N, d_N)$
Convert query & questions into embedding vectors
Retriever R
q_1
q_2
Top-k
Q
q_3
q_6
q_5
q_4
Sentence embedding space
Documents d_i corresponding to the closest questions q_i
d_1
d_2
d_6

Figure 2: The pipeline of retriever.

2 Method

Here we introduce the proposed Eval-RAG framework depicted in Fig. 1, which has two main components: the retriever R and the evaluator E. Once an LLM generates answer A for the query Q, Eval-RAG evaluates whether answer A is proper or not for the query Q. First, the retriever R finds legal documents d that are relevant to the query Q. Second, the evaluator E evaluates the validity of answer based on the related document d.

Retriever Fig. 2 illustrates how retriever R retrieves the documents d related with the query Q. This process contains three steps: data collection, sentence embedding, and query processing. In Step 1, we collect three different types of publicly available legal documents: (1) legal question-answer pairs, (2) legal provisions and (3) legal cases (precedents). Since our focus is given to QA tasks, we convert these documents into (question, document containing the answer) format, denoted by (q, d), in the following manner. Given the legal question-answer (q, a) pairs, we extract related document d from the answer a. For other documents d (legal provisions and legal cases), we use LLMs to generate relevant questions q. Table 1 shows the examples of question-document pair (q, d) for each data source, which are translated from the Korean using *DeepL*[1]. In Step 2, the question-document pairs $\{(q_i, d_i)\}_{i=1}^{N}$ collected in step 1 are stored in a vector database (Johnson et al., 2019) by using a sentence embedding model, which converts a sentence q into an n-dimensional vector. In Step 3, given query Q and the set of questions $\{q_i\}_{i=1}^{N}$ collected in step 2, the retriever R first finds the question q_i that is most similar to Q (in the sentence embedding space) and returns the relevant document d_i.

Evaluator The quality of answer A to a given query Q is evaluated based on the relevant document d, as shown in Fig. 1. Recall that existing LLM-based evaluators (Chan et al., 2023; Liu et al., 2023; Wang et al., 2023b) ask LLMs for the evaluation, where the prompts include Q and A, along with the evaluation criteria and methods for evaluation. The evaluator proposed by us is built on top of these methods, by adding the relevant document d (delivered from retriever R) in the prompt. This allows E to evaluate based on the relevant documents. Note that our method (evaluating LLMs based on related documents) is motivated by the RAG methods (which generates texts with the aid of retrieved documents), and thus we call our evaluation method as Eval-RAG. Table 2 provides an example of how an associated document d is added to the prompt of the existing method called *FairEval* (Wang et al., 2023b). It is worth noting that Eval-RAG can be combined with any existing LLM-based evaluation methods.

3 Experiments

We tested the performances of Eval-RAG and baselines for Korean legal QA tasks, specifically focused on the questions related to divorce.

Experimental Setup The divorce-related QA pairs used for the retriever are collected from *Korea Legal Aid Corporation*[2], while the legal provisions and cases mentioned in the QA pairs are gathered from *Korea Legislation Research Institute*[3]. The number of QA pairs, provisions, and cases used in the experiment were 287, 84, and 240, respectively. we used GPT-4 to generate relevant questions for each provision and case. The query set ($n = 100$) used in the experiment was selected from actual client consultation cases at the law firm *Law&Good*[4]. For each query, we prepared three types of answers: an actual answer from a legal professional (*Lawyer*), an answer generated by assigning the role of a lawyer to ChatGPT (*gpt*), and an answer generated with the inclusion of relevant documents (*rag*). We used OpenAI's 'text-embedding-ada-002' sentence embedding model to store questions in vector database.

---

[1] http://deepl.com/

[2] https://www.klac.or.kr

[3] https://www.klri.re.kr

[4] https://www.lawandgood.com

q: Does divorce evidence have an expiration date?
a: Yes, there is an expiration date for divorce evidence. Article 841 of the Civil Act of Republic of Korea stipulates that divorce evidence is valid for six months from the date the spouse becomes aware of the evidence and two years from the date of the grounds for filing.
d: Article 841 of the Civil Act (Extinction of Right to Apply for Divorce due to Unchastity) The spouse may not file for a divorce when the spouse has given a prior consent or an ex post facto forgiveness, or when six months have passed since the date the spouse becomes aware of the event or two years have passed since the date of the event.

q: I want to know what I can or should do with the items etc. that I gave to and received from the was-to-be-spouse as engagement gifts in case of a disengagement.
provision: Article 806 of the Civil Code (Termination of Engagement and Right to Claim Damages) Upon termination of an engagement, either party may claim compensation for damages from the negligent other party.
d: Article 806 of the Civil Act (Dissolution of Matrimonial Engagement and Claims for Damages) Upon the termination of an engagement, a party may claim compensation for damages from the negligent other party.

q: I would like to raise a claim for division of property on divorce upon the severance pay of the other party/spouse as well. Is this possible?
case: Case No. 2013MU2250 (Main Issues and Holdings) Whether the future retirement benefits are subject to the division of property on divorce when one of the divorcees is still employed at the time of divorce and has not yet retired(affirmative) and the scope of such division.
d: Case No. 2013MU2250 (Summary of Decision) Because it is required to work for a certain period of time to receive retirement benefits, as far as the cooperation of the spouse is recognized to have contributed to such work, the retirement benefits may be subject to a division of property.

Table 1: Examples of (question, relevant document) pair, denoted by (q, d), for three types of data sources: (1) legal question-answer pair (q, a), (2) legal provision, and (3) legal case.

You are a helpful and precise assistant for checking the quality of the answer. Please evaluate the answer based on stated relevant law of South Korea.
[Question]{question}
[The Start of Assistant 1's Answer] {answer_1} [The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer] {answer_2} [The End of Assistant 2's Answer]
[The Start of Relevant Law of South Korea]{document}[The End of Relevant Law of South Korea]

Table 2: The prompt of Eval-RAG based on FairEval (Wang et al., 2023b). Black sentences are originally designed by FairEval, while blue sentences are designed by Eval-RAG.

Methods & Performance Metrics We compare the performances of different methods: (1) the latest LLM-based evaluators, ChatEval (Chan et al., 2023) and FairEval (Wang et al., 2023b), (2) these latest evaluators combined with Eval-RAG, and (3) the human grading evaluated by the lawyers, where all scores are scaled with minimum score 1 and the maximum score 10. Due to the limit of the available prompt length of the evaluator, the retriever returns only one relevant document of the question closest to the query. The human grading for the three types of answers for each query is considered as the benchmark, and we measured the correlation between this benchmark and each method, using *Pearson*, *Spearman*, and *Kendall* correlation coefficients.

Result Table 3 compares different evaluation methods, in terms of the correlation between the score of each method and the human grading. One can confirm that existing evaluators (FairEval and ChatEval) can be improved when combined with Eval-RAG, for both GPT-3.5 and GPT-4.

Table 4 provides qualitative comparison of evaluation methods, for a specific query Q. The table on the top shows query Q, the answer A generated by GPT and RAG (denoted by (gpt) and (rag)), and the related document d. The bottom table compares the evaluation of (gpt) and (rag) for different evaluation methods: Lawyer (human evaluation), ChatEval and ChatEval combined with Eval-RAG.

Here, (gpt) and (rag) have opposing views on the divorce-related capital gains tax. Note that the document d and the lawyer's evaluation indicate that (rag) is correct and (gpt) is incorrect.

Table 3: Correlation between the score of each evaluation method and the human grading. FairEval-RAG means FairEval combined with Eval-RAG.

Evaluator	Pearson	Spearman	Kendall
FairEval (GPT-3.5)	0.0432	0.0515	0.0457
FairEval-RAG (GPT-3.5)	0.1156	0.1264	0.1108
FairEval (GPT-4)	0.4972	0.4999	0.4299
FairEval-RAG (GPT-4)	**0.5923**	**0.5841**	**0.4991**
ChatEval (GPT-3.5)	0.1091	0.0896	0.0768
ChatEval-RAG (GPT-3.5)	0.0905	0.0816	0.0735
ChatEval (GPT-4)	0.4467	0.4541	0.4115
ChatEval-RAG (GPT-4)	**0.5103**	**0.5129**	**0.4618**

However, ChatEval's evaluation, based on LLM's knowledge, fails to recognize that (gpt) is incorrect. In contrast, Eval-RAG, referencing the necessary document d for evaluation, assesses (rag) as the more appropriate answer. This observation is consistent across both the GPT-3.5 and GPT-4 models. More specifically, only the evaluation of ChatEval-RAG (GPT-4) accurately states that (gpt) is incorrect based on document d.

4 Related Work

QA and LLM Question answering (QA) is a task of providing answers to a given question, which is explored in last decades (Rogers et al., 2023; Zhu et al., 2021; Bolotova et al., 2022). Recently, there are studies on QA for specific domains such as law (Ravichander et al., 2019; Wang et al., 2023a; Louis et al., 2023), and studies on long-form question answering (LFQA) which involves searching external documents relevant to a given query, and then generating paragraph-length answers (Krishna et al., 2021; Wang et al., 2022; Xu et al., 2023). The area of QA is greatly advanced using large language models (LLMs) (OpenAI, 2023). However, LLMs have several limitations, such as the generation of hallucinated text (Ji et al., 2023). One way of overcoming the limitations of LLMs is to us the Retrieval Augmented Generation (RAG) method, which integrates the retrieval module (which retrieves information from some data source) and the text generation module (which generates texts based on the retrieved information) in a unified framework (Lewis et al., 2020). RAG methods have recently demonstrated strong performance in QA (Mialon et al., 2023; Lazaridou et al., 2022; Khattab et al., 2023).

NLG Metrics Various types of metrics were developed to measure the quality of natural language generation (NLG). Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002) is a widely used metric for evaluating machine translation systems. It measures the similarity between the machine-generated output and the reference translation based on n-gram precision. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) is a family of metrics for evaluating text summarization systems. ROUGE measures the similarity between the machine-generated summary and the reference summary based on n-gram overlap and sentence-level overlap. BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are also used for QA evaluation, but are not suitable because they are designed as translations and summaries, respectively. BERTScore (Zhang et al., 2020) is a evaluation metric for similarity between two sentences, by using pre-trained BERT models. BERTScore is shown to outperform other metrics in QA tasks. However, BERTscore is not reliable when the texts contain domain-specific knowledge due to the limitations of BERT model (Chalkidis et al., 2020).

LLM evaluation Recently, evaluating generated texts using LLMs is proposed (Fu et al., 2023; Liu et al., 2023; Wang et al., 2023b; Chan et al., 2023). G-Eval (Liu et al., 2023) proposes probability-weighting schemes to calibrate the evaluation score from LLM. FairEval (Wang et al., 2023b) proposes a new evaluation methodology, motivated by the observation that the evaluation results can vary depending on the order of inputs in the prompt. ChatEval (Chan et al., 2023) proposes a novel approach where multiple agents engage in a discussion to evaluate the generated text. Our paper is built on top of these LLM-based evaluations, and outperforms these baselines by using retrieval-based-evaluation.

5 Conclusion

In this paper, we proposed Eval-RAG, a new method of evaluating LLM-generated texts, with the aid of retrieved documents that are related with the target task. Our experiments on Korean Legal QA tasks show that existing LLM-based evaluations can be better aligned with ground-truth human evaluation, when combined with Eval-RAG.

6 Ethical considerations

The example we tested the performance might include sensitive case information of clients. However, it is essential to emphasize that this ethical concern is alleviated, as the samples in question have been withheld from public disclosure.

Q: We would like to know whether the transfer of real estate ownership from one party to the other through a division of property on divorce is subject to the capital gains tax.
(gpt): The transfer of real estate ownership from one party to the other in a divorce constitutes a transfer of property for valuable consideration. Capital gains tax is a tax incurred under the Income Tax Act when properties such as real estate, etc. are transferred. However, the transfer of real estate resulting from the division of property in divorce may be exempted from capital gains tax under Article 9 of the Income Tax Act.
(rag): According to the Supreme Court precedent Case No. 2002DU6422, the division of property constitutes a partition of the article jointly owned and the legal principles on a partition of the article jointly owned shall be applied. The partition of the article jointly owned is understood to be a change in the form of ownership by concentrating the co-owned share which was exercised in proportion to the co-owned share in the article jointly owned into an ownership of a specific part of the article and making it subsist only on that specific part of the article. For this reason, the division of property in a divorce is deemed to be a transfer of property for valuable consideration and thereby is not subject to capital gains tax.
d: Case No. 96NU14401 (Summary of Decision) A division of property in a divorce by agreement to liquidate the de facto marital property is substantially a partition of the article jointly owned in light of its legal nature, the object and scope of such division. Therefore, legal principles on a partition of the article jointly owned shall be apply mutatis mutandis, and the transfer of property by the result of such division is not a transfer of property for valuable consideration which is subject to capital gains tax.

Lawyer : The score of (gpt) = 4, the score of (rag) = 8
ChatEval (GPT-3.5) : After carefully reviewing the responses of both AI assistants, I found that (gpt) provided a more accurate and detailed answer. (gpt) correctly stated that transferring ownership of a property during divorce proceedings is considered a transfer of assets and may be subject to transfer income tax. Moreover, (gpt) provided the specific conditions that must be met for the transfer to be exempt from transfer income tax. **In contrast,** (rag) **incorrectly stated** that transferring ownership of a property during divorce proceedings is not considered a transfer of assets and therefore not subject to transfer income tax. Therefore, the score of (gpt) = 8, The score of (rag) = 5.
ChatEval-RAG (GPT-3.5) : Based on the stated relevant law of South Korea, (rag)'s answer is more accurate and relevant to the question. The law clearly states that the division of property in divorce is not considered a transfer of ownership and therefore not subject to transfer income tax. (rag) correctly cites a relevant court case to support this. (gpt)'s answer is also helpful, but it does not provide as much detail on the specific law and court case. Therefore, the score of (gpt) = 7, the score of (rag) = 9.
ChatEval (GPT-4) : (gpt) provided a detailed and accurate response, explaining the conditions under which the transfer of property due to divorce can be exempt from transfer income tax. This is in line with the Transfer Income Tax Law in South Korea. (rag), while citing a specific case, seems to have misunderstood the nature of property division in a divorce and its tax implications. Therefore, the score of (gpt) = 9, the score of (rag) = 6.
ChatEval-RAG (GPT-4) : Based on the stated relevant law of South Korea, Case 2002DU6422, the division of property during a divorce does not constitute a paid transfer of assets. (rag) correctly pointed this out and provided an accurate interpretation of the law. On the other hand, (gpt) **incorrectly stated** that the division of property during a divorce constitutes a paid transfer of assets and could be subject to capital gains tax. However, (gpt) did provide detailed information about the conditions for exemption from capital gains tax, which could be useful information for the user. The score of (gpt) = 6, the score of (rag) = 9.

Table 4: **(Top):** An example query Q, the answers generated by GPT and RAG, and related document d. **(Bottom):** Evaluation of generated texts by Lawyer, ChatEval (Chan et al., 2023) and ChatEval-RAG, a short-hand notation for combining ChatEval with Eval-RAG. The evaluation with Eval-RAG aligns better with the evaluation of Lawyer.

References

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2022. A non-factoid question-answering taxonomy. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1196–1207. Association for Computing Machinery.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904. Association for Computational Linguistics.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. GPTScore: Evaluate as you desire.

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).

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2023. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering.

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*, volume 33, pages 9459–9474. Curran Associates, Inc.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2023. Interpretable long-form legal question answering with retrieval-augmented large language models.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey.

OpenAI. 2023. GPT-4 technical report.

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*, page 311–318. Association for Computational Linguistics.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10).

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators.

Shufan Wang, Fangyuan Xu, Laure Thompson, Eunsol Choi, and Mohit Iyyer. 2022. Modeling exemplification in long-form question answering via retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2079–2092. Association for Computational Linguistics.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering.