An Empirical Study of Evaluating Long-form Question Answering

Ning Xian*
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
xianning21s@ict.ac.cn

Yixing Fan*
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
fanyixing@ict.ac.cn

Ruqing Zhang
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
zhangruqing@ict.ac.cn

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Jiafeng Guo†
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
guojiafeng@ict.ac.cn

Abstract

Long-form question answering (LFQA) aims to generate lengthy answers to complex questions. This scenario presents great flexibility as well as significant challenges for evaluation. Most evaluations rely on deterministic metrics that depend on string or n-gram matching, while the reliability of large language model-based evaluations for long-form answers remains relatively unexplored. We address this gap by conducting an in-depth study of long-form answer evaluation with the following research questions: (i) To what extent do existing automatic evaluation metrics serve as a substitute for human evaluations? (ii) What are the limitations of existing evaluation metrics compared to human evaluations? (iii) How can the effectiveness and robustness of existing evaluation methods be improved? We collect 5,236 factoid and non-factoid long-form answers generated by different large language models and conduct a human evaluation on 2,079 of them, focusing on correctness and informativeness. Subsequently, we investigated the performance of automatic evaluation metrics by evaluating these answers, analyzing the consistency between these metrics and human evaluations. We find that the style, length of the answers, and the category of questions can bias the automatic evaluation metrics. However, fine-grained evaluation helps mitigate this issue on some metrics. Our findings have important implications for the use of large language models for evaluating long-form question answering. All code and datasets are available at https://github.com/bugtig6351/lfqa_evaluation.

CCS Concepts

• **Information systems → Evaluation of retrieval results**; **Question answering**.

Keywords

Long-form question answering, automatic evaluation

---

*Both authors contributed equally to this research.
†Corresponding author

1 Introduction

LFQA aims to enable generation or retrieval models to answer open-ended questions with long answers at the paragraph level [11, 21]. With the advancement of large language models (LLMs), the ability to generate long-form answers has improved significantly [4, 5]. While such long-form answers could address more complex and diverse questions, their flexibility presents significant challenges for evaluation.

Currently, much of the research on LFQA focuses on designing models and frameworks to address the hallucination of LLM and improve overall performance. For example, Su et al. [34] propose a framework that uses fine-grained, answer-related salient information to enhance the faithfulness of model responses and reduce hallucinations. Tao et al. [35] introduce a chain-of-discussion framework that takes advantage of synergy among multiple open source LLMs to provide more accurate and comprehensive answers. Despite these advancements, most studies still rely on simple question answering (QA) metrics, such as ROUGE [24] and Exact Match[33], which are based on string or n-gram matching, to evaluate long-form answers. However, these metrics often fail to capture the nuanced and flexible evaluation required by the complex structure of long-form answers, demonstrating weak correlation with human judgment [21, 26].

LLM-based evaluators. Recently, there has been growing interest in developing LLM-based evaluators to provide more reliable assessments of long-form answers [8]. For example, Vu et al. [38] introduce LLM autoraters by training the PaLM-2-24B model on a collection of 102 quality assessment tasks comprising more than 5.3M human judgements (i.e., FLAMe). Fan et al. [12] propose an LLM-based metric, EVA-Score, to evaluate the informativeness in abstract lonng-form summarization. Although LLMs have shown promise in providing more comprehensive evaluations of long-form answers, they remain prone to vulnerabilities due to their well-known hallucination issues [39].

Goals and questions. In this paper, we aim to fill the above gap by conducting a comprehensive analysis of existing automatic evaluation metrics, including deterministic metrics and model-based metrics. We compare automatic metrics with human evaluations to examine their strengths and limitations. Given this setup, we explore the following three research questions:

(RQ1) To what extent do existing automatic evaluation metrics serve as substitutes for human evaluations in LFQA, and how accurate are these metrics?

(RQ2) What are the limitations of automatic evaluation metrics compared to human evaluators? What influences their stability and fairness?

(RQ3) How can the effectiveness and robustness of existing evaluation methods be improved?

Main findings. Our study involves both factual and non-factual LFQA tasks, collecting answers generated by seven different LLMs on the ASQA, ANTIQUE and Wikieval datasets, and performing a human evaluation on a subset of these answers (over 2,079 answer pairs with 4,158 ratings and justifications).

- For RQ1, we find that metrics based on large models demonstrate significantly higher consistency with human evaluations than deterministic metrics. They are also more stable in assessing different types of LFQA.

- For RQ2, we conduct an analysis from two perspectives. First, we examine the impact of meaningless minor perturbations to the prompt on evaluation results, assessing the stability of the evaluators. Second, we investigate whether factors such as text length, question type, and non-semantic variations in phrasing introduce evaluator bias, thereby affecting the fairness of the evaluators. Our findings indicate that deterministic metrics are influenced by the length of the reference text and tend to penalize longer answers. In contrast, LLM-based methods do not exhibit any significant biases in this regard.

- For RQ3, our research analyzes the impact of different prompting strategies on outcomes, highlighting that fine-grained evaluation methods exhibit higher accuracy compared to similar evaluation approaches.

2 Related Work

This work aims to analyze the shortcomings of existing evaluations in LFQA and provide guidance for future efforts. Therefore, we review previous work, including (i) automated evaluation methods, and (ii) evaluation of evaluators.

2.1 Automatic Evaluation Methods

To evaluate the performance of large models, numerous studies have used automatic evaluation metrics across multiple independent benchmarks [2, 22, 23]. These metrics are generally categorized into three types: n-gram-based, embedding-based, and LLM-based. Each category offers distinct approaches for assessing model responses, aiming to balance efficiency and accuracy in evaluating various aspects of generated text.

N-gram-based metrics, such as ROUGE [24] and BLEU [31], are widely adopted due to their high efficiency and low cost, primarily verifying the correctness of model responses. However, these metrics can be adversely affected by syntactic errors [32] and may

struggle to capture human preferences when comparing outputs from different models [9, 15], with recent studies further supporting these limitations [21, 43]. Embedding-based methods, such as BERTSCore [46] and MoverScore [48], use pretrained language models to better capture semantic similarities, thereby reducing the impact of superficial textual changes. Despite their advancements, these methods rely heavily on high-quality reference texts, posing challenges for open-ended questions where crafting appropriate references is difficult [6].

While embedding-based approaches improve upon traditional n-gram methods by focusing on semantic content, their dependence on reference texts limits their applicability in more flexible evaluation scenarios. On the other hand, LLM-based evaluators offer a novel approach for reference-free evaluation, demonstrating reasonable performance in zero-shot settings [1, 7, 27]. Examples include GPTScore [13], G-EVAL [26], LLM-EVAL [25], and RAGAS [10], which leverage large language models to provide multi-dimensional evaluations. However, LLMs may inherit biases from their training data [40] and generate hallucinations [18], with studies confirming biases in these evaluators [39, 40, 49]. Although fine-grained evaluations can help mitigate some issues [28, 45], the overall reliability of automatic evaluation methods as substitutes for human evaluators remains an area requiring further investigation.

2.2 Evaluation of Evaluators

Since LLM-as-a-Judge has become a new evaluation paradigm, many studies have assessed the effectiveness of these methods. The basic approach of previous research has been to treat various types of evaluators as virtual annotators and evaluate their consistency with human annotators to achieve alignment with human[16].

MT-bench and the Chatbot arena dataset [49] contain a small amount of human-crafted queries annotated by experts and a large volume of crowdsourced user preferences data from real-world users, analyzing the agreement between LLM evaluators and human annotators and bias of the evaluator. Xu et al. [43] collect the consistency between automatic metrics and human experts in the evaluation of LFQA across seven different knowledge domains, focus on overall answer preference, coherence, and factuality. PandaLM [41] trains a LLM evaluator and constructed a general instruction-tuning benchmark that includes a significant amount of automated scoring and human preference data, using a partial order graph to compare the performance of the models.

In this work, we will adopt previous evaluation approaches for meta-evaluation, assessing the performance of evaluators based on their consistency with human annotators. We will focus on the biases and robustness exhibited by different evaluation methods throughout this process.

3 A Study of LFQA Evaluation Methods

In this section, we examine how existing automatic LFQA evaluation metrics compare to human evaluations from three aspects, namely *accuracy*, *robustness*, and *fairness*. Firstly, the *accuracy* is to test to what extent the automatic metrics match human judgments. Here, we use responses from seven LLMs across three benchmark datasets as testbeds to assess the alignment between automatic metrics and human evaluations. Secondly, the *robustness* is to test

the reliability of the automatic metrics. We assess the stability of the outcome when subjected to minor perturbations in inputs and hyperparameters. Finally, the *fairness* is to test whether existing automatic metrics exhibit biases toward specific attributes, e.g., style, length, or topic.

In the following, we first introduce the setting of the empirical study, including automatic evaluation metrics, models, and the testbed. Then, we show results concerning the *accuracy*, *robustness*, and *fairness* of all metrics.

3.1 Empirical Setup

3.1.1 Testbed. For long-form question answering, we mainly consider two categories: one requires in-depth analysis and detailed explanations, while the other includes opinions, discussions, and other scenarios closely resembling real user interactions, which are often non-factoid. Both types of questions are considered challenging for LLMs, in terms of response generation or evaluation[11, 21, 49]. Specifically, we perform experiments on the following three datasets, each serving as a representative dataset for ambiguous, factoid, and open-ended question answering.

- *ASQA* [33] is an ambiguous factual questions dataset in which each question has multiple disambiguated question-answer pairs and two long-form grounded answers annotated by humans.
- *ANTIQUE* [17] is an open-ended question answering dataset, including 2626 questions asked by users of Yahoo! Answers, and relevant answers annotated by human experts.
- *WikiEval* [10] is a factoid question answering dataset generated from 50 pages from Wikipedia with edits post 2022, annotated by human experts.

Using the datasets mentioned above, we gather responses from seven latest LLMs from five different families for analysis. The models are listed as follows:

- GLM-4-9b-chat [14]: An advanced generative language model designed for dialogue applications with a focus on efficient response generation.
- Llama2-7b-chat [37]: A conversational AI model from the Llama family, optimized for contextual understanding in dialogues with 7 billion parameters.
- Llama2-13b-chat [37]: A more robust version of its predecessor, this model offers enhanced conversational abilities with 13 billion parameters.
- Llama3-8b-instruct [36]: Building on the Llama series, this iteration focuses on instruction-following tasks with an 8 billion-parameter architecture.
- GPT-3.5-instruct [29]: A finetuned version of GPT-3.5, this model excels in following user instructions with improved precision.
- Mistral-7b [19]: A lightweight yet powerful model, designed for streamlined AI tasks.
- Solar-10.7b-instruct [20]: A state-of-the-art model tailored for specific instruction adherence, boasting 10.7 billion parameters for diverse task performance.

To generate answers, we use the recommended generation parameters for each model and guide the models using few-shot settings. For each input, we select up to 8 examples. We selected 500 samples from the dev set of ASQA, 200 samples from the test set of ANTIQUE and 50 samples from WikiEval to generate answers. After
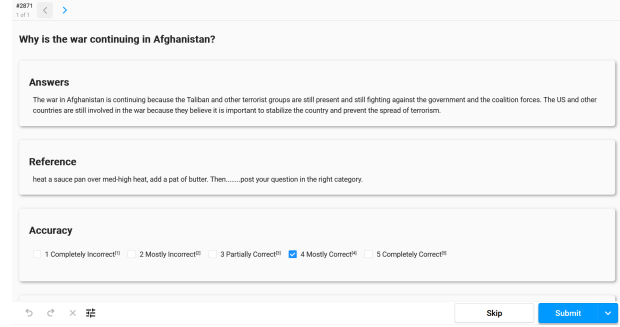

Figure 1: Interface used for collecting human annotations.

filtering out invalid QA pairs-such as those resulting from model refusals-we obtained 3,500 valid samples from ASQA, 1,386 from ANTIQUE, and 350 from WikiEval.

3.1.2 Meta-evaluations. This involves the evaluation of the evaluation, serving as a method to assess the quality of automatic evaluation metrics. Here, we adopt two types of meta-evaluations, as outlined below.

- Correlation coefficient: When comparing the differences between two distributions, a commonly used metric is the correlation coefficient. Following Chen et al. [6], we use the Spearman correlation coefficient to measure the value correlation and Kendall correlation coefficient to measure the rank correlation between human ratings and automatic metrics.
- Win Rate and Agreement: When comparing multiple LLMs, researchers often use the win rate, defined as the fraction of instances in which model A's response outperforms model B's. Furthermore, for two different evaluators, their agreement can be calculated, which represents the proportion of instances where the evaluators made the same judgment regarding the quality of the responses from the two models [49].

3.1.3 Human annotations. In the annotation of model responses, we drew inspiration from previous work [47] and primarily considered two evaluation aspects.

- Correctness aims to determine the accuracy of the answers, determining whether there are any factual errors or inaccuracies in the model's response. A high-quality response should be factually reliable and free from mistakes.
- Informativeness examines whether the response contains sufficient and relevant information, identifying any missing or omitted details, and ensuring there is no excessive redundancy. A high-quality response should should be informative enough without being redundant.

For the results of the seven models on the ASQA, ANTIQUE, and WikiEval datasets, we randomly sampled 50, 200, and 50 questions, respectively, for manual evaluation, with each question having 7 answers from different models. After removing a few cases where models refused to answer or generated incorrect responses, we collected 343, 1386, and 350 valid QA pairs that include both correctness and informativeness. We developed a web-based interface to streamline the annotation process. Figure 1 illustrates the interface for a single question. Each question is paired with a generated answer and a reference answer.

Table 1: Correlation coefficients between automatic metrics and human ratings for ASQA dataset. The best score for each column is highlighted in bold. The second best is underlined.

Spearman/Kendall (%)	Deterministic Metrics			Model-based Metrics					
	RL	EM	DF1	BS	AR	FG	GPT-4o	Claude-3.5	Gemini-2.0
glm4-9b	31.0/22.8	26.1/22.4	31.4/24.0	29.4/22.5	22.4/14.6	**66.5**/**52.2**	56.6/45.0	37.2/30.6	15.2/11.7
gpt-3.5-turbo	-0.3/-0.5	48.6/40.3	18.5/12.9	2.6/2.4	8.3/5.7	50.8/40.6	32.6/25.1	22.3/18.3	15.4/11.9
llama2-13b	21.9/16.3	56.6/47.7	14.4/9.8	31.1/23.7	8.7/6.0	**66.6**/**54.1**	29.1/22.1	13.1/10.5	41.1/30.1
llama2-7b	17.0/12.1	54.3/46.2	14.7/11.1	22.9/16.6	23.7/17.3	43.6/35.6	32.5/24.6	30.8/22.6	22.6/15.4
llama3-8b	-2.4/-1.6	47.5/39.4	59.1/44.8	12.1/8.1	14.0/10.1	**72.1**/**60.5**	43.5/33.8	20.7/17.5	32.1/23.8
mistral-7b	53.3/37.9	57.0/49.9	33.2/24.0	33.5/24.0	17.1/12.0	62.0/49.3	47.8/37.3	59.0/45.7	51.3/39.5
solar-10.7b	4.3/2.6	10.4/8.6	-8.9/-6.1	9.0/8.4	8.9/7.0	22.9/22.3	**41.4**/**32.2**	20.6/17.7	8.5/6.8
Average	11.4/8.0	41.4/34.9	23.0/16.8	16.2/11.7	12.2/8.7	**55.0**/**44.9**	42.0/32.4	33.0/26.3	32.4/24.4

Table 2: Correlation coefficients between automatic metrics and human ratings for ANTIQUE and WikiEval dataset. The best score for each column is highlighted in bold. The second best is underlined.

Spearman/Kendall (%)	ANTIQUE dataset					WikiEval dataset			
	RL	BS	AR	CG	FG	RL	BS	CG	FG
glm4-9b	-14.4/-11.4	-9.8/-7.7	0.5/0.4	36.7/33.1	**54.6**/**51.7**	44.4/34.4	**44.6**/34.4	38.1/34.8	39.1/**35.7**
gpt-3.5-turbo	2.8/2.0	-9.1/-7.0	43.7/35.0	53.2/47.1	**56.5**/**51.7**	**57.1**/**45.1**	56.8/43.1	41.9/36.5	46.9/42.1
llama2-13b	11.0/7.8	0.5/0.5	10.9/7.6	74.9/62.5	**81.5**/**71.2**	61.4/47.2	57.1/44.3	**61.8**/**51.7**	55.4/46.8
llama2-7b	13.8/10.0	9.3/6.9	17.7/13.4	69.3/58.6	**74.6**/**64.7**	74.6/60.4	**75.2**/**60.3**	60.8/52.8	60.1/52.9
llama3-8b	2.4/1.5	8.7/6.3	2.5/1.6	65.6/54.6	**72.2**/**63.1**	51.7/40.6	44.9/34.5	44.1/37.7	35.7/31.9
mistral-7b	29.2/21.0	25.9/18.5	23.3/16.9	73.0/62.5	**83.9**/**73.6**	51.3/40.0	60.7/48.9	47.7/40.0	50.4/42.6
solar-10.7b	-4.6/-3.4	-9.7/-6.6	-3.5/-2.3	54.3/46.2	**70.8**/**62.5**	56.3/43.3	54.3/41.7	41.2/35.5	34.1/31.1
Average	-7.9/-5.9	-24.4/-17.2	28.3/20.6	83.0/70.9	**85.2**/**75.6**	55.8/43.4	55.5/42.6	51.4/**43.9**	49.0/43.1

3.1.4 Automatic metrics. For evaluation metrics, we include 7 widely used metrics for long-form question answering, including traditional deterministic metrics and recently model-based metrics.

Deterministic metrics: The deterministic metrics typically assess results based on lexical overlap between the predicted answer and the reference answer. Early works on LFQA usually take *Exact Match* (EM) and *Rouge-L* (RL) as the evaluation metrics [11, 33]. The EM metric is to evaluate whether the predicted answer matches any of a fraction in the reference answer [33]. The Rouge-L is to evaluate the quality of the predicted answer by measuring the longest common subsequence between the generated text and the reference answer [24]. We also include the *Disambig-F1* (DF1) [33], which is to evaluate the fraction of diambiguated questions answered by the predicted answer, for the evaluation study.

Model-based metrics: The model-based metrics often assess the semantic accuracy of the generated answer using pre-trained models, extending beyond traditional n-gram matching. The *BERTScore* (BS) [46] calculate the cosine similarity between the generated answer and the reference answer. In addition, we considered the answer relevance (AR) from RAGAS [10] to supplement the measurement of responses' informativeness. For LLM-based evaluation, we follow LLM-EVAL [25] to use prompt-driven GPT-4 model as the judge in both coarse-grained (CG) and fine-grained (FG) settings.

3.2 Accuracy of Automatic Metrics

In this section, we try to answer RQ1, that is to what extent do existing automatic evaluation metrics serve as substitutes for human evaluations in LFQA. In this part, we take responses of seven

LLMs on three datasets as the testbed. For each response, we conduct a human rating and compute the scores for each metric. We then assess the correlation coefficient between these metrics and human ratings across the entire dataset. Additionally, for any pair of model-generated responses, we calculate the agreement between each metric and human rating to determine whether they align in their preference between these two answers. The results of correlation coefficients between automatic metrics and human ratings for the three datasets are presented in Table 1 and 2.

Firstly, for deterministic metrics, we can see that exact match demonstrates relatively strong alignment with human evaluation on ASQA dataset, even outperforming LLM-based evaluations such as GPT-4o, Claude-3.5 and Gemini-2.0. This is primarily because exact match focuses solely on assessing short answers in ASQA. However, while evaluating long answers, metrics like Rouge-L and Disambig-F1 exhibit poor consistency with human evaluations. An exception is observed with Rouge-L, which yields significantly different results on the Antique and WikiEval datasets. It shows high consistency with human evaluations on WikiEval but low consistency on Antique. This discrepency arises because Antique features non-factoid QA with open-ended answers, whereas WikiEval focuses on factoid QA with closed-ended answers.

Secondly, for model-based metrics, we can see that BertScore exhibits a trend similar to Rouge-L, showing low consistency with human evaluations on the ASQA and Antique datasets but high consistency on the WikiEval dataset. In contrast, LLM-based evaluations exhibit strong consistency with human evaluations across various styles of LFQA, including the ambiguous QA, fatoid QA,
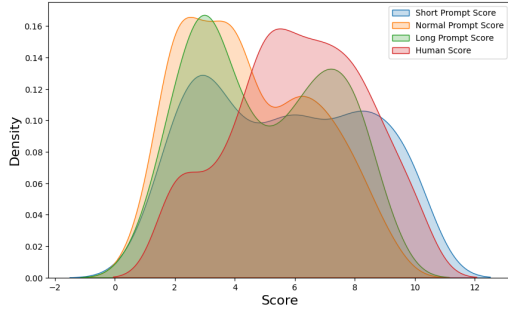
Figure 2: Score distribution of different prompts on ASQA.

and non-factoid QA, highlighting their stability in evaluation. For example, the fine-grained (FG) GPT4 achieves the best performance on both the ASQA and Antique datasets.

Lastly, when comparing different LLM-based evaluations, we find that GPT-4o evaluation demonstrates superior performance over other LLMs. For example, the GPT-4o, Claude-3.5, and Gemini-2.0 achieve spearman correlation score of 42.0, 33.0, and 32.4, respectively, when compared to human evaluations. Moreover, the fine-grained evaluation with GPT-4o improves its score from 42.0 to 55.0, which demonstrates the importance of providing more detailed instructions for LLM-based evaluations.

In summary, LLM-based evaluations demonstrate stability across different types of LFQA, whereas deterministic evaluations are less reliable for open-ended QA. Furthermore, conducting fine-grained evaluations with LLM-based assessments would produce more nuanced results.

3.3 Robustness of Automatic Metrics

In this section, we focus on the first part of RQ2, which investigates the robustness of automatic evaluation metrics. For this purpose, we introduce small perturbations to each evaluation method, and analyze the changes before and after the perturbation. Here, we concentrate on model-based evaluation metrics. Firstly, we collects the scores generated by automatic evaluation metrics under their default configurations. Subsequently, we introduce perturbations to the experimental conditions and obtain the corresponding perturbed scores. To quantify the consistency of evaluators' decisions before and after the introduction of perturbations, we statistically analyzed the model win rates and the distribution of metric scores under different experimental settings.

3.3.1 Perturbation of prompts. The prompts of LLMs have been shown to significantly influence their output, giving rise to a new research area known as prompt engineering [44]. To better understand the impact of prompts on the evaluation capability of LLMs, we conducted experiments comparing different prompts using GPT-4o on the ASQA dataset, including short prompt, normal prompt, and long prompt. More detailed description of prompts are shown in code repository.[1]

The results are depicted in Figure 2: (i) LLM-based evaluations tend to yield lower scores compared to human evaluations, with the majority of LLM-based scores clustering around 3, while human evaluation scores are predominantly around 5. This discrepancy

[1]https://github.com/bugtig6351/lfqa_evaluation/src/prompts.py

(a) Results on the ASQA dataset

(b) Results on the Wikieval dataset

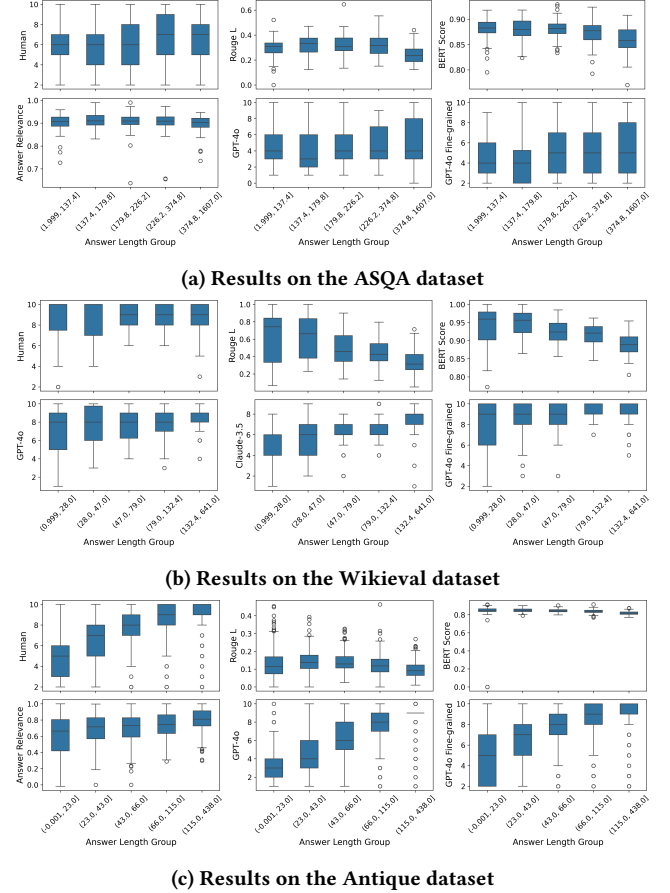**(c) Results on the Antique dataset**

Figure 3: Relationship between answer length and metrics.

may be due to LLM-based evaluations being more stringent than human evaluations. (ii) The length of the prompt has a significant impact on the evaluation scores of LLMs. It is evident that short prompts tend to result in more high scores. For example, the number of samples with scores above 4 for the short prompt is noticeably higher than for the normal and long prompts.

3.3.2 Perturbation of hyper-parameters. Hyperparameters such as sampling temperature, top-k sampling, repetition penalty, and maximum token length all play a role in shaping the LLM's output and overall performance [30]. However, the impact of sampling temperature on LLM-as-a-judge has not been specifically investigated. The selection of sampling temperature is largely based on guesswork and intuition. Here, we take GPT-4o as the judge to analyze the impact of temperature.

All results are summarized in Table 3. If multiple models have the same score, we take the average rank as their shared rank. As we can see: (i) In the ASQA dataset, the evaluation results are highly stable, where minimal changes in score and no changes in the rankings of any LLMs. This could be because the evaluation score for each model vary widely, ranging from 3.03 to 8.66, making it difficult for the rankings to change. (ii) In the WikiEval dataset, the evaluation results vary significantly with changes in the temperature of the LLM judge. For example, the rank of GPT-4o drops from 4th to 6th when the temperature increases from 0.0 to 0.3. In contrast, its rank

Table 3: The relationship between model scores and evaluators' temperature, with the default temperature set to 0.

| | ASQA | | | | | | | | Wikieval | | | | | | | |
| | Scores | Ranks | Scores Gain | | | Rank Gain | | | Scores | Ranks | Scores Gain | | | Rank Gain | | |
	0.0	0.0	0.3	0.7	1.0	0.3	0.7	1.0	0.0	0.0	0.3	0.7	1.0	0.3	0.7	1.0
claude-3.5	8.66	1	0.08	0.05	-0.04	0.0	0.0	0.0	8.26	2	-0.14	-0.04	-0.08	-1.0	-1.0	-1.0
gpt-4o	8.13	2	0.05	0.04	0.04	0.0	0.0	0.0	8.08	4	-0.18	0.24	0.16	-2.0	2.0	2.0
gemini-2.0	6.69	3	0.04	0.03	0.09	0.0	0.0	0.0	7.86	7	-0.40	-0.28	-0.08	-1.0	-1.0	-1.0
glm4-9b	4.87	4	-0.02	0.00	-0.01	0.0	0.0	0.0	8.62	1	-0.16	-0.08	-0.20	0.0	0.0	0.0
solar-10.7b	4.70	5	0.05	0.06	0.07	0.0	0.0	0.0	7.94	5	0.16	0.22	0.18	1.0	0.5	1.0
gpt-3.5-turbo	4.67	6	0.01	0.05	-0.05	0.0	0.0	0.0	7.92	6	0.08	-0.06	0.02	1.0	0.0	0.0
llama3-8b	4.27	7	0.10	0.09	0.09	0.0	0.0	0.0	7.66	8	-0.02	0.18	0.16	1.0	1.0	1.0
llama2-13b	4.01	8	-0.02	0.05	-0.06	0.0	0.0	0.0	6.82	9	0.04	0.08	0.18	0.0	0.0	0.0
llama2-7b	3.60	9	-0.06	-0.02	-0.02	0.0	0.0	0.0	6.60	10	0.04	0.14	-0.04	0.0	0.0	0.0
mistral-7b	3.03	10	0.03	0.02	0.02	0.0	0.0	0.0	8.20	3	-0.04	-0.04	-0.22	1.0	-1.5	-2.0

Table 4: Question types in the ANTIQUE dataset.

| Question Type	Count	Proportion
REASON	427	0.31
INSTRUCTION	357	0.26
EVIDENCE-BASED	266	0.19
EXPERIENCE	154	0.11
DEBATE	105	0.08
COMPARISON	77	0.06

improves from 4th to 2nd when the temperature is set to 0.7 or 1.0. This may be because the evaluation scores of each model are very similar, with the minimum and maximum values being 6.60 and 8.62, respectively. Overall, the above observations indicate that the temperature of the LLM does impact the evaluation, but the extent of this impact depends on the dataset.

3.4 Fairness of Automatic Metrics

In this section, we address the second part of RQ2, which focuses on the fairness of automatic evaluation metrics. Specifically, we conduct a comprehensive analysis of the evaluators' biases across four key dimensions: response length, question type, answer generation models, and language representations.

3.4.1 Length bias. Here, we aim to examine whether answer length influences the evaluation of automatic metrics, particularly LLM-based evaluation methods. To this end, we divide all answers into five bins with equal sample sizes based on their length and analyze the performance of different evaluation metrics across these length intervals. All the results are depicted in Figure 3. Our experimental results reveal distinct patterns in the relationship between answer length and various evaluation metrics.

We observe a negative correlation between answer length and scores on the Rouge-L and BERTScore metrics. As the length of the answer increases, the Rouge-L and BERTScore metrics tend to decrease. This trend suggests that longer answers may introduce redundancy or deviate from the concise, information-dense content that these metrics favor. Rouge-L, which measures the overlap between generated and reference texts, is particularly sensitive to extraneous information that dilutes the precision of the answer. Similarly, BERTScore, which evaluates semantic similarity, may
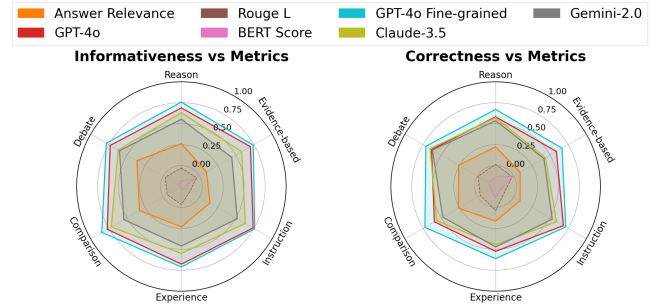
Figure 4: Relationship between different metrics and human evaluations across different question types on the ANTIQUE dataset (left: Informativeness, right: Correctness).

penalize longer answers that include tangential or less relevant content.

On the ASQA and WikiEval dataset, we observe a positive correlation between answer scores and length in Claude-3.5. While the average scores remained relatively stable, the lower bounds of scores assigned by GPT-4 and Gemini-2.0 increased significantly with longer answers. For the Antique dataset, the trend of increasing scores with length is evident, and this trend is more pronounced in LLM-based methods compared to human evaluations. This indicates that LLMs are less likely to assign low scores to longer answers, a phenomenon not observed in human evaluations. These evaluation frameworks may prioritize comprehensiveness and detail over conciseness.

Our analysis reveals that answer length significantly influences the performance of automatic evaluation metrics, with distinct patterns observed across different metrics and models. Combine traditional metrics (e.g., Rouge-L, BERTScore) with LLM-based metrics to balance the strengths and weaknesses of each approach. To address these biases and improve the robustness of automatic evaluation metrics, some strategies could be considered. For example, traditional metrics can ensure precision and conciseness, while LLM-based metrics can capture comprehensiveness and detail. And develop metrics that account for answer length by normalizing scores based on the amount of relevant information. This could involve penalizing redundancy while rewarding additional meaningful content.
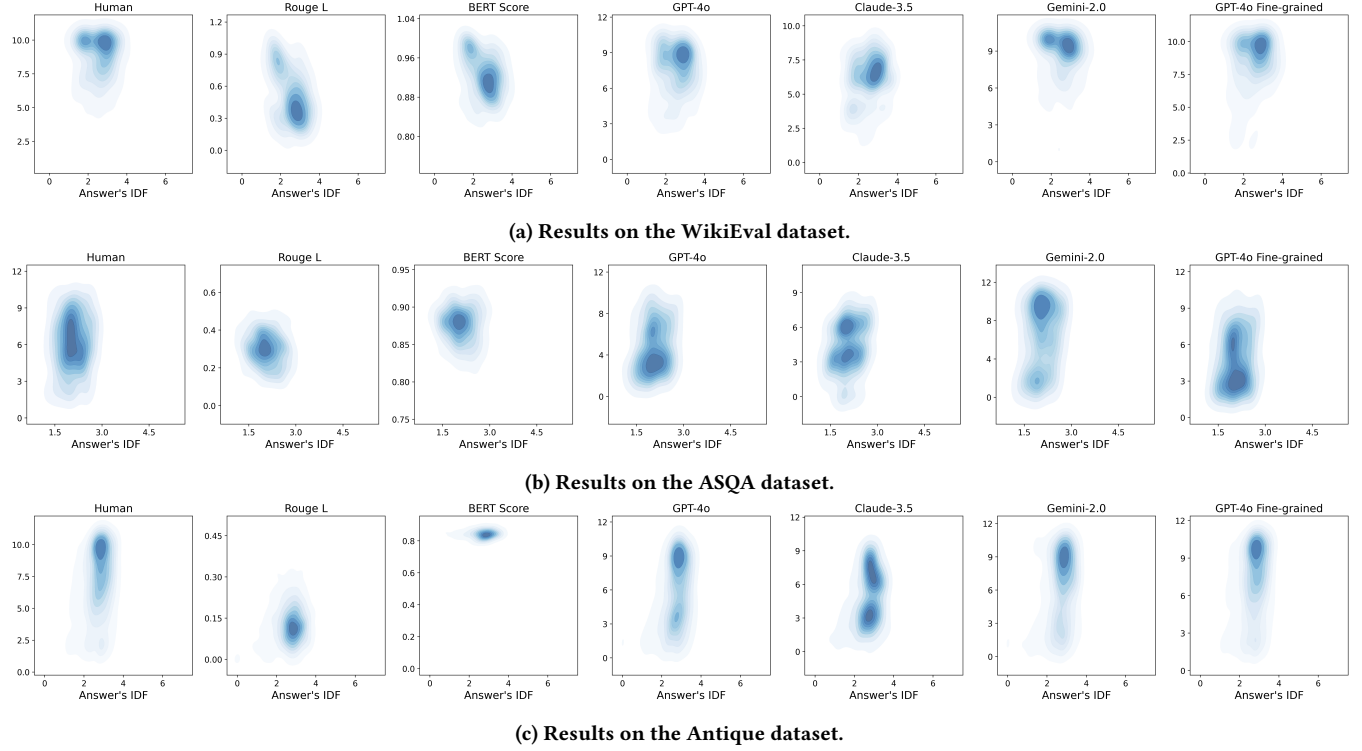
(a) Results on the WikiEval dataset.

(b) Results on the ASQA dataset.

(c) Results on the Antique dataset.

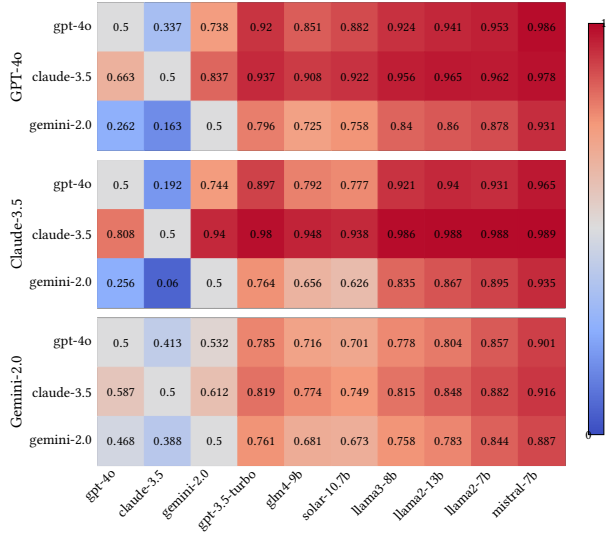Figure 5: Relationship between the IDF of answer and different metrics.

Figure 6: Winrate for different metrics on the ASQA dataset.

3.4.2 Question type. In this section, we investigate the potential biases of different evaluation metrics when assessing various types of questions. Building upon the research conducted by Bolotova et al. [3], we classified 200 questions from the antique dataset, with the specific quantities and proportions detailed in Table 4. For each category of questions, we calculated the kendall correlation consistency between human evaluators' scores for correctness and informativeness and various metrics, as illustrated in Figure 4.

Our analysis reveals several key findings. First, we observe that LLM-based metrics significantly outperform deterministic metrics

(Rouge-L and BERTScore), with the latter showing negative correlation coefficients on this dataset, rendering them practically unreliable for evaluation purposes. Second, LLM-based metrics demonstrate slightly higher consistency in assessing informativeness compared to correctness, particularly for comparison and experience-type questions. This discrepancy may stem from the inherent difficulty of LLMs in identifying factual inaccuracies within generated responses. Finally, evidence-based and experience-type questions emerge as challenging areas for all metrics, with the three primary evaluation models (GPT-4, Claude-3.5, and Gemini-2.0) consistently underperforming relative to their average performance across other question types.

Based on our findings, we can conclude that LLM indeed exhibit evaluation biases across different types of question, particularly for questions relying on precise factual information or personal experiences and recommendations. However, potential improvements can be achieved through the implementation of more granular prompting strategies combined with a diversified evaluation approach.

3.4.3 Self-reinforcing. In this section, we investigate whether LLM-based metrics show a preference for results generated by themselves. We take the widely adopted LLMs, i.e., GPT-4o, Claude-3.5, and Gemini-2.0, as both evaluators and generators on the ASQA dataset. Specifically, we take these three LLM to generate answers for each question, and conduct a pairwise comparison between them and other seven open-sourced LLMs. The results are illustrated in Figure 6.

We can see that: (i) The three closed-source LLMs (GPT-4o, Claude-3.5, and Gemini-2.0) consistently outperform the seven baseline LLMs (Mistral-7b, Llama2-7b, Llama3-8b, Solar-10.7b, GLM4-9b,

Table 5: Variation of correlation coefficients under different prompt words.

Kendall (%)	LLMs			Prompts								
	GPT-4o	Claude-3.5	Gemini-2.0	P1	P2	P3	P4	P5	P6	P7	P8	P9
glm4-9b	34.76	-36.15	-2.72	8.56	-17.45	3.48	9.87	3.91	4.10	2.25	-2.38	7.49
gpt-3.5-turbo	36.47	-2.02	-6.48	16.44	0.61	27.64	22.54	8.36	15.10	12.36	13.24	19.84
llama2-13b	51.66	-23.50	4.93	16.54	10.52	11.73	8.35	11.32	12.11	12.13	12.41	8.62
llama2-7b	52.77	-11.17	-4.88	6.08	2.85	7.78	10.22	4.43	5.69	0.05	6.02	5.79
llama3-8b	37.66	-1.70	-18.25	18.93	18.24	7.69	21.73	15.33	3.63	17.22	15.90	15.67
mistral-7b	39.97	-31.42	-31.06	-6.44	-10.06	-6.78	-16.19	-17.30	-15.88	-6.74	-16.58	-14.19
solar-10.7b	35.45	-8.16	8.56	12.29	15.01	21.72	14.79	9.89	4.65	19.94	14.61	24.10
Average	43.89	-15.91	-4.84	11.68	5.48	12.44	11.15	7.71	5.74	9.49	8.21	11.38

Table 6: Prompt settings.

Prompt	Settings
P1	Task+Data+Output+Criteria
P2	Task+Data+Output
P3	Task+Data+Criteria+Output
P4	Output+Task+Data+Criteria
P5	Task+Output+Criteria
P6	Data+Output+Criteria
P7	Output+Criteria+Task+Data
P8	Data+Task+Criteria(short)+Output
P9	Data+Task+Criteria+Output

and GPT-3.5-turbo). Among them, Claude-3.5 achieves the best performance in all baselines across all three LLM-based evaluations. (ii) Among the three evaluation methods, we observe that GPT-4o and Claude-3.5 evaluations exhibit a strong bias towards their own responses. For example, GPT-4o achieves significantly higher win rates against all baselines when evaluated using GPT-4o compared to evaluations conducted by Claude-3.5 and Gemini-2.0. Similarly, Claude-3.5 demonstrates much higher win rates in its own evaluation than when assessed by GPT-4o and Gemini-2.0. Furthermore, the Gemini-2.0 evaluation also gives much higher scores to itself compared to GPT-4o and Claude-3.5 evaluations (e.g., 0.468 vs. 0.262/0.256, 0.388 vs. 0.163/0.06), when compared with GPT-4o and Claude-3.5.

In summary, LLM-based evaluations tend to assign significantly higher scores to their own outputs, demonstrating a clear evaluation bias. Interestingly, despite this bias, the models' ranking remains consistent across different evaluations. Based on these, we recommend the following: (i) Prioritize using rankings over scores for comparisons, (ii) Employ multiple evaluation methods to achieve more reliable and stable results.

3.4.4 Word expression. Prior work shows that LLMs often produce verbose, formal content [42]. To investigate whether LLM-based evaluations favor answers containing rarer terms, we compute each answer's average inverse document frequency (IDF) and examine its correlation with the evaluation scores. The results are illustrated in Figure 5. As observed: On Antique and WikiEval, human scores are evenly distributed across IDF values, showing no clear link between answer quality and term rarity, while ASQA displays a unique bimodal pattern, suggesting different answer characteristics. Automatic metrics like Rouge-L and BERTScore strongly

correlate with IDF values, following near-normal distributions on Antique and WikiEval. In contrast, GPT-4o and Gemini-2.0 align more closely with human judgments. Notably, under fine-grained settings, GPT-4o tends to favor responses with higher IDF values, indicating a potential bias toward lexically sophisticated responses. These results highlight the importance of dataset-specific evaluation strategies and the need to address IDF-related biases in LLM-based metrics, and this similarity becomes even more pronounced when fine-grained evaluation settings are applied.

4 Method

In this section, we investigate approaches to improve the performance of LLM-based evaluators to address RQ3: "How can the effectiveness and robustness of existing evaluation methods be improved?" To achieve this, we transfer the findings in previous sections into criteria in the prompt of LLM-based evaluation. Specifically, we decompose the prompt architecture into four key components: task description, data specifications, output requirements, and evaluation criteria. Then, we obtain 9 prompts by different combinations of these four components as outlined in Table 6. Due to space limitations, all detailed prompts can be found in our released code.[2] Finally, we measure performance variations by calculating the Kendall correlation coefficient changes between GPT-4o evaluations and human ratings for each prompt configuration.

All results are shown in Table 5. Our main findings are as follows.

(1) **All four components are essential for improving LLM-based evaluations**: Comparing the performance of P1 with P2, P5 and P6, it can be observed that P1 achieves significantly better performance across all LLMs. P1 incorporates all four components, while P2 excludes the *Criteria*, P5 omits the *Data*, and P6 lacks the *Task*. Moreover, the evaluation of glm4-9b and gpt-turbo-3.5 drop significantly when removes the *Criteria* from the prompts (i.e., P2 vs. P3). Besides, the *Task* is important for LLM-based evaluations as performance drops significantly by comparing P3 vs. P6.

(2) **The order of components influences the LLM-based evaluations**: A comparison of P1, P3, P4, P7, and P9 shows that the order of the four components has varying impacts on the evaluation of different LLMs. Notably, the placement of the *Output* appears to have minimal influence on gpt-3.5-turbo, as the performance is very close between P3 and P4—where *Output* is positioned last in P3 and first in P4.

---

[2]https://github.com/bugtig6351/lfqa_evaluation/src/prompts.py

Table 7: Evaluation instructions.

Components	Instructions
Task	Score the following LLM output of a question-answering task with respect to the following aspects using a 1 to 5 star rating system.
Data	The dataset is a Factoid Question-Answering dataset, specifically designed for evaluating factual precision and detailed comparative reasoning in AI-generated answers.
Output	Begin your evaluation by providing a short explanation. Be as objective as possible. After providing your explanation, please provide your evaluation by strictly following the JSON format, such as: [[SCORE]] {"accuracy": 2, "informativeness": 3}.
Criteria	**Accuracy**: Determine the accuracy of the answers, verifying the correctness and reliability of the information provided. 1 star: Incorrect information 2 stars: Partially correct information 3 stars: Half correct information 4 stars: Mostly correct information 5 stars: Perfectly correct information **Informativeness**: Examines whether the answers provide sufficient and meaningful information that is useful to the user and relevant to the question. 1 star: No information or irrelevant information 2 stars: Very little information 3 stars: Some information 4 stars: Enough information 5 stars: Highly informative

(3) **No consistent improvements can be obtained for the evaluation of all LLMs by one strategy**: There is no one strategy can boost the performance of all LLMs evaluations, where glm4-9b, llama2-7b, and llama3-8b performs best with P4, gpt-3.5-turbo performs best with P3, llama2-13b performs best with P1, solar-10.7b performs best with P9. The differences in answer styles between models may significantly affect the evaluator's performance across various prompts.

These results demonstrate the necessity of structured guidance and precise evaluation guidelines for reliable assessments.

5 Conclusion

In this work, we evaluated a range of automatic evaluation metrics, including traditional deterministic metrics and model-based metrics, across three diverse datasets: ASQA, Antique, and WikiEval. For each dataset, we compared the scores generated by these metrics with human evaluations, analyzed their robustness under perturbations, and investigated potential biases related to answer length, question type, self-reinforcement, and language expression. Additionally, we explored methods to improve the performance of LLM-based evaluators through fine-grained evaluation strategies.

Our evaluation indicates that even with relatively high-quality reference answers, deterministic evaluation metrics still perform poorly and often do not exhibit significant correlation with human judgments. Their performance is highly dependent on the dataset and question type, with limited applicability to complex LFQA tasks. LLM-based metrics, such as GPT-4o and Claude-3.5, demonstrate better alignment with human evaluations and greater stability across different types of questions. However, they are susceptible to biases related to answer length, question type, and self-reinforcement. For example, LLM-based evaluators tend to favor longer answers and their own generated responses. The observed

biases in LLM-based evaluations call for the development of more equitable evaluation frameworks that account for factors such as answer length, question type, and language expression.

Although our study has investigated the performance of automatic evaluation metrics across multiple dimensions, there are still some limitations that point to directions for future work: (i) Our 750-question evaluation covers only 3 QA categories, while real-world LFQA involves more diverse formats (e.g., multihop reasoning). Future research should evaluate automatic metrics on broader datasets, including more diverse type and domain of questions. (ii) We adopted a single-point scoring approach, and for human evaluation, we only set two main scoring criteria: correctness and informativeness. This may miss subtle quality differences between answers. A more nuanced evaluation setup would help fully capture various potential responses and edge cases. (iii) Our experiments have revealed the sensitivity of LLM-based models to prompts, but we have not conducted a detailed analysis of how different models respond to this, further investigation is needed.

Acknowledgments

References

[1] Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-yeol Ahn. 2023. Can we trust the evaluation on ChatGPT?. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. Association for Computational Linguistics, Toronto, Canada, 47–54. https://doi.org/10.18653/v1/2023.trustnlp-1.5

[2] Meghana Moorthy Bhat, Rui Meng, Ye Liu, Yingbo Zhou, and Semih Yavuz. 2023. Investigating Answerability of LLMs for Long-Form Question Answering. arXiv:2309.08210 (Sept. 2023). https://doi.org/10.48550/arXiv.2309.08210 arXiv:2309.08210 [cs].

[3] 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 (SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1196–1207. https://doi.org/10.1145/3477495.3531926

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 159, 25 pages.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. arXiv preprint arXiv:2303.12712 (2023).

[6] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*. Association for Computational Linguistics, Nusa Dua, Bali, 361–374. https://doi.org/10.18653/v1/2023.findings-ijcnlp.32

[7] David Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 15607–15631. https://doi.org/10.18653/V1/2023.ACL-LONG.870

[8] David Cheng-Han Chiang and Hung-yi Lee. 2023. A Closer Look into Using Large Language Models for Automatic Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 8928–8942. https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.599

[9] Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On The Evaluation of Machine Translation Systems Trained With Back-Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2836–2846. https://doi.org/10.18653/v1/2020.acl-main.253

[10] 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, EACL 2024 - System Demonstrations, St. Julians, Malta, March 17-22, 2024*. Association for Computational Linguistics, 150–158. https://aclanthology.org/2024.eacl-demo.16

[11] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3558–3567. https://doi.org/10.18653/v1/P19-1346

[12] Yuchen Fan, Xin Zhong, Yazhe Wan, Chengsi Wang, Haonan Cheng, Gaoche Wu, Ning Ding, and Bowen Zhou. 2024. EVA-Score: Evaluating Abstractive Long-form Summarization on Informativeness through Extraction and Validation. *arXiv preprint arXiv:2407.04969* (2024).

[13] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as You Desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Association for Computational Linguistics, Mexico City, Mexico, 6556–6576. https://doi.org/10.18653/v1/2024.naacl-long.365

[14] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793

[15] Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical Power and Translationese in Machine Translation Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Association for Computational Linguistics, 72–81. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.6

[16] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A Survey on LLM-as-a-Judge. arXiv:2411.15594 (Dec. 2024). https://doi.org/10.48550/arXiv.2411.15594 arXiv:2411.15594 [cs].

[17] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. 2020. ANTIQUE: A non-factoid question answering benchmark. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer, 166–173.

[18] 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, Article 248 (March 2023), 38 pages. https://doi.org/10.1145/3571730

[19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).

[20] Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*. Association for Computational Linguistics, 23–35. https://doi.org/10.18653/V1/2024.NAACL-INDUSTRY.3

[21] Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to Progress in Long-form Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (Eds.). Association for Computational Linguistics, Online, 4940–4957. https://doi.org/10.18653/v1/2021.naacl-main.393

[22] Md. Tahmid Rahman Laskar, M. Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. 2023. A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics, 431–469. https://doi.org/10.18653/V1/2023.FINDINGS-ACL.29

[23] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models. arXiv:2211.09110 (Nov. 2022). http://arxiv.org/abs/2211.09110 arXiv:2211.09110 [cs].

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

[25] Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*. Association for Computational Linguistics, Toronto, Canada, 47–58. https://doi.org/10.18653/v1/2023.nlp4convai-1.5

[26] 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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics, 2511–2522. https://doi.org/10.18653/V1/2023.EMNLP-MAIN.153

[27] Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. 2024. LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*. Association for Computational Linguistics, 139–151. https://aclanthology.org/2024.eacl-long.8

[28] 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*. Association for Computational Linguistics, Singapore, 12076–12100. https://doi.org/10.18653/v1/2023.emnlp-main.741

[29] OpenAI. 2023. GPT-3.5-Turbo-Instruct. https://www.openai.com Accessed: 2024-07-18.

[30] OpenAI. 2025. Explore developer resources, tutorials, API docs, and dynamic examples to get the most out of OpenAI's platform. https://platform.openai.com Accessed: 2024-01-24.

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

[32] Ehud Reiter and Anja Belz. 2009. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics* 35, 4 (Dec. 2009), 529–558. https://doi.org/10.1162/coli.2009.35.4.35405

[33] Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid Questions Meet Long-Form Answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 8273–8288. https://doi.org/10.18653/v1/2022.emnlp-main.566

[34] Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before Generate! Faithful Long Form Question Answering with Machine Reading. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 744–756. https://doi.org/10.18653/v1/2022.findings-acl.61

[35] Mingxu Tao, Dongyan Zhao, and Yansong Feng. 2025. Chain-of-Discussion: A Multi-Model Framework for Complex Evidence-Based Question Answering. In *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics, Abu Dhabi, UAE, 11070–11085. https://aclanthology.org/2025.coling-main.734/

[36] Meta LLaMA Team. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/

[37] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[38] Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational Autoraters: Taming Large Language Models for Better Automatic Evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*. Association for Computational Linguistics, 17086–17105. https://aclanthology.org/2024.emnlp-main.949

[39] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*. Association for Computational Linguistics, Singapore, 1–11. https://doi.org/10.18653/v1/2023.newsum-1.1

[40] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 9440–9450. https://doi.org/10.18653/v1/2024.acl-long.511

[41] Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=5Nn2BLV7SB

[42] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics* (2025), 1–65.

[43] Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A Critical Evaluation of Evaluations for Long-form Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 3225–3245. https://doi.org/10.18653/v1/2023.acl-long.181

[44] Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. Prompt Engineering a Prompt Engineer. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, 355–385. https://doi.org/10.18653/v1/2024.findings-acl.21

[45] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=CYmF38ysDa

[46] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkeHuCVFDr

[47] Yue Zhang, Ming Zhang, Haipeng Yuan, Shichun Liu, Yongyao Shi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. LLMEval: A Preliminary Study on How to Evaluate Large Language Models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*. AAAI Press, 19615–19622. https://doi.org/10.1609/AAAI.V38I17.29934

[48] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 563–578. https://doi.org/10.18653/v1/D19-1053

[49] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2020, 29 pages.