LINKAGE: Listwise Ranking among Varied-Quality References for Non-Factoid QA Evaluation via LLMs

Anonymous ACL submission

Abstract

Non-Factoid (NF) Question Answering (QA) is challenging to evaluate due to diverse potential answers and no objective criterion. The commonly used automatic evaluation metrics like ROUGE or BERTScore cannot accurately measure semantic similarities or answers from different perspectives. Recently, Large Language Models (LLMs) have been resorted to for NFQA evaluation due to their compelling performance on various NLP tasks. Common approaches include pointwise scoring of each candidate answer and pairwise comparisons between answers. Inspired by the evolution from pointwise to pairwise to listwise in learning-to-rank methods, we propose a novel listwise NFQA evaluation approach, that utilizes LLMs to rank candidate answers in a list of reference answers sorted by descending quality. Moreover, for NF questions that do not have multi-grade or any golden answers, we leverage LLMs to generate the reference answer list of various quality to facilitate the listwise evaluation. Extensive experimental results on three NFQA datasets, i.e., ANTIQUE, the TREC-DL-NF, and WebGLM show that our method has significantly higher correlations with human annotations compared to automatic scores and common pointwise and pairwise approaches.

1 Introduction

In recent years, studies on various aspects of Large Language Models (LLMs) have been drawing significant attention, a majority of which are based on the task of factoid question answering (QA) (Saad-Falcon et al., 2024; Xu et al., 2024; Lee et al., 2022). New evaluation metrics and benchmarks have also been proposed for assessing the factuality of LLMs (Wang et al., 2023; Min et al., 2023). However, much less research has been conducted on non-factoid question answering (NFQA), which usually requires long-form answers to answer open-ended non-factoid questions (NFQ), such as explanations, opinions, or descriptions. This can be attributed to the inherent difficulty of the NFQA task and the lack of a well-recognized metric to evaluate the generated long-form answers. Effective evaluation of NFQA is the foundation of developing advanced techniques to enhance the quality of LLMs-generated non-factoid answers.

Evaluating NFQA is challenging since non-factoid questions often involve subjective interpretations and the potential answers can be diverse instead of a definite fact. Most prior work used automatic evaluation metrics such as measuring word overlaps (e.g., ROUGE (Lin, 2004) and BLEU(Papineni et al., 2002)) and semantic similarities (e.g., BERTScore (Zhang et al., 2019)) with the ground truth answers. To ensure the evaluation reliability, a small amount of manual annotations are also incorporated to compare the NFQA performance. However, both of them have some limitations: Automatic metrics like ROUGE, BLEU, and BERTScore cannot accurately measure the responses with semantically similar expressions or from a different but reasonable perspective respectively; Human evaluations, although more accurate in measuring various aspects of the long-form answers, often require annotators to have related knowledge to be reliable and are too expensive to apply on a large scale.(Krishna et al., 2021; Liu et al., 2023). Moreover, even for humans, evaluation of NFQA can still be challenging due to the requirement of domain knowledge as well as subjective interpretations of the questions and judgment criterions.

By ingesting large-scale data from multi-tasks, LLMs, such as the GPT series, have achieved compelling performance on numerous Natural Language Processing (NLP) tasks, and sometimes even outperform humans (Zhao et al., 2023). Increasing attention has been drawn to leveraging LLMs as surrogates for large-scale evaluation on model-

generated responses (Min et al., 2023; Saad-Falcon et al., 2024; Fu et al., 2023). Following the routines of human evaluation, approaches that leverage LLMs as judges often adopt the ways of pointwise scoring that grades each candidate answer individually and pairwise comparisons that compare pairs of answers(Zheng et al., 2024). The pair for comparison can be two candidate answers or a candidate answer and a ground truth answer. Figure 1 shows a concrete example of these two approaches.

Pointwise grading is hard since the accurate perception of differences between each grade can be difficult. Subtle differences between candidates may not be discerned and reflected in the final score. Pairwise comparison is relatively easier and can be more accurate but it is not scalable to the large number of candidates when the comparisons are between candidates. In contrast, there is no such issue when comparing the pair of a candidate and a ground truth answer. However, it is not feasible when the ground truth is unavailable. Moreover, when only a single ground truth exists, the evaluation may not be accurate to cover various aspects.

Inspired by the evolution of learning to rank in information retrieval, i.e., from pointwise to pairwise to listwise (Liu et al., 2009; Cao et al., 2007), we propose a listwise NFQA evaluation approach that leverages LLMs to conduct ListwIse raNKing AmonG varied-quality referencEs, abbreviated as LINKAGE. Specifically, we use LLMs to assess a candidate answer by its rank in a list of reference answers sorted by quality descendingly. When there are ground truth answers of multiple grades, they can be used as the varied-quality references. When there is only one or no golden answer, we will construct some examples of multi-grade answers and utilize the in-context learning ability of LLMs to generate more reference answers of different quality. Compared to the pointwise and pairwise approach, listwise ranking can yield more accurate assessment since the LLM judge can take reference answers of various quality into consideration simultaneously. When only one reference answer is used, our method degenerates to pairwise comparisons with a ground truth answer. Additionally, given an ordered reference answer list, LLMs only ingest the reference list and candidate answer once, which costs much less than comparing each reference answer with the candidate pairwise and aggregate the score.

We conduct extensive experiments on three NFQA datasets: ANTIQUE (Hashemi et al., 2020), the non-factoid portion of TREC DL (Craswell et al., 2020, 2021), and WebGLM (Liu et al., 2023). ANTIQUE and TREC DL have multi-grade manual annotations on the candidate answers while WebGLM is a non-factoid QA dataset based on Retrieval Augmented Generation (RAG) that provides retrieval passages and a single ground truth answer. Under the settings where there are multiple, single, or none ground truth answers, our method outperforms the automatic similarity scores, as well as pointwise, and pairwise LLM evaluation methods significantly in terms of the correlation with human judgments. By offering more accurate NFQA evaluation, our work can pave the way for future studies on improving NFQA performance, especially promoting LLMs to become more capable of answering complex questions.

2 Related Work

2.1 Non-factoid Question Answering(QA)

Non-factoid question answering (NFQA) is a complex challenge, characterized by open-ended queries that require complex responses such as descriptions, opinions, or explanations.(Yulianti et al., 2017; Cohen and Croft, 2016). These responses are usually extensive, often requiring paragraph-level answers. The most used benchmark in NFQA is the ELI5 dataset (Fan et al., 2019), which contains 272,000 questions from the "Explain Like I'm Five" Reddit forum. Moreover, multi-document NFQA datasets like WebGLM (Liu et al., 2023), WikihowQA (Bolotova-Baranova et al., 2023) integrate multiple detailed passage-level answers to form long-form answers to NFQ. ANTIQUE (Hashemi et al., 2020) provides a reliable collection with complete relevance annotations of NFQA.

2.2 Non-factoid QA Evaluation

Prior NFQA approaches can be categorized into three categories:

Automatic Evaluation: Before the emergence of LLM, the most commonly used evaluation methods were automatic metrics, such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2019). These metrics evaluate the quality of a generated answer based on text similarity between the answer and human-written answers. However, these automatic metrics calculate scores through n-gram similarity, ignoring semantic information. For instance, Krishna et al. (2021) show that ROUGE is an ineffective

metric in long-form question answer tasks. Another way to implement automatic evaluation is by training a model with human evaluation preferences to conduct automatic assessment, such as QAFactEval(Fabbri et al., 2021) and RankGen(Krishna et al., 2022). However, these methods struggle to generalize to out-of-domain QA evaluation due to limited human annotations.

Human Evaluation: In NFQA tasks, human annotations are usually considered the golden standard. Hurdles (Krishna et al., 2021), WebGPT (Nakano et al., 2021), WikihowQA (Bolotova-Baranova et al., 2023) both ask human annotators to choose their preferred answer between the answer generated by the model and the golden answer. Moreover, to compensate for human lack of understanding in certain domains, they can refer to evidence documents during evaluation. However, human evaluation is expensive and therefore difficult to adopt on a large scale.

LLM Evaluation: As LLMs advance, they are gradually replacing costly human annotations. GPTScore (Fu et al., 2023) uses the generation probability of LLMs to evaluate the model-generated output. LLM-Eval(Lin and Chen, 2023) uses a unique prompt-based evaluation method for open-domain conversations with LLMs. PRD (Li et al., 2023) and CHATEVAL (Chan et al., 2023) integrate different LLMs' evaluation results by ranking, discussing, and debating among LLMs. The advantage of using LLMs as evaluators lies in their explainability and scalability. However, they also encounter issues such as position bias, verbosity bias, and self-enhancement bias. (Zheng et al., 2024) There is a lack of research specifically focused on LLM evaluation for NFQA.

3 Method

In this section, we propose a ListwIse raNKing AmonG varied-quality referencEs method (LINKAGE) for evaluating NFQA. We formally define the task of NFQA evaluation and introduce some basic evaluation approaches, then introduce the details of our LINKAGE.

3.1 Preliminary

Task Definition: Given a non-factoid question q and its corresponding n candidate answers $\mathcal{C} = \{c_1, c_2, ..., c_n\}$ to be evaluated, where c_i represents the i-th candidate answer. The goal is to score each answer with a scorer $Score(c_i)$. The ground truth set of q is $\mathcal{G} = \{g_1, g_2, ..., g_k\}$, in which g_i represents the i-th ground truth. In this paper, the scorer is LLM and we use a prompt $\mathcal{P}$ to query the LLM to get the scoring results.

Currently, the commonly used scoring methods based on LLM are pointwise and pairwise approaches(Zheng et al., 2024).

Pointwise Evaluation: The pointwise evaluation approach assesses an answer c_i only based on its relevance and quality regarding the question q. As shown in Figure 1, the evaluation process may be conducted with or without using ground truth answers as references.

$$Score_{\text{point}}(c_i) = f(\mathcal{P}_{\text{point}}, q, c_i, \mathcal{R}), \quad (1)$$

in which $f(P_{\text{point}}, \cdot)$ represents querying the LLM through prompt $\mathcal{P}_{\text{point}}$. $\mathcal{R} = [r_1, r_2, \ldots, r_m]$ is a reference answer list sorted by quality in descending order, which can be $\mathcal{G}$, a subset of $\mathcal{G}$, or $\emptyset$.

Pointwise grading is easy to conduct but difficult to accurately perceive grade differences. The subtle differences among candidates may not be distinguished and reflected in the final score.

Pairwise Evaluation: As shown in Figure 1, the pairwise evaluation approach performs a pairwise comparison between answers. The pairs can be two candidate answers,

$$Score_{\text{pair}}(c_i) = \sum_{c_j \in \mathcal{C} \setminus \{c_i\}} f(\mathcal{P}_{\text{pair}}, q, c_i, c_j). \quad (2)$$

However, the number of comparisons between candidate answer pairs grows exponentially with the number of candidate answers, and thus cannot be scaled to a large number of candidates. The pair can also be a candidate answer and a reference answer,

$$Score_{\text{pair}}(c_i) = \sum_{r_j \in \mathcal{R}} w_{l_j} * f(\mathcal{P}_{\text{pair}}, q, c_i, r_j), \quad (3)$$

$$f(\mathcal{P}_{\text{pair}}, q, c_i, r_j) = \begin{cases} 1, \text{ if } c_i \text{ is better} \\ -1, \text{ if } r_j \text{ is better} \\ 0, \text{ otherwise} \end{cases}. \quad (4)$$

$\mathcal{R}$ can be $\mathcal{G}$ or a subset of $\mathcal{G}$. w_{l_j} is the weight corresponding to certain grade l_j of answer r_j. In this way, the pairwise approach scores a candidate answer by comparing it with each answer in the reference answer list.

Pairwise comparison is relatively easier and can be more accurate, but when there is only a single

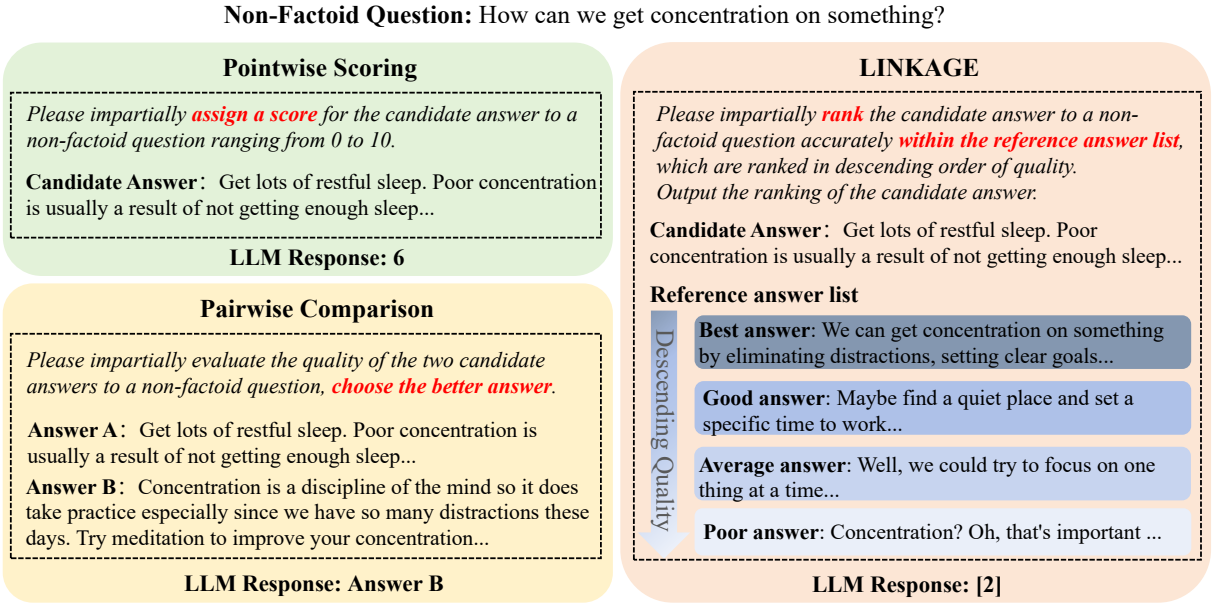

Figure 1: Pointwise scoring evaluation, pairwise comparison evaluation and our LINKAGE evaluation approaches.

ground truth, evaluation becomes less accurate because it is difficult for a single ground truth to cover various aspects of NFQA

3.2 Listwise Ranking Evaluation (LINKAGE)

Figure 1 shows how our LINKAGE works. Specifically, given a reference answer list sorted by descending quality and the answer to be evaluated, the scorer judges its quality by deciding where it should be ranked among the reference answer list,

$$Score_{\text{pair}}(c_i) = f(\mathcal{P}_{\text{list}}, q, c_i, \mathcal{R}). \quad (5)$$

The higher the ranking, the better the quality.

Please note the difference between our method and the pointwise approach with references. Although both methods ask LLMs to directly output a numerical value, in the pointwise approach, references are used to provide a criterion for scoring, and the assignment only focuses on the quality of c_i itself rather than comparisons. The listwise ranking approach relies on comparing it with all reference answers to determine where the answer should be ranked.

3.3 Reference List Construction

Reference answer list $\mathcal{R}$ in LINKAGE is composed of multiple answers ordered in descending quality. Compared to providing LLMs with only one ground truth, more references with different styles and quality enable the LLM evaluators to learn implicit evaluation guidelines from $\mathcal{R}$. The collection method of $\mathcal{R}$ depends on the composition of the ground truth set of the dataset, and we discuss it in three situations:

3.3.1 Multi-grade Ground Truth

When multiple grades of ground truth answers are available, references can be sampled directly from these answers. For instance, ANTIQUE and TREC DL contain multiple answers annotated with four relevant labels.

To reduce bias and ensure the reliability of evaluation results, we randomize the sampling process multiple times. Additionally, the length and the distribution of $\mathcal{R}$ also impact the results. We discuss this in detail in Section 5.2.

3.3.2 Single-grade Ground Truth

Some NFQA datasets, such as WebGLM, only contain a single grade of ground truth. For this scenario, we prompt LLMs to generate answers of varying quality to serve as references. Specifically, we first prompt LLMs to answer the question based on the original golden answer, thus obtaining a new high-quality golden answer. The prompt is in Figure 8 (Appendix A.2). This step ensures that both the golden reference and other reference answers are generated by LLMs, avoiding the introduction of style bias between human and machine writing. We then use the prompt in Figure 9 (Appendix A.2) to obtain other lower-quality reference answers. To ensure the diversity of references, we use three LLMs to generate separate lists of reference answers. Then we randomly sample reference answers from three lists to form $\mathcal{R}$ for each grade.

Table 1: Statistics of ANTIQUE and TREC-DL-NF we use in experiments.

Number statistics	ANTIQUE	TREC-DL-NF
#Question	500	55
#Avg doc labeled 3	5.8	9.6
#Avg doc labeled 2	4.5	18.1
#Avg doc labeled 1	6.5	24.9
#Avg doc labeled 0	3.6	48.0
#Avg total documents	20.4	100.7

3.3.3 Absence of Ground Truth

In real-world scenarios, non-factoid questions may not have reference answers. To tackle the problem of ground truth missing, considering the powerful capabilities of LLMs like GPT-4 (OpenAI, 2022a), we get a quality-assured answer from GPT-4 directly. The ways of generating reference answers of other quality are the same as described in Section 3.3.2.

4 Experimental Settings

4.1 Datasets

We evaluate the effectiveness of baseline methods and our proposed LINKAGE using the following three datasets.

ANTIQUE (Hashemi et al., 2020) dataset contains 2,626 open-domain non-factoid questions asked by real users in a community question answering service, i.e., Yahoo! Answers. Similar to TREC-DL, all passages are graded into four levels (3: reasonable and convincing, 2: not sufficiently convincing, 1: unreasonable, 0: make no sense). We merge the 200 questions from the test set and the 300 questions randomly sampled from the training set, yielding a total of 500 queries as our experiment dataset.

TREC-DL-NF (Craswell et al., 2020, 2021) In our experiments, we use TREC-DL 2019, 2020 datasets, which comprise 43 and 54 MS MARCO queries respectively. Each question has multiple passages labeled with four levels of relevance (3: perfectly relevant, 2: highly relevant, 1: related, 0: irrelevant). Not all questions are NF questions, so we filter factoid questions with a non-factoid question category classifier (Bolotova et al., 2022). This leaves us a total of 55 non-factoid questions, denoted as TREC-DL-NF.

The statistics of ANQIQUE and TREC-DL-NF can be found in Table 1.

WebGLM (Liu et al., 2023) is a high-quality quoted long-formed retrieval-augmented QA dataset. Each question is accompanied by 5 top-ranked documents retrieved by a vanilla Contriever (Izacard et al., 2021). Question and corresponding candidate references are fed together to OpenAI text-davinci-003 (Ye et al., 2023) to generate long-formed answers by 1-shot in-context learning. To obtain candidate answers of different styles and quality, we use gpt-3.5-turbo-16k (OpenAI, 2022b) to generate two answers with 5 relevant and 3 relevant plus 2 irrelevant documents respectively. The third answer is generated by Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) with 5 relevant documents. We sample 50 cases and manually label three candidate answers with three levels (3,2,1). Details about manual annotation are in the Appendix D.

4.2 Methods for Comparison

We compare the following NFQA evaluation baselines and our LINKAGE under different situations.

4.2.1 Baselines

Automatic Metrics:

ROUGE(Lin, 2004), **BERTScore**(Zhang et al., 2019), **BLEU**(Papineni et al., 2002) are all reference-based metrics based on text similarity. ROUGE and BLEU focus on exact n-gram matching, while BERTScore evaluates the semantic similarity of embeddings.

LLM Evaluation Baselines:

- **Pointwise**$^{R=\emptyset}$: This method asks LLMs to directly assign a quality score from 1 to 10 to the candidate answer without any reference answers.
- **Pointwise**$^{R\neq\emptyset}$: Based on the basic pointwise method, this method also provides a list of reference answers sorted in descending order of quality for LLMs to refer to when scoring.
- **Pairwise**: This method scores a candidate answer based on comparing it with each answer in the reference list. To eliminate position bias, i.e., the LLM judge might favor the forward-positioned one when comparing two answers, we randomly permute the positions of the candidate answer and ground truth answer during evaluation.

4.2.2 LINKAGE

LINKAGE: To ensure that $\mathcal{R}$ uniformly contains answers of varying quality, we randomly select the same number of reference answers from the answer set of each level to create the reference answer list. For TREC-DL-NF, the grades of answers in the reference list are $\mathcal{L} = (3, 2, 1, 0)$. For ANTIQUE,

Table 2: The performance of different methods on ANTIQUE and TREC-DL-NF. K, S, and P represent Kendall's tau, Pearson's r, and Spearman's rho coefficient respectively. The best results of each evaluator model are in bold.

	Method	ANTIQUE			TREC-DL-NF		
		K	S	P	K	S	P
Automatic Metrics	ROUGE-1	0.2088	0.2563	0.2878	0.2442	0.3060	0.3412
	ROUGE-2	0.1807	0.2089	0.2281	0.2064	0.2441	0.2808
	ROUGE-L	0.2012	0.2463	0.2708	0.2171	0.2721	0.3178
	BERTScore	0.1562	0.1938	0.1950	0.2258	0.2824	0.2842
	BLEU	0.1808	0.2153	0.2063	0.2106	0.2650	0.2208
LLM Evaluation with Mistral	Pointwise$^{R=\emptyset}$	0.2202	0.2499	0.2519	0.2366	0.2773	0.2677
	Pointwise$^{R\neq\emptyset}$	0.2229	0.2516	0.2547	0.3138	0.3382	0.3302
	Pairwise	0.1827	0.2134	0.2132	0.2501	0.2967	0.2939
LINKAGE on Mistral	LINKAGE^{0_shot}	0.3585	0.3790	0.3893	0.3287	0.3539	0.3401
	LINKAGE^{few_shot}	**0.3742**	**0.4200**	**0.4373**	**0.4312**	**0.4725**	**0.4958**
LLM Evaluation on ChatGPT	Pointwise$^{R=\emptyset}$	0.2777	0.3118	0.3244	0.3176	0.3640	0.3660
	Pointwise$^{R\neq\emptyset}$	0.2752	0.3112	0.3224	0.3746	0.4288	0.4449
	Pairwise	0.2979	0.3494	0.3756	0.3204	0.3692	0.3749
LINKAGE on ChatGPT	LINKAGE^{0_shot}	0.3070	0.3404	0.3514	0.3923	0.4315	0.4376
	LINKAGE^{few_shot}	**0.3096**	**0.3543**	**0.3688**	**0.3993**	**0.4325**	**0.4481**

Table 3: Results for the situation of single-grade ground truth. The best results of each model are in bold.

Model	Method	ANTIQUE		TREC-DL-NF	
		K	S	K	S
Mistral	Pointwise$^{R=\emptyset}_{1GT}$	22.02	24.99	23.66	27.73
	Pointwise$^{R\neq\emptyset}_{1GT}$	25.26	28.31	33.28	38.25
	Pairwise$_{1GT}$	20.89	23.41	30.43	36.62
	LINKAGE$^{0_shot}_{1GT}$	32.92	35.80	36.60	39.93
	LINKAGE$^{few_shot}_{1GT}$	**42.89**	**47.06**	**42.13**	**46.18**
ChatGPT	Pointwise$^{R=\emptyset}_{1GT}$	27.77	31.18	31.76	36.40
	Pointwise$^{R\neq\emptyset}_{1GT}$	27.91	30.71	39.75	44.66
	Pairwise$_{1GT}$	29.88	32.32	30.28	34.14
	LINKAGE$^{few_shot}_{1GT}$	**32.93**	**33.54**	**44.83**	**48.51**

Table 4: Results for the situation of absence of ground truth. The best results of each model are in bold.

Model	Method	ANTIQUE		TREC-DL	
		K	S	K	S
Mistral	Pointwise$^{R=\emptyset}_{0GT}$	22.02	24.99	23.66	27.73
	LINKAGE$^{0_shot}_{0GT}$	30.05	32.87	34.28	37.65
	LINKAGE$^{few_shot}_{0GT}$	**39.51**	**43.48**	**42.35**	**46.39**
ChatGPT	Pointwise$^{R=\emptyset}_{0GT}$	27.77	31.18	31.76	36.40
	LINKAGE$^{few_shot}_{0GT}$	**36.57**	**40.43**	**43.77**	**46.96**

$\mathcal{L} = (3, 3, 2, 2, 1, 1, 0, 0)$, which are the best settings in our experiment.

LINKAGE-1GT: We also test the case where there is only one ground truth. For questions with multi-grade answers, we randomly sample one answer from the highest-grade ground truth set as the only ground truth to simulate this situation.

LINKAGE-0GT: In this case, we do not use any labeled ground truth to simulate the situation where no ground truth is available.

4.3 Evaluation Metrics

To evaluate the effectiveness of NFQA evaluation, we use **Kendall's tau**, **Pearson's r** and **Spearman's rho coefficient** to calculate the extent of consistency between the resulting sorted sequences and the manually labeled sequences. Spearman's rho coefficient is chosen as our primary metric due to its balance between robustness and sensitivity to monotonic relationships.

4.4 Implementation Details

The evaluation experiments are based on two representative LLMs: (i) The open-source model Mistral (Mistral-7B-Instruct-v0.2 [1]) (Jiang et al., 2023). (ii) The close-source model ChatGPT (gpt-3.5-turbo-16k) (OpenAI, 2022b), for which results are obtained through API. The temperature for all experiments is set to 0.8.

When only one or no ground truth exists,

[1] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Table 5: Results on WEBGLM based on Mistral. RL, BS, and B represent ROUGE-L, BERTScore, and BLUE, respectively. Acc(b) means the accuracy of finding the best answer. Acc(b+w) means the accuracy of finding both the best and the worst answers.

	RL	BS	B	Point$^{R\neq\emptyset}$	Pair	LINKAGE
Acc(b)	0.42	0.48	0.50	0.46	0.54	**0.76**
Acc(b+w)	0.32	0.38	0.44	0.22	0.32	0.34

Table 6: Different composition of $\mathcal{R}$ on ANTIQUE and TREC-DL-NF and using Mistral. The settings we use in LINKAGE are in bold.

Dataset	$\lvert\mathcal{R}\rvert$	$\mathcal{R}$	0-shot		3-shot	
			K	S	K	S
ANTIQUE	4	3210	29.00	31.01	36.23	40.74
	8	33321000	31.73	33.64	35.46	39.96
		33221100	**35.85**	**37.90**	**37.42**	**42.00**
		32221110	27.12	29.52	37.29	42.03
TREC-DL-NF	4	**3210**	**32.87**	**35.39**	**43.12**	**47.25**
	8	33321000	29.84	32.29	36.78	40.98
		33221100	30.73	33.54	35.67	39.68
		32221110	31.66	34.37	37.35	41.35

we use gpt-4-1106-preview (OpenAI, 2022a) to generate the golden answer. For generating other references with descending quality, we use three different LLMs in 3-shot setting: (i) Mistral-7B-Instruct-v0.2, (ii) gpt-3.5-turbo-16k, (iii) Meta-Llama-3-8B-Instruct [2] (Meta, 2024). All our experiments are done on a single Tesla A100 80G GPU.

5 Experimental Results

5.1 Overall Results

The results on the multi-grade ground truth situation, single-grade ground truth situation, and absence of ground truth situation are shown in Table 2, Table 3, and Table 4 respectively. The results on WebGLM are shown in Table 5. It can be seen that our method always shows better consistency with human evaluation.

Additionally, we have the following observations:

LLM-based methods perform generally better than automatic metrics. This indicates that automatic metrics have limitations in NFQA evaluation, therefore should be used with caution in future research. Among LLM-based methods, our proposed LINKAGE outperforms all other baselines by a significantly large margin leveraging both Mistral and ChatGPT. This confirms the the superiority of listwise approach over the pointwise and pairwise approaches on NFQA evaluation.

Few-shot in-context Learning can enhance the performance of LINKAGE. Comparing with results under few-shot and zero-shot, providing LLMs with a few examples can help demonstrate the evaluation task more clearly. Compared to Mistral, the enhancement of few-shot ICL on ChatGPT is less. We think that it is because ChatGPT has a much better understanding of instructions so the few-shot example does not help it much. More experiment details about few-shot learning are in the Appendix B.1.

Reference answer list is important for understanding NFQ evaluation criteria. By analyzing the pointwise method results with and without reference, we find Pointwise$^{R\neq\emptyset}$ always performs better. In some cases, it can even exceed the performance of pairwise methods. This indicates that providing the reference answer list helps LLMs understand NFQ evaluation criteria so that Pointwise$^{R\neq\emptyset}$ can assign a more reliable score than Pointwise$^{R=\emptyset}$. This further illustrates that providing $\mathcal{R}$ in evaluating NFQA can lead to significant gains.

LINKAGE is applicable in various of situations. Table 3 and Table 4 show that LINKAGE-1GT and LINKAGE-0GT both perform the best among all LLM evaluation methods. This illustrates that our method is still effective when generalized to other evaluation scenarios, i.e., when there is only one ground truth or no ground truth.

5.2 Study on the Reference List Composition

We conduct experiments on different reference distributions to analyze their impact. As shown in Table 6, varying length and distribution of $\mathcal{R}$ affects the performance of LINKAGE.

The impact of length depends on the quality of the dataset. ANTIQUE is collected from web data and contains more noise, so increasing the number of references can help LLMs better build evaluation criteria. The conclusion on TREC-DL-NF is the opposite. For quality assurance datasets, increasing the number of references, however, exacerbates the burden of understanding long texts, thereby impairing evaluation performance. For the grade distribution of reference answers, uniform

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Non-Factoid Question: What is wifi vs bluetooth ?

Reference Answer List:

- **Best Answer 4**: Wi-Fi and Bluetooth are to some extent complementary in their applications and usage..
- **Good Answer 3**: "Bluetooth vs. WiFi - Range: Maximum range for Bluetooth based wireless connections is 30m while for Wi-Fi, it can extend well upto 100m...
- **Average Answer 2**: Bluetooth and WiFi are different standards for wireless communication. ...
- **Poor Answer 1**: Headphones use over 90% of available Bluetooth bandwidth...

Candidate Answer 1: Learn about Bluetooth and Wi-Fi for your Apple Watch, and why you should use both. To enjoy every feature on your Apple Watch, you need to turn on Wi-Fi and Bluetooth on your paired iPhone. Swipe up on your iPhone to open Control Center.

Human Label: 0 (0-3) ☹

Evaluation Results (0-10):

- **Pointwise：** 5 (0-10)
- **Pairwise:** 2.5 (Answer 4/3/2: Lose; Answer 1:Win)
- **LINKAGE:** 0 (Rank: 5)

Candidate Answer 2: You can also share a smartphone mobile data connection with other devices via the wireless Bluetooth radio. This is known as a Bluetooth personal area network, or PAN. Devices that include Bluetooth radios can connect to the smartphone via Bluetooth and access the Internet through it.

Human Label: 2 (0-3) ☺

Evaluation Results (0-10):

- **Pointwise：** 6 (0-10)
- **Pairwise:** 2.5 (Answer 4/3/2: Lose; Answer 1:Win)
- **LINKAGE:** 10 (Rank: 1)

Figure 2: An example of our LINKAGE compared with pointwise and pairwise approaches. We standardized the score range of all methods to $[0, 10]$ for easy comparison and understanding.

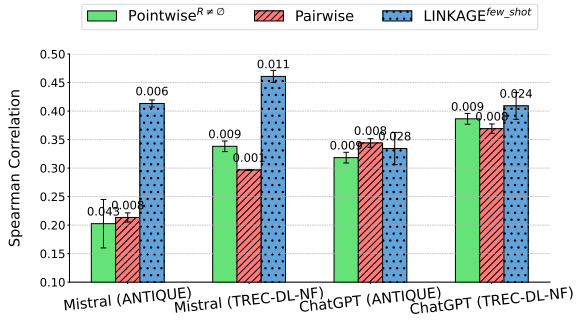

Figure 3: Comparison of Spearman Correlation for Mistral and ChatGPT on ANTIQUE and TREC-DL-NF. The error bars denote the standard deviation, illustrating the variability in the results.

sampling always brings the best results, as it allows LLMs to understand all grades of answers while avoiding introducing grade preference bias.

5.3 Study on the Reference List Randomness

Our experiments involve random sampling of the ground truth set, so we evaluate the results under 3 randomizations to analyze the impact of randomness on performance. Details about experiments are in Appendix B.2. From Figure 3, we can observe that for all LLMs on all datasets, the standard deviations of the experiments are always small. This indicates that the randomness of the selection of reference answers has little impact on the evaluation results, which proves that the improvement brought by our method is significant.

6 Case Study

We conduct case studies to qualitatively compare the results of different methods. As shown in the Figure 2, because candidate answer 1 contains many matching keywords, even though it does not effectively answer the question, pointwise method and pairwise method both assign it a high score. As a result, the two candidate answers cannot be effectively distinguished. In contrast, our LINKAGE can better distinguish the fine-grained quality differences between candidate answers and obtain results that are more consistent with humans.

7 Conclusion

In this paper, we propose a listwise NFQA evaluation approach (LINKAGE), which leverages LLMs to assess a candidate answer by its rank in a list of sorted reference answers. Our approach is capable of considering reference answers of various quality simultaneously. Therefore, it can enable LLMs to establish a better evaluation system and yield more accurate assessments. Extensive experiments on three datasets, i.e., ANTIQUE, TREC-DL-NF, and WebGLM, demonstrate the effectiveness of our method, whether it is in situations with multi-grade ground truth answers, single-grade ground truth answers, or no ground truth. Hoping this more accurate evaluation method can promote future research on NFQA.

Limitations

There are two primary limitations: (i) Our method demands multiple grading labels when constructing the reference answer list. When grading labels are missing, utilizing LLMs to generate reference answers increases the cost of inference. How to reduce the computational cost requires further research in the future. (ii) Compared with the pointwise and pairwise methods, the listwise method considers the relationship between all documents, so it requires the scoring model to have a good long-text understanding ability.

References

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

Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. 2023. Wikihowqa: A comprehensive benchmark for multi-document non-factoid question answering. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5314.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. *Proceedings of the 24th international conference on Machine learning*, pages 129–136.

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. *arXiv preprint arXiv:2308.07201*.

Daniel Cohen and W Bruce Croft. 2016. End to end long short term memory networks for non-factoid question answering. *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pages 143–146.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the trec 2020 deep learning track. *Preprint*, arXiv:2102.07662.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the trec 2019 deep learning track. *Preprint*, arXiv:2003.07820.

Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

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*, pages 166–173. Springer.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

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*.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. Rankgen: Improving text generation with large ranking models. *arXiv preprint arXiv:2205.09726*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. *arXiv preprint arXiv:2103.06332*.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Ruosen Li, Teerth Patel, and Xinya Du. 2023. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*.

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.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *Preprint*, arXiv:2305.13711.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient

web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4549–4560.

Meta. 2024. Welcome llama 3 - meta's new open llm.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI. 2022a. Introducing chatgpt.

OpenAI. 2022b. Introducing chatgpt.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. *Preprint*, arXiv:2311.09476.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. *Preprint*, arXiv:2304.14732.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *Preprint*, arXiv:2303.10420.

Evi Yulianti, Ruey-Cheng Chen, Falk Scholer, W Bruce Croft, and Mark Sanderson. 2017. Document summarization for answering non-factoid queries. *IEEE transactions on knowledge and data engineering*, 30(1):15–28.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Instruction Details

A.1 Instruction for Evaluation

The evaluation prompts are adopted in LINKAGE and LLM baselines (detailedly introduced in Sec4.2). These prompts are fed to LLMs, allowing them to generate scores, preferences or rankings.

Please impartially assign a score for the answer to a non-factoid question by comprehensively considering the answer's fluency, accuracy, truthfulness, objectivity and redundancy, within the range of 0-10. Higher scores means better quality.
Fluency measures the language smoothness and quality of the given answer.
Truthfulness measures whether the text of the answer is factually sound, including the factual consistency. of the answer and whether the answer contains contradictions or hallucinate information.
Objectivity measures whether the information of an answer is from provided references.
Redundancy measures the duplication of content within the limited text length. Repetitive content will reduce informativeness. The lower redundancy, the higher score of the answer.
Below are the non-factoid question and the candidate answer for evaluation.
Assign a score for the answer ranging from 0 to 10.
Output your final verdict by strictly following this format: \"[[8]]\" if score is 8.

Question:{#question}
Candidate answer:{#candidate}

Figure 4: Instruction for pointwise scoring without references.

Please impartially assign a score for the answer to a non-factoid question by comprehensively considering the answer's fluency, accuracy, truthfulness, objectivity and redundancy, within the range of 0-10. Higher scores means better quality.I will give you a reference answer list, which are ranked in descending order of quality.
Correctness measures the coherence of the answer and its corresponding question.
Fluency measures the language smoothness and quality of the given answer.
Truthfulness measures whether the text of the answer is factually sound, including the factual consistency. of the answer and whether the answer contains contradictions or hallucinate information.
Objectivity measures whether the information of an answer is from provided references.
Redundancy measures the duplication of content within the limited text length. Repetitive content will reduce informativeness. The lower redundancy, the higher score of the answer.
Below are the non-factoid question and the candidate answer for evaluation.
Assign a score for the answer ranging from 0 to 10.
Output your final verdict by strictly following this format: \"[[8]]\" if score is 8.
Question:{#question}
Reference answer list:{#reference}
Candidate answer:{#candidate}

Figure 5: Instruction for pointwise scoring with references.

Please impartially judge and evaluate the quality of the two candidate answers to a non-factoid question and choose the better answer.
Your evaluation should consider factors such as the correctness, fluency, truthfulness and redundancy.
Correctness measures the coherence of the answer and its corresponding question.
Fluency measures the language smoothness and quality of the given answer.
Truthfulness measures whether the text of the answer is factually sound, including the factual consistency of the answer and whether the answer contains contradictions or hallucinate information.
Redundancy measures the duplication of content within the limited text length. Repetitive content will reduce informativeness. The lower redundancy, the higher score of the answer.
Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision.Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants.Be as objective as possible.
After providing your explanation, output your final verdict by strictly following this format: \"[[A]]\" if assistant A is better, \"[[B]]\" if assistant B is better, and \"[[C]]\" for a tie.
[Question]: {#question}
[The Start of Assistant A's Answer]: {#answer_a}
[The End of Assistant A's Answer]
[The Start of Assistant B's Answer]: {#answer_b}
[The End of Assistant B's Answer]

Figure 6: Instruction for pairwise comparison.

Please impartially rank the given candidate answer to a non-factoid question accurately within the reference answer list, which are ranked in descending order of quality. The top answers are of the highest quality, while those at the bottom may be poor or unrelated.
Determine the ranking of the given candidate answer within the provided reference answer list. For instance, if it outperforms all references, output [[1]]. If it's deemed inferior to all four references, output [[5]].
Your response must strictly following this format: \"[[2]]\" if candidate answer could rank 2nd.
Below are the user's question, reference answer list, and the candidate answer.
Question: {#question}
Reference answer list: {#reference}
Candidate answer: {#candidate}

Figure 7: Instruction for our proposed LINKAGE.

A.2 Instruction for Generating Reference List

Given a non-factoid question:{#question} and its answer:{#ground0}
Use your internal knowledge to rewrite this answer.

Figure 8: Instruction for generating the highest standard reference answer.

Generate three different answers to a non-factoid question from good to bad in quality, each inferior to the golden answer I give you. Ensure that the quality gap from good to bad is very significant among these three answers. Golden answer is the reasonable and convincing answer to the question. Answer 1 can be an answer to the question, however, it is not sufficiently convincing. Answer 2 does not answer the question or if it does, it provides an unreasonable answer. Answer 3 is completely out of context or does not make any sense.

Here are 3 examples for your reference.
1.Non-factoid Question: how can we get concentration on something?
Golden Answer: To improve concentration, set clear goals, create a distraction-free environment, use time management techniques like the Pomodoro Technique, practice mindfulness, take regular breaks, stay organized, limit multitasking, practice deep work, maintain physical health, and seek help if needed.
Output:
Answer1:Improve focus: set goals, quiet space, Pomodoro Technique, mindfulness, breaks, organization, limit multitasking, deep work, health, seek help if needed.
Answer2:Just like and enjoy the work you do, concentration will come automatically.
Answer3:If you are student, you should concentrate on studies and don't ask childish questions.

2.Non-factoid Question: Why doesn't the water fall off earth if it's round?
Golden Answer: Earth's gravity pulls everything toward its center, including water. Even though Earth is round, gravity keeps water and everything else anchored to its surface. Gravity's force is strong enough to counteract the Earth's curvature, preventing water from falling off.
Output:
Answer1:This goes along with the question of why don't we fall off the earth if it is round. The answer is because gravity is holding us (and the water) down.
Answer2:Same reason the people don't.
Answer3:When rain drops fall through the atmosphere CO2 becomes dissolved in the water. CO2 is a normal component of the Earth's atmosphere, thus the rain is considered naturally acidic.

3.Non-factoid Question: How do I determine the charge of the iron in FeCl3?
Golden Answer: Since chloride ions (Cl-) each carry a charge of -1, and there are three chloride ions in FeCl3, the total negative charge from chloride ions is -3. To balance this, the iron ion (Fe) must have a charge of +3 to ensure the compound has a neutral overall charge. Therefore, the charge of the iron ion in FeCl3 is +3.
Output:
Answer1: Charge of Fe in Fecl3 is 3. Iron has either 2 as valancy or 3. in this case it bonds with three chlorine molecules. therefore its valency and charge is three.
Answer2:If two particles (or ions, or whatever) have opposite charge, then one has positive charge and one has negative charge.
Answer3:take a piece of iron. Wrap a copper wire around the iron in tight close coils. run a charge through the wire.

Below are the non-factoid question, and the golden answer.
Non-factoid Question: {#question}
Golden Answer: {#ground0}

Figure 9: Instruction for generating other reference answers in $\mathcal{R}$ sorted by quality descendingly.

B Experiment Details

B.1 N-shot LINKAGE performance

We conduct several sets of few-shot experiments of LINKAGE on TREC-DL-NF and ANTIQUE using Mistral. Results are in Table 7 and Table 8. We find the number of samples cannot be too large. When the number exceeds a certain value, the performance will deteriorate. This is because the shot number increasing leads to a significant increase in the input length, which will make the LLMs difficult to understand.

n-shot	Kendal	Spearman	Pearson
n=0	0.3287	0.3539	0.3401
n=1	0.4246	0.4656	0.4724
n=3	**0.4312**	**0.4752**	**0.4958**
n=5	0.4339	0.4725	0.4958

Table 7: The performance of LINKAGE under different few shot setting on TREC-DL-NF using Mistral.

B.2 Experiments on Randomness of Reference List

Three independent experiments on randomly selecting $\mathcal{R}$ on TREC-DL-NF and ANTIQUE using ChatGPT and Mistral are shown in Table 9 and Table 10.

n-shot	Kendal	Spearman	Pearson
n=0	0.2900	0.3101	0.3172
n=1	**0.3850**	**0.4183**	**0.4256**
n=3	0.3696	0.4012	0.4122
n=5	0.3654	0.3934	0.4041

Table 8: The performance of LINKAGE under different few shot setting on ANTIQUE using Mistral.

Table 9: Average Spearman coefficient and standard deviation of randomly selecting $\mathcal{R}$ in three dependent experiments on TREC-DL-NF using Mistral and ChatGPT.

Model	Method	Spearman 1	Spearman 2	Spearman 3	Average	Std
Mistral	Pointwise$^{R\neq\emptyset}$	0.3382	0.3463	0.3567	0.3471	0.0093
	Pairwise	0.2967	0.2783	0.2912	0.2887	0.0094
	LINKAGE^{few_shot}	0.4725	0.4579	0.4520	0.4608	0.0105
ChatGPT	Pointwise$^{R\neq\emptyset}$	0.2777	0.4288	0.4526	0.3864	0.0948
	Pairwise	0.3692	0.3687	0.3544	0.3641	0.0083
	LINKAGE^{few_shot}	0.4094	0.3854	0.4325	0.4091	0.0235

Table 10: Average Spearman coefficient and standard deviation of randomly select $\mathcal{R}$ in three dependent experiments on ANTIQUE using Mistral and ChatGPT.

Model	Method	Spearman 1	Spearman 2	Spearman 3	Average	Std
Mistral	Pointwise$^{R\neq\emptyset}$	0.2516	0.1781	0.1778	0.2025	0.0425
	Pairwise	0.2210	0.2059	0.2134	0.2134	0.0082
	LINKAGE^{few_shot}	0.4200	0.4078	0.4122	0.4133	0.0062
ChatGPT	Pointwise$^{R\neq\emptyset}$	0.3118	0.3250	0.3180	0.3182	0.0094
	Pairwise	0.3495	0.3387	0.3402	0.3428	0.0076
	LINKAGE^{few_shot}	0.3543	0.3143	0.3339	0.3340	0.0283

C Case Study

The details of case in Section 6 of the main paper is in Figure 10.

<table>
<tr><td colspan="2">Non-Factoid Question: What is wifi vs bluetooth ?
Reference Answer List:
·Best Answer 4: Wi-Fi and Bluetooth are to some extent complementary in their applications and usage. Wi-Fi is usually access point-centered, with an asymmetrical client-server connection with all traffic routed through the access point, while Bluetooth is usually symmetrical, between two Bluetooth devices.
·Good Answer 3: "Bluetooth vs. WiFi - Range: Maximum range for Bluetooth based wireless connections is 30m while for Wi-Fi, it can extend well upto 100m. In Wi-Fi, range depends on the version of Wi-Fi protocol applied and addition of antennas in the communication system while no such concerns of range or extra antenna are much known in Bluetooth. . Bluetooth vs. WiFi - Devices Connected: In Bluetooth, upto 7 devices can be connected to each other (piconet) while in Wi-Fi, the maximum connections depend on Wi-Fi router which can accommodate 1 to several communicating devices at a time.",
·Average Answer 2: "Bluetooth and WiFi are different standards for wireless communication. Bluetooth technology is useful when transferring information between two or more devices that are near each other when speed is not an issue, such as telephones, printers, modems and headsets.",
·Poor Answer 1: "Headphones use over 90% of available Bluetooth bandwidth. If you initiate any other Bluetooth activity (view devices in range, or try to use any other Bluetooth services), the music may play intermittently, skip, or the headphone's synchronization with the audio source may disconnect."</td></tr>
<tr><td>Candidate Answer 1: Learn about Bluetooth and Wi-Fi for your Apple Watch, and why you should. use both. To enjoy every feature on your Apple Watch, you need to turn on Wi-Fi and Bluetooth on your paired iPhone. Swipe up on your iPhone to open Control Center. Then make sure Wi-Fi and Bluetooth are on.
Human Label: 0 (0-3)
Pointwise Scoring： 5 (0-10)
Pairwise Comparison: Answer 4/3/2: Lose; Answer 1: Win
LINKAGE Rank:[5]</td><td>Candidate Answer 2: You can also share a smartphone mobile data connection with other devices via the wireless Bluetooth radio. This is known as a Bluetooth personal area network, or PAN. Devices that include Bluetooth radios can connect to the smartphone via Bluetooth and access the Internet through it.
Human Label: 2 (0-3)
Pointwise Scoring： 6 (0-10)
Pairwise Comparison: Answer 4/3/2: lose; Answer 1: Win
LINKAGE Rank: [1]</td></tr>
</table>

Figure 10: An example of our LINKAGE compared with Pointwise and Pairwise approach.

D Human Annotation

We recruit one domain expert who has earned at least a bachelor's degree in Computer Science to annotate WEBGLM candidate answer's quality label. The instruction is shown in Figure 11.

I will give you a non-factoid question and three candidate answers.
Please label each answer according to their quality, giving labels of 3, 2, 1. The best answer is labelled 3, the worst answer is labelled 1. If there are two answers that you think are close in quality, you can give the same label.

Non-Factoid Question: Why is driving into mild to heavy snowfall at night so disorienting?

Answer 1: The reason driving into mild to heavy snowfall at night is disorienting is because the snow obstructs your view and reflects your headlights. This makes it difficult to see where you are going.....
Answer 2: Driving into mild to heavy snowfall at night can be disorienting due to several factors. Firstly, the snowflakes in the air can reflect the headlights, creating a glare that obstructs visibility.....
Answer 3: Driving into mild to heavy snowfall at night can be disorienting because the snowflakes can reflect the headlights.....

Figure 11: Instructions for labeling WEBGLM for human annotators.