RE-RAG: Improving Open-Domain QA Performance and Interpretability with Relevance Estimator in Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-augmented generation (RAG) framework is showing state-of-the-art performance on open-domain question answering tasks by referencing external knowledge. However, the RAG system faces challenges with performance degradation when it is fed contexts of low relevance or when the relative relevance among the input contexts is inaccurately assessed. In this work, we propose a RE-RAG framework that injects an explicit context relevance estimator (RE) into the RAG system. RE-RAG re-evaluates the retrieved contexts with the proposed context RE and passes the more relevant contexts along with their measure importance to the generator. To train context RE, we propose an unsupervised learning method, which does not utilize any labeled document ranking data to train the context RE. To examine the efficacy of RE-RAG, we examine its performance on Natural Questions and TriviaQA datasets. RE-RAG achieves on-par performance compared to the FiD variants while utilizing fewer contexts (0.25x). We show that the proposed context RE, which was trained with the T5 model, is also applicable to RAG with LLMs(ChatGPT) by improving the performance on NQ (+6.4EM) and TQA (+2.8EM), respecitvely. Lastly, we display that RE can add interpretability to RAG framework as RE score highly correlates with the RE-RAG accuracy. Consequently, RE can be utilized to filter out unanswerable scenarios where context does not contain answers with 38.9%-51.3% accuracy just by examining a set of retrieved contexts.

1 Introduction

In recent years, the retrieval augmented generation (RAG) framework has shown promising progress in natural language generation, specifically on knowledge-intensive tasks (Lewis et al., 2020b). This approach enhances the model's faithfulness and reliability by leveraging not only limited parametric memory but also additional non-parametric knowledge (Lewis et al., 2020b; Luo et al., 2023). In particular, Open-domain question answering (ODQA) is a knowledge-intensive question-answering task where the task requires a model to provide an answer based on factual information when no specified context is provided along with the question. The RAG framework has shown great success in the ODQA problem and motivated many new research endeavors under the RAG framework.

One of the most prominent models among the RAG variants is the Fusion-in-Decoder (FiD) (Izacard and Grave, 2021b) which showed generally higher performance than original RAG. In turn, since the proposal of FiD, most of the research endeavors focused on the FiD architecture (Izacard and Grave, 2021a; Jiang et al., 2022; Fajcik et al., 2021; Asai et al., 2022). However, despite its superior performance, FiD is hard to interpret as it relies on hundreds of documents through soft cross-attention. Futhermore, FiD is incompatible with black-box LLMs such as GPT and PaLM (Brown et al. (2020), Chowdhery et al. (2023)) that do not disclose their parameters.

On the other hand, RAG, although generally lower-performing than FiD, is more interpretable because it generates answers by each context and marginalizes them. Additionally, RAG decoding steps, which marginalizes answers obtained by providing the generator with integrated questions and contexts, can be easily integrated with recent LLMs (Yang et al. (2023), Xu et al. (2023)). For these reasons, this work revisits the classic RAG framework to leverage its advantages, interpretability and applicability to LLMs, with a modification that can uplift the performance at the level of FiD.

In this work, we propose the RE-RAG framework, which extends the RAG with a context relevance estimator (RE) that re-ranks the retrieved context and provides a precise relevance measure. Our RE-RAG framework retains the interpretable decoding struc-

ture of the existing RAG, while achieving higher performance through a context relevance estimator.

The main contributions of our work are as follows:

1. We propose a new framework **RE-RAG** by expanding RAG with a **Relevance Estimator** (RE). We further suggest a training method that can train RE without any labeled data on question-context compatibility.

2. We demonstrate that RE-RAG, enhanced with RE, significantly improves upon the existing RAG and achieves performance on par with FiD that utilizes many more (4x) contexts.

3. We show that RE can improve LLMs such as GPT, even when it was trained on a much smaller language model.

4. We explore methods to filter out low-relevance context in advance by having the RE pre-evaluate the context set before passing it to the generator.

2 Method

In this section, after reviewing the basic RAG framework, we present the RE-RAG model combined with our relevance estimator.

2.1 Basic RAG overview

Retriever Retriever searches for information in an external knowledge base and returns a related context set $\mathbf{C}_i$. In general, RAG systems use a bi-encoder type retriever such as DPR (Karpukhin et al., 2020), which is effective and fast in retrieving information. A question $\mathbf{q}_i \in \mathbf{Q}$ and a context $\mathbf{c}_j \in \mathbf{C}_i$ are input to the encoder independently to obtain an embedding of $\mathbf{Emb}_q = Encoder(q_i)$, $\mathbf{Emb}_c = Encoder(c_j)$. The similarity score $\mathbf{S}_{i,j} = \mathbf{Emb}_q \cdot \mathbf{Emb}_c$ is calculated from the obtained embedding and then used to perform top-k context retrieval.

Generator Generators that utilize the sequence-to-sequence model typically take a question and context as input and produce an answer $\mathbf{y}_{i,j}$ with probability $\mathbf{P}_G(\mathbf{y}_{i,j}|\mathbf{q}_i, \mathbf{c}_j)$.

Answer marginalization RAG (Lewis et al., 2020b) introduced the answer generation models of RAG-sequence and RAG-token. We focus on the RAG-sequence model which marginalizes probability of $y_l \in \mathcal{Y}_i$ where $\mathcal{Y}_i$ is an aggregated set of $\mathbf{y}_{i,j}$. which achieves higher performance than the RAG-token model and ensures the interpretability of the answer generation process. Individually generated answers $\mathbf{y}_{i,j}$ per $\mathbf{c}_j$ are marginalized as $\mathbf{y}_l$ using the similarity score $\mathbf{S}_{i,j}$ as shown in eq.(2).

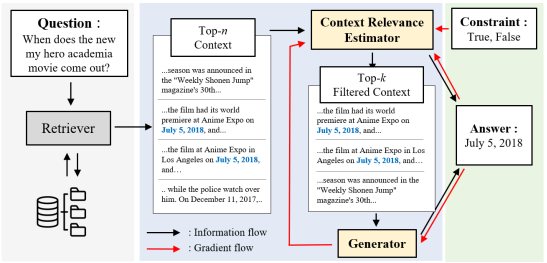

Figure 1: Overview of our proposed RE-RAG framework. The black lines represent the flow of information and the red lines represent the flow of gradients.

$$\mathbf{P}_R(\mathbf{S}_{i,j}) = \frac{e^{\mathbf{S}_{i,j}}}{\sum_k e^{\mathbf{S}_{i,k}}} \quad (1)$$

$$\mathbf{P}_a(\mathbf{y}_l|\mathbf{q}_i, \mathbf{C}_i) = \sum_j \mathbf{P}_R(\mathbf{S}_{i,j}) \cdot \mathbf{P}_G(\mathbf{y}_l|\mathbf{q}_i, \mathbf{c}_j) \quad (2)$$

2.2 RE-RAG framework

The retriever similarity score $\mathbf{S}_{i,j}$ is trained to achieve high recall when retrieving multiple contexts, however, it was not initially designed to provide fine-grained relevancy score $\mathbf{P}_R(\mathbf{S}_{i,j})$ for aiding RAG generation steps in eq.(2). To address this issue, we propose a context relevance estimator (RE) that re-ranks contexts and provides precise relevance scores to the generator.

Relevance Estimator Context relevance estimator (RE) measures the relevance between a question and context. We utilize a similar architecture to Nogueira et al. (2020) which utilizes a sequence-to-sequence model as a passage reranker.

Our context RE receives the same input of question and context as the generator, but is trained to generate a **classification token** ("true" or "false") based on the relevance of the context to the input question. We normalize the probability of generating "true" and "false" tokens to get the final probability of generating the classification token. The obtained probability of a "true" token can independently be an indicator of the relevance of a single context to a given question. When comparing between multiple contexts, the "true" token probability can be converted to logit and used as the relevance score of the retrieved context.

$$\mathbf{RE}_{i,j} = \frac{\mathbf{P}(\text{“}true\text{”}|\mathbf{q}_i, \mathbf{c}_j)}{\mathbf{P}(\text{“}true\text{”}|\mathbf{q}_i, \mathbf{c}_j) + \mathbf{P}(\text{“}false\text{”}|\mathbf{q}_i, \mathbf{c}_j)} \quad (3)$$

Reranking of contexts by relevance With the trained relevance estimator RE, we can rerank contexts in the initial retrieved set $\mathbf{C}_i$ by their relevance and only take top-k contexts to redefine $\mathbf{C}_i$ before the answer-generation step. With a precise relevance score from RE, we can expect the RE-RAG to be more efficient, i.e. stronger performance with lower computation (see §4.2).

Answer marginalization with context RE The question and context are concatenated and input to the generator model, and the generator generates $\mathbf{P}_G(\mathbf{y}_{i,j}|\mathbf{q}_i, \mathbf{c}_j)$ per question. We replace the probability distribution $\mathbf{P}_R(\mathbf{S}_{i,j})$ in eq.(2) with the relevance scores from context RE to form eq.(6) as following:

$$\sigma(\mathbf{RE}_{i,j}) = \log\left(\frac{\mathbf{RE}_{i,j}}{1 - \mathbf{RE}_{i,j}}\right) \quad (4)$$

$$\mathbf{P}_{\mathbf{RE}}(\mathbf{q}_i, \mathbf{c}_j) = \frac{e^{\sigma(\mathbf{RE}_{i,j})}}{\sum_k e^{\sigma(\mathbf{RE}_{i,k})}} \quad (5)$$

$$\mathbf{P}_a(\mathbf{Y}_i|\mathbf{q}_i, \mathbf{C}_i) = \sum_j \mathbf{P}_{\mathbf{RE}}(\mathbf{q}_i, \mathbf{c}_j) \cdot \mathbf{P}_G(\mathbf{y}_{i,j}|\mathbf{q}_i, \mathbf{c}_j). \quad (6)$$

We can expect higher performance with the marginalized answer $\mathbf{y}_l$ if RE can provide an accurate relevance distribution $\mathbf{P}_{\mathbf{RE}}$ (see §5.5).

2.3 Joint training of RE-RAG

We propose to utilize three different types of losses to train RE-RAG with our proposed relevance estimator. First, to train the generator model, we use a loss that combines the commonly used negative likelihood loss for ground truth $\mathbf{a}_i$ with a probability that represents the relevance of the question and context.

$$\mathbf{L}_{\text{gen}} = -\sum_{i,j} \log\left(\mathbf{P}_{\mathbf{RE}}(\mathbf{q}_i, \mathbf{c}_j) \cdot \mathbf{P}_G(\mathbf{a}_i|\mathbf{q}_i, \mathbf{c}_j)\right) \quad (7)$$

$\mathbf{L}_{\text{gen}}$ simultaneously adjusts the probability of generating the classification token for the relevance estimator while training the generator.

Second, to obtain a learning signal for training the context relevance estimator, we calculate the log-likelihood loss of the generator per retrieved context and compute its distribution across contexts as follows:

$$\mathbf{F}_{i,j} = \log(\mathbf{P}_G(\mathbf{a}_i|\mathbf{q}_i, \mathbf{c}_j)) \quad (8)$$

$$\mathbf{Q}_G(\mathbf{q}_i, \mathbf{c}_j) = \frac{e^{\mathbf{F}_{i,j}}}{\sum_k e^{\mathbf{F}_{i,k}}}. \quad (9)$$

The log-likelihood loss varies depending on whether an answer can be inferred from the input context. Therefore, applying the softmax function to the log-likelihood loss values yields a probability distribution that represents the relevance between the given set of contexts and the question. We do not leverage any labeled data that entails the relevance of questions and contexts.

$\mathbf{Q}_G(\mathbf{q}_i, \mathbf{c}_j)$ represents relative relevance between $\mathbf{q}_i$ and $\mathbf{c}_j$

We calculate the KL-divergence loss between the probability distributions of the generator and the context RE, and use this loss to train the model.

$$\mathbf{L}_{\text{re}} = D_{\text{KL}}(\mathbf{P}_{RE}(\mathbf{q}_i, \mathbf{c}_j)||\mathbf{Q}_G(\mathbf{q}_i, \mathbf{c}_j)) \quad (10)$$

Lastly, in addition to applying a training loss on the probability of generating the classification token, we need to set an additional loss to prevent the context RE from generating tokens other than the classification token. To do this, we utilize the additional loss as the sum of the probability of context RE of generating all tokens other than classification token.

$$\mathbf{L}_{\text{tok}} = \sum_{t \in T \setminus \{\text{"true","false"}\}} \mathbf{P}(t|\mathbf{q}_i, \mathbf{c}_k) \quad (11)$$

To train an effective system, the two models are trained jointly utilizing all three losses as follows:

$$\mathbf{L}_{\text{tot}} = \mathbf{L}_{\text{gen}} + \alpha_1 \mathbf{L}_{\text{re}} + \alpha_2 \mathbf{L}_{\text{tok}} \quad (12)$$

where α_1 and α_2 are hyperparameters that act as scaling factors to balance the impact of each loss.

3 Experimental Setup

We evaluated the performance of our model on an open-domain QA dataset. In this section, we describe the dataset we used in our experiments and the details of our experiments.

Model	Extra	Generator	NQ	TQA	# Contexts
RAG (Lewis et al., 2020b)	-	445M	44.5	56.8	50
FiD$_{base}$ (Izacard and Grave, 2021b)	-	220M	48.2	65.0	100
FiD$_{large}$ (Izacard and Grave, 2021b)	-	770M	51.4	67.6	100
FiD-KD$_{base}$ (Izacard and Grave, 2021a)	-	220M	50.1	69.3	100
FiD-KD$_{large}$ (Izacard and Grave, 2021a)	-	770M	54.4	**72.5**	100
ReAtt (Jiang et al., 2022)	-	770M	**54.7**	-	100
FiD-KD$_{base}$ (Izacard and Grave, 2021a)	-	220M	48.6	67.4	25
FiD-KD$_{large}$ (Izacard and Grave, 2021a)	-	770M	53.9	**71.2**	25
R2-D2 (Fajcik et al., 2021)	125M	1.04B	**55.9**	69.9	25
RE-RAG$_{base}$	223M	223M	49.9	68.2	25
RE-RAG$_{mixed}$	770M	223M	51.4	69.5	25
RE-RAG$_{large}$	770M	770M	54.0	70.2	25

Table 1: EM scores on Natural Questions and TriviaQA datasets. The parameters of the generator and the extra module that evaluates a given context are listed separately. # Contexts refers to the number of contexts utilized for inference. We divided the groups based on the number of contexts utilized for inference to enable an effective comparison. For FiD-KD, we used model[1] to calculate the score when utilizing 25 contexts. The bold is the best score in each group, and the underline is the second best.

3.1 Dataset

We evaluate our performance on two open-domain QA datasets:Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017). To train and evaluate our model, we utilize the context datasets retrieved for each question from NQ and TQA, as used in FiD-KD (Izacard and Grave, 2021a) and Akari (Asai et al., 2022). The dataset includes the top 20 training contexts, while the dev and test sets contain the top 100 contexts retrieved by the retriever. We used 20 contexts for training and the top-25 contexts extracted by the context RE from the top-100 retrieved contexts for inference.

Natural Questions Natural Questions (Kwiatkowski et al., 2019) is a dataset of real questions asked by users on the web. The dataset consists of questions collected from the web, a long answer that can be viewed as gold context for the question, and a short answer with a short span. The open-domain QA version dataset of Natural Questions is a dataset that collects only questions where the answer span of the short answer is 5 tokens or less in length. We use the NQ-open dataset.

TriviaQA TriviaQA (Joshi et al., 2017) is a dataset of question-answer pairs collected from trivia enthusiasts. Each question and answer in the dataset has been reviewed by human annotators. We want to use the unfiltered version of TriviaQA dataset.

3.2 Evaluation Metric

The predicted answers are evaluated using **EM score**, a commonly used metric as in Izacard and Grave (2021b), Rajpurkar et al. (2016). The generated answers are normalized (e.g., lowercase, punctuation, article stripping) and compared to the correct answers in the dataset. We consider a generated answer to be correct if it exactly matches one of the correct answers in the given dataset after normalization.

3.3 Baseline

We investigate whether the performance of RE-RAG is competitive with that of the FiD (Izacard and Grave, 2021b)-based system. FiD has achieved excellent performance on the Question-Answering task, and the FiD-based application system also outperforms the RAG (Lewis et al., 2020b)-based system on the QA task. Most of the models under comparison involve additional training of the retriever (Izacard and Grave, 2021a) or improvements to the retrieve system (Jiang et al., 2022). Therefore, we use the FiD-KD improved retriever's dataset for baseline comparisons with other models.

3.4 Model

The two components of our framework, context RE and the generator, utilize the T5 model (Raffel et al., 2020). We utilize the T5-base, T5-large models, and explore three different model sizes depending on the combination of the two models.

4 Experiment Results

We investigate the QA performance of the RAG system with our newly proposed context relevance estimator (RE). In addition to the QA performance of the whole system, we also examine the performance of the context RE independently.

Dataset	Model	Recall@k			
		R@1	R@5	R@10	R@20
NQ	FiD-KD	49.4	73.8	79.6	84.3
	RE-RAG$_{base}$	59.5	77.8	82.7	85.5
	RE-RAG$_{large}$	61.9	79.4	83.6	86.4
TQA	FiD-KD	60.1	77.0	80.9	83.6
	RE-RAG$_{base}$	67.0	81.5	83.6	85.4
	RE-RAG$_{large}$	70.4	82.2	84.4	86.1

Table 2: Performance of RE as a re-ranker. The table displays recall@k ranked by the FiD-KD retriever and context RE, out of the top-100 contexts from the FiD-KD retriever. The recall@k performance for these top-100 contexts is 89.3 on NQ and 87.7 on TQA.

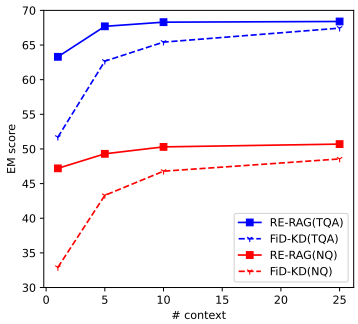

Figure 2: Performance of RE-RAG$_{base}$ and FiD-KD$_{base}$ as a function of the number of input contexts.

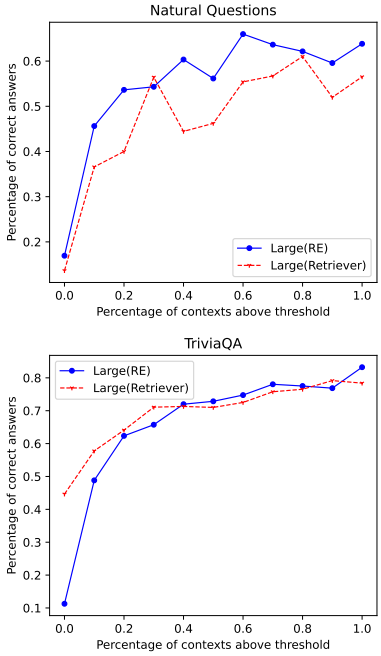

Figure 3: The figure shows the relationship between the quality of the context set used and the accuracy level of the model. The quality of a context set is expressed as the percentage of contexts with context REscores above the threshold. See Appendix D for similar analysis for the RE-RAG base model.

4.1 Main Results

The overall accuracy of our system on the two datasets we evaluated, NQ and TQA, is shown in Table 1. Compared to basic RAG, our system RE-RAG shows better performance despite having the same number of total parameters. Our proposed context relevance estimator (RE) leverages the RAG system while providing a more accurate measure of the relevance between question and context, improving the overall reliability of the system. Despite using top-20 contexts as training dataset and top-25 contexts for inference, our model shows competitive performance compared to FiD-KD (Izacard and Grave, 2021a)and ReAtt (Jiang et al., 2022), which used top-100 contexts for training and inference. The context RE enhances the system's performance, independent of the generator model's parameter size.

4.2 Effect of using less document

Table 2 shows the performance of our proposed RE-RAG's context RE as a reranker. Table 2 presents the performance of re-ranking using retriever's similarity score and context RE's relevance score in the top-100 context of retriever. For the Recall@k metric, we use the retrieval accuracy used by DPR (Karpukhin et al., 2020), FiD-KD (Izacard and Grave, 2021a), and ColbertQA (Khattab et al., 2021). Although the comparison retriever has been enhanced through knowledge distillation methods using FiD attention scores, our proposed context RE still demonstrated superior performance. In particular, the performance improvement of the context RE over the retriever becomes more pronounced as the number of contexts decreases. This suggests that the proposed context RE is more effective when the number of contexts that can be fed into the generator is limited.

We examined how QA performance changes when inferring answers using fewer documents. Figure 1 shows that our proposed system performs more robustly when the number of available contexts decreases compared to FiD-KD. In particular, the performance degradation is limited for both models up to 10 contexts, but the difference increases when the number of utilized contexts decreases dramatically.

[1] https://github.com/facebookresearch/FiD

Question	Context	Gold Answer	"True" prob
who played mark on the show the rifleman	...**Mark McCain is the son of fictitious rancher Lucas McCain in the ABC Western television series "The Rifleman,"** starring Chuck Connors, which ran from 1958 to 1963. Singer/actor and former Mouseketeer **Johnny Crawford was cast in the role** and...	John Ernest Crawford	0.987
when does the cannes film festival take place	...2017 **Cannes Film Festival The 70th Cannes Film Festival took place from 17 to 28 May 2017**, in Cannes, France ...	Cannes, France, usually in May	0.994
how many strong verbs are there in german	...**Germanic strong verbs are commonly divided into 7 classes**, based on the type of vowel alternation. This is in turn based mostly...	more than 200, more than 200 strong	0.949
how many episodes of corrie has there been	...The show airs six times a week: Monday, Wednesday and Friday 7:30-8 pm and 8:30-9 pm. Since 2017, **ten sequential classic episodes** of the series from 1986...	9,436	0.147

Table 3: The relevance measure of the question and context output by the context RE. The first two show relevant contexts that contain the correct answer even if the context does not include exactly the same surface form compared to the true answer. The last two examples show irrelevant contexts that actually have high overlap with question tokens, however, without pertaining the correct answer.

5 Analysis

5.1 Relationship between context relevance quality and answer confidence

We examine how the quality of the context set input to the generator relates to the confidence in the answer generation. We set the threshold for the generation probability of "true" tokens to 0.5, meaning that the generation probability of "true" tokens produced by the context RE is greater than that of "false" tokens. Then, we classify the input text as having high relevance quality if the context RE's probability of generating a "true" token exceeds the threshold. For comparison to the baseline retriever, we use the cosine similarity of the hidden representation between the question and context in the retriever.

Figure 2 shows how the accuracy of answers varies with the proportion of high relevance quality among the top-25 input contexts. In both datasets, accuracy is increasing with the percentage of high relevance quality as measured by the context RE. In particular, accuracy decreases significantly in the absence of high relevance quality context. For the baseline retriever, we notice different behavior in two datasets. The baseline shows higher-than-expected performance for contexts with low relevance in TQA and lower-than-expected accuracy for high relevance contexts. This shows that the context RE can estimate the reliability of the final answer that will be generated by the system by measuring the context RE for the retrieved context in advance while baseline cannot.

5.2 Effectiveness of the context RE

We perform a qualitative analysis to see if our proposed context relevance estimator (RE) is effectively classifying relevant contexts. Table 3 shows a few contexts in the NQ test set.

Some of the contexts that the context RE predicts are highly relevant to the question even when they do not contain the exact ground truth answer. The first few examples in Table 3 are examples that are categorized as true context because they contain phrases that are semantically equivalent to the correct answer albeit not having the exact same form in the context. This shows that although the context RE is trained to measure the relevance of a question to a context through a limited set of ground truth answers, it is actually capable of measuring a broader range of relevance.

In addition to the examples above, there are cases where the context RE misclassified contexts as containing the correct answer. As shown in the example in Table 3, the context RE classified the context containing "the number of classes of strong verbs in German" as the correct context for the question about "the number of strong verbs in German", which means that our context RE is still limited in its ability to capture the fine-grained meaning of the question in the retrieved context. On the other hand, in the last example, for the question about "the number of episodes", it succeeded in classifying the context containing "the number of classical episodes" as an incorrect context.

5.3 RE for classifying "irrelevant" context set

Table 4 shows the performance of the context relevance estimator (RE) as a "irrelevant" set classifier.

Dataset	Model	Recall	Precision	F1
NQ	FiD-KD	73.2	21.9	33.7
	RE-RAG$_{base}$	51.3	33.9	40.9
	RE-RAG$_{large}$	45.9	38.3	41.7
TQA	FiD-KD	64.3	24.5	35.5
	RE-RAG$_{base}$	38.9	46.7	42.5
	RE-RAG$_{large}$	39.0	43.2	41.0

Table 4: Classification results for context sets that do not contain an answer within the top-25 context set. We used cosine similarity for FiD-KD's retriever and "true" token probability for our method. The threshold of each model was varied from 0.5 to 0.9 in increments of 0.1 to find the optimal value.

"irrelevant" set means that the context set of the top-25 contexts input to the generator does not contain an answer in any context. For classification, we used the cosine similarity score of the hidden representation of the question and context for retriever and the probability of generating a "true" token by the model for context RE. For the optimal threshold, we searched for the value that maximizes F1 score in steps of 0.1 from 0.5 to 0.9.

Our context RE showed better "irrelevant" set classification performance than FiD-KD's improved retriever based on F1 score. Looking at the detailed performance, we found that the retriever performed better for recall, but the context RE performed better for precision. This is because the retriever classified a large number of context sets as all "irrelevant" sets, while our proposed context RE showed a good balance between classification precision and recall.

Table 5 shows the accuracy changes after making the model respond with "unanswerable" to "irrelevant" sets, divided into context sets with and without answers. Model set to respond "unanswerable" if no context exceeds threshold set in Table 4. The accuracy in sets where answers can be found slightly decreases due to incorrect "unanswerable" responses. Conversely, in sets where answers cannot be found, accuracy increases (from 0) by responding with "unanswerable" to "irrelevant" sets, thereby improving accuracy in cases where answers are unattainable.

5.4 Plugging the context RE into the LLM

We investigate whether the context relevance estimator (RE), which has been effective in relatively small-sized models, is also effective in improving LLM's performance. We follow the method proposed in REPLUG (Shi et al., 2023) to marginalize the answers generated by the LLM according to the input contexts. We use OpenAI "gpt-3.5-turbo-0125" (Brown et al., 2020), which generates answers using the top-k contexts evaluated by our context RE and evaluated by FiD-KD's retriever. The score for each context was calculated using the "true" token logit from the context RE and the cosine similarity of the hidden representation of the question and context generated by the retriever. We used 8-shot prompts for NQ and 2-shot prompts for TQA. The detailed prompts are shown in Appendix C.

In both datasets, our proposed context RE outperforms the baseline retriever on both datasets. The difference in EM score was 6.4 for NQ and 2.8 for TQA. This indicates that context RE's improved ability to re-rank contexts and its ability to calculate more accurate relevance scores can improve overall answer quality by simply combining context RE with LLM's RAG system.

Dataset	Model	relevant context set	
		O	**X**
NQ	RE-RAG$_{base}$	58.3 → 54.9	51.3
	RE-RAG$_{large}$	61.5 → 57.9	45.9
TQA	RE-RAG$_{base}$	78.7 → 77.0	38.9
	RE-RAG$_{large}$	80.4 → 77.9	39.0

Table 5: We examine whether RE can successfully identify unanswerable scenarios where retrieved contexts do not hold true answers. **O** refers to the retrieval context set that contains true answers and **X** refers to the set without which we dim as *unanswerable*. Under the **X**, we display the accuracy of RE-thresholding in classifying unanswerable instances. Under the **O**, we denote the accuracy change as the RE thresholding will inevitably classify the context sets with answers as unanswerable. Left of the arrow denotes original accuracy on **O** and the right denotes accuracy after RE score thresholding.

Model	top-5		top-10	
	NQ	TQA	NQ	TQA
GPT w/ Retriever	41.7	67.3	42.9	69.0
GPT w/ RE	48.8	70.7	49.3	71.8

Table 6: RAG on GPT-3.5 model. The table displays EM score on Natural Question and TriviaQA for using FiD-KD retriever alone (w/ Retriever) and with the addition of the context RE (w/ RE).

5.5 Ablation Study

We perform an ablation study to investigate the effectiveness of the added context RE in RE-RAG. The

Model	NQ	TQA
Baseline	39.5	54.9
Baseline w/ RE score	43.1	60.1
Baseline w/ RE context	46.8	63.9
Baseline w/ RE context, score	49.6	67.8
RE-RAG$_{base}$	49.9	68.2

Table 7: An ablation study to decompose the effect of RE in RE-RAG. We compared the basic RAG model without RE, with reranking of context RE(RE context), with RE score in answer generation (RE score), and with both (RE context, score).

effect of our proposed context RE is twofold. First, it performs better re-ranking than the retriever, selecting more accurate context and passing it to the generator. Second, it calculates a more accurate relevance score than retriever's similarity score and uses it in the answer marginalization process. In Table 7, the performance of methods with each component of the context RE added is presented, using a model that was trained with only the T5-base generator, after removing the context RE, as the baseline.

We construct the following experiment to isolate the two effects. First, we apply the top 25 contexts from retriever and their similarity scores to the baseline model. Next, there are the top-25 contexts from the retriever with the context RE's score applied (RE score) and the top-25 contexts from the context RE with the retriever's similarity score applied (RE context). Finally, we compare the performance of applying the context RE's top-25 contexts and score to the baseline model (RE context, score).

Both effects of the context RE are found to be significant in improving the performance of the final model. This shows that not only the quality of the context input to the generator plays an important role, but also the score, which means the importance of each context. The effect of context RE is 7.3 EM for NQ and 9.0 EM for TQA. The impact of Context score is 3.6 EM for NQ and 5.2 EM for TQA.

6 Related Works

Previous research has shown that the performance of Question Answering systems can be improved by utilizing external knowledge about questions (Chen et al., 2017). Methods for more accurate retrieval of external knowledge (Karpukhin et al. (2020); Khattab et al. (2021); Gao and Callan (2022)) have been studied to make these systems more efficient. In open-domain QA, models that extract and use answers from retrieved documents have been studied (Karpukhin et al. (2020); Khattab et al. (2021); Cheng et al. (2021)), but studies that utilize generative models such as T5 (Raffel et al., 2020) or BART (Lewis et al., 2020a) have become more common (Lewis et al. (2020b); Izacard and Grave (2021b)). RAG and FiD achieved state-of-the-art performance in open-domain QA using different methods. Subsequently, models (Izacard and Grave (2021a); Jiang et al. (2022); Fajcik et al. (2021)) that leverage and improve upon the structural advantages of FiD have been proposed. For Atlas (Izacard et al., 2022), state-of-the-art performance was achieved through an improved retriever (Izacard et al., 2021) and scaling up the model. In the case of RAG, there is a study that improved performance by introducing a BERT (Devlin et al., 2019)-based reranker (Glass et al., 2022), but it utilized additional data and high-quality label data when training the reranker. Recently, LLMs such as GPT (Brown et al., 2020) and Llama2 (Touvron et al., 2023), which have been developed in recent years, face limitations with FiD methods that require encoded data. Consequently, research on RAG models, which can directly input context, has received renewed attention. (Shi et al. (2023); Asai et al. (2023); Lin et al. (2023)) These approaches have achieved performance improvements by training a retriever, which can also be applied to LLM, or by performing the review of questions and context within the model itself.

7 Conclusion

We propose RE-RAG, which adds a context RE to the RAG system to evaluate the relevance between question and context. We show that our proposed RE-RAG achieves competitive performance on NQ and TQA datasets, and that the context RE can be combined into LLM independently to improve performance. Furthermore, RE-RAG is relatively easy to train as it does not utilize label data such as whether the context contains an answer or not to train the context RE.

8 Limitation

Our research has explored how to improve performance and interpretability by evaluating the relevance of questions and context in a original RAG model. Our work has focused on RAG systems,

with limited exploration of how our methods can improve FiD-based systems. We believe that such research could be conducted in the future.

References

Akari Asai, Matt Gardner, and Hannaneh Hajishirzi. 2022. Evidentiality-guided generation for knowledge-intensive nlp tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2226–2243.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2021. UnitedQA: A hybrid approach for open domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3080–3090, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Martin Fajcik, Martin Docekal, Karel Ondrej, and Pavel Smrz. 2021. R2-d2: A modular baseline for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 854–870.

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

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*.

Gautier Izacard and Edouard Grave. 2021a. Distilling knowledge from reader to retriever for question answering. In *ICLR 2021-9th International Conference on Learning Representations*.

Gautier Izacard and Edouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 874–880. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Zhengbao Jiang, Luyu Gao, Zhiruo Wang, Jun Araki, Haibo Ding, Jamie Callan, and Graham Neubig. 2022. Retrieval as attention: End-to-end learning of retrieval and reading within a single transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2336–2349.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided supervision for OpenQA with ColBERT. *Transactions of the Association for Computational Linguistics*, 9:929–944.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Augmented large language models with parametric knowledge guiding. *arXiv preprint arXiv:2305.04757*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

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*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023. PRCA: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5364–5375, Singapore. Association for Computational Linguistics.

A Dataset Statistics

Table 8 shows the statistics for the Natural Questions and TriviaQA unfilitated datasets we used.

Dataset	Train	Dev	Test
Natural Questions	79,168	8,757	3,610
TriviaQA	78,785	8,837	11,313

Table 8: Dataset statistics for Natural Questions and TriviaQA

B Training Details

We used T5-base with a parameter size of 223M and T5-large model with a parameter size of 770M as modulators in all experiments. We trained the RE-RAG$_{base}$ system on 4 A6000 GPUs, while RE-RAG$_{mixed}$ and RE-RAG$_{large}$ were trained on 2 A100 and 4 A100 GPUs, respectively.

We used a constant learning rate of 10^-4 for all sizes of RE-RAG systems. We used AdamW as the optimizer and weight decay was 10^-3. For batch size, we used gradient accumulation for all sizes of models, resulting in an effective batch size of 64. For the hyperparameters that balance the proposed losses, we utilized the default value of 1 for both α_1 and α_2. We did not explore hyperparameters that achieve better performance due to time and limited computing resources.

For model selection, we evaluated every 1 epoch and selected the case with the highest answer accuracy of the dev set. The dev set answer accuracy

was measured using the top-10 context of the context RE. Since the answer accuracy of the top-10 context of the context RE is similar to the answer accuracy of the top-25 context, this helped to save computational resources and time while still producing valid results.

C Prompts utilized in GPT-3.5

Table 9 and Table 10 show the prompts provided in GPT-3.5 used in our experiments. We provided an 8-shot example for Natural Questions and a 2-shot example for TriviaQA. We performed a simple normalize on the sequence generated by the GPT model with the following prompts, and then treated as correct the answer that exactly matched the answer in the dataset.

D Relationship between context relevance quality and answer confidence at base model

Figure 3 illustrates the relationship between context relevance quality and answer confidence described in Section 5.1 in RE-RAG$_{base}$. The overall trend is not significantly different from the large model, but we can see that the difference in accuracy is caused by the difference in the number of parameters in the generator model.

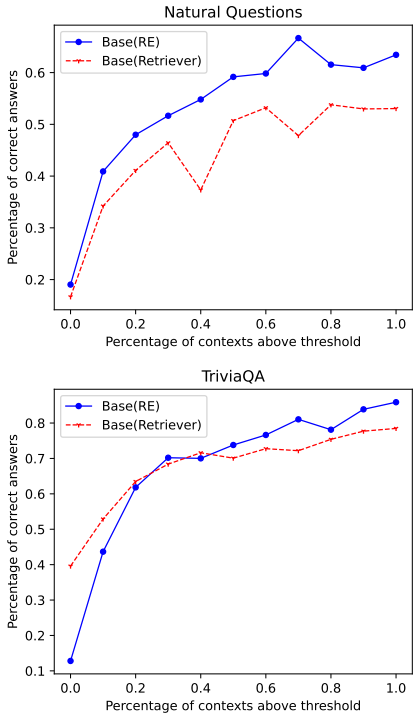

Figure 4: The relationship between the quality of the context set used, as measured by the context RE, and the correctness rate of the base model.

###Description : Below are some examples of question and answer formats. Use these examples as a guide to help you come up with the right answer to the question you'll eventually be asked.

Example 1

Context: 'title': Service club, 'text': "the services", a common expression for the military or uniformed forces. In the Americas, these types of clubs are commonly known as veteransórganizations or veteransfraternal groups. The worldś first service club, the Rotary Club of Chicago, was formed in 1905 by Paul P. Harris, an attorney who wanted to create in a professional club with the same friendly spirit he had felt in the small towns of his youth. The Rotary name derived from the early practice of rotating meetings among membersóffices. Many of these service clubs were started early in the 20th century, such as Kiwanis,
Question: In which city were Rotary Clubs set up in 1905?
Answer: Chicago

Example 2

Context: 'title': Jason Schwartzman, 'text': he played a writer who moonlights as an unlicensed private detective by advertising himself on Craigslist. He currently releases music through his solo project Coconut Records, and was formerly the drummer of rock band Phantom Planet. Schwartzman was born in Los Angeles, California, the son of actress Talia Shire (née Coppola) and the late producer Jack Schwartzman. Schwartzman's brother is actor and musician Robert Schwartzman, and his paternal half-siblings are Stephanie and cinematographer John Schwartzman. Many other members of Schwartzman's family are involved in film: he is the nephew of Francis Ford Coppola, cousin of Nicolas Cage, Sofia Coppola, Roman
Question: Which famous brother of Talia Shire does not share her last name?
Answer: Francis Ford Coppola

###Instructions: Provide the correct answer to the given question. The given question is accompanied by context related to the question. Be sure to refer to the context provided and enter your answer based on the context after "Answer:". The question and the answer you provide must be relevant. Write your answer in a short "short answer".

###Context: {context}

###Question: {question}

###Answer:

Table 9: Example GPT-3.5 serving prompt for TriviaQA dataset

###Description : Below are some examples of question and answer formats. Use these examples as a guide to help you come up with the right answer to the question you'll eventually be asked.

- Example 1

Context: 'title': Sports in the United States, 'text': Erving (won MVP awards in both the ABA and NBA), Kareem Abdul-Jabbar (6 time MVP), Magic Johnson (3 time MVP), Larry Bird (3 time MVP), Michael Jordan (6 time finals MVP), John Stockton (1 in career assists and steals), Karl Malone (14 time all NBA team), Kobe Bryant (NBA's third all-time leading scorer), Tim Duncan (15-time NBA all-star), Shaquille O'Neal (3 time finals MVP) and Jason Kidd (2 in career assists and steals). Notable players in the NBA today include LeBron James (4 MVP awards), Stephen Curry (2 time MVP), Dwyane Wade (10 time all-star), and Kevin Durant (MVP, 4
Question: who are the top 5 leading scorers in nba histor
Answer: Kobe Bryant

Example 2

Context: 'title': My Hero Academia: Two Heroes, 'text': would be joining the cast as Melissa Shield and Katsuhisa Namase would play David Shield, both original characters. On June 11, 2018, "Weekly Shōnen Jump" announced that Rikiya Koyama had been cast as the filmś villain, Wolfram. Masaki Suda performs the filmś theme song , which was written and composed by Hiromu Akita of amazarashi. Funimation and Toho premiered the film at Anime Expo in Los Angeles on July 5, 2018, and it was later released in Japan on August 3 of that year. The first one million audience members to see the movie will receive a special book containing
Question: when does the new my hero academia movie come out
Answer: July 5, 2018

...

Example 8

Context: 'title': King Kong, 'text': and the subsequent appeal. Since the court case, Universal still retains the majority of the character rights. In 1986 they opened a King Kong ride called "King Kong Encounter" at their Universal Studios Tour theme park in Hollywood (which was destroyed in 2008 by a backlot fire), and followed it up with the Kongfrontation ride at their Orlando park in 1990 (which was closed down in 2002 due to maintenance issues). They also finally made a King Kong film of their own, "King Kong" (2005). In the summer of 2010, Universal opened a new 3D King Kong ride called at
Question: when did the king kong ride burn down
Answer: 2008

###Instructions: Provide the correct answer to the given question. The given question is accompanied by context related to the question. Be sure to refer to the context provided and enter your answer based on the context after "Answer:". The question and the answer you provide must be relevant. Write your answer in a short "short answer" of 5 words or less.

###Context: {context}

###Question: {question}

###Answer:

Table 10: Example GPT-3.5 serving prompt for Natural Questions dataset