REAR: A Relevance-Aware Retrieval-Augmented Framework for Open-Domain Question Answering

Yuhao Wang[1*] **Ruiyang Ren**[1*] **Junyi Li**[3] **Wayne Xin Zhao**[1†]
Jing Liu[4†] **Ji-Rong Wen**[1,2]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Information, Renmin University of China
[3]Department of Computer Science, National University of Singapore [4]Baidu Inc.
{yh.wang, reyon.ren}@ruc.edu.cn, batmanfly@gmail.com

Abstract

Considering the limited internal parametric knowledge, retrieval-augmented generation (RAG) has been widely used to extend the knowledge scope of large language models (LLMs). Despite the extensive efforts on RAG research, in existing methods, LLMs cannot precisely assess the relevance of retrieved documents, thus likely leading to misleading or even incorrect utilization of external knowledge (*i.e.,* retrieved documents). To address this issue, in this paper, we propose **REAR**, a **RE**levance-**A**ware **R**etrieval-augmented approach for open-domain question answering (QA). As the key motivation, we aim to enhance the self-awareness regarding the reliability of external knowledge for LLMs, so as to adaptively utilize external knowledge in RAG systems. Specially, we develop a novel architecture for LLM-based RAG systems, by incorporating a specially designed assessment module that precisely assesses the relevance of retrieved documents. Furthermore, we propose an improved training method based on bi-granularity relevance fusion and noise-resistant training. By combining the improvements in both architecture and training, our proposed REAR can better utilize external knowledge by effectively perceiving the relevance of retrieved documents. Experiments on four open-domain QA tasks show that REAR significantly outperforms previous a number of competitive RAG approaches. Our codes can be accessed at https://github.com/RUCAIBox/REAR.

1 Introduction

Despite the progressive capacities, large language models (LLMs) (Brown et al., 2020; Zhao et al., 2023) still struggle with knowledge-intensive tasks like open-domain question answering (QA), lacking in real-time and domain knowledge (Cheng et al., 2024; Li et al., 2023a). To mitigate this issue, retrieval-augmented generation (RAG) provides LLMs with potentially relevant documents through a retrieval module (Gao et al., 2023), aiding in generating more precise content.

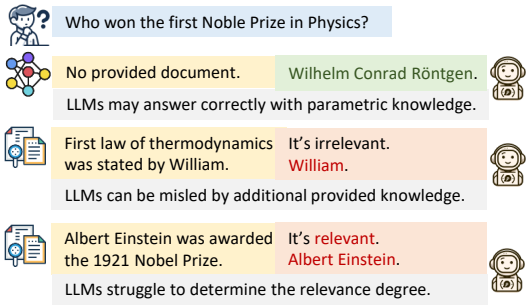

Figure 1: LLMs may be misled by irrelevant documents, and struggle to determine the relevance of a document (Ren et al., 2023; Zhang et al., 2024).

While RAG offers clear benefits, it also introduces several technical challenges for effectively improving LLMs. Firstly, the retrieved results likely contain irrelevant content or documents, which may mislead LLMs and even cause them to respond incorrectly (Mallen et al., 2023; Ren et al., 2023). Moreover, it has become common to incorporate multiple reference documents to boost the overall reliability of retrieved documents. However, this approach potentially amplifies the impact of the noise present in the retrieved documents (Liu et al., 2023; Shi et al., 2023). Thus, LLMs face difficulties in filtering irrelevant documents and integrating their internal knowledge (Dong et al., 2023), which needs to avoid potential interference with noisy content.

Recently, several studies (Asai et al., 2023; Luo et al., 2023; Yoran et al., 2023) have attempted to enhance the robustness of RAG systems. For instance, Self-RAG (Asai et al., 2023) allows the model to introspect its outputs by generating special tokens to discriminate if the documents are relevant, and RobustLM (Yoran et al., 2023) prompts LLMs to first discriminate if the documents are rel-

[*]Equal contributions.
[†]Corresponding authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5613–5626
November 12-16, 2024

evant and then generate answers. However, these approaches perform the assessment of document relevance solely based on binary labels, which are highly sparse and not precise to capture the fine-grained relevance. In addition, they seldom consider the varied relevance degree of reference documents, making the utilization of external knowledge somehow blind.

To this end, in this paper, we propose **REAR**, a **RE**levance-**A**ware **R**etrieval-augmented generation approach for open-domain question answering (QA). Our key idea is to develop robust self-awareness regarding the reliability of external knowledge (*i.e.,* retrieved documents) within RAG systems, so that the LLM can learn to adaptively utilize the internal and external knowledge for solving complex QA tasks. To achieve this goal, we make two major contributions in both model architecture and training. First, we propose relevance-aware RAG architecture by incorporating explicit assessment modules in LLMs' generation architecture to perform an additional relevance assessment task. In our architecture, the assessment module effectively captures relevance signals, and feeds them back to avoid distractions from irrelevant external knowledge during generation. Secondly, to support the relevance-aware RAG architecture, we further propose two training strategies. Bi-granularity relevance fusion strategy further integrates both coarse and fine-grained relevance supervision to overcome the limitations of binary discriminative methods, while noise-resistant training strategy enhances the discrimination ability of the LLM by incorporating negatives in the training procedure.

To the best of our knowledge, we are the first to introduce the idea of incorporating explicit assessment modules in the generation architecture of LLMs to aid in irrelevance-resistant generation. Extensive experiments on public open-domain QA benchmarks attest to the effectiveness of our REAR framework. Notably, we also demonstrate the strong generalization capability of REAR by conducting out-of-domain evaluation on multiple open-domain QA benchmarks.

2 Related Work

Open-domain Question Answering. Modern open-domain QA systems combine traditional IR techniques with neural reading comprehension models (Chen et al., 2017). After retrieving documents (Ren et al., 2021a; Zhang et al., 2021), an extractive or generative reader is typically used for answer generation (Zhu et al., 2021). Models like REALM (Guu et al., 2020), RAG (Lewis et al., 2020), RETRO (Borgeaud et al., 2022) and In-context RALM (Ram et al., 2023) have demonstrated improved factual generation capabilities. However, these readers make generation quality more prone to noise impact, for lacking explicit relevance discernment. We propose an architecture that explicitly generates relevance scores to assist in subsequent generation tasks.

Retrieval-augmented LLMs. Several research aims at aligning the retriever outputs with the preferences of the LLMs (Izacard and Grave, 2021a; Sachan et al., 2021). And works like Atlas (Izacard et al., 2022), RA-DIT (Lin et al., 2023) jointly train the language model and the retriever for advanced performance on RAG. Some other work improves the quality of retrieved documents by expanding the knowledge sources (Li et al., 2023b) or query rewriting (Zheng et al., 2023). However, we focus on a scenario where the irrelevant documents from retrieval could mislead LLMs. Several recent studies (Asai et al., 2023; Luo et al., 2023; Yoran et al., 2023) attempt to adopt a paradigm in which an initial judgment on relevance is made by generating a statement or special token before proceeding to content generation. However, these methods still lack accuracy in relevance discrimination and LLMs are still vulnerable to irrelevant document interference. Therefore, we propose a framework that can accurately assess the relevance degree, and is more robust to irrelevant content.

3 Task Formulation

In this work, we focus on the task of open-domain question answering (QA) (Chen et al., 2017; Zhao et al., 2024), aiming at answering questions using a large collection of documents. Typically, open-domain QA tasks are often tackled with a *retriever-reader* approach (Chen and Yih, 2020), where the retriever finds relevant evidence and the reader generates the answer based on the retrieved evidence.

Formally, given a query q, the retriever outputs top-k documents $\mathcal{D} = \{d_i\}_{i=1}^{k}$ from a document collection (can be refined by an optional *reranker*) at the first stage. Different from prior studies that combine the entire set of retrieved documents as a unified reference for answer generation (Hofstätter et al., 2023; Luo et al., 2023; Xu et al., 2023), our approach emphasizes individual document uti-

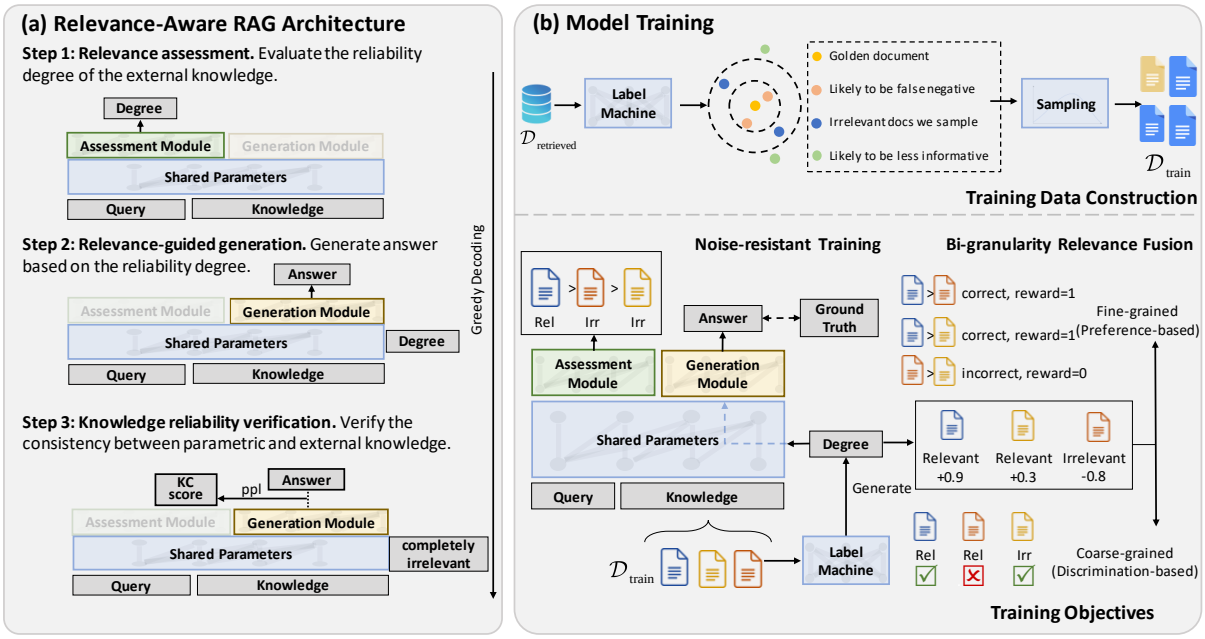

Figure 2: The overview of the proposed REAR framework.

lization, which can be also extended to a multi-document setting. Given the input query q and reference documents $\mathcal{D} = \{d_i\}_{i=1}^{k}$, the reader (*i.e.*, the LLM) generates an answer a_i based on each reference document d_i, forming an answer set $\mathcal{A}$:

$$\mathcal{A} = \{a_i\}_{i=1}^{k} = \{\text{LLM}(q, d_i) \mid d_i \in \mathcal{D}\}. \quad (1)$$

Subsequently, we can choose the final answer from $\mathcal{A}$ based on some specific ways, ensuring it aligns best with the query q.

Based on this task formulation, we consider enhancing two key aspects: precise evaluation of relevance between queries and documents (*identifying relevant references*), and leveraging relevance signal for noise-resistant generation (*reducing the influence of irrelevant content*). Therefore, we introduce a relevance-aware approach designed specifically for these challenges.

4 Methodology

In this section, we present the proposed **RE**levance-**A**ware **R**etrieval-augmented generation framework (**REAR**), which is capable of precisely assessing the relevance degree during the generation process by incorporating explicit assessment modules within the LLM. Furthermore, we propose optimized training methods that are compatible with the REAR framework to support efficient operation, including bi-granularity relevance fusion and noise-resistant training.

4.1 Relevance-Aware RAG Architecture

In this part, we propose a novel architecture that augments the LLM with a relevance-assessing module for enhancing the awareness of irrelevant interference. As shown in Fig. 2 (a), the inference of the REAR architecture encompasses three steps, including relevance assessment, relevance-guided generation, and knowledge reliability verification.

4.1.1 Relevance Assessment

Instead of treating all the retrieved documents equally, we first aim to assess the relevance degrees of the documents. Drawing from the success of LLM-based decoder in achieving precise relevance assessment (Ma et al., 2023; Sun et al., 2023), we first map the input query-document pair into the relevance embedding $\boldsymbol{v}_{\text{rel}}$ by the LLM:

$$\boldsymbol{v}_{\text{rel}} = \text{LLM}(q, d)[-1]. \quad (2)$$

Subsequently, $\boldsymbol{v}_{\text{rel}}$ is quantified into a score s_{rel} by the assessment module:

$$s_{\text{rel}} = \text{Assess}(\boldsymbol{v}_{\text{rel}}), \quad (3)$$

where $\text{Assess}(\cdot)$ is the assessment module implemented as a linear projection layer.

4.1.2 Relevance-guided Generation

Different from previous works that ignore the relevance of document (Cuconasu et al., 2024), we aim to integrate the relevance score of each document

into LLMs to assess document reliabilities and subsequently guide the generation process. Since the relevance score s_{rel} (in Eq. 3) is a scalar, which may not be fully utilized by LLMs, we further incorporate an embedding layer to map it into a dense vector $\boldsymbol{v}_{\text{guide}}$ as:

$$\boldsymbol{v}_{\text{guide}} = \text{Embedding}(s_{\text{rel}}). \quad (4)$$

This embedding vector serves as a cue for the LLM to generate an answer a based on either the internal knowledge of LLM (the relevance score is low) or external evidence (the relevance score is high) as:

$$a = \text{LLM}(q, d, \boldsymbol{v}_{\text{guide}}). \quad (5)$$

4.1.3 Knowledge Reliability Verification

Based on the generated answers, we finally verify the correctness of the answers by considering two factors: (a) Is the provided document reliable enough to trust the corresponding answer? (b) Without referring to the documents, to what degree will the LLM adhere to its original response? Specially, we propose two strategies, namely source-reliability and knowledge-consistency.

• *Source-reliability*: This strategy primarily emphasizes the quality of external knowledge. If an LLM assigns a high relevance score to a document, then the answer derived from it is considered more reliable.

• *Knowledge-consistency* : This approach further verifies if the provided knowledge conflicts with the parametric knowledge. Specifically, inspired by the success of self-consistency in Chain-of-Thought reasoning (Wang et al., 2022; Wei et al., 2022), we inform the LLM that the document is irrelevant by setting the relevance score to zero (denoted by $\hat{s}_{\text{rel}}$) and calculate the inverse of perplexity c (Meister and Cotterell, 2021) of generating the answer a:

$$c = \frac{1}{\text{PPL}(a \mid q, d, \hat{s}_{\text{rel}} = 0)}, \quad (6)$$

which evaluates the extent of LLM to stand by its original answer based on the parametric knowledge. Then, we linearly combine the knowledge-consistency score c_i with the relevance score $s_{\text{rel}}(q, d_i)$ to select the final answer.

4.2 Model Training

In this part, we will introduce the training pipeline for optimizing our approach, As shown in Fig. 2 (b).

4.2.1 Bi-granularity Relevance Fusion

Precise relevance assessment is crucial for the reliable utilization of retrieved documents. Previous work often adopts the coarse-grained binary discrimination task (Yoran et al., 2023) , which cannot provide sufficient evidence for solving complex QA tasks. Therefore, we consider further incorporating a preference-based fine-grained task. Specifically, for the fine-grained supervision, we utilize the estimated relevance scores (See Section 4.2.3) for deriving relevance preference constraints:

$$\mathcal{L}_{\text{fine}} = -\sum_i \sum_j (s_i > s_j) \log\left(\sigma_i - \sigma_j\right), \quad (7)$$

where σ_i denotes the normalized probability of assessing (q, d) as relevant by the LLM. Furthermore, we combine it with the coarse-grained binary loss $\mathcal{L}_{\text{coarse}}$, as the objective of the bi-granularity relevance fusion:

$$\mathcal{L}_{\text{bi-granularity}} = \mathcal{L}_{\text{coarse}} + \mathcal{L}_{\text{fine}}. \quad (8)$$

4.2.2 Noise-resistant Training

In addition to improving the capability of identifying relevant documents, we further consider enhancing the discrimination ability when reference documents contain irrelevant content or even noise, such that the LLM can adaptively use external evidence for task solving. Specially, we further incorporate negative example documents $\mathcal{D}^-$ into the original corpus $\mathcal{D}$ for optimizing LLMs:

$$\mathcal{L}_{\text{noise-resistant}} = \sum_{d \in \mathcal{D} \cup \mathcal{D}^-} \log P(a \mid q, d, s_{\text{rel}}). \quad (9)$$

Through noise-resistant training, the LLM can learn to discern the incorporation of irrelevant documents, without being encumbered by extraneous information.

4.2.3 Training Data Construction

To optimize our model, we need high-quality training samples (both positive and negative samples) and labels.

Relevance Labels Acquisition. To obtain fine-grained relevance labels used in Section 4.2.1, we employ a small-scale reranker to acquire the continuous relevance score s_{ce}. We adopt rerankers with the cross-encoder architecture, since they are regarded as effective for assessing relevance degree (Khattab and Zaharia, 2020; Zhao et al., 2024).

Aspect	Self-RAG	CoN	SAIL	REAR (ours)
Assess	Gen	Gen	Gen	Explicit Module
Train	SFT	SFT	SFT	SFT+ Contrastive Loss
Data	GPT	GPT	GPT	Free Model (110M)

Table 1: The difference between REAR and previous work. Assess, Train and Data are short for relevance assessment method, training loss, and data construction methods respectively. REAR utilizes an explicit module for relevance assessment, and adopts bi-granularity (involving contrastive loss) for training. Furthermore, we label the data without access to GPT APIs.

In combination with the traditional method of binary annotating label y, the estimated score is given as:

$$s_{\text{rel}} = \frac{1}{2}\left(s_{\text{ce}} + y\right). \tag{10}$$

This labeling approach combines lexical and semantic similarity, allowing for the acquisition of high-quality labels without accessing GPT APIs.

Irrelevant Documents Sampling. The training method necessitates the use of irrelevant (negative) documents. It has been shown that negative sampling has a large impact on relevance assessment (Xiong et al., 2020). Specially, as shown in Fig. 2 (b), we refine SimANS (Zhou et al., 2022) that ensures negatives are neither too difficult (false negatives) nor too trivial (uninformative):

$$p_i \propto \begin{cases} \exp(-a(s_i - \hat{s}^+ - b)^2), & s_i < \hat{s}^+ - b, \\ \exp(-ak(s_i - \hat{s}^+ - b)^2), & s_i \geq \hat{s}^+ - b, \end{cases} \tag{11}$$

where the sampling probability for the hard negative document is p_i, s_i and $\hat{s}^+$ respectively denote the relevance scores of document d_i and the positive document, and a, b, and k are hyperparameters. By incorporating a decay scaler k into the sampling probability when relevance scores are high, we reduce the chance of sampling false negatives.

Finally, we define the overall loss function for our REAR framework by combining the bi-granularity loss by Eq. 8 and noise-resistant loss by Eq. 9:

$$\mathcal{L}_{\text{REAR}} = \mathcal{L}_{\text{bi-granularity}} + \mathcal{L}_{\text{noise-resistant}}. \tag{12}$$

4.3 Discussion

Distinctions from Existing Methods. As shown in Table 1, our primary contribution lies in the architecture design, which differs significantly from existing studies. Under our optimized architecture, LLMs can generate more fine-grained relevance signals to aid in the following generation process. Besides, LLMs can further calculate the consistency between parametric and external knowledge to evaluate the reliability of answers. Moreover, this architecture makes it easy to adopt the proposed preference-based and noise-resistant loss functions. Furthermore, our label machine makes good use of smaller models and traditional labels, and our sampling strategy improves training data quality, eliminating the need for GPT APIs. As a result, REAR achieves more precise relevance evaluation and better generation performance (Table 3).

Methods	T.C.	Training	Inference
CoN	$\mathcal{O}((p+nd)^2)$	10.34s/step	0.82s/sample
Self-RAG	$\mathcal{O}(n(p+d)^2)$	6.52s/step	1.41s/sample
REAR (ours)	$\mathcal{O}(n(p+d)^2)$	6.33s/step	0.45s/sample

Table 2: The efficiency analysis of REAR and previous work. T.C. is short for time complexity. d, p and n denote the length of the document, the length of the prompt, and the number of documents respectively.

Efficiency. We further discuss the efficiency of our REAR, as shown in Table 2. First, we compare REAR with other RAG frameworks that employ different task formulations, such as Chain-of-Note (CoN) (Yu et al., 2023). CoN processes extensive paragraphs and generates in-depth analyses to identify usable parts of document collections. This methodology leads to increased training and inference times due to the quadratic time complexity associated with transformers (Dong et al., 2024), where time is proportional to the square of the input sequence length. Besides, compared to Self-RAG, which follows a similar approach, REAR achieves a reduction in inference time. This improvement is primarily due to our integration of PagedAttention (Kwon et al., 2023). By using PagedAttention, we ensure that calculations performed during the relevance assessment phase are preserved, thereby eliminating the need for redundant recalculations. The comparisons of actual training and inference times in Table 2 further illustrate the computational efficiency of our method.

5 Experiments

In this section, we detail the experimental setup and then report the main findings of our results.

LLMs	NQ		TriviaQA		WebQ		SQuAD		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Direct Retrieval-Augmented QA										
LLaMA2-Chat $_{7B}$	30.47	41.39	53.92	62.70	22.79	38.29	21.09	31.67	32.07	43.51
Mistral-It $_{7B}$	10.83	31.77	44.59	62.55	8.71	30.79	13.78	34.25	19.48	39.84
Baichuan2-Chat $_{7B}$	33.49	45.61	61.17	69.98	23.87	40.78	26.55	38.97	36.27	48.84
ChatGLM3 $_{6B}$	13.27	20.48	24.57	33.76	5.61	18.38	8.31	15.98	12.94	22.15
RobustLM prompting (4-shot)										
LLaMA2-Chat $_{7B}$	30.53	42.57	53.27	63.52	21.01	38.29	21.83	33.45	31.66	44.46
Mistral-It $_{7B}$	19.11	32.80	48.31	59.87	13.63	30.76	15.98	28.28	24.26	37.93
Baichuan2-Chat $_{7B}$	27.42	39.72	52.07	62.27	18.90	36.13	19.24	30.92	29.41	42.26
ChatGLM3 $_{6B}$	24.65	32.67	46.57	54.23	20.37	34.60	18.71	25.90	27.58	36.85
Fine-tuned RALMs										
Self-RAG $_{7B}$ †	41.02	46.78	52.38	39.15	31.40	26.41	35.28	19.33	40.02	32.92
RobustLM $_{7B}$	44.40	53.08	62.86	70.88	32.48	46.89	27.52	36.75	41.82	51.90
REAR $_{7B}$ w/ Source Rel.	51.33	**60.53**	65.36	74.14	33.02	47.67	36.78	46.64	46.62	57.25
REAR $_{7B}$ w/ Knowledge Con.	**51.41**	60.50	**66.26**	**74.87**	**33.51**	**48.14**	**37.21**	**47.19**	**47.10**	**57.68**

Table 3: A comparison between REAR and baselines on NQ, TriviaQA, WebQ and SQuAD datasets and the averaged performance. our REAR approach surpasses all the other baselines in QA performance. The best and second-best results are in **bold** and underlined fonts respectively. Self-RAG† is evaluated using accuracy (Acc) instead of EM, which is a less strict metric that measures whether the responses contain the answers. The last two lines are our REAR with different verification strategies: "w/ Source Rel." means the source-reliability strategy, and "w/ Knowledge Con." means the knowledge-consistency strategy.

5.1 Experimental Setup

Datasets. We collect the training data from the Natural Questions (NQ) (Kwiatkowski et al., 2019) training set. To ensure the model's adaptability, we also test its performance on three additional open-domain datasets, including TriviaQA (Joshi et al., 2017), WebQuestions (WebQ) (Berant et al., 2013), and SQuAD (Rajpurkar et al., 2016), showing its generalization capabilities to out-of-domain data. We follow the test split in prior work (Karpukhin et al., 2020). The details are in Appendix B.

Baselines. We consider the following two lines of baselines for comparison.

(1) Retrieval augmentation based prompt methods: we design different prompting strategies based on open-source LLMs (without tuning tailored to RAG tasks) to support RAG, including

• *Direct Retrieval-Augmented QA*: We concatenate the top 10 retrieved documents as a single reference document for RAG. To enhance EM metric accuracy, we further incorporate several answer examples within the prompts, as illustrated in Fig. 5.

• *RobustLM prompting*: We following the approach of the prompting strategy (Yoran et al., 2023). The LLMs are required to determine document relevance before generating responses. It employs 4-shot demonstrations (Fig. 6), and provides the top 1 retrieved document.

For open-source LLMs, we consider LLaMA2-Chat (Touvron et al., 2023), Mistral-It (Jiang et al., 2023), Baichuan2-Chat (Yang et al., 2023), and ChatGLM3 (Du et al., 2022).

(2) Specially designed RAG methods: we also consider fine-tuned RobustLM (Yoran et al., 2023) and Self-RAG (Asai et al., 2023) as baselines, which have been specially optimized for the RAG tasks. To ensure a fair comparison, the two frameworks above are evaluated with the same set of retrieved documents as used for REAR.

Metrics. We employ three metrics to evaluate the model's capability of QA accuracy. Exact match (**EM**) (Lee et al., 2019) and **F1** are widely adopted for open-domain QA evaluation. EM calculates whether responses exactly match the gold truth answers and calculates the precision-recall overlap of predicted and true answers. We further evaluate the accuracy in determining the relevance of the given document for LLMs with another two metrics. **Hit@1** evaluates if the document referenced for the model's final answer generation is relevant. **JAcc**, short for judgmental accuracy, measures the proportion of documents correctly evaluated by LLMs as relevant or not.

Implementation Details. To implement our REAR approach, we fine-tune LLaMA2-Base$_{7B}$ (Touvron et al., 2023) on the NQ training set for 1 epoch.

LLaMA2$_{7B}$	NQ		TriviaQA		WebQ		SQuAD	
	JAcc	Hit@1	JAcc	Hit@1	JAcc	Hit@1	JAcc	Hit@1
+ RobustLM-prompting (Yoran et al., 2023)	25.04	-	43.36	-	28.84	-	16.84	-
+ RobustLM-training (Yoran et al., 2023)	56.59	-	56.09	-	49.61	-	56.99	-
+ Self-RAG (Asai et al., 2023)	19.81	51.11	35.69	64.47	25.69	47.98	10.73	38.73
+ REAR (ours)	**74.04**	**66.79**	**80.79**	**74.98**	**65.99**	**56.69**	**59.36**	**53.26**

Table 4: The relevance discrimination and comparison capabilities of REAR and previous approaches. Generative LLMs struggle to determine the relevance degree of the given document, while our REAR overcomes it with the well-designed assessment module.

We set the learning rate to 1e-6. For evaluation, all retrieval documents are sourced from the top 10 documents retrieved by dense retrievers (detailed in Appendix C).

5.2 Main Results

Table 3 shows the results of REAR and baselines on four open-domain QA tasks.

First, our REAR approach surpasses all the other baselines in QA performance. REAR not only performs well on the trained dataset, but also achieves good results on non-training datasets. This demonstrates that our precise signals for capturing relevance effectively guide the generation process. Thus, LLM can generate with good use of both parametric and external knowledge.

Besides, the result shows the efficiency of data construction method, even without access to GPT APIs. Self-RAG labels the relevance degree with GPT-4, while RobustLM and REAR utilize our proposed label machine and sampling method. The result indicates that our data construction strategy is effective and less costly.

Third, generative LLMs struggle to determine the reliability degree of the given document, while our REAR overcomes it with the well-designed assessment module. As shown in Table 4, even the fine-tuned generative approaches (RobustLM-training and Self-RAG) adequately discriminate relevance. In comparison, REAR significantly enhances this capability, highlighting its effectiveness in architectural design.

5.3 Detailed Analysis

In this part, we further present the analysis of the ablation study and impacts of retrieved documents.

5.3.1 Ablation Study

We analyze how each of the proposed components affects final performance. Table 5 shows the performance of our default method and its five variants

Methods	Aspect	Hit@1	EM	F1
REAR	-	66.79	**53.13**	**61.84**
w/o Assessment	Arch.	13.80	38.14	47.44
w/o Consistency	Arch.	**67.48**	52.91	61.49
w/o Bi-granularity	Obj.	66.54	51.88	59.91
w/o Noise-resistant	Obj.	49.25	25.54	33.05
w/o Sampling	Sam.	61.99	49.00	53.62

Table 5: Ablation study on our REAR. The "aspect" denotes the affected aspect. Arch., Obj. and Sam. denote architecture, training objective and sampling strategy.

in three aspects, including the architecture, training objectives and sampling strategy.

(1) *w/o Assessment*: the variant without the integration of the rating module. We utilize language generation to assess relevance degrees instead of the rank head. The document is selected based on the probability of generating judgmental statements. There is a notable drop in the comparison accuracy (see Hit@1 metric), similar shortfall is also observed in Self-RAG (Table 3). This demonstrates the effectiveness of our architectural design, which not only minimizes interference between language generation and relevance discrimination, but also facilitates the incorporation of various loss functions.

(2) *w/o Consistency*: using the path-reliability strategy instead of the knowledge consistency strategy. The path-reliability approach achieves higher Hit@1 rates, yet falls behind in EM and F1 scores compared to the knowledge-consistency strategy. The latter conducts a self-verification of outputs based on its generation ability, effectively integrating inherent knowledge in relevance assessment, which enhances the accuracy of RAG.

(3) *w/o Bi-granularity*: the variant without bi-granularity fusion in relevance assessment training. We replace the bi-granularity loss with the coarse-grained loss function. The results indicate that the fine-grained relevance training could enhance the LLMs in relevance comparison among documents,

LLM	Settings	Rel Doc (EM/Acc)	Irr Doc (EM/Acc)	Overall (EM/Acc)
$LLaMA2_{7B}$	4-shot	54.41	6.40	30.53
$LLaMA2_{13B}$	4-shot	53.36	6.40	30.00
$Mistral_{7B}$	4-shot	36.05	2.00	19.11
$Baichuan2_{7B}$	4-shot	48.68	5.96	27.42
$ChatGLM3_{6B}$	4-shot	46.97	2.12	24.65
$Self\text{-}RAG_{7B}$	fine-tuned	73.48	6.23	40.03
$REAR_{7B}$	fine-tuned	**73.84**	**20.09**	**46.79**

Table 6: Results of factual generation accuracy provided with top-1 retrieved documents on the test set of NQ. Categorized by performance when providing relevant (Rel) and irrelevant (Irr) documents.

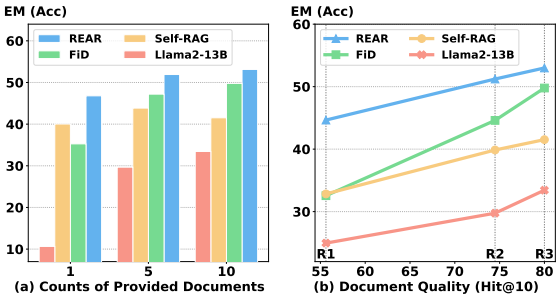

Figure 3: Results of RAG performance vary in overall document count and quality. The left one presents RAG performance with varying numbers of retrieved documents. The right one is the results of RAG with different retriever engines. R1, R2, and R3 represent BM25, Contriever-msmarco, and the FiD-distilled retriever, R1 < R2 < R3 (Table 9 of the Appendix) .

and result in better performance.

(4) *w/o Noise-resistant*: the variant without noise-resistant training. We exclude the gold-noise data pairing, using the similar training construction approach of Self-RAG and RobustLM, with one document per query. We observe a notable decline, underscoring the effectiveness of noise-resistant training to enhance generation against irrelevant document interference.

(5) *w/o Sampling*: the variant training with random hard negatives for training. We can observe a significant drop in relevance assessment capability, further illustrating the effectiveness of our method.

5.3.2 Impact of Retrieved Documents

In this part, we further analyze the impact of retrieved documents in both single-document and multi-document settings.

Single-Document Setting. We first examine the impact of external evidence in single document setting, where only the top first retrieved document is taken for reference. Table 6 shows the factual accuracy of different LLMs. We can see that both Self-RAG and REAR, after fine-tuning, perform well in relevant document utilization. However, REAR significantly outperforms other LLMs in generating accurate responses when the reference document is irrelevant, highlighting its robust resistance to interference from noisy documents.

Multi-Document Setting. In the second setting, we assume that multiple retrieved documents can be used for reference. Specially, we mainly examine the impact of the *total number* and *relevance degree* of reference documents. For this purpose, we vary the number of provided documents (Fig. 3(a)) and the retriever's capabilities (Fig. 3(b)). From Fig. 3(a), we can see that our REAR approach performs well when provided with a single document (*i.e.,* the top retrieved one), while base models without fine-tuning suffer from significant degradation in this case. Furthermore, as shown in Fig. 3(b), our approach is very robust to external retrievers of varied retrieval capacities. Especially, when equipped with the weakest retriever BM25, it yields a large improvement over the other baselines, which further demonstrates that our approach can effectively perceive the relevance of external evidence for more suitable utilization.

6 Conclusion

In this paper, we aimed to enhance the self-awareness of source relevance in RAG systems, and proposed **REAR**, a **RE**levance-**A**ware **R**etrieval-augmented approach for open-domain question answering (QA). For model architecture, we explicitly integrate an assessment module to precisely capture the relevance signals, and employ it to guide the utilization of external knowledge. For model training, we designed an improved training method with bi-granularity relevance fusion and noise-resistant training, which enhance the capacities of fine-grained relevance assessment and adaptive use of retrieved documents. Our data construction strategy collects high-quality data without access to GPT APIs. Extensive experiments on four datasets demonstrate the effectiveness and generalization of REAR's knowledge utilization.

As future work, we will extend the proposed approach REAR to deal with more fine-grained source utilization (*e.g.,* passage or sentence level augmentation), and also consider applying REAR to other knowledge-intensive tasks.

Limitations

For LLMs, the challenge of being misled by irrelevant retrieved documents is a significant obstacle, underscoring the crucial need for enhancing LLMs' ability to adaptively utilize retrieved documents. In response to this issue, our work has concentrated on refining the architecture and training methods to bolster the effective use of retrieved documents by LLMs. We have implemented document-level relevance assessment and dynamic utilization strategies, significantly boosting the factual accuracy of generated content by LLMs. However, our current approach has not delved into guiding LLMs to focus more granularly on key sentences or tokens within the retrieved documents.

Moreover, the applicability of our methods across a broader spectrum of RAG tasks, such as those encompassed by the KILT benchmark, remains to be thoroughly evaluated. This gap presents a pivotal area for our future investigations.

7 Acknowledgments

This work was partially supported by National Natural Science Foundation of China under Grant No. 62222215, Beijing Natural Science Foundation under Grant No. L233008 and 4222027.

References

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*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

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)*. Association for Computational Linguistics.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. *ACL 2020*, page 34.

Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, Hongzhi Zhang, Fuzheng Zhang, Di Zhang, Kun Gai, and Ji-Rong Wen. 2024. Small agent can also rock! empowering small language models as hallucination detector. *arXiv preprint arXiv:2406.11277*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. *arXiv preprint arXiv:2401.14887*.

Zican Dong, Junyi Li, Xin Men, Wayne Xin Zhao, Bingbing Wang, Zhen Tian, Weipeng Chen, and Ji-Rong Wen. 2024. Exploring context window of large language models via decomposed positional vectors. *arXiv preprint arXiv:2405.18009*.

Zican Dong, Tianyi Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

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

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2023. Fid-light: Efficient and effective retrieval-augmented text generation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1437–1447.

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

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*.

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*.

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)*. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

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

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jingyuan Wang, Jian-Yun Nie, and Ji-Rong Wen. 2023b. The web can be your oyster for improving large language models. *arXiv preprint arXiv:2305.10998*.

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*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. *arXiv preprint arXiv:2106.00085*.

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.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021a. Pair: Leveraging passage-centric similarity relation for improving dense passage retrieval. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2173–2183.

Ruiyang Ren, Peng Qiu, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2024. Bases: Large-scale web search user simulation with large language model based agents. *arXiv preprint arXiv:2402.17505*.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021b. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers 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 6648–6662.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *ArXiv*, abs/2304.09542.

Xinyu Tang, Xiaolei Wang, Wayne Xin Zhao, Siyuan Lu, Yaliang Li, and Ji-Rong Wen. 2024. Unleashing the potential of large language models as prompt optimizers: An analogical analysis with gradient-based model optimizers. *arXiv preprint arXiv:2402.17564*.

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*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

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

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Adversarial retriever-ranker for dense text retrieval. In *International Conference on Learning Representations*.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are large language models good at utility judgments? *arXiv preprint arXiv:2403.19216*.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H Chi, Quoc V Le, and Denny Zhou. 2023. Take a step back: evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.

Kun Zhou, Yeyun Gong, Xiao Liu, Wayne Xin Zhao, Yelong Shen, Anlei Dong, Jingwen Lu, Rangan Majumder, Ji-Rong Wen, and Nan Duan. 2022. Simans: Simple ambiguous negatives sampling for dense text retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 548–559.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

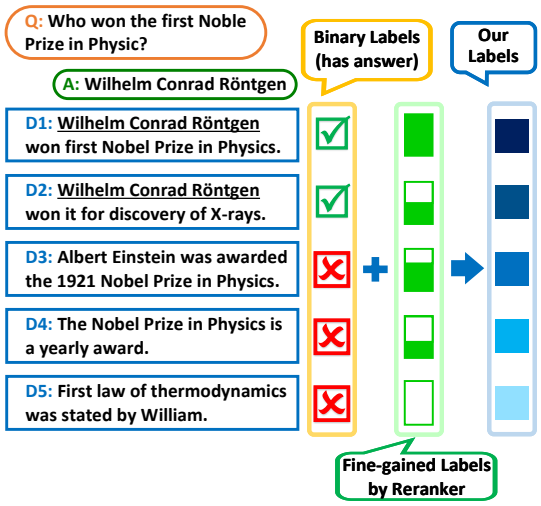

Figure 4: The illustration of different retrieved documents and different labeling metrics.

A Details on Fine-Gained Relevance Optimization

We first illustrate why to design the fine-grained optimization for the assessment module. Traditional annotation methods always use a binary labeling method (Karpukhin et al., 2020), which is based on the presence of an answer within a document. As shown in Fig. 4, both D1 and D2 are labeled as "*relevant*". However, while D1 allows for direct answer derivation, D2 requires additional external knowledge for induction. Training models solely on simple binary classification fails to distinguish the superiority of D1 over D2, potentially leading to inaccuracies in finer relevance judgments.

Previous work has achieved success in relevance assessment by distilling the ranking results from GPT-4 (Sun et al., 2023). Inspired by it, we propose a less costly solution by labeling with small-scale, well-trained cross-encoder rerankers RocketQAv2 (Ren et al., 2021b). Despite the good relevance evaluation performance, it still may get wrong. We adopt three strategies to reduce the negative impact of annotation errors on training. Firstly, we design the sampling method (Eq. 11), which reduces the likelihood of potentially false negatives being sampled. Besides, we linearly combine the binary label with cross-encoder scores. Thirdly, to mitigate noise from rerankers, we disregard differences smaller than 0.1 in fine-gained relevance training in Eq. 7. These strategies enhance the quality of training data, which in turn improves the performance of REAR.

B Details on Dataset.

We utilize four open-domain QA datasets, Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQuestions (WebQ) (Berant et al., 2013) and SQuAD (Rajpurkar et al., 2016)).

Dataset	NQ	TriviaQA	WebQ	SQuAD
Num. Data	3,610	11,313	2,032	10,570

Table 7: Dataset statistics of the test set.

• **NQ:** a dataset designed to support comprehensive QA systems. It includes questions sourced from actual Google search queries. The corresponding answers are text spans within Wikipedia articles, meticulously identified by human annotators.

• **TriviaQA:** a compilation of trivia questions paired with answers, both of which were initially extracted from online sources.

• **WQ:** constructed from questions proposed via the Google Suggest API, with answers being specific entities listed in Freebase.

• **SQuAD:** a dataset for evaluating reading comprehension, and also is used for training and testing open-domain QA engines.

NQ is used for both training and inference, while the other three are only used for inference. We use the same split as previous work (Karpukhin et al., 2020). The training set of NQ contains 58,880 samples.

C Details on Document Collection

In this part, we introduce the retrievers we used to collect documents. We employ task-specific retrievers to acquire the retrieved document. For inference, we utilize FiD-distilled retrievers (Izacard and Grave, 2021a) for NQ and TriviaQA datasets. And we implement a strategy incorporating in-batch negatives and joint retriever-ranker training, starting from the Contriever-msmarco (Izacard et al., 2021) checkpoint for SQuAD and WQ datasets. The recall and MRR rates of the retrieved documents for inference are shown in Table. 8.

Metrics	NQ	TriviaQA	WQ	SQuAD
Hit@1	50.25	62.91	50.64	37.53
Hit@10	80.00	81.78	75.89	68.51

Table 8: Retrievers we use for testing Hit rates (Recall rates) across datasets on the test sets.

Metric	R1	R2	R3
Hit@10	55.62	74.49	80.00
MRR@10	32.35	51.45	60.32

Table 9: Performance of three retrievers on the NQ test set. Hit@10 measures the percentage of correct answers within the top 10 results, indicating the precision of the retriever. MRR@10 (Mean Reciprocal Rank at 10) calculates the average of the reciprocal ranks of the first correct answer within the top 10 results, reflecting the effectiveness and rank of correct answers by the system. R1, R2 and R3 denote BM25 (Robertson et al., 1995), Contriever-msmarco (Izacard et al., 2021) and the dense retriever (Izacard and Grave, 2021a) trained by distilling attention scores of FiD reader (Izacard and Grave, 2021b)

Knowledge:
{retrieved document 1}
{retrieved document 2}
.....
{retrieved document 9}

Answer the following question with a very short phrase, such as "1998", "May 16th, 1931", or "James Bond", to meet the criteria of exact match datasets.

Question: {question}

Answer:

Figure 5: Prompts for "*direct RAG QA*".

D Details on Implementation

Following the previous work (Asai et al., 2023), we apply joint optimization combining relevance assessment and relevance-guided generation, as specified in Eq. 12. The training utilizes a learning rate of 1e-6, a warm-up ratio of 0.03, a batch size of 64 and a cosine scheduler for 1 epoch. Our experiments leverage the computational power of 8 NVIDIA Tesla A100 GPUs, each with 40G of memory.

E Details on Baselines

In this part, we detail the prompt design and inference details for baselines. For the prompt-based inference, we utilize the instruction-tuned open-source models obtained from Hugging Face. Following the existing work (Asai et al., 2023; Ren et al., 2024; Tang et al., 2024), we use the greedy decoding strategy for inference. The specific instruction formats used in our tests are illustrated in Fig. 5 and Fig. 6.

Given a passage and a query, predict whether the passage includes an answer to the query by producing either 'Yes' or 'No'. And then answer with the given passage if 'yes', or answer with your external knowledge if 'No'.

Passage: {positive document example 1}.
Query: {question 1}
Judge: Yes.
Answer: {answer 1}

Passage: {negative document example 2}.
Query: {question 2}
Judge: No.
Answer: {answer 2}

Other 2 examples...

Passage: {one of the retrieved document}.
Query: {question}
Judge: (calculate the difference in log perplexity for "Yes" and "No" and fill accordingly).
Answer:

Figure 6: Prompts for "*RobustLM based prompting*".