

# Enhancing Noise Robustness of Retrieval-Augmented Language Models with Adaptive Adversarial Training

Feiteng Fang<sup>1,2\*</sup>, Yuelin Bai<sup>2\*</sup>, Shiwen Ni<sup>2†</sup>, Min Yang<sup>2†</sup>, Xiaojun Chen<sup>3</sup>, Ruifeng Xu<sup>4</sup>

<sup>1</sup>University of Science and Technology of China

<sup>2</sup>Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

<sup>3</sup>Shenzhen University <sup>4</sup>Harbin Institute of Technology (Shenzhen)

feitengfang@mail.ustc.edu.cn, {yl.bai, sw.ni, min.yang}@siat.ac.cn,

xjchen@szu.edu.cn, xuruifeng@hit.edu.cn

## Abstract

Large Language Models (LLMs) exhibit substantial capabilities yet encounter challenges, including hallucination, outdated knowledge, and untraceable reasoning processes. Retrieval-augmented generation (RAG) has emerged as a promising solution, integrating knowledge from external databases to mitigate these challenges. However, inappropriate retrieved passages can potentially hinder the LLMs' capacity to generate comprehensive and high-quality responses. Prior RAG studies on the robustness of retrieval noises often confine themselves to a limited set of noise types, deviating from real-world retrieval environments and limiting practical applicability. In this study, we initially investigate retrieval noises and categorize them into three distinct types, reflecting real-world environments. We analyze the impact of these various retrieval noises on the robustness of LLMs. Subsequently, we propose a novel RAG approach known as Retrieval-augmented Adaptive Adversarial Training (RAAT). RAAT leverages adaptive adversarial training to dynamically adjust the model's training process in response to retrieval noises. Concurrently, it employs multi-task learning to ensure the model's capacity to internally recognize noisy contexts. Extensive experiments demonstrate that the LLaMA-2 7B model trained using RAAT exhibits significant improvements in F1 and EM scores under diverse noise conditions. For reproducibility, we release our code and data at: <https://github.com/calubkk/RAAT>.

## 1 Introduction

Large language models (LLMs) have garnered substantial attention in both academic and industrial research within the domain of artificial intelligence due to their remarkable capabilities (Brown et al., 2020; Bubeck et al., 2023). Despite their immense



Figure 1: An illustrative example of the RAG process applied to question answering. The model predicts the correct answer with accurate retrieved text. However, it fails to produce the right answer when the retrieved text contains misleading or inaccurate information.

power, LLMs face challenges such as hallucinations and outdated knowledge (Gao et al., 2023). Moreover, a lack of domain knowledge may hinder their performance on domain-specific tasks (Kandpal et al., 2023). To mitigate these challenges, recent studies improve LLMs by retrieving passages from external databases and pretending them in context, constituting a framework known as retrieval-augmented language models (RALMs) (Mao et al., 2020; Lewis et al., 2020).

However, RALMs also present significant limitations. Previous studies (Yoran et al., 2023; Yu et al., 2023; Shi et al., 2023) have empirically demonstrated that retrieved noisy passages are problematic for LLMs, resulting in performance degradation. We term this issue as the noise robustness problem of RALMs. As illustrated in Figure 1, the model can provide correct answers when the re-

\*Equal contribution.

†Corresponding author.

trieving context is accurate and related to the query. However, when the retrieved context contains misleading or inaccurate information, the model may yield incorrect answers. As the retriever inherently cannot achieve complete accuracy, the presence of noise in the retrieved context is inevitable. Therefore, designing robust algorithms against retrieved noises is of great practical importance.

Recently, several studies (Yoran et al., 2023; Li et al., 2022) have attempted to enhance the noise robustness of RALMs through noisy training, which involves incorporating retrieved noisy contexts into fine-tuning data. While noisy training exhibits promise, its effectiveness heavily relies on the composition of the training dataset. Incorrectly introducing noises to the training data can lead to model overfitting, adversely affecting generalization. In practice, meticulous adjustment of the type and intensity of noises is essential to ensure the model’s proficiency across various tasks and datasets. This demands significant experimentation and tuning, adding complexity to the development process. Moreover, the lack of clear classification for retrieval noises in current studies stands in contrast to the diverse range of noises encountered in real retrieval environments.

This paper systematically explores three types of retrieval noises: (i) contexts that are superficially related to the query but lack the correct answer (*Relevant retrieval noise*), (ii) contexts that are irrelevant to the query (*Irrelevant retrieval noise*), and (iii) contexts that are topically related to the query but contain incorrect information (*Counterfactual retrieval noise*). Our empirical study indicates that LLMs exhibit varying robustness to these three types of noise. Compared to entirely irrelevant texts, texts that are superficially related to the query or those containing counterfactual details often lead to more misinformation.

In response to diverse types of noises, we propose a novel approach named Retrieval-augmented Adaptive Adversarial Training (RAAT), which employs adaptive adversarial training to dynamically regulate the model’s training process in response to retrieved noisy texts. Concretely, RAAT generates adversarial samples (noises) by considering the model’s sensitivity to different types of noises during training, which aligns with the min-max paradigm of adversarial training (Morris et al., 2020; Ivgi and Berant, 2021). Moreover, RAAT utilizes multi-task learning (Ruder, 2017) to encourage the LLMs to generate tokens that are aware of

noises, thereby enabling the model to internally recognize retrieved noisy contexts and improve the overall generation performance.

The main contributions of this paper can be summarized as follows:

- We systematically explore three types of retrieval noises and investigate the sensitivity of LLMs to these diverse types of noises.
- We propose a novel adaptive adversarial training method (called RAAT) to enhance the robustness of RALMs against various retrieval noises. RAAT dynamically adjusts the training process of the model in diverse noise environments. In addition, it integrates multi-task learning to encourage the model to improve its ability to discern different types of noises.
- We set up a benchmark (named RAG-Bench) for assessing the noise robustness problem of RALMs based on three open-domain question-answering datasets. Experimental results demonstrate that our RAAT method enhances robustness across diverse retrieval noise environments.

## 2 Related Work

**Retrieval-Augmented Generation with Noisy Context** Retrieval-Augmented Language Models (RALMs) have shown impressive performance in various NLP tasks (Gao et al., 2023; Zhu et al., 2023). However, limited by the capabilities of the retriever, retrieval-augmented systems inevitably introduce irrelevant or partially relevant knowledge to the models (Yin et al., 2023). Recent studies (Yu et al., 2023; Yoran et al., 2023; Chen et al., 2023) have increasingly focused on the impact of noisy information on retrieval-augmented generation. For example, Jia and Liang (2017); Creswell et al. (2022) observed that adding irrelevant noise to the context could detrimentally affect model performance. Chen et al. (2023) demonstrated that as the proportion of noise in the retrieval context increases, the performance of LLMs experiences a notable decline. Similar phenomena have been reported by Yoran et al. (2023) and Thakur et al. (2023).

**Adversarial Training** Adversarial training is recognized as a crucial method for enhancing model robustness, initially proposed by Goodfellow et al. (2014). Early studies widely investigated adversarial training in the computer vision domain (Kurakin

et al., 2016; Madry et al., 2017). In the NLP domain, Miyato et al. (2016) applied perturbations to word embeddings, making the model less prone to overfitting. Similarly, perturbations on different granularities have been extensively studied, encompassing various aspects of NLP tasks (Yasunaga et al., 2017; Wu et al., 2017; Zhu et al., 2019; Wang et al., 2020; Ni et al., 2023; Liang et al., 2023).

Recently, several studies have concentrated on generating adversarial examples designed to induce LLMs to generate harmful or non-factual content (Zou et al., 2023; Shen et al., 2023) instead of merely causing the model to make inaccurate predictions. Shen et al. (2023) employed decision-based perturbation at different levels to craft adversarial examples, revealing vulnerabilities in ChatGPT to both sentence-level and character-level adversarial attacks. Shi et al. (2023) added irrelevant context to an arithmetic reasoning dataset, finding that including irrelevant information distracted the model’s predictions. Zou et al. (2023) proposed a method that could reliably generate adversarial attack suffixes, yielding adversarial prompts that exhibit high transferability.

In this study, we investigate adversarial training concerning LLMs in response to various retrieval noises, aiming to efficiently obtain adversarial examples that enhance model robustness while reducing training overhead. We construct noisy adversarial examples by sampling or paraphrasing the original dataset. This approach ensures more dependable and precise outputs even when confronted with imperfect retrieved contexts.

### 3 Methodology

#### 3.1 Problem Setup

In the standard RALM, given input query  $x$ , a retriever  $r$  is designed to retrieve relevant contexts  $C = \{c_1, c_2, \dots\}$  from an external database. During inference, the content of retrieval context is concatenated with  $x$  to form  $d$ , which is then fed into the pre-trained language model  $M$ , yielding a response  $\hat{y}$  regarding  $x$ . If the retrieved context  $c$  contains the correct answer  $y$  about  $x$ , we can denote  $c$  as  $c_{golden}$ , representing the golden retrieval context. However, if  $c$  does not contain the correct answer  $y$  or contains partially irrelevant content, we can denote  $c$  as  $c_{noisy}$ .

In our study, we transform open-domain question answering (QA) into a reading comprehension task to meet the RAG settings. Formally, given

the objective  $f$  of an open domain question answering task is  $f : \{x\} \rightarrow y$ , we can formulate the objective of the reading comprehension task as  $f : \{c_{golden}, x\} \rightarrow y$ . In examining the challenge of the retrieval noise robustness problem of RALM, we aim to obtain a fine-tuned model  $M'$  that can not only fulfill the function  $f : \{c_{golden}, x\} \rightarrow y$  but also produce accurate answers even in the presence of additional retrieval noise  $c_{noisy}$  and achieve function  $f : \{c_{golden}, c_{noisy}, x\} \rightarrow y$ .

#### 3.2 Diverse Retrieval Noises

We systematically classify the retrieval noise present in  $c_{noisy}$  to closely mimic real-world conditions. Existing studies (Yoran et al., 2023; Yu et al., 2023) on retrieval noise robustness often dichotomize noise into relevant and irrelevant categories. However, we contend that such a classification may not fully align with the retrieval noise robustness of RALMs. In this work, we propose a more nuanced classification of retrieval noise, differentiating it into three distinct types: *Relevant retrieval noise*, *Irrelevant retrieval noise*, and *Counterfactual retrieval noise*. Specifically, *Relevant retrieval noise* (denoted as  $c_r$ ) pertains to contexts that exhibit superficial relevance to the query  $x$  but lack the information necessary for the correct answer  $y$ . These contexts may appear relevant at first glance but ultimately mislead the model. *Irrelevant retrieval noise* (denoted as  $c_i$ ) encompasses contexts with low relevance to the query  $x$ , often arising from erroneous retrievals and generally being off-topic. *Counterfactual retrieval noise* (denoted as  $c_c$ ) encompasses contexts that are topically related to  $x$  but contain incorrect and misleading information, often attributed to inaccuracies in the retriever’s database.

To examine the influence of three distinct types of retrieval noise on LLMs, we establish a benchmark for assessing retrieval noise robustness in LLMs by employing three open-domain question-answering datasets: Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and WebQ (Berant et al., 2013). Leveraging this benchmark, we evaluated the susceptibility of various open-source large language models to the effects of the three identified types of noise. The details of the construction of this benchmark can be found in Section 4.1. Leveraging this benchmark, we evaluate the sensitivity of various LLMs to the effects of the three types of noise. Specifically, we conduct experiments on six LLMs,

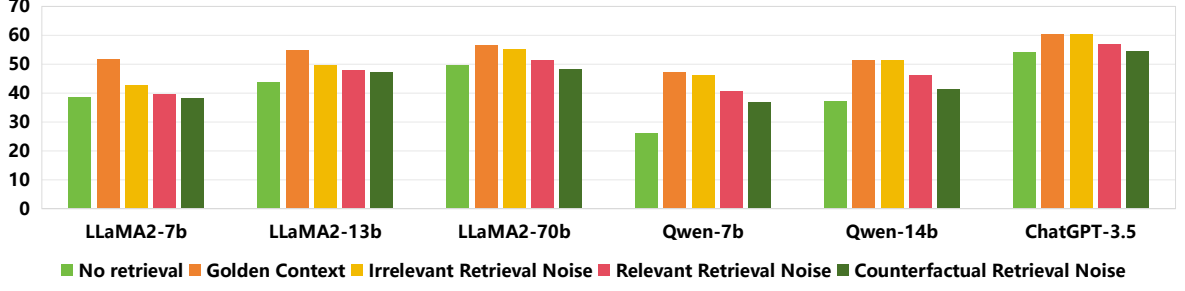


Figure 2: Exact match (EM) scores of various models under different types of retrieval noises. “Golden Context” denotes instances where LLMs respond to questions with reference to the golden retrieval context. “No Noise” indicates instances where LLMs answer questions without any retrieval. The experimental configurations of other models involve the introduction of different types of noises on the foundation of the “Golden Context”.

including ChatGPT<sub>3.5</sub>, LLaMA2<sub>7B</sub> (Touvron et al., 2023), LLaMA2<sub>13B</sub> (Touvron et al., 2023), LLaMA2<sub>70B</sub> (Touvron et al., 2023), Qwen<sub>7B</sub> (Bai et al., 2023), and Qwen<sub>14B</sub> (Bai et al., 2023). For each model, our experiments encompass two distinct settings: one with the exclusive presence of the golden retrieval context  $c_{golden}$  and another incorporating the introduction of three different types of retrieval noise  $c_{noisy}$ . As shown in Figure 2, all LLMs experience varying degrees of impact from the three types of noise. The performance of LLMs exhibits a decline ranging from 0.2% to 13.43%. Through a comparative analysis of the effects of the three types of noise, we observe that irrelevant retrieval noise has a comparatively minor impact on LLMs with substantial capabilities.

### 3.3 Retrieval-augmented Adaptive Adversarial Training

Recently, several studies (Yoran et al., 2023; Li et al., 2022) attempted to enhance the noise robustness of LLMs through noisy training, which involves incorporating retrieved noisy contexts into fine-tuning data. The essence of noisy training involves the exploration of offline data augmentation, while in contrast, adversarial training leverages online data augmentation for a similar purpose (Ivgi and Berant, 2021). The core idea of adversarial training is to fortify the models against adversarial conditions by introducing adversarial perturbations (Jain et al., 2023). In the construction of adversarial samples, also known as noise samples, the min-max optimization strategy assumes a pivotal role, encompassing two fundamental steps. Initially, the maximization process involves adjusting the input data to intentionally mislead the model, inducing the maximum prediction error.

Then, the minimization process entails fine-tuning the model’s parameters to enhance its resistance against these meticulously crafted input perturbations (Bai et al., 2021). This strategy seeks to strike a balance, allowing the model to accurately identify normal data while robustly defending against potential attacks from adversarial examples.

In this study, we aim to refine the objective of adversarial training while exploring the noise robustness challenges of RALMs. Considering a given query  $x$ , we assume the existence of four types of data augmentation, namely, golden retrieval context only ( $da_g$ ), additional relevant retrieval noise ( $da_r$ ), additional irrelevant retrieval noise ( $da_i$ ), and additional counterfactual retrieval noise ( $da_c$ ). The space of data augmentation is denoted as  $DA = \{da_g, da_r, da_i, da_c\}$ . Then, the optimization problem can be formulated as follows:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \max_{da \in DA} \mathcal{L}(\theta, da(x), y) \right] \quad (1)$$

where  $\mathcal{D}$  denotes training data,  $\mathcal{L}$  is the loss function,  $\theta$  denotes the parameters of LLMs, and  $da(x)$  represents the data augmentation of  $x$ .

Building upon the optimization problem outlined above, we introduce adaptive adversarial training as a tailored approach to enhance the robustness of RALMs against retrieval noise. Within adaptive adversarial training, the model refrains from updating parameters across all adversarial samples. Instead, it initiates the process by computing the generation loss for each adversarial sample, quantifying its adaptability to varying noise environments. Notably, a higher generation loss implies reduced adaptability of the model to the noisy environment. Given that each query involves one sample with a golden retrieval context and three



How much money did the film "Titanic" make?

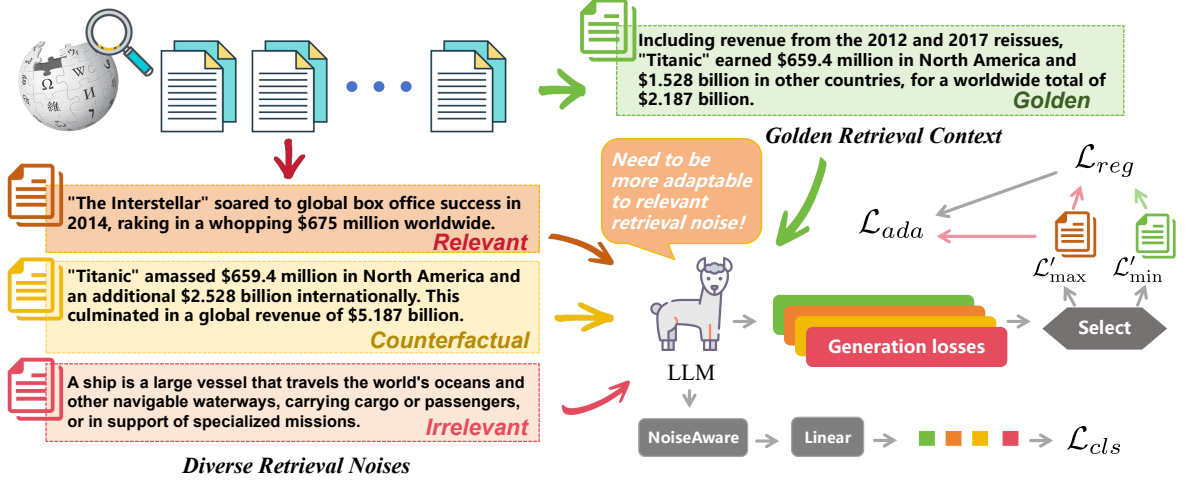


Figure 3: The overview of our proposed RAAT method, which incorporates three distinct types of retrieval noises and the golden retrieval context during the training process.

adversarial samples, the model generates four distinct generation losses in each iteration. Following a min-max optimization strategy, the model prioritizes the selection of the largest loss to guide subsequent parameter update. Formally, we define the generation loss function  $\mathcal{L}'$  for the augmented input  $x'$  as:

$$\mathcal{L}'(\theta, x', y) = -\frac{1}{|y|} \sum_{t=1}^{|y|} \log P_{\theta}(y_t | x', y_{<t}) \quad (2)$$

where  $x' = da(x)$  represents the augmented noise context of  $x$ .

To effectively enhance performance across diverse noise environments, adaptive adversarial training incorporates a regularization term into its loss function. This incorporation of a regularization term is designed to mitigate the risk of the model overfitting to a particular type of noise. The regularization term acts as a stabilizing factor, promoting generalization and preventing the model from becoming overly specialized in its response to a specific noise profile. To achieve this goal, we introduce a regularization term specifically designed to mitigate the variance between these generation losses. This regularization term operates by identifying the largest  $\mathcal{L}'_{max}$  and the smallest  $\mathcal{L}'_{min}$  of the four generation losses at each training step.  $\mathcal{L}'_{max}$  is the generation loss with the highest numerical value among four losses being considered. Conversely,  $\mathcal{L}'_{min}$  is the loss function with the lowest numerical value. Here, an increased loss value indicates a greater magnitude of error or dispar-

ity in the aspect of the model's performance being assessed. This suggests that the model exhibits heightened sensitivity to adversarial examples reflecting retrieval noise. These adversarial examples are designed to probe and exploit weaknesses in the model's processing capabilities, especially in how it deals with noisy information in its input data. The regularization term, calculated as the square of the difference between  $\mathcal{L}'_{max}$  and  $\mathcal{L}'_{min}$ , aims to reduce the model's sensitivity to retrieval noise by encouraging a more balanced optimization. Formally, we design the regularization term  $\mathcal{L}_{reg}$  as:

$$\mathcal{L}_{reg} = \|\mathcal{L}'_{max} - \mathcal{L}'_{min}\|_2^2 \quad (3)$$

Subsequently, we define the adaptive adversarial training loss function  $\mathcal{L}_{ada}$  as follows:

$$\mathcal{L}_{ada} = \mathcal{L}'_{max} + w_{reg} \cdot \mathcal{L}_{reg} \quad (4)$$

where  $w_{reg}$  is a pre-defined hyperparameter to control the weight of  $\mathcal{L}_{reg}$ .

### 3.4 Incorporating Noise Awareness

Accurately identifying retrieval noise plays a pivotal role in fortifying the robustness of RALMs against the retrieval noise. Models endowed with the ability to discern different types of noise can more effectively choose and utilize training data, leading to an improvement in the overall quality of their generated outputs. This capacity to distinguish between various noise types contributes significantly to the model's adaptive learning process, enabling it to optimize performance in the

presence of diverse noise scenarios. Inspired by the above motivation, we propose an auxiliary task designed to autonomously recognize the types of noisy retrieval texts, aiming to significantly bolster the retrieval robustness of RALMs. This auxiliary task serves as a valuable augmentation, contributing to the overall adaptability and effectiveness of the model in scenarios involving retrieval noise.

Specifically, we attempt to enable the model to generate tokens that are sensitive to noise, thereby improving the model’s capacity to discern various types of retrieval noise internally. Specifically, we first incorporate a linear layer beneath LLMs. Subsequently, a classification loss  $\mathcal{L}_{\text{cls}}$  is computed for each of the golden retrieval context and the three adversarial samples corresponding to each input  $x$ . One-hot encoding is employed in classification tasks, assigning values from 1 to 4 as labels to train the classifiers, where each classifier is tailored to a different retrieval noise type. The loss function  $\mathcal{L}_{\text{cls}}$  is computed using cross-entropy.

Finally, we formulate the final RAAT loss  $\mathcal{L}_{\text{RAAT}}$  by combining the adaptive adversarial training loss and the classification loss in the context of multi-task learning:

$$\mathcal{L}_{\text{RAAT}} = w_{\text{ada}} \cdot \mathcal{L}_{\text{ada}} + w_{\text{cls}} \cdot \mathcal{L}_{\text{cls}} \quad (5)$$

where  $w_{\text{ada}}$  and  $w_{\text{cls}}$  represent pre-defined hyperparameters used to balance the importance of these two different tasks.

## 4 Experiments

### 4.1 Dataset Construction

We have formulated a benchmark named RAG-Bench that is specifically designed to evaluate the retrieval noise robustness of LLMs. RAG-Bench is established upon three widely available datasets that center around open-domain question answering (QA): Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and WebQ (Berant et al., 2013). For each dataset, we employ the retrieval model DPR (Karpukhin et al., 2020) as our retriever, which retrieves ten passages from Wikipedia for each query. Then, we apply filtering to the queries, ensuring that each query in the filtered subset contains at least two golden retrieval contexts, indicating the presence of correct answers. Detailed statistics for both the full set and the filtered subset can be found in Table 1.

Each sample in our dataset contains a golden retrieval context and is deliberately designed to in-

Datasets	Train		Test	
	#Full	#Subset	#Full	#Subset
NQ	79,168	40,551	3,610	1,833
TriviaQA	78,785	51,202	11,313	7,010
WebQ	3,778	2,316	2,032	1,057

Table 1: The statistics of the three QA datasets.

corporate three types of augmented retrieval noise. To introduce **relevant retrieval noise**, we choose the context most pertinent to the query from the set of ten retrieval texts, excluding the golden retrieval context. In the case of **irrelevant retrieval noise**, no selection is made from the retrieval texts associated with the current query. Instead, a passage is randomly chosen from the retrieval contents of other queries, ensuring its complete irrelevance to the current query. For the **counterfactual retrieval noise**, we randomly select one passage from the two golden retrieval contexts and substitute its answer entity with an incorrect one.

The test set of RAG-Bench comprises 1000 randomly chosen samples from the test sets of three QA datasets, resulting in a total of 3000 samples. The training set consists of 1500 samples randomly selected from the training sets of the three datasets, totaling 4500 samples. The validation set, drawn from the training sets of three QA datasets, contains 300 samples. Notably, careful measures were taken to ensure no overlap with the training data of RAG-Bench.

### 4.2 Evaluation Metrics

We evaluate the effectiveness of our method using two metrics: exact match (EM) and F1 score (Chen et al., 2017). Concretely, EM assesses the extent to which the answer generated by the system aligns precisely with the standard answer without any disparities at the character level. In contrast, the F1 score incorporates precision and recall, accounting for the equilibrium between correctly identifying answers and avoiding omitting correct answers.

### 4.3 Baseline Methods

We conduct a comparison of our RAAT method against zero-shot LLMs, as well as finetuning approaches applied to LLaMA2<sub>7B</sub>, which shares a common backbone with RAAT.

**Zero-Shot Methods** Within the open-source community, many foundational and supervised fine-

Method	Golden Only		Golden & $c_i$		Golden & $c_r$		Golden & $c_c$		Avg	
	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
LLaMA2 <sub>7B</sub>	65.56	51.80	56.14	42.87	53.10	39.73	51.81	38.37	56.68	43.19
Qwen <sub>7B</sub>	62.57	47.07	61.48	46.06	55.50	40.50	53.26	36.90	58.20	42.63
LLaMA2 <sub>13B</sub>	69.27	55.00	63.25	49.47	62.27	47.97	62.07	47.17	64.22	49.90
Qwen <sub>14B</sub>	67.45	51.43	66.71	51.20	61.88	46.16	58.65	41.30	63.67	47.52
LLaMA2 <sub>70B</sub>	71.43	56.56	70.05	55.13	65.97	51.33	63.91	48.27	67.84	52.82
ChatGPT <sub>3.5</sub>	73.98	60.50	72.24	60.30	70.65	56.89	69.00	54.64	71.47	58.10
RALM <sub>golden</sub>	80.31	74.03	79.33	72.73	73.26	66.33	73.08	65.40	76.50	69.62
RetRobust	80.10	73.80	79.25	72.97	74.81	68.30	75.46	68.43	77.41	70.88
RALM <sub>retrieved</sub>	80.04	73.40	81.09	74.80	75.99	69.10	73.10	65.67	77.55	70.74
RALM <sub>multiple</sub>	85.47	80.17	85.27	81.20	83.07	78.33	83.25	79.23	84.27	79.73
RAAT	<b>87.15</b>	<b>83.07</b>	<b>86.80</b>	<b>82.73</b>	<b>85.14</b>	<b>81.00</b>	<b>86.29</b>	<b>82.10</b>	<b>86.35</b>	<b>82.23</b>

Table 2: Experimental results on our RAG-Bench benchmark. “Golden Only” denotes a scenario where LLMs only consult the golden retrieval context. In “Golden &  $c_i/c_r/c_c$ ”, LLMs consider both the golden retrieval context and *irrelevant retrieval noise/relevant retrieval noise/counterfactual retrieval noise*.

tuning (SFT) models have emerged. In our experiments, we select six renowned LLMs as baselines: ChatGPT<sub>3.5</sub>, LLaMA2<sub>7B</sub> (Touvron et al., 2023), LLaMA2<sub>13B</sub>, LLaMA2<sub>70B</sub>, Qwen<sub>7B</sub> (Bai et al., 2023), and Qwen<sub>14B</sub>.

**Fine-tuning Methods** We further compare RAAT with various fine-tuning methods.

- **RALM<sub>golden</sub>** This is a RALM with instruction tuning (Lin et al., 2023). It prepends a golden retrieval text  $c_{golden}$  in context to fine-tune LLaMA2<sub>7B</sub>.
- **RetRobust** To ensure that the model is exposed to both golden retrieval texts and various retrieval noise during training, Yoran et al. (2023) proposes RetRobust. For each query, RetRobust selects top-1, low-ranked, or random retrieved passages with equal probability for training.
- **RALM<sub>retrieved</sub>** This variant is a RALM incorporating instruction tuning. In contrast to RALM<sub>golden</sub>, RALM<sub>retrieved</sub> does not manually design retrieval noise in the training set but directly uses the top-2 retrieved passages. This training method is more aligned with real retrieval environments.
- **RALM<sub>multiple</sub>** This approach closely resembles RetRobust, differing only in the construction of the training dataset. In RALM<sub>multiple</sub>, rather than introducing one type of retrieval noise randomly for each query, each type of retrieval noise

is combined with the sample and incorporated into the dataset. That is, each query is associated with four augmented noisy samples.

#### 4.4 Implementation Details

Our RAAT method relies on LLaMA2-7B as the foundational model. We set the weight parameters as follows:  $w_{reg}$  to 0.1,  $w_{raat}$  to 2, and  $w_{cls}$  to 1. The sequence length, epoch, and learning rate are configured to 512, 2, and 5e-6, respectively. Our experiments are conducted on a computational cluster equipped with 4 NVIDIA A100 GPUs, each boasting a capacity of 80GB.

### 5 Experimental Results

#### 5.1 Main Results

Table 2 illustrates the efficacy of our RAAT method compared to the baselines in terms of F1 and EM scores. We observe that all models are affected by three different types of retrieval noise attacks. The influence of *irrelevant retrieval noise* is marginal, while *counterfactual retrieval noise* exerts the most significant impact. For the models sharing the same architecture, larger parameter sizes correlate with superior performance and better robustness against retrieval noise. For instance, LLaMA2<sub>7B</sub> exhibits a 12.46% reduction in F1 score when confronted with *relevant retrieval noise*, whereas LLaMA2<sub>13B</sub> only experiences a 7% decrease under identical conditions. This trend is also evident in Qwen.

Method	Golden Only		Golden & $c_i$		Golden & $c_r$		Golden & $c_c$		Avg	
	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
RAAT	87.15	83.07	86.80	82.73	85.14	81.00	86.29	82.10	86.35	82.23
RAAT w/o $\mathcal{L}_{cls}$	86.76	82.77	86.45	82.27	84.69	80.63	85.54	81.20	85.86	81.71
RAAT w/o $\mathcal{L}_{reg}$	86.87	83.03	83.92	79.86	84.69	80.57	87.02	82.80	85.63	81.57

Table 3: Ablation test results in terms of EM and F1 score.

From Table 2, we can also observe that fine-tuning enables LLMs to better utilize information from the retrieval texts. Fine-tuned models significantly outperform the zero-shot LLMs with varying parameter sizes. Moreover,  $\text{RALM}_{multiple}$  shows a significant improvement over  $\text{RALM}_{golden}$ ,  $\text{RALM}_{retrieved}$  and RetRobust, reflecting the sensitivity of retrieval noise to the training dataset and the importance of diversity in noise attacks during training. Our RAAT method achieves even better performance than  $\text{RALM}_{multiple}$  in all four environments, with an average increase of 2.08% in the F1 score and 2.5% in the EM score, demonstrating its superior ability to handle diverse retrieval noise.

## 5.2 Ablation Study

To gain a comprehensive understanding of the individual contribution of each component within RAAT to the overall performance, we conducted an ablation study by removing the regularization term loss (denoted as w/o  $\mathcal{L}_{reg}$ ) and the noise-aware classification loss (denoted as w/o  $\mathcal{L}_{cls}$ ). The experimental results are shown in Table 3. After removing the classification loss, we observe that the average performance of the model decreased by 0.49% and 0.52% in terms of F1 score and EM score, respectively. While removing the regularization term, there was a significant performance decrease in handling *irrelevant retrieval noise*.

## 5.3 Further Discussion

**What types of adversarial samples does RAAT employ during training?** To gain a comprehensive understanding of the underlying mechanisms of RAAT, particularly its utilization of specific retrieved data to augment model robustness, we undertook an in-depth examination of its training process, involving meticulously tracking the training iterations and conducting a thorough statistical analysis to quantify the number of different types of adversarial examples incorporated during the training phase. The statistical results are illustrated

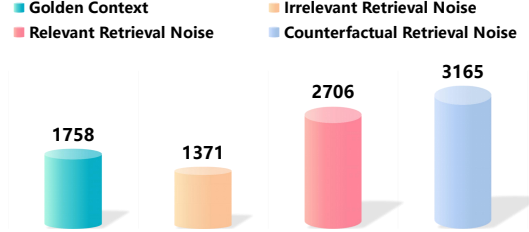


Figure 4: The number of queries and parameter updates are 4,500 and 9,000, respectively. The statistical content in this table pertains to different types of retrieval noises selected by RAAT each time the model parameters undergo an update.

in Figure 4. We observe that RAAT prioritizes the selection of adversarial examples that can significantly improve model robustness, as reflected in its tendency to choose certain types of adversarial examples. This is consistent with our empirical findings described in Section 3.2. RAAT tends to select adversarial examples associated with relevant retrieval noise and counterfactual retrieval noise for training.

## 6 Conclusion

This work initially investigated retrieval noises in RALMs and categorized them into three distinct types, reflecting real-world environments. In addition, we introduced RAAT as a solution to address the noise robustness challenges faced by RALMs, which leveraged adaptive adversarial learning and multi-task learning to enhance the model’s capability. Moreover, we established a benchmark to verify the effectiveness of RAAT based on three open-domain QA datasets. Experimental results demonstrate substantial improvements in F1 and EM scores for the LLaMA2 7B model fine-tuned with RAAT across diverse noise conditions.



## 7 Limitations

In this section, we delve into the limitations inherent in our work, with the objective of pinpointing areas for refinement and bolstering the performance of our model in future endeavors. Two principal limitations have been identified. Firstly, the benchmark constructed for our experiments relies exclusively on datasets sourced from three open-domain question answering repositories. Going forward, we intend to compile additional high-quality datasets from varying NLP tasks and endeavor to retrieve texts from a more extensive array of knowledge bases. This strategic expansion aims to facilitate the creation of a more diversified and expansive benchmark tailored for evaluating the retrieval noise robustness of large language models. Secondly, within the framework of RAAT, our efforts have been singularly concentrated on fortifying the retrieval noise robustness at the LLM end. However, the prospect of jointly training large language models and retrieval models emerges as a promising avenue for enhancing the overall robustness of RALMs. Although this dimension was not the primary focal point of our current work, in our subsequent investigations into retrieval noise robustness, we plan to delve into this avenue. This approach would facilitate the synchronized progress of both the large language model and the retrieval model, contributing to an overall improvement in their robustness.

## Acknowledgments

Min Yang was supported by National Key Research and Development Program of China (2022YFF0902100), National Natural Science Foundation of China (62376262), the Natural Science Foundation of Guangdong Province of China (2024A1515030166), Shenzhen Science and Technology Innovation Program (KQTD20190929172835662), Shenzhen Basic Research Foundation (JCYJ20210324115614039). This work was supported by Alibaba Group through Alibaba Innovative Research Program, Postdoctoral Fellowship Program of CPSF (GZC20232873), Guangdong Basic and Applied Basic Research Foundation (2023A1515110718 and 2024A1515012003).

## References

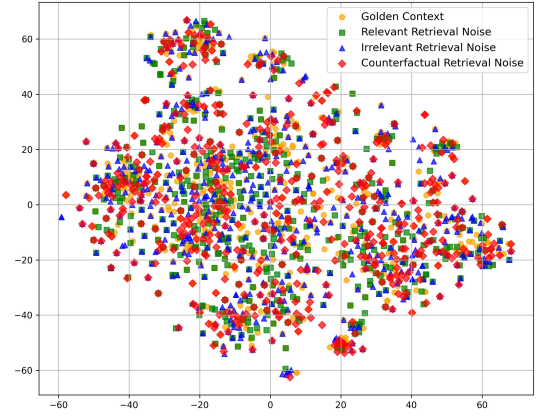
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*.
- 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.
- 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.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrmke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation. *arXiv preprint arXiv:2309.01431*.
- Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.
- 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*.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2023. Knowledge is a region in weight space for fine-tuned language models. *arXiv preprint arXiv:2302.04863*.
- Maor Ivgi and Jonathan Berant. 2021. Achieving model robustness through discrete adversarial training. *arXiv preprint arXiv:2104.05062*.

- Neel Jain, Ping-yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R Bartoldson, Bhavya Kaikhura, Avi Schwarzschild, Aniruddha Saha, et al. 2023. Neftune: Noisy embeddings improve instruction finetuning. *arXiv preprint arXiv:2310.05914*.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.
- Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.
- Tom Kwiakowski, 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.
- 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.
- Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2022. Large language models with controllable working memory. *arXiv preprint arXiv:2211.05110*.
- Ke Liang, Yue Liu, Sihang Zhou, Wenxuan Tu, Yi Wen, Xihong Yang, Xiangjun Dong, and Xinwang Liu. 2023. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Transactions on Knowledge and Data Engineering*.
- 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*.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.
- Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*.
- Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.
- John X Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*.
- Shiwen Ni, Jiawen Li, Min Yang, and Hung-Yu Kao. 2023. Dropattack: A random dropped weight attack adversarial training for natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.
- Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979*.
- 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.
- Nandan Thakur, Luiz Bonifacio, Xinyu Zhang, Odunayo Ogundepo, Ehsan Kamaloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, et al. 2023. Nomiracle: Knowing when you don’t know for robust multilingual retrieval-augmented generation. *arXiv preprint arXiv:2312.11361*.
- 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*.
- Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2020. Infobert: Improving robustness of language models from an information theoretic perspective. *arXiv preprint arXiv:2010.02329*.

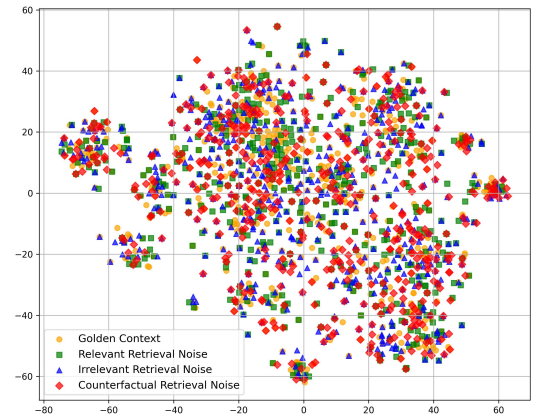
- Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783.
- Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2017. Robust multilingual part-of-speech tagging via adversarial training. *arXiv preprint arXiv:1711.04903*.
- Xunjian Yin, Baizhou Huang, and Xiaojun Wan. 2023. Alcuna: large language models meet new knowledge. *arXiv preprint arXiv:2310.14820*.
- 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*.
- Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.
- Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.
- Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

## A Has the Model Truly Attained Noise Awareness?

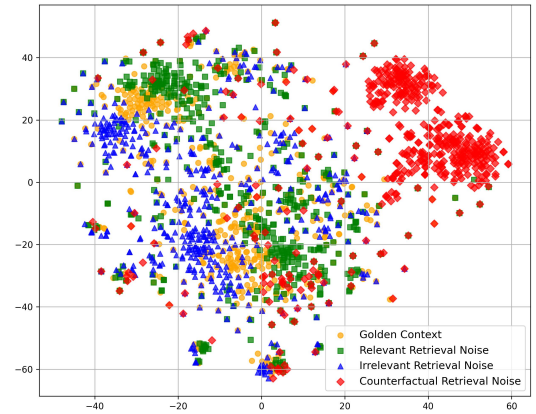
Our preliminary investigation focused on the intrinsic capability of **RALM<sub>golden</sub>** and **RetRobust** to classify the types of retrieval noises. Drawing inspiration from previous work (Gueta et al., 2023), we approached this matter through the application of clustering algorithms. The results, illustrated in Figure 5, reveal suboptimal clustering of text vectors from **RALM<sub>golden</sub>** and **RetRobust**, suggesting that the internal representations for noise classification in these models may lack clarity. Consequently, we introduced a noise classification loss  $\mathcal{L}_{cls}$  into our RAAT method. The experimental results demonstrated tangible benefits with the incorporation of the classification loss. Additionally, we assessed the clustering effectiveness in models fine-tuned with RAAT, observing minimal distances among samples of irrelevant, relevant, and no retrieval noises, in contrast to the considerable distance from counterfactual retrieval noise samples. In particular, counterfactual retrieval noise posed the most significant challenge to LLMs; however, after RAAT tuning, it exhibited superior clustering and representation learning outcomes, indirectly validating the efficacy of RAAT.



a) RALM<sub>golden</sub>



b) RetRobust



c) RAAT

Figure 5: The results of T-SNE visualization. Following the introduction of four types of adversarial samples (i.e., retrieval noises) into models tuned by various methods, the hidden state of the last token is extracted. Subsequently, dimensionality reduction using t-SNE, clustering, and visualization are performed. This visual representation includes three methods, namely RALM<sub>golden</sub>, RetRobust, and RAAT.