Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models

Anonymous ARR submission

Abstract

Retrieval-augmented language model (RALM) represents a substantial advancement in the capabilities of large language models, notably in reducing factual hallucination by leveraging external knowledge sources. However, the reliability of the retrieved information is not always guaranteed. The retrieval of irrelevant data can lead to misguided responses, and potentially causing the model to overlook its inherent knowledge, even when it possesses adequate information to address the query. Moreover, standard RALMs often struggle to assess whether they possess adequate knowledge, both intrinsic and retrieved, to provide an accurate answer. In situations where knowledge is lacking, these systems should ideally respond with "unknown" when the answer is unattainable. In response to these challenges, we introduces CHAIN-OF-NOTE (CON), a novel approach aimed at improving the robustness of RALMs in facing noisy, irrelevant documents and in handling unknown scenarios. The core idea of CON is to generate sequential reading notes for retrieved documents, enabling a thorough evaluation of their relevance to the given question and integrating this information to formulate the final answer. We employed ChatGPT to create training data for CON, which was subsequently trained on an LLaMa-2 7B model. Our experiments across four open-domain QA benchmarks show that RALMs equipped with CON significantly outperform standard RALMs.

1 Introduction

Retrieval-augmented language models (RALMs) represent a novel framework that significantly advances large language models (Touvron et al., 2023; OpenAI, 2023) by addressing key limitations such as reducing factual hallucinations (Ji et al., 2023; Zhang et al., 2023a), injecting up-to-date knowledge in a plug-and-play manner (Dhingra et al., 2022; Vu et al., 2023), and enhancing domain-specific expertise (Li et al., 2023; Qin et al., 2023).

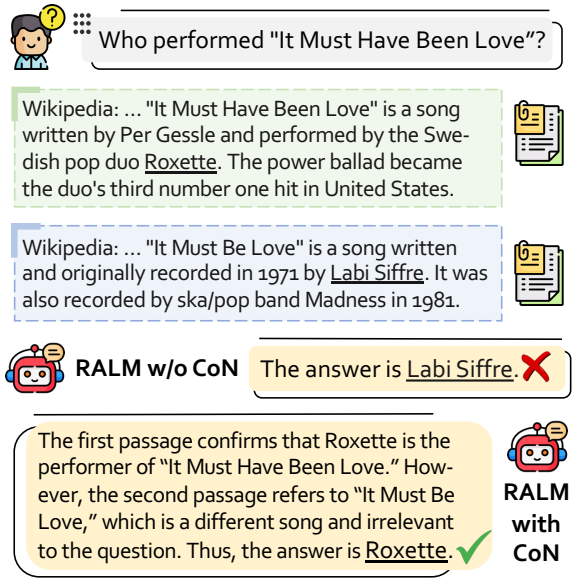


Figure 1: Compared with the current RALMs, the core idea behind Chain-of-Note (CoN) is to generate sequential reading notes for the retrieved documents, ensuring a systematic assessment of their relevance to the input question before formulating a final response.

These enhancements primarily stem from integrating large language models (LLMs) with external knowledge sources (Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022; Shi et al., 2023c). In a typical RALM setup, a query is first processed by a retriever that searches a vast evidence corpus for pertinent documents. A reader then examines these documents, extracting useful information and formulating the final output answer.

However, there exist several issues with the current RALM framework. First, there is no guarantee that the information retrieval (IR) system will always yield the most pertinent or trustworthy information. The retrieval of irrelevant data can lead to misguided responses (Shi et al., 2023a; Yoran et al., 2023), and potentially causing the model to overlook its inherent knowledge, even when it possesses adequate information to address the query (Mallen et al., 2023). Secondly, state-of-the-art LLMs of-

ten hallucinate when addressing fact-oriented questions, a deficiency that can be risky and may discourage users (Ji et al., 2023; Zhang et al., 2023a). Ideally, an intelligent system should be capable of determining whether it has enough knowledge, both intrinsic and retrieved, to provide an accurate answer. In cases where knowledge is insufficient, the system should respond with "unknown" when the answer cannot be determined. Based on the shortcomings of the standard RALM system, in this paper, we aims to improve the robustness of RALMs, mainly focusing on two pivotal aspects:

(1) Noise Robustness: The ability of a RALM to discern and disregard noisy information present in irrelevant retrieved documents, while appropriately leveraging its intrinsic knowledge.

(2) Unknown Robustness: The capacity of a RALM to acknowledge its limitations by responding with "unknown" when given a query it does not have the corresponding knowledge to answer, *and* the relevant information is not found within the retrieved documents.

In this work, we introduce a novel framework named CHAIN-OF-NOTE (CON), designed to enhance the robustness of RALMs. The cornerstone of CON is to generate a series of reading notes for retrieved documents, enabling a comprehensive assessment of their relevance to the input query. This approach not only evaluates each document's pertinence but also pinpoints the most critical and reliable information therein. This process effectively filters out irrelevant or less credible content, leading to responses that are more precise and contextually relevant, as exemplified in Figure 1. Besides, CON enhances the capability of RALM to handle queries fall outside the scope of training data. In cases where the retrieved documents do not provide any relevant information, CON can guide the model to acknowledge its limitations and respond with an "unknown" or provide possible explanation based on available data, enhancing reliability.

To validate the effectiveness of the CON idea, we first prompt ChatGPT (OpenAI, 2023) to generate a 10K training data based on questions collected from Natural Questions (NQ) (Kwiatkowski et al., 2019). Subsequently, we trained a LLaMa-2 7B model to incorporate the note-taking ability integral to CON. Our evaluation of the RALM, integrated with CON and compared to the standard RALM system, focused on three major aspects: (1) overall QA performance using DPR-retrieved documents, (2) noise robustness, assessed by introducing noisy information to the system, and (3) unknown robustness, evaluated through queries not covered in the LLaMa-2 pre-training data, i.e., real-time questions. The evaluations were conducted on the NQ and three additional out-of-domain open-domain QA datasets, namely TriviaQA (Joshi et al., 2017), WebQ (Berant et al., 2013), and RealTimeQA (Kasai et al., 2023). Our experiments show that (CON not only improves overall QA performance when employed with DPR-retrieved documents but also significantly enhances robustness in both noise and unknown aspects. This includes a +7.9 increase in accuracy (measured by the exact match score) with noisy retrieved documents, and a +10.5 increase in the rejection rate for real-time questions[1] that are beyond the pre-training knowledge scope.

2 Related Work

2.1 Robustness of Retrieval-Augmented Language Models

Retrieval-Augmented Language Models (RALMs) represent a significant advancement in natural language processing, combining the power of large language models with the specificity and detail provided by external knowledge sources (Guu et al., 2020; Lewis et al., 2020; Izacard et al., 2022). Recent studies highlight the impact of context relevance on language model performance (Creswell et al., 2022; Shi et al., 2023a; Yoran et al., 2023). Notably, Creswell et al. (2022) demonstrated that incorporating random or irrelevant contexts could adversely affect QA performance. In contrast, Shi et al. (2023a) discovered that adding irrelevant context to exemplars or task-specific instructions can sometimes enhance model performance, implying that models might intrinsically possess capabilities, developed during pre-training, to manage such scenarios. Most pertinent to our research is the study by Yoran et al. (2023), which focused on training RALMs to disregard irrelevant contexts. This approach, while distinct from our proposed solution, underscores the importance of context relevance in enhancing the effectiveness of RALMs.

2.2 Chain-of-X Approaches in Large Language Models

Recent research shows that large language models (LLMs) are capable of decomposing complex prob-

[1]We use real-time questions collected from RealTimeQA after May 2023, which was not trained by LLaMa-2.

lems into a series of intermediate steps, pioneered by the concept of Chain-of-Thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022). The CoT approach mirrors human problem-solving methods, where complex issues are broken down into smaller components. By doing so, LLMs can tackle each segment of a problem with focused attention, reducing the likelihood of overlooking critical details or making erroneous assumptions. This sequential breakdown makes the reasoning process more transparent, allowing for easier identification and correction of any logical missteps.

The CoT methodology has been effectively applied in various contexts, including multi-modal reasoning (Zhang et al., 2023b), multi-lingual scenarios (Shi et al., 2023b), and knowledge-driven applications (Wang et al., 2023b). And additionally, there has been a surge in the development of other chain-of-X methods, addressing diverse challenges in LLM applications. These include chain-of-explanation (Huang et al., 2023), chain-of-knowledge (Wang et al., 2023a), chain-of-verification (Dhuliawala et al., 2023) and IR chain-of-thought (Trivedi et al., 2023). For instance, Chain-of-Verification (Dhuliawala et al., 2023) generates an initial response, formulates verification questions, and revises the response based on these questions, reducing factual errors and hallucinations in the response. Closely related to our work is IR chain-of-thought (Trivedi et al., 2023), which employs CoT to infer and supplement unretrieved information, thereby improving the accuracy of complex reasoning tasks. While chain-of-**X** approaches have shown promise in enhancing LLMs' performance across various domains, their application in RALMs, particularly for improving robustness in noisy and unknown scenarios, is relatively unexplored. This gap signifies further research in applying these strategies to augment RALMs, thereby enhancing their robustness and reliability.

3 Proposed Method

3.1 Overview

In this section, we introduce CHAIN-OF-NOTE, an innovative advancement for retrieval-augmented language models (RALMs). Specifically, CON framework generates sequential reading notes for the retrieved documents, which enables a systematic evaluation of the relevance and accuracy of information retrieved from external documents. By creating sequential reading notes, the model not only assesses the pertinence of each document to the query but also identifies the most critical and reliable pieces of information within these documents. This process helps in filtering out irrelevant or less trustworthy content, leading to more accurate and contextually relevant responses.

3.2 Background of Existing RALMs

RALMs signify a transformative development in language models, enhancing their output by incorporating external knowledge. These models operate by introducing an auxiliary variable, denoted as d, which represents retrieved documents. This inclusion allows them to consider a range of possible documents, thereby producing responses that are more informed and precise (Lazaridou et al., 2022; Shi et al., 2023c). The RALM models can be represented as $p(y|x) = \sum_i p(y|d_i, x)p(d_i|x)$. Here, x represents the input query, and y signifies the model's generated response. In practice, it is infeasible to compute the sum over all possible documents due to the vast number of potential sources. Consequently, the most common approach involves approximating the sum over d using the k highest ranked documents, and providing all these documents as part of the input. We assume, w.l.o.g., that these documents are $[d_1, \ldots, d_k]$, yielding $p(y|x) = \sum_{i=1}^{k} p(y|d_i, x)p(d_i|x)$.

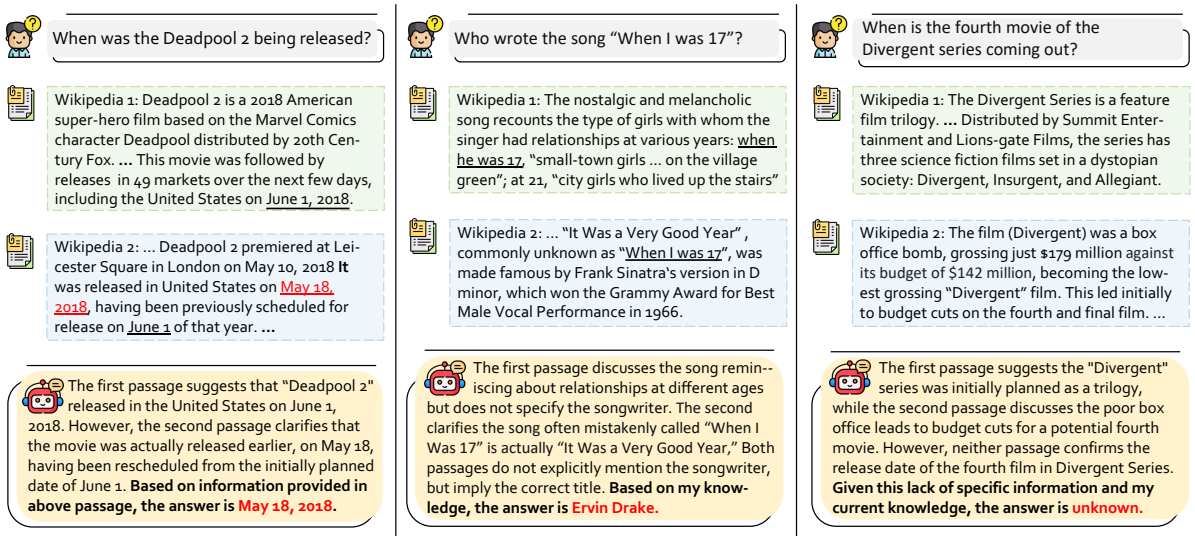
However, the existing RALMs suffer from several limitations:

• Risk of Surface-Level Processing: When directly generating an answer, language models might rely on surface-level information without deep comprehension. Thus, they could easily overlook the nuances of question or documents, particularly in complex or indirect questions.

• Difficulty in Handling Contradictory Information: When faced with documents containing contradictory information, directly generating an answer becomes challenging. The model may struggle of these contradictions or to determine which piece of information is more credible or relevant.

• Reduced Transparency and Interpretability: Direct answer generation offers limited insight into how the model arrived at its conclusion. This lack of transparency makes it challenging for users to understand the basis of the model's conclusions.

• Overdependence on Retrieved Documents: Direct generation can lead to an overreliance on the content of the retrieved documents (i.e. tendency to extract information from retrieved documents (Shi



(a) CoN: Relevant -> Find the answer (b) CoN: Irrelevant -> Infer the answer (c) CoN: Irrelevant -> Answer *Unknown*

Figure 2: Illustration of the CHAIN-OF-NOTE (CON) framework with three distinct types of reading notes. Type (a) depicts the scenario where the language model identifies a document that directly answers the query, leading to a final answer formulated from the retrieved information. Type (b) represents situations where the retrieved document, while not directly answering the query, provides contextual insights, enabling the language model to integrate this context with its inherent knowledge to deduce an answer. Type (c) illustrates instances where the language model encounters irrelevant documents and lacks the necessary knowledge to respond, resulting in an "unknown" answer. This figure exemplifies the CoN framework's capability to adaptively process information, balancing direct information retrieval, contextual inference, and the recognition of its knowledge boundaries.

et al., 2023a)), ignoring the model's inherent knowledge base. This can be particularly limiting when the retrieved documents are noisy or out-of-date.

3.3 The Chain-of-Note Framework

The CHAIN-OF-NOTE (CON) framework presents a solution to the challenges faced by retrieval-augmented language models (RALMs). This framework significantly enhances the ability of RALMs to critically assess retrieved documents through a structured note-taking process. Specifically, it involves generating concise and contextually relevant summaries or notes for each document. This method allows the model to systematically evaluate the relevance and accuracy of information drawn from external documents. By creating sequential reading notes, CON not only assesses the pertinence of each document to the query but also pinpoints the most reliable information and resolves conflicting information. This approach effectively filters out irrelevant or less trustworthy content, leading to responses that are both more accurate and contextually relevant.

Given an input question x and k retrieved documents $[d_1, \cdots, d_k]$, the model aims to generate textual outputs comprising multiple segments $[y_{d_1}, \cdots, y_{d_k}, y]$. Here, y_{d_i} signifies the tokens for the i-th segment, representing the reading note for the corresponding document d_i, as shown in Figure 2. After generating individual reading notes, the model synthesizes the information to create a consolidated final response y. The implementation of the CHAIN-OF-NOTE (CON) involves three key steps: (1) designing the notes y_{d_i}, (2) collecting the data, and (3) training the model.

3.3.1 Chain-of-Note Format Design

The framework primarily constructs three types of reading notes, as shown in Figure 2 , based on the relevance of the retrieved documents to the input question: First, when a document directly answers the query, the model formulates the final response based on this relevant information, as shown in Figure 2(a). Second, if the retrieved document does not directly answer the query but provides useful context, the model leverages this information along with its inherent knowledge to deduce an answer, as shown in Figure 2(b). Third, in cases where the retrieved documents are irrelevant, and the model lacks sufficient knowledge to answer, it defaults to responding with "unknown", as shown in Figure 2(c). This nuanced approach mirrors human information processing, striking a balance between direct retrieval, inferential reasoning, and the ac-

knowledgment of knowledge gaps.

3.3.2 Data Collection

To equip the model with the ability to generate such reading notes, it's essential to gather appropriate training data. Manual annotation for each reading note is resource-intensive, so we employ a state-of-the-art language model – ChatGPT – to generate the notes data. This method is both cost-effective and enhances reproducibility. We initiate this process by randomly sampling 10k questions from the NQ (Kwiatkowski et al., 2019) training dataset. ChatGPT is then prompted with specific instructions and in-context examples to the three distinct types of note generation (detailed in Appendix A.4). The quality of ChatGPT's predictions is subsequently assessed through human evaluations on a small subset of the data before proceeding to the entire set. The NQ dataset is chosen as our primary dataset due to its diverse range of real user queries from search engines. However, to ensure the model's adaptability, we also test its performance on three additional open-domain datasets, including TriviaQA, WebQ, and RealTimeQA, showing its generalization capabilities to out-of-domain (OOD) data.

3.3.3 Model Training

After collecting 10K training data from ChatGPT, the next step is to use them to train our CHAIN-OF-NOTE (CON) model, which is based on an open-source LLaMa-2 7B (Touvron et al., 2023) model. To do this, we concatenate the instruction, question and documents as a prompt and train the model to generate notes and answer in a standard supervised way. Our in-house LLaMa-2 7B model learns to sequentially generate reading notes for each document to assess their relevance to the input query. Responses are generated based on the document's relevance, enhancing accuracy and reducing misinformation. If all documents are irrelevant, the model either relies on inherent knowledge for an answer or responds with "unknown" if the answer cannot be determined accurately.

Weighted Loss on Notes and Answers. A unique aspect of our training approach is the implementation of a weighted loss strategy. This involves varying the loss weights assigned to reading notes and answers. In our preliminary studies, we observed that assigning equal loss to both components can reduce the quality of the final answer and prolong the training time for convergence. This issue arises mainly because notes, being lengthier, contribute disproportionately to the loss. To overcome the drawback, we alternate the focus of the loss function: 50% of the time, the next token prediction loss is computed on the entire notes and answer sequence $[y_{d_1}, \cdots, y_{d_k}, y]$, and the remaining 50% of the time, the next token prediction loss is calculated solely on the answer y. For more mathematical details, please refer to Appendix A.5. This strategy is designed to ensure that while the model learns to generate contextually rich reading notes, the primary focus remains on the accuracy and reliability of the final answer.

Datasets	Full size	IR Recall	Subset size
NQ	3,610	73.82	2,086
TriviaQA	7,993	89.95	7,074
WebQ	2,032	64.22	1,231

Table 1: Dataset statistics. The recall evaluation is based on DPR retrieval on the full test set..

4 Experiments

4.1 Experimental Settings and Evaluations

4.1.1 Datasets and Splits

We conducted comprehensive experiments using three benchmark datasets in open-domain question answering (QA): NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and WebQ (Berant et al., 2013), with further details provided in Appendix A.2. Additionally, we employed RealTimeQA (Kasai et al., 2023) as a special case to evaluate "unknown" robustness.

The evaluation was conducted based on two evaluations sets: **full set and subset evaluation.** Firstly, akin to traditional open-domain QA evaluation, we assessed the models using all questions from the test set to evaluate the **overall QA performance**. The documents were retrieved using DPR, and the top-k documents were fed into the generator. We adhered to the same test splits for the open-domain QA setting as used by Izacard and Grave (2021); Karpukhin et al. (2020). For TriviaQA, evaluations from LLaMa-2 (Touvron et al., 2023) were conducted on the Wikipedia dev set comprising 7,993 examples. Therefore, we also follow the same evaluation on this dev set to facilitate comparisons with their performance. Secondly, to assess the model's **noise robustness and unknown robustness**, we extracted subsets from the above test sets that contained relevant documents

Models	NQ		TriviaQA		WebQ		Average	
	EM	F1	EM	F1	EM	F1	EM	F1
LLaMa-2 w/o IR	28.80	37.53	63.19	68.61	28.30	42.77	35.98	44.27
DPR + LLaMa-2	47.39	55.81	74.92	81.53	29.58	43.51	48.49	56.97
+ CHAIN-OF-NOTE	48.92	57.53	76.27	82.25	32.33	46.68	50.46	58.78
	(+1.53)	**(+1.72)**	**(+1.35)**	**(+0.72)**	**(+2.75)**	**(+3.17)**	**(+1.97)**	**(+1.81)**

Table 2: Overall QA Performance on the entire test sets. Equipped with the same retrieved documents, our CHAIN-OF-NOTE outperforms the standard RALM system on three open-domain QA datasets.

Models	Noise Ratio	NQ		TriviaQA		WebQ		Average	
		EM	F1	EM	F1	EM	F1	EM	F1
LLaMa-2 w/o IR	-	42.89	49.44	67.76	72.80	40.29	56.44	50.31	59.56
DPR + LLaMa-2	100%	34.28	41.74	55.30	61.67	29.58	46.34	39.72	49.92
+ CHAIN-OF-NOTE		41.83	49.58	64.30	70.00	36.85	53.07	47.66	57.55
		(+7.55)	**(+7.84)**	**(+9.00)**	**(+8.33)**	**(+7.27)**	**(+6.73)**	**(+7.94)**	**(+7.63)**
DPR + LLaMa-2	80%	54.28	61.03	73.83	80.02	35.46	52.70	54.52	64.58
+ CHAIN-OF-NOTE		56.63	63.23	75.89	81.24	40.60	56.54	57.70	67.00
		(+2.35)	**(+2.20)**	**(+2.06)**	**(+1.22)**	**(+5.14)**	**(+3.84)**	**(+3.18)**	**(+2.42)**
DPR + LLaMa-2	60%	61.44	67.94	78.44	83.65	37.01	54.16	58.96	68.58
+ CHAIN-OF-NOTE		63.43	69.33	78.79	84.07	41.26	56.91	61.16	70.10
		(+1.99)	**(+1.39)**	**(+0.35)**	**(+0.42)**	**(+4.25)**	**(+2.75)**	**(+2.20)**	**(+1.52)**
DPR + LLaMa-2	40%	64.62	71.12	80.56	86.76	38.40	55.60	61.19	71.16
+ CHAIN-OF-NOTE		65.91	72.22	81.72	87.11	42.16	58.15	63.26	72.49
		(+1.29)	**(+1.10)**	**(+1.16)**	**(+0.35)**	**(+3.76)**	**(+2.55)**	**(+2.07)**	**(+1.33)**
DPR + LLaMa-2	20%	67.21	73.69	81.73	87.89	39.95	56.66	62.96	72.75
+ CHAIN-OF-NOTE		70.00	76.08	82.86	88.24	44.36	60.13	65.74	74.82
		(+2.79)	**(+2.39)**	**(+1.13)**	**(+0.35)**	**(+4.41)**	**(+3.47)**	**(+2.78)**	**(+2.07)**
DPR + LLaMa-2	0%	69.23	75.57	83.34	89.44	42.24	58.59	64.93	74.53
+ CHAIN-OF-NOTE		73.28	79.86	83.52	88.94	46.16	62.38	67.65	77.06
		(+4.05)	**(+4.29)**	**(+0.18)**	**(-0.50)**	**(+3.92)**	**(+3.79)**	**(+2.72)**	**(+2.53)**

Table 3: Evaluation on Noise Robustness. The CHAIN-OF-NOTE framework shows superior performance compared to the standard RALM system, particularly notable at higher noise ratios.

in the retrieved list. We then enumerated each retrieved document to determine if it was a golden document for the given question. Based on the noise ratio r, for instance, if the top-k documents are needed for the generator, then $k \cdot r$ would be the number of noisy documents, and $k \cdot (1 - r)$ would be the number of relevant documents. For example, when noise ratio is 20% and top-5 documents are needed, then 4 are relevant documents, and 1 is irrelevant documents. During the enumeration of the retrieved documents, we populated two lists; when one list reached its limit, we stopped adding more documents to that list until both lists were complete. In instances where no relevant documents are retrieved by the DPR for certain questions, we exclude these from robustness evaluation. Therefore, the subset is smaller than the original test set, as shown in Table 1.

Models	Noise type	RealTimeQA EM	F1	RR
DPR + LLaMa-2	Retrieval	15.62	19.85	6.07
+ CHAIN-OF-NOTE	noise	15.72	20.31	12.95
DPR + LLaMa-2	Random	14.52	18.69	7.16
+ CHAIN-OF-NOTE	noise	15.53	20.22	17.65

Table 4: Evaluation on Unknown Robustness. The CON shows better performance than standard RALM system.

4.1.2 Baseline Methods

For fair comparability, we trained all models using the same training set, with the main difference being in the input and output formats. As outlined in the methods section, we denote an input question as x and its corresponding answer as y. Besides, d_i represents the i-th retrieved document, and y_{d_i} is the associated reading note for that document. Here we show the difference of methods to compare.

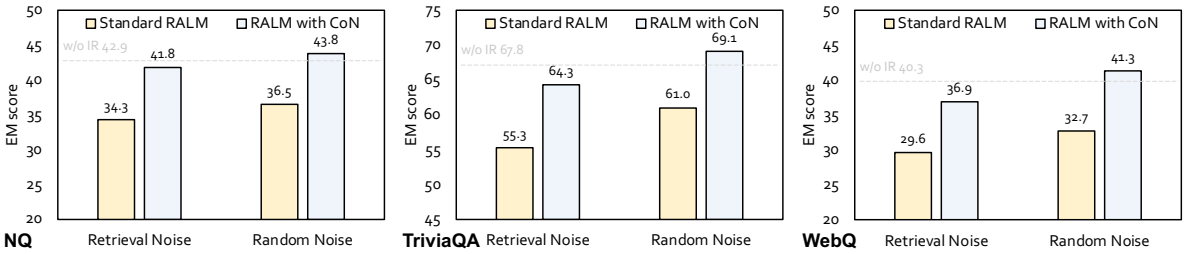


Figure 3: Evaluation on Noise Robustness with two different scenarios: noisy documents obtained through retrieval and completely random documents sampled from the entire Wikipedia.

LLaMa-2 w/o IR: This model is trained to directly generate an answer from the input question, without relying on any external retrieved information. Essentially, it learns the function $f : x \rightarrow y$, transforming a given question x directly to answer y.

DPR + LLaMa-2: This approach trains the model to generate an answer not only from the question but also by incorporating retrieved documents. It learns the function $f : \{x, d_1, \cdots, d_k\} \rightarrow y$, meaning it transforms the question x and a set of retrieved documents $\{d_1, \cdots, d_k\}$ into an answer y.

DPR + LLaMa-2 with CHAIN-OF-NOTE: In this model, the training process involves generating reading notes for each retrieved document before formulating the final answer. It learns the function $f : \{x, d_1, \cdots, d_k\} \rightarrow \{y_{d_1}, \cdots, y_{d_k}, y\}$, thereby enabling the model to process the question x and retrieved documents $\{d_1, \cdots, d_k\}$ to produce reading notes $\{y_{d_1}, \cdots, y_{d_k}\}$ and the final answer y.

4.1.3 Evaluation Metrics

For the evaluation of open-domain QA performance, we have employed two widely recognized metrics: Exact Match (EM) and F1 score, as suggested by prior work in the Chen et al. (2017); Karpukhin et al. (2020); Zhu et al. (2021). For EM score, an answer is deemed correct if its normalized form – obtained through the normalization procedure delineated by (Karpukhin et al., 2020) – corresponds to any acceptable answer in the provided list. Similar to EM score, F1 score treats the prediction and ground truth as bags of tokens, and compute the average overlap between the prediction and ground truth answer (Chen et al., 2017). Besides, we use reject rate (RR) to evaluate the unknown robustness when given questions beyond a language model's knowledge scope.

4.2 Evaluation on Overall QA Performance

We compared our method and various baselines across three open-domain QA benchmarks, as detailed in Table 2. We noted that RALM (DPR + LLaMa-2) with retrieval functionality consistently outperformed LLaMa-2 without retrieval. This improvement is closely tied to the effectiveness of the retrieval process. As indicated in Table 1, DPR demonstrates markedly superior retrieval performance on the NQ and TriviaQA datasets compared to WebQ. Consequently, the benefits of retrieval are more pronounced on NQ and TriviaQA.

Furthermore, when comparing our enhanced RALM, which integrates CoN, with the standard RALM, our method persistently shows better performance. There is an average improvement of +1.97 in EM scores across all three datasets. Delving deeper, we find that this improvement varies depending on whether DPR successfully retrieves relevant documents. Specifically, the average improvement is +1.2 when DPR retrieves relevant documents and +2.3 when it does not on the NQ dataset. This disparity suggests that our CoN improve RALM's in scenarios where more noisy documents are fetched in the first retrieval stage. This observation aligns with our findings on noise robustness, which are elaborated upon in the subsequent sections detailing our experimental results.

4.3 Evaluation on Noise Robustness

Our evaluation of noise robustness was carried out under two scenarios: using noisy documents obtained through retrieval (by removing relevant documents from the retrieved sets and retaining the top-ranked irrelevant ones) and using completely random documents sampled from the entire Wikipedia. Noisy retrieved documents often contain misleading information due to their semantic similarity to the input question, contrasting with random documents which represent total noise.

Table 3 shows that RALM enhanced with CoN consistently outperforms the standard RALM, especially in scenarios with exclusively noisy documents. An average improvement of +7.9 in EM

Question: When was the last Summer Olympics in the USA **Answer:** 1996
Wikipedia document #1: In Chicago's 2016 Summer Olympics bid, the games would have been held from July 22 to August 7, with the Paralympics held between August 12 and 28. The plan emphasized use of Chicago Park District ... (Wikipedia page: Chicago bid for the 2016 Summer Olympics)
Wikipedia document #2: The 1996 Olympics are the most recent edition of the Summer Olympics to be held in the United States. Los Angeles will host the 2028 Summer Olympics, 32 years after the games were held in Atlanta. ... (Wikipedia page: 1996 Summer Olympics)
(Other retrieved Wikipedia documents are omitted.)
Standard RALM: 2016 ✘
RALM with CHAIN-OF-NOTE: Wikipedia passage #1 discusses Chicago's bid for the 2016 Summer Olympics, but does not mention whether it is a successful bid. Wikipedia passage #2 confirms that the 1996 Summer Olympics were the most recent edition of the Summer Olympics to be held in the United States. Based on the information given in above passages, the answer is 1996 ✔

Table 5: Case Study. Our RALM with CHAIN-OF-NOTE shows a deeper understanding of document relevance to questions compared to Standard RALM, going beyond surface-level terms for more accurate responses.

score on fully noisy documents is observed on three open-domain QA datasets, in average. Experiments with lower noise ratios also consistently demonstrate the improvements brought by CoN, aligning the overall performance with that presented in Table 2. We observed that when presented with entirely noisy documents, both the standard RALM and our CoN performed worse than the original LLaMa-2 without IR. This suggests that RALMs can be misled by noisy information, leading to more hallucinations. However, our model can perform almost as well as the original LLaMa-2 without IR, indicating its noise robustness and its capability to ignore irrelevant information.

Furthermore, our comparison with random noise revealed several important observations. Figure 3 illustrates that both standard RALM and RALM with CoN perform better with random documents than with noisy retrieved ones. This indicates that semantically relevant noisy documents are more likely to mislead the language model into producing incorrect information. Moreover, in both noisy scenarios, our method shows enhanced robustness compared to the standard RALM.

4.4 Evaluation on Unknown Robustness

Table 4 illustrates that our RALM equipped with CoN exhibits superior robustness in handling unknown scenario, particularly evident in the RealTimeQA benchmark. This benchmark falls completely outside the model's domain and contains real-time information that was not part of the LLaMa-2 pre-training data. Despite this, models are still capable of providing correct answers in some cases, as the answers remain consistent over time. In comparison to the standard RALM system, our method shows a significant improvement, exceeding +10.5 in its ability to reject to answer questions in unknown scenario. The evaluation is based on reject rate (RR), i.e., number of rejected questions / total questions. This highlights our model's enhanced capability to discern and disregard information that is unfamiliar or not learned during its initial training phase.

4.5 Case Studies

In Table 5, the question pertains to the most recent Summer Olympics held in the USA. The standard RALM is misled by the mention of "Chicago's bid for the 2016 Summer Olympics." Lacking a deep comprehension of the content, it incorrectly focuses on the more recent year (2016), resulting in an inaccurate answer. In contrast, the RALM with CoN carefully analyzes the information. It notes that while Chicago bid for the 2016 Olympics, there's no confirmation of it being a successful bid. This leads to the correct conclusion that the most recent Olympics in the USA were held in 1996.

5 Conclusion

In this paper, we introduce the CHAIN-OF-NOTING (CoN) framework, a novel methodology designed to enhance the robustness of RALMs. The central concept of CoN revolves around the generation of sequential reading notes for each retrieved document. This process allows for an in-depth assessment of document relevance to the posed question and aids in synthesizing this information to craft the final answer. We utilized ChatGPT to generate the initial training data for CoN, which was further refined using an LLaMa-2 7B model. Our tests across various open-domain QA benchmarks reveal that RALMs integrated with CoN considerably surpass traditional RALMs in performance.

6 Limitations

One significant limitation of the CHAIN-OF-NOTE (CON) approach is the increased inference cost attributed to the generation of sequential notes. This process, while beneficial for assessing the relevance and integrating external knowledge, inevitably leads to prolonged response times. This delay is particularly noticeable in time-sensitive applications, where speed is as crucial as accuracy. Additionally, the efficiency of the CoN system heavily relies on the succinctness and relevance of the notes generated, which may vary depending on the complexity of the retrieved documents.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, 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.

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

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

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Proceedings of the ACM Web Conference 2023*, pages 90–93.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*, pages 874–880.

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*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, pages 1601–1611.

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

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime qa: What's the answer right now? *Advances in Neural Information Processing Systems*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

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. *TACL*, pages 452–466.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

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.

Xianzhi Li, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? an examination on several typical tasks. *arXiv preprint arXiv:2305.05862*.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088.

Kaixin Ma, Hao Cheng, Yu Zhang, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. 2023. Chain-of-skills: A configurable model for open-domain question answering. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistic*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness and limitations 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.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Devendra Singh Sachan, Mike Lewis, Dani Yogatama, Luke Zettlemoyer, Joelle Pineau, and Manzil Zaheer. 2022. Questions are all you need to train a dense passage retriever. *arXiv preprint arXiv:2206.10658*.

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

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2023b. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

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

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

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*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Jianing Wang, Qiushi Sun, Nuo Chen, Xiang Li, and Ming Gao. 2023a. Boosting language models reasoning with chain-of-knowledge prompting. *arXiv preprint arXiv:2306.06427*.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023b. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

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

Donghan Yu, Chenguang Zhu, Yuwei Fang, Wenhao Yu, Shuohang Wang, Yichong Xu, Xiang Ren, Yiming Yang, and Michael Zeng. 2022. Kg-fid: Infusing knowledge graph in fusion-in-decoder for open-domain question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4961–4974.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023a. Generate rather than retrieve: Large language models are strong context generators. *International Conference for Learning Representation (ICLR)*.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023b. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023a. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Training language models with memory augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5657–5673.

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*.

A Appendix

A.1 More Related Work of Retrieval-Augmented Language Model

Retrieval-Augmented Language Models (RALMs) represent a significant advancement in natural language processing, combining the power of large language models with the specificity and detail provided by external knowledge sources (Guu et al., 2020; Lewis et al., 2020; Izacard et al., 2022). These models first leverage a retriever to scan a vast evidence corpus, such as Wikipedia, to identify a set of documents pertinent to the user's query. Following this, a reader component is employed to meticulously analyze these documents and formulate a response. This two-pronged approach ensures both relevance and depth in the generated answers. Recent follow-up work has mainly focused on improving the retriever (Karpukhin et al., 2020; Qu et al., 2021; Sachan et al., 2022; Ma et al., 2023) or the reader (Izacard and Grave, 2021; Cheng et al., 2021; Yu et al., 2022), training the system end-to-end (Lewis et al., 2020; Singh et al., 2021), and integrating the retrieval systems with large-scale black-box language models (Yu et al., 2023a; Shi et al., 2023c; Yu et al., 2023b; Trivedi et al., 2023). Another line of RALMs such as kNN-LM (Khandelwal et al., 2020; Zhong et al., 2022) retrieves a set of tokens and interpolates between the next token distribution and kNN distributions computed from the retrieved tokens at inference. The evolution has also led to the emergence and popularity of retrieval-augmented products, such as ChatGPT plugin, Langchain, and New Bing.

A.2 Dataset Information

– TriviaQA (Joshi et al., 2017) contains a set of trivia questions with answers originally scraped from trivia and quiz-league websites.
– WebQ (Berant et al., 2013) consists of questions selected using Google Suggest API, where the answers are entities in Freebase.
– NQ (Kwiatkowski et al., 2019) were collected from real Google search queries and the answers are one or multiple spans in Wikipedia articles identified by human annotators.

A.3 Implementation Details

In the retrieval phase, we employed DPR (Karpukhin et al., 2020) to retrieve documents from Wikipedia. We accessed the model via direct loading from the official DPR repository hosted on GitHub. Subsequent to retrieval, our fine-tuning process for the LLaMA-2 (Touvron et al., 2023) model runs for 3 epochs with a batch size set to 128, leveraging the DeepSpeed library (Rasley et al., 2020) and the ZeRO optimizer (Ma et al., 2021), with bfloat16 precision. The learning rates are set to $\{1e\text{-}6, 2e\text{-}6, 5e\text{-}6, 1e\text{-}5, 2e\text{-}5\}$, and the empirical results indicated that $5e\text{-}6$ yielded the best model performance, hence we standardized the learning rate for all reported numbers. Greedy decoding is applied during inference on all experiments to ensure deterministic generations.

A.4 Instruction Prompts

(1) For standard RALM, the instruction is:

Task Description: The primary objective is to briefly answer a specific question.

(1) For RALM with CoN, the instruction is:

Task Description:
1. Read the given question and five Wikipedia passages to gather relevant information.
2. Write reading notes summarizing the key points from these passages.
3. Discuss the relevance of the given question and Wikipedia passages.
4. If some passages are relevant to the given question, provide a brief answer based on the passages.
5. If no passage is relevant, direcly provide answer without considering the passages.

A.5 Weighted Loss

In the weighted loss function: 50% of the time, the next token prediction loss is computed on the entire notes and answer sequence $[y_{d_1}, \cdots, y_{d_k}, y]$, and the remaining 50% of time, the next token prediction loss is calculated solely on answer y.

$$\mathcal{L}(\theta) = -[\sum_{t=1}^{|Y|} \log\left(p_\theta(y_t|y_{<t}, X)\right) + \sum_{t=d_k}^{|Y|} \log\left(p_\theta(y_t|y_{<t}, X)\right)],$$

where $|Y|$ is the number of all tokens in the answer sequence $[y_{d_1}, \cdots, y_{d_k}, y]$.