

# QPaug: Question and Passage Augmentation for Open-Domain Question Answering of LLMs

Minsang Kim<sup>1</sup>, Cheoneum Park<sup>2</sup>, Seungjun Baek<sup>1\*</sup>

<sup>1</sup>Korea University, <sup>2</sup>Hanbat National University,  
kmswin1@korea.ac.kr

## Abstract

Retrieval-augmented generation (RAG) has received much attention for Open-domain question-answering (ODQA) tasks as a means to compensate for the parametric knowledge of large language models (LLMs). While previous approaches focused on processing retrieved passages to remove irrelevant context, they still rely heavily on the quality of retrieved passages which can degrade if the question is ambiguous or complex. In this paper, we propose a simple yet efficient method called question and passage augmentation (QPaug) via LLMs for open-domain QA. QPaug first decomposes the original questions into multiple-step sub-questions. By augmenting the original question with detailed sub-questions and planning, we are able to make the query more specific on what needs to be retrieved, improving the retrieval performance. In addition, to compensate for the case where the retrieved passages contain distracting information or divided opinions, we augment the retrieved passages with self-generated passages by LLMs to guide the answer extraction. Experimental results show that QPaug outperforms the previous state-of-the-art and achieves significant performance gain over existing RAG methods. The source code is available at <https://github.com/kmswin1/QPaug>.

## 1 Introduction

Large language models (LLMs) have shown remarkable in-context learning capability (Brown et al., 2020; Touvron et al., 2023) for various real-world applications such as assistant chatbot (Achiam et al., 2023; Team et al., 2023), robot planning (Wang et al., 2023b), search ranking (Sun et al., 2023), and code generation (Chen et al., 2021). However, the knowledge of LLMs is limited to the pre-training corpus, making it difficult

to provide answers to questions on up-to-date information. To overcome such limitations via external knowledge sources, recent works have focused on the retrieval-augmented-generation (RAG) (Guu et al., 2020; Lewis et al., 2020b). The RAG systems consist of *retrievers* which search and retrieve related information from knowledge sources and *readers* which generate responses based on the retrieved information. The RAG approach has been proven effective in leveraging external knowledge to complement the parametric knowledge of LLMs.

Open-domain Question Answering (ODQA) (Yang et al., 2018; Kwiatkowski et al., 2019; Ho et al., 2020; Karpukhin et al., 2020) is one of the NLP tasks most relevant to RAG systems. In the ODQA, retrievers search for relevant passages from questions, and readers answer the questions based on retrieved contexts. Research efforts have been put into enhancing retrievers (Xiong et al., 2020; Izacard et al., 2021) and readers (Kenton and Toutanova, 2019; Lewis et al., 2020a). For example, RAPTOR (Sarathi et al., 2023) is an advanced retriever which recursively captures multiple levels of details of a text using a tree. However, even advanced retrievers may fetch poor passages when *the question is ambiguous or complex*, requiring question clarification (Zamani et al., 2020; Lee et al., 2023; Kim et al., 2023a) or multi-step retrieval (Feldman and El-Yaniv, 2019; Welbl et al., 2018; Trivedi et al., 2022), rendering the retrieval inaccurate and inefficient.

Thus, for tough questions, it is unclear whether the retrieved passages are of high quality, i.e., they contain context relevant to the question, which makes the design of readers challenging. The problem can be alleviated if the parametric knowledge of LLMs can be tuned to extract answers given the relevant retrieved passages (Guu et al., 2020; Izacard and Grave, 2021). However, fine-tuning is often infeasible due to the sheer scale of LLMs, or some LLMs are essentially black-box APIs (Team

\*Corresponding Author

et al., 2023; Achiam et al., 2023). Recently, (Lazaridou et al., 2022; Kim et al., 2023b) have proposed to process retrieved passages via LLM prompting without fine-tuning, but with Internet retrieval and reranking (Lazaridou et al., 2022), or with summarization and verification (Kim et al., 2023b). However, these approaches are still limited in that they heavily rely on the quality of contexts provided by retrieved passages.

In this paper, we deal with question complexity and complement the quality of retrieval by in-context learning. The goal is to harmoniously combine the parametric and non-parametric knowledge of LLMs through prompting. We propose QPaug (pronounced *cue-pug*) which stands for question and passage augmentation. Firstly, the question augmentation is based on the hypothesis that the LLMs can decompose complex questions into multiple easier sub-questions (Kojima et al., 2022). The sub-questions contain fine-grained information and planning as to what knowledge should be retrieved from external sources. The question for retrieval is composed by augmenting the original question with the sub-questions. Secondly, the passage augmentation is based on the hypothesis that the LLMs’ parametric knowledge is the most pragmatic alternative source to non-parametric (retrieved) knowledge to deal with the degraded quality of retrieval. A self-generated passage is composed with respect to the augmented question, leveraging extensive knowledge of LLMs. Also, self-generation can prevent irrelevant retrieved passages from overriding the LLMs’ factual knowledge (Zheng et al., 2023; Wei et al., 2023). We augment the retrieved passages with the self-generated passages. Experiments show that the question and passage augmentation through LLMs complement the retrieved generation well, achieving state-of-the-art performance on ODQA benchmark datasets.

Our contributions are summarized as follows. (i) We propose a simple yet effective framework for the LLM-based question and passage augmentation method (QPaug) based on the prompting of LLMs. (ii) QPaug successfully decomposes and augments questions improving the performance of retrieval. In addition, the self-generated passage by QPaug, when combined with retrieved passages, strengthens the factual knowledge for the reader, achieving large performance gains. (iii) Extensive ablation studies show that QPaug can be integrated with various LLMs and retrievers, exhibiting excellent performances on several ODQA benchmarks.

## 2 Related Work

### 2.1 Open-Domain Question Answering

(Chen et al., 2017) first proposed retrieve-and-read system for solving open-domain question answering tasks. Following conventional lexical-based sparse retriever systems like BM25 (Robertson et al., 2009), DPR (Karpukhin et al., 2020) proposed a *dense passage retrieval* for a semantic retriever system. The semantic retriever is based on sentence embeddings, and there have been a number of works for improving embeddings (Reimers and Gurevych, 2019; Xiong et al., 2020; Izacard et al., 2021; Wang et al., 2022; Gao et al., 2023). By contrast, the reader system which extracts answers from retrieved documents consists of extractive methods such as BERT (Kenton and Toutanova, 2019) or RoBERTa (Liu et al., 2019) and generative methods like BART (Lewis et al., 2020a) or T5 (Raffel et al., 2020).

### 2.2 Retrieval-Augmented Generation

Augmenting language models with retrieved information from external knowledge sources have proven effective for a wide range of NLP tasks (Guu et al., 2020; Lewis et al., 2020b; Borgeaud et al., 2022). In the LLM era, (Lazaridou et al., 2022; Izacard et al., 2023) proposed an in-context learning-based retrieval-augmented generation methods. (Asai et al., 2023) proposed Self-RAG which generates and reflects on retrieved passages and own generated text using reflection tokens. (Sarathi et al., 2023) proposed recursive retrieval methods using embedding, clustering, and summarizing chunks of text, where they construct a tree with differing levels of summarization. (Kim et al., 2023b) proposed summarizing retrieved passages conditioned on candidate answers to select the more relevant context.

### 2.3 Prompting of Large Language Models

GPT-3 (Brown et al., 2020) opened a few-shot learning era of language models. (Si et al., 2022) extensively studied about prompting of GPT-3 using manually designed prompts on diverse tasks. They showed that GPT-3 is more reliable with proper prompts. (Wei et al., 2022; Kojima et al., 2022) proposed chain-of-thoughts (CoT), which decomposes a problem into multi-step subproblems. In addition, there have been modified works of CoT (Wang et al., 2023c; Dhuliawala et al., 2023; Yao et al., 2024; Besta et al., 2024). (Wang et al.,

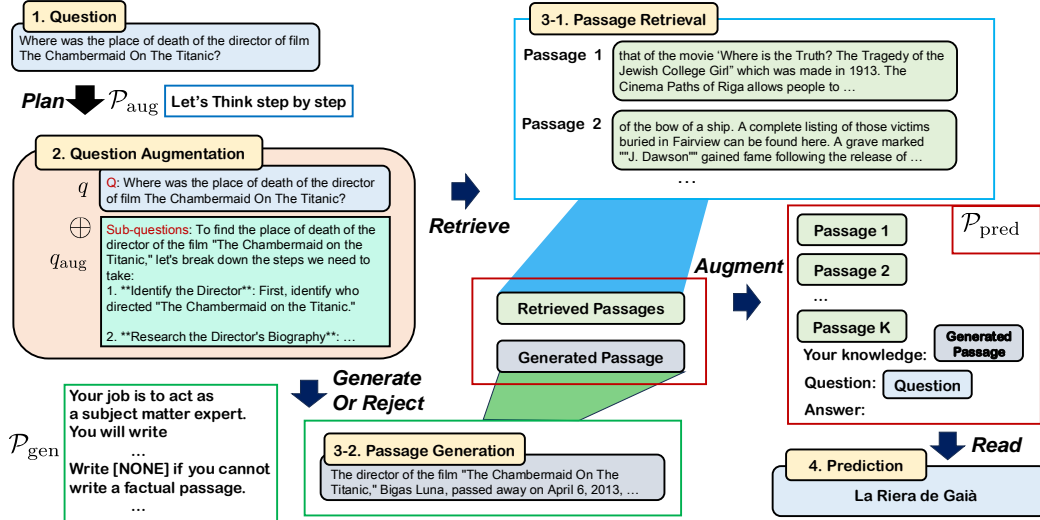


Figure 1: **Step 1.** LLMs receive questions. **Step 2.** The original question is decomposed into multi-step sub-questions on what information should be retrieved. **Step 3-1.** Retrieve passages from augmented questions. **Step 3-2.** Self-generate as a passage from the augmented questions using factual knowledge. The LLM is asked to generate [NONE] if it does not have the required knowledge. **Step 4.** Augment retrieved passages with the self-generated passage, then predict answers.  $\mathcal{P}_x$  denotes the prompt at each step.

2023a) extensively studied the properties of CoT, and illustrated that even incorrect reasoning paths can improve performances of LLM reasonings.

### 3 Method

#### 3.1 Problem Statement and Notations

Open domain question answering (ODQA) is the extension of QA tasks in which explicit evidence or context is not provided to the model. Thus, ODQA requires other knowledge sources such as an external knowledge base. The basic strategy to solve ODQA tasks is *retrieve-and-read*. Specifically, *retriever*  $\mathcal{R}$  obtains top- $K$  related passages  $\mathcal{C}_K$  from knowledge source  $\mathcal{Z}$ :

$$\mathcal{C}_K = \mathcal{R}(q, \mathcal{Z}, K) \quad (1)$$

Then *reader* extracts predicted answer  $\hat{a}$  about question  $q$  using given the retrieved passages  $\mathcal{C}_K$ :

$$\hat{a} = \mathcal{M}(\mathcal{P}(q, \mathcal{C}_K)) \quad (2)$$

where we denote the LLM-generated text as  $\mathcal{M}(\cdot)$ , and  $\mathcal{P}$  denotes certain prompt.

#### 3.2 Question Augmentation

We propose *question augmentation* to enhance the retrieval capability. The original question is decomposed into multiple sub-questions which can arise during the reasoning process. The motivation is that a better retrieval is possible if we augment the original question with those sub-questions

which can provide fine-grained hints for resolving the question. We utilize a zero-shot chain of thoughts (CoT) (Kojima et al., 2022) for composing sub-questions. Specifically, we use prompt  $\mathcal{P}_{\text{aug}}(\cdot)$  for question  $q$ , where  $\mathcal{P}_{\text{aug}}(q)$  adds the sentence “Let’s think step-by-step” (Kojima et al., 2022) to  $q$ .  $\mathcal{P}_{\text{aug}}(q)$  is then passed to LLM to generate augmented question  $q_{\text{aug}}$ , i.e.,

$$q_{\text{aug}} = \mathcal{M}(\mathcal{P}_{\text{aug}}(q)) \quad (3)$$

An example is given in Fig. 1: see Steps 1 and 2. In Step 2, we observe that  $q_{\text{aug}}$  contains fine-grained instructions to tackle the question. Notably, although prior zero-shot CoT (Kojima et al., 2022) is shown to achieve large performance gains on arithmetic reasoning tasks, its effectiveness was less examined on ODQA tasks. However, we observe that augmenting the question with reasoning steps planned out by CoT boosts the retrieval capability, leading to improved performances on ODQA tasks.

Next, the augmented question is used for retrieval. Specifically, we create query  $\hat{q}$  by concatenating  $q$  and  $q_{\text{aug}}$ :

$$\hat{q} = q \oplus q_{\text{aug}} \quad (4)$$

where  $\hat{q}$  is used for the retrieval. We retrieve top- $K$  passages  $c_1, \dots, c_K$  using Approximated Max Inner Product Search (MIPS) (Ram and Gray, 2012; Auvolat et al., 2015).

$$\mathcal{C}_K = \mathcal{R}(\hat{q}, \mathcal{Z}, K) = \{c_1, \dots, c_K\}. \quad (5)$$

### 3.3 Passage Self-Generation

We propose to utilize the vast knowledge of LLMs to complement the contextual information provided by retrieved passages. Specifically, the LLM is asked to self-generate a passage regarding the question. We augment the retrieved passages with the generated passage. Let  $\mathcal{P}_{\text{gen}}(q)$  denote a prompt for passage generation given question  $q$ . We have

$$\hat{c} = \mathcal{M}(\mathcal{P}_{\text{gen}}(\hat{q})) \quad (6)$$

where  $\hat{c}$  denotes the LLM-generated passage. For example, see Step 1 and 2-2 in Fig. 1, and see the outline of prompt  $\mathcal{P}_{\text{gen}}$  for passage generation in Step 2-2.

Finally, we extract predicted answers from the LLM using  $\mathcal{C}_K$  and  $\hat{c}$ . The final prompt  $\mathcal{P}_{\text{pred}}$  for prediction contains  $q$ ,  $\mathcal{C}$ , and  $\hat{c}$ , which is input to LLM for the predicted answer, i.e.,

$$\hat{a} = \mathcal{M}(\mathcal{P}_{\text{pred}}(q, \mathcal{C}_K, \hat{c})) \quad (7)$$

An important consideration is that, the LLM may not have sufficient knowledge for the question, and should refrain from generating plausible but incorrect passages. To derive as truthful passages as possible, we explicitly instruct LLM to base its passage on factual knowledge, and to generate [NONE] in case it does not have sufficient knowledge: see the detailed prompt  $\mathcal{P}_{\text{gen}}$  for passage generation in Table 7 of Appendix A.1. There are two cases as a result.

**LLM admits lack of knowledge.** In this case, LLM returns [NONE] passage, and only the retrieved passages are used at the final *read* step to predict the answer.

**LLM hallucinates.** LLM provides a bogus passage, believing that it is from factual knowledge. To prevent possible hallucinations from mixing with retrieved passages, we label the LLM-generated passage as “Your Knowledge:” in composing the final prompt  $\mathcal{P}_{\text{pred}}$  for the answer prediction: see the box of  $\mathcal{P}_{\text{pred}}$  above Step 4 in Fig. 1. Presumably, if a majority of  $K$  retrieved passages provide relevant information, LLM is likely to recover the correct answer in spite of hallucination. Examples of the above two cases are provided in Table 15 and Table 16 of Appendix A.2.2. Overall, the proposed passage augmentation is more beneficial than harmful, as demonstrated by experiments in the following section.

Dataset	# samples
Passages	21,015,324
NQ	4,289
2wiki	12,576
Hotpot	7,405

Table 1: Dataset Statistics of passages and test sets of ODQA benchmark datasets.

## 4 Experiment

### 4.1 Experimental settings

**Evaluation datasets and metrics.** We experiment with zero-shot QA tasks on three ODQA benchmarks: Natural Questions (NQ) (Kwiatkowski et al., 2019), 2wiki hop questions (2wiki) (Ho et al., 2020), and HotpotQA (Hotpot) (Yang et al., 2018). The test sets from those datasets are used for the experiments. 2wiki and HotpotQA are multi-hop QA datasets that typically require multi-step retrieval for traditional methods. As the knowledge sources for retrieval, we use 21M passages of Wikipedia dump proposed by DPR (Karpukhin et al., 2020). Table 1 shows the dataset statistics of passages and test sets. Since the datasets include both short and long answers in benchmark datasets, answers are not only words but also phrases or sentences. Thus, we use Rouge-L and F1 scores as evaluation metrics to compute fine-grained scores the same as (Nguyen et al., 2016; Kwiatkowski et al., 2019; Ho et al., 2020; Yang et al., 2018). We normalize answers and predictions before computing both metrics similar to (Rajpurkar et al., 2016). All the baseline methods use  $K = 10$  retrieved passages as the input to the reader. For a fair comparison, QPaug uses  $K = 9$  retrieved passages combined with one LLM-generated passage.

**Baseline models.** We experiment with three retrievers: SBERT (Reimers and Gurevych, 2019), ANCE (Xiong et al., 2020), and Contriever (Izacard et al., 2021). In addition, we experiment with three LLMs as readers: Llama-2 (Touvron et al., 2023), GPT-3.5 (Ouyang et al., 2022) and GPT-4 (Achiam et al., 2023). We use the same LLM as the reader model for the proposed question and passage augmentation. We compare QPaug with no retrieval methods, e.g., chain-of-thoughts (Kojima et al., 2022) and Self-verification (Weng et al., 2023) as well as context-augmentation methods such as Rerank (Lazaridou et al., 2022) and SuRE (Kim et al., 2023b).

**Implementation Details.** Greedy decoding is used



Method	NQ	2wiki	Hotpot
No retrieval	37.9	27.1	35.3
Chain-of-thoughts	38.2	28.2	35.5
Rerank	38.0	26.6	33.2
Self-verification	38.4	30.8	35.9
SuRE	40.4	32.6	33.6
<b>QPaug (Proposed)</b>	<b>44.6</b>	<b>35.5</b>	<b>45.1</b>

Table 2: Comparison between QPaug and baseline methods. We use retriever as Contriever for Rerank, SuRE, and QPaug, and use GPT-3.5 as the LLM of all methods. The evaluation metric is the F1 score between answers and predictions. **Bold** indicates the best performance.

for LLM generations in all the experiments. We evaluate zero-shot QA tasks with retrieval where the retrieval is implemented based on (Thakur et al., 2021) with Faiss indexing (Johnson et al., 2019). We use Langchain<sup>1</sup> for LLM inference with manually-designed prompt templates. Detailed prompt templates are provided in Appendix A.1.

## 4.2 Main Results

Table 2 shows the comparison between QPaug and other baseline methods, where Contriever is used as the retriever and GPT-3.5 as the LLM. Experimental results demonstrate that QPaug achieves a performance gain of 10.4% on NQ, 8.9% on 2wiki, and 34.2% on HotpotQA over SuRE which is the current state-of-the-art. In addition, QPaug achieves average performance improvements of 22.6% and 27.9% over Chain-of-thoughts (Kojima et al., 2022) and Self-Verification (Weng et al., 2023) respectively. The results demonstrate that QPaug successfully can retrieve relevant context and generate factual knowledge supporting retrieved passages.

Next, we examine the compatibility of QPaug with various types of retrievers and LLMs. Table 3 summarizes the experimental results on ODQA benchmark datasets with three retrievers and LLMs. Each element indicates the combinations of retrievers (column) and LLMs (row). We first apply QPaug to proprietary LLMs (Achiam et al., 2023). GPT-4 with QPaug achieves performance gain on average from 14.8% to 23.8% in Rouge and from 20.1% to 27.9% in F1 score. In addition, GPT-3.5 achieves a similar performance gain on average of up to 24.5% in Rouge and 27.5% in F1 score when they are combined with QPaug. As an open-source LLM, LLaMA-2-7b-chat (Touvron et al., 2023) achieves slightly lower average gains up to 20.1%

<sup>1</sup><https://www.langchain.com/>

and 23.9% in Rouge and F1 scores respectively. We observe that QPaug consistently improves the performances of ODQA irrespective of the types of LLMs and retrievers.

In particular, QPaug exhibits notable improvements in the multi-hop QA datasets of 2wiki and HotpotQA. GPT-4 with SBERT obtains performance gains of 2.1% in Rouge and 17.4% in F1 score on NQ; however, the same model achieves gains of 34.2% / 37.9% on 2wiki, and 14.5% / 20.1% on HotpotQA. The results show that QPaug achieves particularly large performance gains in multi-hop QA datasets, illustrating its effectiveness in solving complex questions.

## 5 Ablation study

We conduct ablation study on two main components, i.e., question augmentation and passage generation, denoted by *Qaug* and *Pgen* in the following subsections.

### 5.1 Question Augmentation (Qaug)

In this section, we investigate the effect of question augmentation (Qaug). Table 4 shows the change in the retrievers’ performance measured in Recall@10 by adding only Qaug component. We experimented with various LLMs: GPT-4, GPT-3.5 and LLaMA-2. Table 4 demonstrates that Qaug boosts the retrieval performance by up to 30% on average with GPT-4, where the improvements are greater than GPT-3.5. In addition, we experiment with both LLaMA-2 models of 7b and 70b sizes which achieve performance gains in all datasets where 70b size model obtains higher performance gains by a large margin.

Fig. 2 shows the performance gains achieved by Qaug with the varying number of retrieved passages  $K$ . We observe that the performance gain is more pronounced in  $K = 10$  than 50 or 100. This is because, if the number of passages to be retrieved is limited, the relevance and fine-granularity of queries composed by Qaug become more important for better retrieval. We also observe that Qaug improves the retrieval performances with a relatively large number of retrieved passages as well, i.e., with  $K = 50$  or 100.

### 5.2 Passage Self-Generation (Pgen)

In this section, we examine the effect of the passage self-generation (Pgen) component. Table 5, 6 and 7 show the change in performance by adding the self-

LLM	Datasets	No retrieval	SBERT	+ QPaug	ANCE	+ QPaug	Contriever	+ QPaug
GPT-4	NQ	37.0 / 42.0	42.8 / 46.0	<b>43.7 / 54.0</b>	40.8 / 45.0	<b>43.9 / 52.0</b>	39.4 / 43.7	<b>42.4 / 52.0</b>
	2wiki	26.8 / 37.6	27.8 / 36.5	<b>37.3 / 49.1</b>	25.0 / 33.6	<b>36.7 / 47.2</b>	24.6 / 32.6	<b>35.2 / 47.2</b>
	Hotpot	33.9 / 42.2	39.2 / 49.0	<b>44.9 / 54.6</b>	34.6 / 43.6	<b>42.4 / 52.0</b>	34.0 / 43.2	<b>43.7 / 53.5</b>
	Average	32.6 / 40.6	36.6 / 43.8	<b>42.0 / 52.6</b>	33.5 / 40.7	<b>41.0 / 50.4</b>	32.7 / 39.8	<b>40.4 / 50.9</b>
GPT-3.5	Datasets	No retrieval	SBERT	+ QPaug	ANCE	+ QPaug	Contriever	+ QPaug
	NQ	35.1 / 37.9	35.5 / 40.3	<b>40.7 / 44.2</b>	35.5 / 39.8	<b>40.9 / 45.4</b>	35.2 / 37.9	<b>41.6 / 44.6</b>
	2wiki	19.7 / 27.1	20.2 / 30.2	<b>24.0 / 35.3</b>	20.0 / 26.0	<b>24.6 / 35.7</b>	20.3 / 26.4	<b>24.5 / 35.5</b>
	Hotpot	27.4 / 35.3	28.3 / 38.3	<b>38.0 / 44.9</b>	29.0 / 34.8	<b>37.3 / 44.4</b>	28.2 / 33.7	<b>38.4 / 45.1</b>
Average	27.4 / 33.4	28.0 / 36.1	<b>34.2 / 41.8</b>	28.2 / 33.5	<b>34.3 / 41.5</b>	27.9 / 32.7	<b>34.8 / 41.7</b>	
LLaMA-2-7b-chat	Datasets	No retrieval	SBERT	+ QPaug	ANCE	+ QPaug	Contriever	+ QPaug
	NQ	13.0 / 15.5	24.0 / 27.5	<b>29.5 / 34.1</b>	23.3 / 26.5	<b>28.1 / 32.4</b>	20.8 / 24.7	<b>27.9 / 32.0</b>
	2wiki	16.7 / 20.3	20.9 / 23.2	<b>22.0 / 26.9</b>	20.6 / 23.0	<b>22.5 / 25.8</b>	20.3 / 22.5	<b>22.3 / 25.9</b>
	Hotpot	16.1 / 18.6	22.5 / 27.5	<b>24.9 / 33.0</b>	21.5 / 26.9	<b>22.2 / 29.3</b>	20.1 / 24.4	<b>23.3 / 31.0</b>
Average	15.3 / 18.1	22.5 / 26.1	<b>25.5 / 31.3</b>	21.8 / 25.5	<b>24.3 / 29.1</b>	20.4 / 23.9	<b>24.5 / 29.6</b>	

Table 3: Performance (Rouge/F1 score) comparison between no-retrieval, base RAG and QPaug across various retrievers and readers (LLMs). For example, the column labelled “SBERT” means a base RAG with SBERT as the retriever, and column labelled “+QPaug” on the right means we use QPaug as “add-on” to the RAG with SBERT retriever. GPT-4 is used for the question augmentation in QPaug.

Model / Dataset	NQ	2wiki	Hotpot
SBERT	63.41	27.90	47.47
+Qaug by LLaMA-2-7b-chat	63.74 (+0.33)	29.21 (+1.31)	52.07 (+4.60)
+Qaug by LLaMA-2-70b-chat	66.10 (+2.69)	33.56 (+5.66)	56.89 (+9.42)
+Qaug by GPT-3.5	64.98 (+1.57)	30.37 (+2.47)	51.06 (+3.59)
+Qaug by GPT-4	70.20 (+6.79)	38.68 (+10.8)	62.08 (+14.6)
ANCE	60.43	22.57	38.81
+Qaug by LLaMA-2-7b-chat	62.91 (+2.48)	24.04 (+1.47)	43.88 (+5.07)
+Qaug by LLaMA-2-70b-chat	64.58 (+4.15)	26.18 (+3.61)	46.63 (+7.82)
+Qaug by GPT-3.5	62.55 (+2.12)	22.91 (+0.34)	40.85 (+2.04)
+Qaug by GPT-4	66.68 (+6.25)	27.31 (+4.74)	49.59 (+10.8)
Contriever	52.48	21.26	40.54
+Qaug by LLaMA-2-7b-chat	58.10 (+5.62)	25.00 (+3.74)	47.78 (+7.42)
+Qaug by LLaMA-2-70b-chat	62.18 (+9.70)	29.04 (+7.78)	53.19 (+12.7)
+Qaug by GPT-3.5	52.76 (+0.28)	21.51 (+0.26)	42.74 (+2.20)
+Qaug by GPT-4	60.38 (+7.90)	28.95 (+7.69)	53.69 (+13.2)

Table 4: Answer Recall @ 10 of retrieved passages question augmentation.

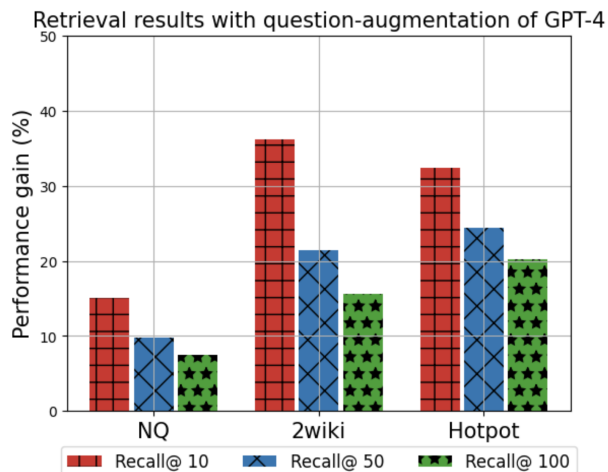


Figure 2: Performance gains of recall @ K with question augmentation of GPT-4. K is 10, 50, and 100. The base retriever is Contriever.

generated passage to retrieved passages, with GPT-4, GPT-3.5 and LLaMA-2-7b-chat respectively.

Table 5 shows that the addition of Pgen with GPT-4 improves performances in all datasets as compared to the baseline using only retrieved passages. In particular, the performance gain is significantly large at 35.1% on 2wiki dataset for which the search/retrieval results tend to be relatively poor due to multi-hop questions. Pgen is particularly effective when the retrieval performance degrades. This demonstrates that the self-generated passage is able to complement the missing context in the retrieved passages. We observe similar trends with other LLMs, e.g., see Table 6 and Table 7 for the effects of Pgen with GPT-3.5 and LLaMA-2-7b-chat. The performance gains with GPT-3.5 are on average from 14.2% to 25.9%, and are on average from 11.6% to 25.9% with LLaMA-2-7b-chat.

Finally, we examine the effect of Pgen by varying the number of retrieved passages  $K$ . Fig. 3 shows the F1 scores per top- $K$  grounded passages. Similar to Fig. 2, experimental results show that Pgen achieves performance gains when the number of retrieved passages increases. Thus, LLM can extract factual knowledge even though there are a number of possibly irrelevant retrieved passages. Also, the results show that the passage generation method can be plugged into various LLMs since it illustrates the same trend for GPT-4, GPT-3.5, and LLaMA-2-7b-chat.

## 6 Qualitative Analysis

**Effectiveness of Qaug.** Table 8 shows the comparison between the retrieval results using the original question versus the augmented question (Qaug)

Model / Dataset	NQ	2wiki	Hotpot
SBERT	46.0	36.5	49.0
+Pgen by GPT-4	<b>47.6</b>	<b>47.8</b>	<b>53.0</b>
ANCE	45.0	33.6	43.6
+Pgen by GPT-4	<b>47.8</b>	<b>45.6</b>	<b>47.4</b>
Contriever	43.7	32.6	43.2
+Pgen by GPT-4	<b>47.2</b>	<b>45.2</b>	<b>47.7</b>

Table 5: F1 score between predictions and answers of GPT-4.

Model / Dataset	NQ	2wiki	Hotpot
SBERT	40.3	30.2	38.3
+Pgen by GPT-3.5	<b>44.0</b>	<b>34.2</b>	<b>43.1</b>
ANCE	39.8	26.0	34.8
+Pgen by GPT-3.5	<b>44.2</b>	<b>35.2</b>	<b>44.1</b>
Contriever	37.9	26.4	33.7
+Pgen by GPT-3.5	<b>43.0</b>	<b>34.8</b>	<b>44.6</b>

Table 6: F1 score between predictions and answers of GPT-3.5.

Model / Dataset	NQ	2wiki	Hotpot
SBERT	27.5	23.2	27.5
+Pgen by LLaMA-2-7b-chat	<b>31.2</b>	<b>26.2</b>	<b>30.1</b>
ANCE	26.5	23.0	26.9
+Pgen by LLaMA-2-7b-chat	<b>28.9</b>	<b>25.6</b>	<b>29.0</b>
Contriever	24.7	22.5	24.4
+Pgen by LLaMA-2-7b-chat	<b>29.7</b>	<b>26.2</b>	<b>29.0</b>

Table 7: F1 score between predictions and answers of LLaMA-2-7b-chat.

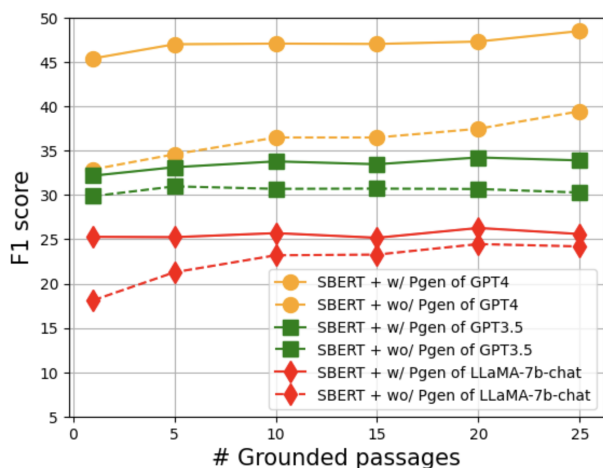


Figure 3: Comparison of F1-scores between RAG and QPaug per top-K grounded passages on 2wiki. Retriever is vanilla SBERT.

on 2wiki. The retrieval by Qaug successfully obtains relevant information containing the answer, while the retrieved passages from the original ques-

tion do not contain relevant clues. As a result, the LLM outputs the correct answer, *Catherine Robbe-Grillet* with Qaug, but outputs *Not mentioned* with the original question (for details, refer to Table 11 in Appendix A.2.1). Thus, Qaug enhances the retrieval of passages with relevant contexts, leading to significant improvement on the performance of ODQA.

**Effectiveness of Pgen.** Table 9 shows a qualitative analysis of passage generation (Pgen) using LLMs. In this example, the Contriever fails to retrieve relevant passages. The top-1 passage provides a description of *Mole Men Against the Son of Hercules*. The passage likely received the highest score due to the exact match of the phrase *Mole Men Against the Son of Hercules* with the question. The top-2 passage contains information about *The Fury of Hercules*. It seems to have been retrieved based on its focus on *Hercules*. Lastly, the top-3 passage is also completely unrelated to the correct answer. It is due to getting a high score due to the presence of the words *director* and *Against the Son of Hercules*. The rest of the passages in top-10 did not include content relevant to the correct answer. However, the LLM-generated passage accurately contained a short biography of *Antonio Leonviola*, the director of *Mole Men Against The Son Of Hercules*, which revealed a crucial hint on the place of death. Results show that even though the retriever does not retrieve relevant context, Pgen complements the retrievers by their knowledge. Conversely, as mentioned earlier, there are cases where the LLM-generated passage did not provide relevant context to the question. Nevertheless, a correct answer was recovered in case the retrieved passages provided relevant information. Some examples are shown in Table 15 and Table 16 of Appendix A.2.2.

## 7 Conclusion

In this paper, we propose a simple yet effective question and passage augmentation (QPaug) method via LLMs. QPaug harmoniously combines parametric and non-parametric knowledge of LLMs by leveraging the massive knowledge of LLMs for designing queries for retrieval and for guiding answer extraction from passages. Experiments show that both proprietary and open-source LLMs successfully decompose questions into multiple sub-questions by the proposed question augmentation, which significantly improved the performances of passage retrieval. In addition,

Question	Answer
Who is the spouse of the director of film Eden And After?	<a href="#">Catherine Robbe-Grillet</a>
<b>Augmented Question by Qaug</b>	
1. <b>Identify the Director</b> : First, we need to determine who directed "Eden And After". 2. <b>Research the Director</b> : Once we have the name of the director, we can research more about his or her personal life, focusing on marital status to find out who their spouse is (or was, depending on current status). 3. <b>Verify Information</b> : It's important to cross-check the information from multiple sources to ensure accuracy. ...	
<b>Retrieved Passage w/o Qaug</b>	
<i>Top-1</i> and during "I Dream of Jeannie". Also covered are her three marriages to Michael Ansara (1958–1974), Charles Fegert (1977–1982) and Jon Eicholtz (1991–present) and her "emotional breakdown" following the death of her only child, ...	
<b>Retrieved Passage w/ Qaug</b>	
<i>Top-1</i> ... <a href="#">Catherine Robbe-Grillet</a> (; "née" Rstakian; born 24 September 1930) is a French theatre and cinema actress, photographer, and writer of Armenian descent who has published sadomasochistic writings under ...	

Table 8: Qualitative comparison of retrieved passage with or without question augmentation (Qaug) on 2wiki dataset. The question-augmentation method is GPT-4. The first passage results are obtained by searching using the original question, while the second passage results are obtained by searching using the augmented question. The complete search results and examples are provided in Appendix A.2.1.

Question	Answer
Where was the place of death of the director of film Mole Men Against The Son Of Hercules?	<a href="#">Rome</a>
<b>Retrieved Passages</b>	
<i>Top-1</i> ... "Against the Son of Hercules" was released on 10 October 1961. Mole Men Against the Son of Hercules Mole Men Against <i>Top-2</i> ... poisoned by his own men (the "sers felons" Antipater and Divinuspater), as was another key figure of the work, Darius... <i>Top-3</i> at Dubrava Film in Zagreb, Croatia and on location in Zagreb. "The Fury of Hercules" as released in Italy on 21 March 1962. The Fury of Hercules The Fury of Hercules () is a 1962 peplum film written and directed by Gianfranco Parolini...	
<b>Generated passage</b>	
The director of the film "Mole Men Against The Son Of Hercules" was Antonio Leonviola. He passed away on December 14, 1971, in <a href="#">Rome</a> , Italy...	

Table 9: Qualitative analysis of generated passage (Pgen) compared to retrieved passages. The passage generation method is GPT-4. While retrieved passages do not include relevant context with the answer, generated passage includes the answer. The entire search results and additional examples are provided in Appendix A.2.2.

the self-generated passage was shown to achieve a large performance gain when it is augmented with the retrieved passages.

## 8 Limitations

Although we showed that the proposed question and passage augmentation method highly improves the performances on ODQA tasks, the detailed planning of questions and self-generation of passages rely much on LLMs' knowledge. However, the capability of LLMs are growing at an unprecedented rate, and we believe our work is timely in that it is reported that even LLM can generate synthetic data to train other language models and achieve good performance (Li et al., 2023; Gunasekar et al., 2023; Ben Allal et al., 2024; Abdin

et al., 2024), and such massive knowledge of LLMs should be more actively explored for many tasks, including RAG. In addition, although we explicitly instructed LLMs not to generate bogus passages, hallucinations still can occur. However, we believe that alignment techniques reducing LLMs' hallucinations are actively researched, which can further benefit our method in the future.

## 9 acknowledgement

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2022R1A5A1027646), the MSIT (Ministry of Science and ICT), Korea, under the ICT Creative Consilience program (IITP-2020-0-01819) supervised



by the IITP (Institute for Information & communications Technology Planning & Evaluation).

## References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.
- Alex Auvolat, Sarath Chandar, Pascal Vincent, Hugo Larochelle, and Yoshua Bengio. 2015. Clustering is efficient for approximate maximum inner product search. *arXiv preprint arXiv:1507.05910*.
- Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. *Cosmopedia*.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- 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*.
- Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2296–2309.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.
- Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 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. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- 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*.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joon-suk Park, and Jaewoo Kang. 2023a. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 996–1009.
- Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2023b. Sure: Improving open-domain question answering of llms via summarized retrieval. In *The Twelfth 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. *Advances in neural information processing systems*, 35:22199–22213.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- 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*.
- Dongryeol Lee, Segwang Kim, Minwoo Lee, Hwanhee Lee, Joonsuk Park, Sang-Woo Lee, and Kyomin Jung. 2023. Asking clarification questions to handle ambiguity in open-domain qa. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11526–11544.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. *Lost in the Middle: How Language Models Use Long Contexts*. *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.
- Parikshit Ram and Alexander G Gray. 2012. Maximum inner-product search using cone trees. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 931–939.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2023. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

- Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal 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. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.
- Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.
- Lirui Wang, Yiyang Ling, Zhecheng Yuan, Mohit Shridhar, Chen Bao, Yuzhe Qin, Bailin Wang, Huazhe Xu, and Xiaolong Wang. 2023b. Gensim: Generating robotic simulation tasks via large language models. In *The Twelfth International Conference on Learning Representations*.
- Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, et al. 2023c. Chain-of-table: Evolving tables in the reasoning chain for table understanding. In *The Twelfth International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.
- Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.
- Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.
- Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020*, pages 418–428.
- Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

## A Appendix

### A.1 Prompt templates

#### Prompts for answering the questions without passages.

```
Question: {question} Do not exceed 3 words.  
Answer:
```

Figure 4: Prompt template design for retrieval-augmented answering questions.

#### Prompts for answering the questions only using retrieved passages.

```
{passages}  
Question: {question} Do not exceed 3 words.  
Answer:
```

Figure 5: Prompt template design for retrieval-augmented answering questions.

#### Prompts for answering the questions using augmented passages.

```
{passages}  
Your knowledge: {generated passage}  
Question: {question} Do not exceed 3 words.  
Answer:
```

Figure 6: Prompt template design for retrieval-augmented answering questions.

#### Prompt template of passage generation.

```
Your job is to act as a subject matter expert. You will write a good-quality passage that  
can answer the question based on your factual knowledge. Do not write a passage if you  
don't know accurate information about the question.  
  
Now, let's start. After you write, please write [DONE] to indicate you are done.  
Write [NONE] if you cannot write a factual passage.  
Question: {question}  
Passage:
```

Figure 7: Prompt template design for generating knowledgeable passages.

### A.2 Additional Qualitative Results

#### A.2.1 Retrieval Results from Original Question and Augmented Question

Table 11, Table 12, and 10 show search results from the question and augmented question on NQ, 2wiki, and Hotpot QA datasets.

Table 10 demonstrates that the retrieval from original question retrieves the passage containing the answer in the top-3 results, while the augmented question retrieves it at the top-1. This example demonstrates that using the augmented question improves the search process, leading to higher-quality search



Question	Answer
who was the buccaneers qb when they won the superbowl	<b>Brad Johnson</b>
<b>Retrieved Passages (w/o Qaug)</b>	
<p><i>Top-1</i> in as QB due to Mark Sanchez suffering an apparent shoulder injury. His first game didn't start off on the right foot, though. He fumbled after taking a sack from Buccaneers MLB Mason Foster, and threw an interception to Buccaneers linebacker Lavonte David. Smith eventually rebounded by throwing his first TD pass to Kellen Winslow Jr. to make the game close by halftime. The second half was a defensive struggle between both teams that wound up in both teams scoring in field goals. The biggest moment of the game came when Geno Smith scrambled out-of-bounds and Lavonte David (the same</p> <p><i>Top-2</i> later became head coach of the Seattle Seahawks where they played in Super Bowl XL after the 2005 season. However, the Seahawks lost to the Pittsburgh Steelers. One of Holmgren's former assistants, Jon Gruden, has had reasonable success running the West Coast offense in his own right. He started his head coaching career with the Oakland Raiders, leading them from 1998-2001, and turned the Raiders into a strong playoff contender. Gruden then went on to become head coach of the Tampa Bay Buccaneers, winning Super Bowl XXXVII after the 2002 season. Gruden coached the Buccaneers from 2002-2008. After several years</p> <p><i>Top-3</i> <b>Johnson</b> again had another great season in Tampa Bay and won a Super Bowl the following year. In 2001, <b>Brad Johnson</b> was pursued by the Baltimore Ravens in the off season (coached by Brian Billick, former offensive coordinator for the Vikings) but he spurned them to join the Bucs. <b>Johnson</b> was reunited with former Vikings assistant-coach Tony Dungy for his first season with the Tampa Bay Buccaneers. That year, he broke Tampa Bay team records for passing yards with 3,406, completions with 340, and attempts with 540. In the 2002 season he led the Buccaneers to their first ever Super</p> <p><i>Top-4</i> QB Richard Todd in 1976. In his tenure with the Jets, he threw for more interceptions than he did touchdowns. In the '81 season, the Jets played vs the Miami Dolphins in the AFC Championship Game. Todd threw for 5 interceptions and the Jets lost the game. A year later, Todd would be traded to the New Orleans Saints. The most recent bust, Dee Milliner, was drafted by the team in 2013. Milliner played his college career at the University of Alabama and had high expectations after being drafted. Lasting just 3 years with the team, Milliner's career was plagued</p> <p><i>Top-5</i> when he served as head coach of the San Diego Chargers from 2007-2012. His Chargers teams showcased the talents of QB Philip Rivers, RB LaDainian Tomlinson and TE Antonio Gates. The St. Louis Rams ran the Coryell system successfully under coordinator and then head coach Mike Martz. Martz served as St. Louis's offensive coordinator under head coach Dick Vermeil in the 1999 season, when the Rams won Super Bowl XXXIV. Martz then served as the Rams head coach 2000-2005. His teams were anchored offensively by QB Kurt Warner and RB Marshall Faulk, both of whom are Hall of Famers. Earlier</p>	
<b>Answer prediction w/o Qaug: Brad Johnson</b>	
<b>Retrieved Passages (w/ Qaug)</b>	
<p><i>Top-1</i> <b>Johnson</b> again had another great season in Tampa Bay and won a Super Bowl the following year. In 2001, <b>Brad Johnson</b> was pursued by the Baltimore Ravens in the off season (coached by Brian Billick, former offensive coordinator for the Vikings) but he spurned them to join the Bucs. <b>Johnson</b> was reunited with former Vikings assistant-coach Tony Dungy for his first season with the Tampa Bay Buccaneers. That year, he broke Tampa Bay team records for passing yards with 3,406, completions with 340, and attempts with 540. In the 2002 season he led the Buccaneers to their first ever Super</p> <p><i>Top-2</i> XXXVII in 2003 under coach Jon Gruden. Tampa has hosted four Super Bowls: Super Bowl XVIII (1984), Super Bowl XXV (1991), Super Bowl XXXV (2001), and Super Bowl XLIII (2009). The first two events were held at Tampa Stadium, and the other two at Raymond James Stadium. Tampa will be the host for Super Bowl LV in 2021. Originally the Pittsburgh Gladiators and a charter member of the Arena Football League (AFL), the Tampa Bay Storm relocated from Pittsburgh in 1991 and won ArenaBowl V that year. They later won 4 more ArenaBowls (VII, IX, X, and XVII, and also</p> <p><i>Top-3</i> Super Bowl ring following the Buccaneers' victory in Super Bowl XXXVII. He signed with the Arena Football League's Orlando Predators in 2004 and guided the team to a 9-5 record and the playoffs, despite suffering another knee injury and missing two and a half games. He was then signed by the Indianapolis Colts in 2004, reuniting with former Buccaneers coach Tony Dungy, but only saw limited action in one game before being released during the season. He returned to the Orlando Predators where he was the starting quarterback through the 2006 season. He has a 32-15 record as the Predators'</p> <p><i>Top-4</i> Play of season QB Rating = Passer rating W/L Record = Super Bowl/Postseason Won/Loss Record After his retirement from professional football as a player, Aikman joined Fox's NFC telecasts as a color commentator for the 2001 season. A year later, he was named to the network's lead announcing crew, teaming with Joe Buck and (from 2002-2004) Cris Collinsworth. Aikman received an Emmy Award nomination for his television work in 2004 and has helped broadcast five Super Bowls (XXXIX, XLII, XLV, XLVIII, and LI) to date. It was revealed in 2016 that in 2004, Aikman nearly came out of retirement to</p> <p><i>Top-5</i> championship games (XVII, XXV, XXVIII, XXXIV, and XXXVI). It was also the last Super Bowl played in the month of January. This was the first Super Bowl in which the league's number one-ranked offense (Raiders) faced the league's number one-ranked defense (Buccaneers). The game sometimes is referred to as the ""Gruden Bowl"", because the primary storyline surrounding the game revolved around Jon Gruden. Gruden was the Raiders' head coach from 1998 to 2001, and then became the Buccaneers coach in 2002. Tampa Bay, ""Gruden's ""new"" team"", made their first Super Bowl appearance in team history after posting a regular season</p>	
<b>Answer prediction w/ Qaug: Brad Johnson</b>	

Table 10: Result of retrieved passages from the original question and augmented question on NQ.

Question	Answer
Who is the spouse of the director of film Eden And After?	<b>Catherine Robbe-Grillet</b>
<b>Retrieved Passage (w/o Qaug)</b>	
<p><i>Top-1</i> and during ""I Dream of Jeannie"". Also covered are her three marriages to Michael Ansara (1958–1974), Charles Fegert (1977–1982) and Jon Eicholtz (1991–present) and her ""emotional breakdown"" following the death of her only child, Matthew Ansara. On November 17, 1988, <i>Eden</i> received a star on the Hollywood Walk of Fame for her contributions to television. In 1990, the University of West Los Angeles School of Law granted <i>Eden</i> an honorary Doctor of Laws degree. Barbara Eden Barbara <i>Eden</i> (born Barbara Jean Morehead, August 23, 1931) is an American film, stage, and television actress, and singer, best known for her starring</p> <p><i>Top-2</i> to his lawyer, the filmmaker informed French tax authorities in the month preceding the release of the papers about his offshore holdings. Mareva Grabowski is listed in the Paradise Papers. She is the wife of Kyriakos Mitsotakis, who is leader of the opposition and president of the New Democracy political party. She owns 50% of an offshore company, Eterna Capital Management in the Cayman Islands. This match is verified in Appleby and on listed in Cayman records on 30 March 2010. In total, 2.829 Greek names are listed in the papers. U2 lead singer Bono is listed in the papers</p> <p><i>Top-3</i> Lauren and Michael are having a hard time not playing those guardian roles. I like that <i>Eden</i> wants to be her own person and I really enjoy bringing that feistiness to her. Since being in a relationship with Noah, living in Paris and then going to rehab for an eating disorder, she has to be a different person."" River Baldwin, ex-husband of Gloria Bardwell and absentee father of Michael Baldwin, was first seen onscreen in 2008, and was rumored to have a daughter in which he raised on an Ashram. Rivers' daughter arrived in town, operating under the alias <i>Eden</i></p> <p><i>Top-4</i> with Sergio Leone's restored Once Upon a Time in the West, in the presence of Claudia Cardinale. Lívia Gyarmathy is one of the greatest figures of Hungarian cinema. Even her first film, Do You Know Sunday-Monday? (Ismeri szandi-mandit?), was well-perceived both by the audience and the critics. The film, starring local legends Ila Schütz, Margit Dajka and Manyi Kiss is still a classic and repeatedly screened. The effects of the political changes of 1989 are pictured in Rapture of Deceit, starring Rita Tushingham. Besides her fine feature films she is also well known for her documentaries, like the touching The</p> <p><i>Top-5</i> she went on to answer her own questions through film: ""The war lingers in my head, and I always search for it traces. So I wanted to ask my questions on the screen"". She began her socio-political documentary work around the year 2000. When asked about her objectivity in ""Sleepless Nights"" Eliane responded: ""In the field of cinema, objectivity is a big lie, I tried to be as objective as possible, that is why I did not make a film about the Lebanese civil war, rather I made a film about two people who lived through the war. I began</p> <p>...</p>	
<b>Answer prediction of w/o Qaug: Not mentioned</b>	
<b>Retrieved Passage (w/ Qaug)</b>	
<p><i>Top-1</i> writer and filmmaker <b>Alain Robbe-Grillet</b> in Paris on 23 October 1957; he died in February 2008. In 2014, she was the subject of a documentary film entitled ""The Ceremony"", which examines her life as a member of the BDSM (sodomasochistic) community. <b>Catherine Robbe-Grillet Catherine Robbe-Grillet</b> (; ""née"" Rstakian; born 24 September 1930) is a French theatre and cinema actress, photographer, and writer of Armenian descent who has published sadomasochistic writings under the pseudonyms Jean de Berg and Jeanne de Berg. She was born in Paris, where she attended secondary school and high school. ""L'Image"", a sadomasochistic novel published in 1956</p> <p><i>Top-2</i> a decade before the appearance of his next feature film, ""La belle captive"" (""The Beautiful Captive"" (1983), but <b>Alain Robbe-Grillet</b> was fortunate enough to enlist the services of Henri Alekan as cinematographer, the visionary master of cinematography for the films of Jean Cocteau. Subsequently, more than a decade passed before <b>Alain Robbe-Grillet</b> got behind the lens again, this time filming a mystery thriller on a small Greek island with Fred Ward starring as the confused Frank in ""Un bruit qui rend fou"". <b>Robbe-Grillet</b> (""A Maddening Noise"", aka: ""The Blue Villa"" (1995). Before his death in 2008 <b>Robbe-Grillet</b> was to direct</p> <p><i>Top-3</i> 1982); references to works are often accompanied by their ""H"" (for Hitchcock) number. The following lists show selected pieces, not his entire production in each genre. Marc-Antoine Charpentier Marc-Antoine Charpentier (; 1643 – 24 February 1704) was a French composer of the Baroque era. Exceptionally prolific and versatile, Charpentier produced compositions of the highest quality in several genres. His mastery in writing sacred vocal music, above all, was recognized and hailed by his contemporaries. Any family relationship between him and Gustave Charpentier, the late-nineteenth and early-twentieth century French opera composer, is highly unlikely. Charpentier was born in or near Paris</p> <p><i>Top-4</i> of the male protagonist, who becomes the narrative voice, instead of the female protagonist, Violette (Catherine Jourdan). N. a pris les dés... N. a pris les dés... (, French for ""N. has taken the dice..."" is a 1971 French experimental independent underground drama art film directed by <b>Alain Robbe-Grillet</b>. <b>Alain Robbe-Grillet</b> had signed for a production of two separate films from the same shot, with different editings of the same scenes, so to create two totally different plots: the first was ""Eden and After"", the second ""N. a pris les dés..."", whose title is indeed an anagram of the other</p> <p><i>Top-5</i> ""Captive"". The book is referred to as a ""roman"" (novel) and is illustrated with 77 paintings by Magritte interspersed with discourse written by <b>Robbe-Grillet</b>. The eponymous film ""La Belle captive"", written and directed by Robbe-Grillet, was released in 1983. <b>Alain Robbe-Grillet Alain Robbe-Grillet</b> (; 18 August 1922 – 18 February 2008) was a French writer and filmmaker. He was one of the figures most associated with the ""Nouveau Roman"" (new novel) trend of the 1960s, along with Nathalie Sarraute, Michel Butor and Claude Simon. <b>Alain Robbe-Grillet</b> was elected a member of the Académie française on 25 March 2004, succeeding Maurice</p> <p>...</p> <p><i>Top-10</i> ... <b>Alain Robbe-Grillet Alain Robbe-Grillet</b> (; 18 August 1922 – 18 February 2008) was a French writer and filmmaker. He was one of the figures most associated with the ""Nouveau Roman"" (new novel) trend of the 1960s, along with Nathalie Sarraute, Michel Butor and Claude Simon. <b>Alain Robbe-Grillet</b> was elected a member of the Académie française on 25 March 2004, succeeding Maurice Rheims at seat No. 32. <b>He was married to Catherine Robbe-Grillet (née Rstakian)</b>. <b>Alain Robbe-Grillet</b> was born in Brest (Finistère, France) to a family of engineers and scientists. He was trained as an agricultural engineer. During the years 1943</p>	
<b>Answer prediction of w/ Qaug: Catherine Robbe-Grillet</b>	

Table 11: Result of retrieved passages from the original question and augmented question 2wiki dataset.

results. Also, when passages containing clues about the correct answer are placed at the beginning of the prompt, it can be more helpful in generating the answer, as observed by (Liu et al., 2024).

In the example of the 2wiki dataset (Table 11), the question is asking about the spouse of the director of the film *Eden And After*. The search results using only the original question mainly focus on the word *Eden*, and we observe that the overall content of the retrieved passages is disconnected from the correct answer. Based on these search results, the output answer is *Not mentioned*. By contrast,

Question	Answer
Where is the stadium at which 1964 Georgia Tech Yellow Jackets football team played their home game located?	<a href="#">North Avenue at Techwood Drive</a>
<b>Retrieved Passage (w/o Qaug)</b>	
<p><i>Top-1</i> field fence and wall could be subject to damage from long home runs. The O'Keefe lot and others nearby (Architecture and Van Leer Electrical Engineering lots) are no longer available due to construction or other campus projects. Consult The Georgia Tech Athletic Association, Rusty C, or Beesball.com website links provided in the ""External links"" section below for maps and more detailed information on recommended parking. Russ Chandler Stadium Russ Chandler Stadium is a college baseball stadium in Atlanta, Georgia. It has been the home field of the Georgia Tech Yellow Jackets college baseball team since 1930. The current stadium opened</p> <p><i>Top-2</i> Tech blockers 78 yards down the gridiron into the endzone as time expired, giving Georgia Tech its third win of the 2015 season, and third Top 10 win under Paul Johnson's guidance. <b>Bobby Dodd Stadium Bobby Dodd Stadium</b> at Historic Grant Field is the football stadium located at the corner of <b>North Avenue at Techwood Drive</b> on the campus of the Georgia Institute of Technology in Atlanta. It has been home to the Georgia Tech Yellow Jackets football team, often referred to as the ""Ramblin'Wreck"", in rudimentary form since 1905 and as a complete stadium since 1913. The team</p> <p><i>Top-3</i> the Yellow Jackets home field. A new facility being constructed adjacent to historic O'Keefe High School was completed February 2009, supplanting the twenty-year-old Glenn Field. The new facility will hold 1,000 spectators and make it possible for the Jackets to host postseason games with the presence of stadium lights. Glenn Field Glenn Field is the former home field for the Georgia Tech Yellow Jackets fast pitch softball team located in Midtown Atlanta, Georgia. Glenn Field was built in 1987 and has a capacity of 500 spectators. Glenn Field is one of the two Tech sports facilities off campus being located</p> <p><i>Top-4</i> in Atlanta and Georgia in Athens. ""We simply could not stop them."" - Bobby Bowden Pregame Line: Virginia Tech -3.5 GT vs. Virginia Tech Hokies Josh Nesbitt scored three touchdowns and Georgia Tech ran for 306 yards to upset #4 Virginia Tech on a cold, memorable night in Atlanta. For the second straight week, head coach Paul Johnson led the #19 Yellow Jacket squad to a victory, earning a win that ended a 17-game losing skid to top five opponents played at <b>Bobby Dodd Stadium</b> since 1962. Jason Peters batted down and intercepted Tyrod Taylor in the 1st quarter. Josh</p> <p><i>Top-5</i> where they still are today. Piedmont Park served the Atlanta Crackers, the city's original professional baseball team, before they moved to a stadium on Ponce de Leon Avenue in 1904. The two baseball teams have met 345 times since 1898. Georgia Tech has 148 wins, Georgia has 195 wins, and there are 2 ties in the series. Three baseball games are played between the two institutions every year. Two of the three games are played at the respective colleges baseball stadiums while the finale is played at SunTrust Park, home of the Atlanta Braves. The 2004 Georgia Tech vs. Georgia</p>	
<b>Answer prediction w/o Qaug: Bobby Dodd Stadium</b>	
<b>Retrieved Passage (w/ Qaug)</b>	
<p><i>Top-1</i> Tech blockers 78 yards down the gridiron into the endzone as time expired, giving Georgia Tech its third win of the 2015 season, and third Top 10 win under Paul Johnson's guidance. <b>Bobby Dodd Stadium Bobby Dodd Stadium</b> at Historic Grant Field is the football stadium located at the corner of <b>North Avenue at Techwood Drive</b> on the campus of the Georgia Institute of Technology in Atlanta. It has been home to the Georgia Tech Yellow Jackets football team, often referred to as the ""Ramblin'Wreck"", in rudimentary form since 1905 and as a complete stadium since 1913. The team</p> <p><i>Top-2</i> of the 2017 MLS season at <b>Bobby Dodd Stadium</b> until Mercedes-Benz Stadium was ready in mid-2017. The soccer club paid the university and the athletic association over \$1 million for the usage of <b>Bobby Dodd Stadium</b>. October 7, 1916: Georgia Tech 22, Cumberland College 0&lt;br&gt;In the most lopsided game in American football history, Georgia Tech, under coach John Heisman, defeated Cumberland College, 22-0. It has been said that Coach Heisman was repaying the Bulldogs for a 22-0 baseball defeat the previous year in which Cumberland had allegedly used professional players to ensure victory or conversely that he was showing how</p> <p><i>Top-3</i> in Atlanta and Georgia in Athens. ""We simply could not stop them."" - Bobby Bowden Pregame Line: Virginia Tech -3.5 GT vs. Virginia Tech Hokies Josh Nesbitt scored three touchdowns and Georgia Tech ran for 306 yards to upset #4 Virginia Tech on a cold, memorable night in Atlanta. For the second straight week, head coach Paul Johnson led the #19 Yellow Jacket squad to a victory, earning a win that ended a 17-game losing skid to top five opponents played at <b>Bobby Dodd Stadium</b> since 1962. Jason Peters batted down and intercepted Tyrod Taylor in the 1st quarter. Josh</p> <p><i>Top-4</i> field fence and wall could be subject to damage from long home runs. The O'Keefe lot and others nearby (Architecture and Van Leer Electrical Engineering lots) are no longer available due to construction or other campus projects. Consult The Georgia Tech Athletic Association, Rusty C, or Beesball.com website links provided in the ""External links"" section below for maps and more detailed information on recommended parking. Russ Chandler Stadium Russ Chandler Stadium is a college baseball stadium in Atlanta, Georgia. It has been the home field of the Georgia Tech Yellow Jackets college baseball team since 1930. The current stadium opened</p> <p><i>Top-5</i> <b>Bobby Dodd Stadium Bobby Dodd Stadium</b> at Historic Grant Field is the football stadium located at the corner of <b>North Avenue at Techwood Drive</b> on the campus of the Georgia Institute of Technology in Atlanta. It has been home to the Georgia Tech Yellow Jackets football team, often referred to as the ""Ramblin'Wreck"", in rudimentary form since 1905 and as a complete stadium since 1913. The team participates in the NCAA Division I Football Bowl Subdivision as a member of the Atlantic Coast Conference. It is the oldest stadium in the FBS and has been the site of more</p>	
<b>Answer prediction w/ Qaug: Bobby Dodd Stadium</b>	

Table 12: Result of retrieved passages from the original question and augmented question on HotpotQA.

Question	Answer
Where was the place of death of the director of film Mole Men Against The Son Of Hercules?	<b>Rome</b>
<b>Retrieved passages</b>	
<p><i>Top-1</i> Trapped in the caverns, the now-freed slaves break out onto the surface, thanks to Maciste's strength. Halis Mosab, who has become aware that she was captured as a child and somehow became queen of the 'mole men', is overjoyed when she comes to the surface and finds that she can live in the sunlight. But she accidentally falls off a cliff and into the waters of the Sacred Waterfall. The fate of her people is not revealed. ""Mole Men Against the Son of Hercules"" was released on 10 October 1961. Mole Men Against the Son of Hercules Mole Men Against</p> <p><i>Top-2</i> Alexander will not realize this goal: poisoned by his own men (the ""sers felons"" Antipater and Divinuspater), as was another key figure of the work, Darius. The cause of the death of the hero is shrouded in ambiguity: prophetic in nature, it poses the problem of divine retribution, patricide and political mistakes. Was Alexander killed because of his desire to elevate himself to the godslevel? Was he killed by the magician Nectanabo, who is his father in the Greek and Roman tradition, and who also presided over his birth (Alexander kills him in a spite of rage)? Or was</p> <p><i>Top-3</i> at Dubrava Film in Zagreb, Croatia and on location in Zagreb. ""The Fury of Hercules"" as released in Italy on 21 March 1962. The Fury of Hercules The Fury of Hercules () is a 1962 peplum film written and directed by Gianfranco Parolini. Hercules is reached by the slave Daria, who informs him that his country has fallen into the hands of the tyrant Meniste. Hercules follows Daria in his homeland, where he discovers that Meniste enslaves the citizens of the city, and that a group of rebels are trying an insurrection. Meniste fears the power of Hercules, however he</p> <p><i>Top-4</i> at Dubrava Film in Zagreb, Croatia and on location in Zagreb. ""The Fury of Hercules"" as released in Italy on 21 March 1962. The Fury of Hercules The Fury of Hercules () is a 1962 peplum film written and directed by Gianfranco Parolini. Hercules is reached by the slave Daria, who informs him that his country has fallen into the hands of the tyrant Meniste. Hercules follows Daria in his homeland, where he discovers that Meniste enslaves the citizens of the city, and that a group of rebels are trying an insurrection. Meniste fears the power of Hercules, however he</p> <p><i>Top-5</i> his middle initial ""M"" as the name Anthony Dawson was used by the English actor. He was the only Italian director who worked directly for American production companies like MGM, United Artists, 20th Century Fox, Columbia Pictures, etc. with films like: ""Yor, the Hunter from the Future"", ""Take a Hard Ride"", ""Killer Fish"", etc. Margheriti said his action/adventure/war films were his favorites, and his least favorite movies were the sword-and-sandal ""peplum"" films he made in the early 1960s (such as ""Devil of the Desert Against the Son of Hercules"" and ""Giants of Rome""). For years, director Paul Morrissey disputed Margheriti's ...</p>	
<b>Generated passage</b>	
The director of the film "Mole Men Against The Son Of Hercules" was Antonio Leonviola. He passed away on December 14, 1971, in <b>Rome</b> , Italy. Antonio Leonviola was a prominent Italian director known for his work in the peplum genre, including films like "Mole Men Against The Son Of Hercules." His contributions to Italian cinema have left a lasting impact on the industry.	
<b>Answer Prediction</b>	
<b>Rome</b>	

Table 13: Qualitative result of retrieved passages with Pgen (Ours) compared to Base RAG for 2wiki.

augmented question retrieves passages containing *director of film Eden And After* throughout the searched passages. The top-1 document includes the correct answer, **Catherine Robbe-Grillet**. Moreover, the top-10 passages contain clear sentences that indicate the relationship between the director and their spouse, providing the necessary information to answer the question.

The last example, Table 12 shows that both retrieved passages from the original question and retrieved passages from the augmented question predict the same answer, **Bobby Dodd Stadium**. The address of this output is *177 North Avenue NW, Atlanta, GA 30332, USA*, which is the same location as the correct answer. Consequently, we observe that even if the LLM's answer does not exactly match the correct answer, it can respond with a semantically equivalent entity. Comparing the search results, the augmented question retrieves passages containing the correct answer as more relevant passages compared to retrieval from original questions. It seems that the augmented question can find one more relevant passage, providing sufficient information for grounding the LLM.

### A.2.2 Qualitative Results for generated passage

Table 13 shows more detailed passages of the summarized version in Table 9. Even though the retrieved passages cannot obtain the relevant context, the generated passage contain relevant information. Thus, LLM can extract correct answer based on Pgen.

Table 14 shows that all the retrieved passages are irrelevant to the correct answer, and LLM refuses the passage generation since they do not have high-confident factual knowledge. Instead, LLM generates a sentence *I do not have access to real-time information* to avoid hallucination. In this case, Pgen cannot avoid predicting wrong answers. In addition, Table 15 demonstrates that even though the generated passage is [None] due to lack of knowledge, LLMs can successfully extract the correct answer from retrieved passages if they contain the relevant contexts. Similarly, Table 16 illustrates that if LLMs generate irrelevant passages, they still extract relevant context from retrieved passages and predict the



Question	Answer
Where did the director of film Playing It Wild die?	<a href="#">Hollywood</a>
<b>Retrieved passages</b>	
<p><i>Top-1</i> the trailer for which included the title shot of a deer being killed. This scene was not, however, in the finished film. A song with the same name is also featured in the film, sung by Edward Tudor-Pole. Additional footage appeared in the 2000 documentary ""The Filth and the Fury"". In April 2010, Roger Ebert posted the screenplay of ""Who Killed Bambi?"" (originally titled ""Anarchy in the UK"" ) on his blog. Who Killed Bambi? (unfinished film) Who Killed Bambi? was to be the first film featuring the punk rock band the Sex Pistols, and was due to be released in</p> <p><i>Top-2</i> a success. ""Buck Wild"" premiered at the Dallas International Film Festival in April 2013. It was released on home video on March 18, 2014. Patrick Naugle of DVD Verdict wrote, """"Buck Wild"" is like sitting through a root canal without any anesthetic, being performed by a blind orangutan."" Lizzie Duncan of HorrorNews.Net wrote, ""Whilst there were definitely some scenes that gave away the amateur status of this film, on the whole I thought that it was well made and well edited."" Buck Wild Buck Wild is a zombie comedy horror film directed by Tyler Glodt, written by Glodt and producer</p> <p><i>Top-3</i> after Wild was born, the family moved to Vancouver where Wild would live the majority of her life. While studying at the University of British Columbia, Wild gained a Bachelor of Fine Arts (BFA) with a major in creative writing along with a minor in film and theatre. Alongside her studies, Wild co-founded Touchstone Theatre and Headlines Theatre with David Diamond, a fellow student. Wild worked with the Touchstone Theatre in 1975-1976 and Headlines Theatre during 1980-1985. In 1991, she founded the Canada Wild Production with producer Betsy Carson. The production company was named in part after Wild's full name</p> <p><i>Top-4</i> girlfriend. Phil gets them to be friends again. He tells them of his illness and all three decide to go to Los Angeles for Phil's dying wish: to be a contestant on ""Jeopardy!"" On the way there they meet an attractive wild woman with a heart of gold (Annie Potts). Roger Ebert savaged the film in his contemporary review, damning its inauthentic script wherein ""one appalling scene follows another"". In an exasperated tone, he asked: ""Was there no one to cry out, ""Stop this madness?"" No one to read the script and see that it was without sense or sensibility?</p> <p><i>Top-5</i> again encounters Wild. The film centers on Brian Slade, a sexually fluid and androgynous glam rock icon who was patterned after David Bowie, Jobriath and, to a lesser extent, Marc Bolan. Bowie initially disapproved of the film and its many similarities with his life story, and threatened to sue, resulting in substantial rewrites to create more distance between the character and the real man. Ewan McGregor co-stars in the role of Curt Wild, a genre-defying performer who doesn't back down from sex, nudity, or drugs on or off stage, and whose biographical details are based on Iggy Pop (who grew</p>	
<b>Generated Passage</b>	
I'm sorry, but I cannot provide information on the location of the director of the film Playing It Wild's death as I do not have access to real-time information.	
<b>Answer Prediction</b>	
<b>I cannot predict.</b>	

Table 14: The case of predicting correct answers when LLM cannot generate the passage. However, LLMs extract correct answer from retrieved passages.

correct answer. Thus, Pgen benefits the RAG systems in most cases since current LLMs can extract the relevant information from diverse and heterogeneous contexts.

Question	Answer
Where was the place of death of James Adam (Architect)'s father?	<a href="#">Edinburgh</a>
<b>Retrieved passages</b>	
<p><i>Top-1</i>  Architecture of Robert and James Adam"" (in 1773–1778 and 1779; a third volume was published posthumously, in 1822). James Adam (architect) James Adam (21 July 1732 – 20 October 1794) was a Scottish architect and furniture designer, but was often overshadowed by his older brother and business partner, Robert Adam. They were sons of architect William Adam. In 1755 James worked on Gunsgreen House in the Berwickshire town of Eyemouth. In 1758, Robert, James, and their younger brother William Adam started their business in London (living in Lower Grosvenor Street), focusing on designing complete schemes for the decoration and furnishing</p> <p><i>Top-2</i>  as well as having a villa at Merchiston. Adam Square was demolished in the 1870s, and the site is now occupied by Adam House, a building of the University of <a href="#">Edinburgh</a>. Upon his death in 1792, he was succeeded as laird of Blair Adam by his only surviving son, the politician and judge William Adam. He is buried in his father's mausoleum in Greyfriars Kirkyard in <a href="#">Edinburgh</a>. It is the largest monument in the graveyard and stands just south-west of the church. He was married to Jean Ramsay of Abbotshall in Fife (d.1795). John Adam (architect) John Adam (5 March</p> <p><i>Top-3</i>  court from 1815 until his death. In 1832-3 his home address was 31 Charlotte Square in <a href="#">Edinburgh</a>. He died in <a href="#">Edinburgh</a> on 17 February 1839 and was buried at Greyfriars Kirkyard He lies in the huge family vault built for his grandfather, William Adam the architect, facing his father, John Adam. The vault lies south-west of the church, in front of the Covenanters Prison. On 7 May 1777, William Adam married Eleanora Elphinstone (d. 4 February 1800), daughter of Charles, 10th Lord Elphinstone. They had six children: William Adam of Blair Adam The Right Hon. William Adam of Blair Adam</p> <p><i>Top-4</i>  James Adam (architect) James Adam (21 July 1732 – 20 October 1794) was a Scottish architect and furniture designer, but was often overshadowed by his older brother and business partner, Robert Adam. They were sons of architect William Adam. In 1755 James worked on Gunsgreen House in the Berwickshire town of Eyemouth. In 1758, Robert, James, and their younger brother William Adam started their business in London (living in Lower Grosvenor Street), focusing on designing complete schemes for the decoration and furnishing of houses. Palladian design was popular, but Robert had evolved a new, more flexible style incorporating elements of</p> <p><i>Top-5</i>  Adam succumbed to illness in late 1747, dying the following summer. He was buried in Greyfriars Kirkyard, <a href="#">Edinburgh</a>, where John Adam designed the family mausoleum built in 1753. This was restored by <a href="#">Edinburgh</a> City Council and Historic Scotland in 1997 to mark the 250th anniversary of his death. Adam used a wide variety of sources for his designs, and created an inventive personal style of decoration. His chief influences were from English Palladianism, and several of his houses have been likened to designs reproduced in Colen Campbell's ""Vitruvius Britannicus"", but Adam mixed these with English Baroque motifs from Gibbs and ...</p>	
<b>Genrated passage</b>	
[None]	
<b>Answer Prediction</b>	
<a href="#">Edinburgh</a>	

Table 15: The case of predicting correct answers when the generated passage is None. LLMs extract correct answer from retrieved passages despite the generated passage being [None].

Question	Answer
Where did Oleg Kerensky's father die?	<a href="#">New York City</a>
<b>Retrieved passages</b>	
<p><i>Top-1</i> and where he taught graduate courses. He wrote and broadcast extensively on Russian politics and history. Kerensky died of arteriosclerotic heart disease at St. Luke's Hospital in <a href="#">New York City</a> in 1970, one of the last surviving major participants in the turbulent events of 1917. The local Russian Orthodox Churches in <a href="#">New York City</a> refused to grant Kerensky burial, because of his association with Freemasonry and because they saw him as largely responsible for the Bolsheviks seizing power. A Serbian Orthodox Church also refused burial. Kerensky's body was flown to London, where he was buried at the non-denominational Putney Vale</p> <p><i>Top-2</i> had two sons, Oleg and Gleb, whom both went on to become engineers. Kerensky's grandson (also named Oleg) played his grandfather's role in the 1981 film "Reds". Kerensky and Olga were divorced in 1939 and soon after he settled in Paris, and while visiting the United States he met and married 1939 the Australian former journalist Lydia Ellen "Nell" Tritton (1899–1946). The marriage took place in Martins Creek, Pennsylvania. When Germany invaded France in 1940, they emigrated to the United States. After the Nazi-led invasion of the Soviet Union in 1941, Kerensky offered his support to Joseph Stalin. When his</p> <p><i>Top-3</i> the 1981 film "Reds" portraying his grandfather when he was the head of the Russian Provisional Government. Oleg Kerensky Oleg Aleksandrovich Kerensky CBE FRS (, (16 April 1905 – 25 June 1984) was a Russian civil engineer, one of the foremost bridge designers of his time. Kerensky was born in St. Petersburg, Russia, the son of future Russian prime minister Alexander Kerensky, who survived the events of the Russian Civil War and emigrated to Paris in 1918. Both Oleg and his younger brother Gleb graduated as engineers in 1927, and both settled in England. As an associate of Dorman Long</p> <p><i>Top-4</i> in the October Revolution. He spent the remainder of his life in exile, in Paris and <a href="#">New York City</a>, and worked for the Hoover Institution. Alexander Kerensky was born in Simbirsk (now Ulyanovsk) on the Volga River on 4 May 1881 and was the eldest son in the family. His father, Fyodor Mikhailovich Kerensky, was a teacher and director of the local gymnasium and was later promoted to Inspector of public schools. His maternal grandfather was head of the Topographical Bureau of the Kazan Military District. His mother, Nadezhda Aleksandrovna (née Adler), (the first-name Nadezhda meaning "Hope"; her patronymic last</p> <p><i>Top-5</i> Oleg Kerensky Oleg Aleksandrovich Kerensky CBE FRS (, (16 April 1905 – 25 June 1984) was a Russian civil engineer, one of the foremost bridge designers of his time. Kerensky was born in St. Petersburg, Russia, the son of future Russian prime minister Alexander Kerensky, who survived the events of the Russian Civil War and emigrated to Paris in 1918. Both Oleg and his younger brother Gleb graduated as engineers in 1927, and both settled in England. As an associate of Dorman Long, Kerensky assisted on the landmark 1932 Sydney Harbour Bridge. As an associate, and then a partner, in</p> <p>...</p>	
<b>Generated passage</b>	
<p>Oleg Kerensky's father, Oleg Kerensky Sr., died in Paris, France. He was a Russian engineer and a member of the White Russian community in exile after the Bolshevik Revolution. The Kerensky family, including Oleg Kerensky Jr., who became a notable British civil engineer, had to flee Russia due to the political upheaval and the rise of the Soviet Union. Paris was a common destination for many Russian émigrés during this period, offering a refuge for those escaping the communist regime.</p>	
<b>Answer Prediction</b>	
<a href="#">New York City</a>	

Table 16: The case of predicting correct answers when the LLM-generated passage is irrelevant context. LLMs extract the correct information from retrieved passages despite the generated passage being irrelevant.