An Information Bottleneck Perspective for Effective Noise Filtering on Retrieval-Augmented Generation

Kun Zhu[1*], **Xiaocheng Feng**[1,2*], **Xiyuan Du**[1], **Yuxuan Gu**[1], **Weijiang Yu**[3], **Haotian Wang**[1], **Qianglong Chen**[4], **Zheng Chu**[1], **Jingchang Chen**[1], **Bing Qin**[1,2]
[1]Harbin Institute of Technology [2] Peng Cheng Laboratory
[3] Sun Yat-sen University [4] Zhejiang University
{kzhu,xcfeng,xydu,yxgu,zchu,jcchen,qinb}@ir.hit.edu.cn
{weijiangyu8, wanght1998, chenqianglong.ai}@gmail.com

Abstract

Retrieval-augmented generation integrates the capabilities of large language models with relevant information retrieved from an extensive corpus, yet encounters challenges when confronted with real-world noisy data. One recent solution is to train a filter module to find relevant content but only achieve suboptimal noise compression. In this paper, we propose to introduce the information bottleneck theory into retrieval-augmented generation. Our approach involves the filtration of noise by simultaneously maximizing the mutual information between compression and ground output, while minimizing the mutual information between compression and retrieved passage. In addition, we derive the formula of information bottleneck to facilitate its application in novel comprehensive evaluations, the selection of supervised fine-tuning data, and the construction of reinforcement learning rewards. Experimental results demonstrate that our approach achieves significant improvements across various question answering datasets, not only in terms of the correctness of answer generation but also in the conciseness with 2.5% compression rate.

1 Introduction

Large language models represent a significant advancement in natural language understanding and generation, with the capability to process and produce human-like language at an unprecedented scale and complexity (Achiam et al., 2023; Touvron et al., 2023; Team et al., 2023). Nonetheless, large language models have several drawbacks, such as hallucination (Huang et al., 2023) and lacking knowledge for specific domains or highly specialized queries (Kandpal et al., 2023). Retrieval-augmented generation (Lewis et al., 2020) has gained attention for its ability to incorporate information from external knowledge sources during the inference stage. By combining retrieval-based methods with generative models, this approach can improve the relevance, coherence, and factual accuracy of text generation (Gao et al., 2023).

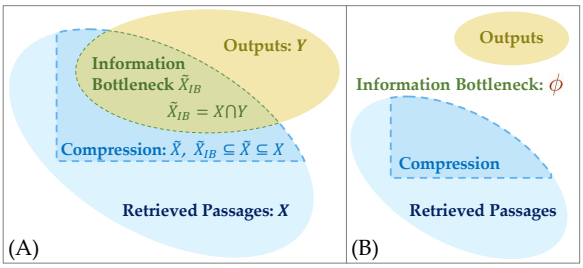

Figure 1: In retrieval-augmented generation, passages X are retrieved to enhance the generation of output Y. (A) Recent Noise filtering approaches obtain the compression $\tilde{X} \subseteq X$ with log likelihood objective to outputs Y. Our information bottleneck objective enables a precise delineation of the intersection $\tilde{X}_{\text{IB}} = X \cap Y$. (B) Information bottleneck explicitly compresses $\tilde{X}_{\text{IB}} = \phi$, when retrieved passages are irrelevant to outputs.

Retrieval-augmented generation also presents problems. On the one hand, the retriever's efficacy may be suboptimal in practical use (Izacard et al., 2022; Shi et al., 2023b; Cheng et al., 2023; Lin et al., 2023). On the other hand, the internet data is often of low quality, with redundancy and noise. Indeed, the retrieved content can be completely irrelevant to the query, leading to the model producing incorrect results (Shi et al., 2023a). Recent solutions to mitigate noise in retrieval evidence often involve the adoption of a filtering module (Liu et al., 2023; Yang et al., 2023a; Xu et al., 2023). However, these methods encounter several issues: (1) The inability to ensure that the annotated filtering results can effectively support the generation model in accurately answering questions. (2) The difficulty in directing the filter to refrain from answering when confronted with retrieval evidence that does not support question resolution. (3) The lack of adaptation to the compression extent of the filtering results, impeding the achievement of an optimal solution in terms of cost performance.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1044–1069

We observe that issues above originate from suboptimal objectives. As shown in Figure 1A, the intersection between retrieved passages X and outputs Y denotes the precise information in X which is useful for Y. The noise filter extracts compression $\tilde{X}$ from retrieved passages X, where the filter is optimized with log likelihood objective to output Y. The noise filter trained with this objective can obtain a compression $\tilde{X}$ containing the intersection $X \cap Y$, but is incapable of realizing its exact area, which means the filter cannot in principle eliminate the interference of noise for subsequent generation. Therefore, we propose to utilize the information bottleneck theory (Tishby et al., 1999) to optimize the noise filter from a comprehensive perspective, via simultaneously maximizing the useful information while minimizing the noise, thus facilitating a precise delineation of the intersection $\tilde{X}_{\text{IB}} = X \cap Y$. Furthermore, in cases (Figure 1B) where retrieval is not necessitated for content generation or exhibits limited efficacy, the information bottleneck objective enables noise filters to compress the retrieved passages into empty $\tilde{X}_{\text{IB}} = \phi$.

Specifically, we consider information bottleneck as a principle for retrieval augmentation. We first theoretically derive the formula of information bottleneck for retrieval-augmented generation, which integrates large language models. Then we introduce information bottleneck as a new comprehensive evaluation metric for noise filtering, assessing both conciseness and correctness of compressed contents. Next we derive information bottleneck version of supervised fine-tuning and reinforcement learning objectives to train the noise filter.

We conduct experiments on the open-domain question answering datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and a more complex multi-hop HotpotQA (Yang et al., 2018). Using Llama2 as filtering and generation model, our approach is proved to be effective compared with strong baseline models, including RankGPT, LongLLMLingua and Lllama2 on all three datasets. We achieve a 2.5% significant compression rate and a 3.2 improvement of exact answer match at most.

Our paper presents three main innovations:

- We are the first, to the best of our knowledge, to introduce the information bottleneck theory into retrieval-augmented generation, which illustrates the optimum of filtration.
- We propose to apply the information bottleneck on evaluation metrics, supervised fine-tuning objectives, and reinforcement learning rewards for retrieval-augmented generation.
- Experimental results reveal the effectiveness of our approach in terms of generation correctness and compression conciseness.

2 Related Work

Information Bottleneck The Information Bottleneck (IB) (Tishby et al., 1999; Fischer, 2020) is a rather simplistic concept: when facing a task, one should attempt to accomplish it using minimal information. The Information Bottleneck theory characterizes learning as a delicate balance between data compression and information retention. When applied to specific tasks, the idea is to extract all the essential informative features for the task while discarding redundant information (Shwartz-Ziv and LeCun, 2023). It has been applied in the study of representation learning (Wu et al., 2020; Federici et al., 2020; Lee et al., 2021), deep learning (Tishby and Zaslavsky, 2015; Saxe et al., 2019; Kawaguchi et al., 2023a), document clustering (Slonim and Tishby, 2000), speech recognition (Hecht et al., 2009), summarization (West et al., 2019), etc.

Noise Filtering Retrieval-augmented generation usually concatenates retrieved passages with their queries as the input of language models. However, this can potentially exceed the context window limit, introduce extra noise and redundancy, and increase computing resource requirements, leading to a decrease in model performance.

Flare (Jiang et al., 2023c) and Self-Rag (Asai et al., 2023) are dedicated to training models to have the capability of actively retrieving and filtering retrieval content on their own. Replug (Shi et al., 2023b) improves the retriever by the KL divergence between the retriever and the LLM.

Post-processing techniques such as noise filtering can help alleviate these issues. (Bai et al., 2023) focuses on re-ranking retrieved articles to filter out noise. Some methods like Selective Context (Li, 2023) and LLMLingua (Jiang et al., 2023a) make use of small language models to measure prompt mutual information or perplexity, finding the highest-scoring elements.

There are also some methods employ summarization techniques to design compressors (Xu et al., 2023; Wang et al., 2023). Tcra-LLM (Liu et al., 2023) and LongLLMLingua (Jiang et al., 2023b)

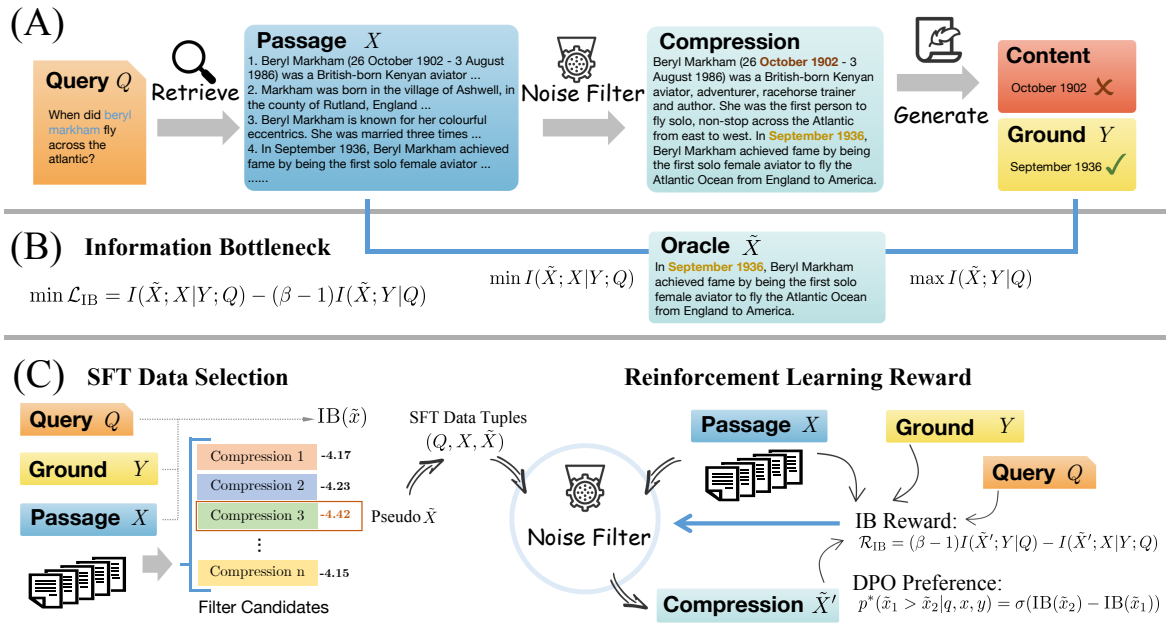

Figure 2: Overview of our approach. (A) Traditional retrieval-augmented generation with noise filtering. Although the compression contains relevant information, the unavoidable noise continues to disrupt subsequent generation processes. (B) The information bottleneck theory provides an objective to oracle compression $\tilde{X} = X \cap Y$ via eliminating the influence of noise to the greatest extent, $\min I(\tilde{X}; X|Y; Q)$. (C) We derive the formula for applying information bottleneck theory to retrieval-augmented generation, which can be utilized for the selection of supervised fine-tuning dataset and the construction of reinforcement learning reward.

combine summarization and semantic compression techniques. PRCA (Yang et al., 2023b) also incorporates reinforcement learning algorithms in model training. However, these compression methods do not have a unified evaluation of the compression results. RECOMP (Xu et al., 2023) achieves a compression rate of 6%, but it comes at the cost of degraded performance. Our approach is to find the optimal balance between compression rate and performance with the information bottleneck theory.

3 Methodology

In this section, we first introduce the background of information bottleneck (§3.1), and then convert the information bottleneck into forms for noise filters of retrieval augmented generation (§3.2). Next, we provide details of information bottleneck objectives for retrieval augmented generation (§3.3).

3.1 Preliminary

The information bottleneck principle (Tishby et al., 1999) has been a great concept in finding a compression $\tilde{X}$ for signals X that preserves the maximum information relevant to signals Y. Given a joint probability distribution $p(X, Y)$ between a random variable X and an observed relevant variable Y (with support sets $x \in \mathcal{X}$ and $y \in \mathcal{Y}$), the amount of information about Y in compressed representation $\tilde{X}$ ($\tilde{x} \in \tilde{\mathcal{X}}$) is given by mutual information:

$$I(\tilde{X}; Y) = \int_{\tilde{x}} \int_{y} p(\tilde{x}, y) \log \frac{p(\tilde{x}, y)}{p(\tilde{x})p(y)} \mathrm{d}\tilde{x}\mathrm{d}y, \tag{1}$$

where $I(\tilde{X}; Y) \leq I(X; Y)$ since compressed data can not convey more information than original ones. The information bottleneck is obtained through

$$\min \mathcal{L}_{\text{IB}} = I(\tilde{X}; X) - \beta I(\tilde{X}; Y), \tag{2}$$

where β is the Lagrange multiplier for the trade-off between preserving meaningful information and compressing at various resolutions.

3.2 Noise Filtering

Retrieval-augmented generation involves in generating contents conditioned on input queries $q \in Q$, where relevant passages $x \in X$ are retrieved to advance the content generation. As illustrated in Figure 2A, recent noise filtering approaches for retrieval-augmented generation learn to compress retrieved passages $\tilde{x} \in \tilde{X} \subseteq X$ for large language models with a log likelihood objective $-\log p_{\text{LM}}(y|[q, \tilde{x}])$ to ground outputs $y \in Y$ (Liu

et al., 2023; Yang et al., 2023a; Xu et al., 2023; Wang et al., 2023), which are special cases of conditional mutual information $I(\tilde{X}, Y|Q)$ and still unable to avoid irrelevant information.

Now we introduce our information bottleneck approach for retrieval-augmented generation and we derive the information bottleneck between retrieved passages X and ground outputs Y conditioned on queries Q given Equation 2. As demonstrated in Figure 2B, the noise filter is required to simultaneously maximizing the mutual information of its compression with ground outputs while minimizing the mutual information with retrieved passages:

$$\min \mathcal{L}_{\text{IB}} = \underbrace{I(\tilde{X}, X|Q)}_{\textit{conciseness}} - \beta \underbrace{I(\tilde{X}; Y|Q)}_{\textit{correctness}}. \quad (3)$$

The former term $I(\tilde{X}; X|Q)$ serves not only to enhance efficiency, which is a common application, but also to promote conciseness by minimizing the inclusion of irrelevant information. As the filtered information becomes increasingly precise, language models are able to reduce the computational resources allocated to input extraction, thereby enhancing their capacity to concentrate on producing higher-quality contents. It's worth noting that this term guarantees that the filtered information will be rendered null when the retrieved content is entirely irrelevant to the output.

Then we provide the details of each term in the information bottleneck. The *conciseness* is:

$$\begin{aligned}
&I(\tilde{X}; X|Q) = \\
&\quad \mathbb{E}_q\Big[\int p(\tilde{x}, x|q) \log \frac{p(\tilde{x}, x|q)}{p(\tilde{x}|q)p(x|q)} \mathrm{d}\tilde{x}\mathrm{d}x\Big] \\
&= \mathbb{E}_q\Big[\int p(x|q)p(\tilde{x}|x, q) \\
&\qquad \log \frac{p(x|\tilde{x}, q)\cancel{p(\tilde{x}|q)}}{p(x|q)\cancel{p(\tilde{x}|q)}} \mathrm{d}\tilde{x}\mathrm{d}x\Big] \\
&\leq \mathbb{E}_q\Big[\int p(x|q) \log \frac{\int p(x|\tilde{x}, q)p(\tilde{x}|x, q)\mathrm{d}\tilde{x}}{p(x|q)} \mathrm{d}x\Big] \\
&= -\mathbb{E}_q\Big[D_{\text{KL}}\Big[p(x|q)||\mathbb{E}_{\tilde{x}\sim p(\tilde{x}|x,q)}p(x|\tilde{x}, q)\Big]\Big],
\end{aligned} \quad (4)$$

where we get an upper bound of *conciseness* based on Jensen's inequality. Therefore, $I(\tilde{X}; X, Q)$ can be converted to the form of the Kullback–Leibler divergence between the retrieval probability distribution $p(x|q)$ and the expectation of the probability to recover retrieved passages from compression $p(x|\tilde{x}, q)$, where $\tilde{x}$ is required to be integrated over the representation space of noise filter $p(\tilde{x}|x, q)$.

In the scenario of offline retrievers, where the retrieved passages and queries are jointly sampled from training datasets $\{(q, x, y)\}$, $p(x|q)$ becomes a constant number and we can simplify:

$$\min I(\tilde{X}; X|Q) \simeq \min \mathbb{E}_{(q,x,\tilde{x})}\big[p(x|\tilde{x}, q)\big]. \quad (5)$$

Hence, in cases where training the retriever jointly is not necessary, minimizing the conditional mutual information $I(\tilde{X}; X|Q)$ can be elucidated as *the process of selectively filtering out information to such an extent that it becomes unfeasible to reconstruct the original content, regardless of the strength of generative language models employed.*

Next, the *correctness* is derived as:

$$\begin{aligned}
&I(\tilde{X}; Y|Q) = H(Y|Q) - H(Y|X, Q) \\
&= -\int p(y, q) \log p(y|q) \mathrm{d}y\mathrm{d}q - H(Y|X, Q)
\end{aligned} \quad (6)$$

where the former term $\mathbb{E}_{(q,y)}\big[\log p(y|q)\big]$ is considered as a constant independent of the noise filter. We can ignore this term and simplify as:

$$\begin{aligned}
&I(\tilde{X}; Y|Q) \simeq -H(Y|\tilde{X}, Q) \\
&= \int p(y, \tilde{x}, q) \log p(y|\tilde{x}, q) \mathrm{d}y\mathrm{d}\tilde{x}\mathrm{d}q \\
&= \mathbb{E}_{(q,\tilde{x},y)}\big[\log p(y|\tilde{x}, q)\big].
\end{aligned} \quad (7)$$

When generative language models are fixed, query-answer pairs $\{(q, y)\}$ are sampled from datasets and $\tilde{x}$ is pre-obtained with noise filter. Therefore, maximizing $I(\tilde{X}; Y|Q)$ approximates maximizing log likelihood $\log p(y|\tilde{x}, q)$, which is explained as *the process of selectively retaining as much useful information as possible to enable the language model generate target outputs.*

Besides, recent studies on the information bottleneck (Fischer, 2020; Federici et al., 2020; Lee et al., 2021; Kawaguchi et al., 2023b) suggest to replace $I(X; \tilde{X})$ with $I(X; \tilde{X}|Y)$, because $I(X; \tilde{X})$ can not be zero while maintaining the target-relevant information. Therefore, we follow Federici et al. (2020) and decompose $I(\tilde{X}; X|Q)$ into two components by using the chain rule as:

$$I(\tilde{X}; X|Q) = I(\tilde{X}; X|Y; Q) + I(\tilde{X}; Y|Q). \quad (8)$$

Finally, our information bottleneck for noise filtering in retrieval-augmented generation is:

$$\begin{aligned}
\mathcal{L}_{\text{IB}} &= I(\tilde{X}; X|Y; Q) - (\beta - 1)I(\tilde{X}; Y|Q) \\
&\simeq \mathbb{E}_{(q,x,\tilde{x},y)}\big[p(x|\tilde{x}, q, y)\big] \\
&\quad - (\beta - 1)\, \mathbb{E}_{(q,\tilde{x},y)}\big[\log p(y|\tilde{x}, q)\big],
\end{aligned} \quad (9)$$

where the Lagrange multiplier $\beta - 1 > 0$.

3.3 Information Bottleneck as a Principle

The information bottleneck represents not merely a methodological approach, but a fundamental principle to be applied in retrieval-augmented generation. In this section, we will delineate three distinct applications of information bottlenecks, encompassing the establishment of an evaluation metric for noise filtering, the creation of supervised-fine-tuning (SFT) training datasets, and the formulation of reward functions in reinforcement learning.

3.3.1 Evaluation Metric

Prior to describing the method for training a noise filter, it is imperative to establish criteria assessing the efficacy of filtration outcomes, where the information bottleneck serves as an important evaluation metric. Given $\{(q, x, y)\}$ from dataset and the compression generated by the noise filter $p(\tilde{x}|x, q)$, based on Equation 9, we define the IB score as:

$$\mathrm{IB}(\tilde{x}) = p_{\mathrm{LM}}(x|[q, \tilde{x}, y]) - \alpha\, p_{\mathrm{LM}}(y|[q, \tilde{x}]), \quad (10)$$

where large language models are employed to estimate probability distributions, with the inputs to the language model comprising concatenated conditional variables. In addition, we balance the magnitude of values by implying logarithms into the hyperparameter α. The range of IB score is $[\text{-}\alpha, 1]$ and smaller IB means better filtration performance.

3.3.2 Supervised Fine-tuning

Training the noise filter from scratch is challenging since there is no ground truth compression of retrieved passages. Although Equation 9 provides a way for achieving the oracle compression, we have to search the optimal one $\tilde{x}$ from all potential subsequences of the retrieved passage x, which is to calculate the integral of $\tilde{x}$ over the language space:

$$\mathbb{E}_{(q,x,y)}\left[\int \mathrm{IB}(\tilde{x})\; p(\tilde{x}|x, q)\mathrm{d}\tilde{x}\right]. \quad (11)$$

The integral is obviously intractable, but we can estimate it with the Monte Carlo sampling strategy. We utilize different existing compression or filtering approaches $\{p_{\theta_1}(\tilde{x}|x, q), \ldots, p_{\theta_n}(\tilde{x}|x, q)\}$ to generate candidate compression outputs, as an approximation of sampling from $p(\tilde{x}|x, q)$, and the candidate of the best IB score is considered as the pseudo $\tilde{x}$. As illustrated in the left part of Figure 2C, we collect pseudo $\tilde{x}$ over the retrieval-augment generation datasets and construct the $\{(q, x, \tilde{x})\}$ tuples as training data for supervised learning of our noise filter $p_\theta(\tilde{x}|x, q)$. In addition, since the noise filter is required to possess capabilities of input understanding and instruction following, we choose pretrained language models as backbones and fine-tune the model to a silver noise filter. The optimization objective is commonly used negative log likelihood $\mathcal{L}_{\mathrm{SFT}} = -\sum_{(q,x,\tilde{x})} \log p_\theta(\tilde{x}|x, q)$.

It's worth noting that our approach shows a strong capability to handle the situation when retrieved passages X are irrelevant to ground outputs Y via minimizing $I(\tilde{X}; X|Y; Q) \to 0$, which usually makes $\tilde{x} \to \phi$. Despite the fact that our information bottleneck objective inherently encompasses optimization objectives for addressing issues related to low-quality information, such as retrieval-free questions, noisy retrieval, and high-loss compression, we demonstrate the incorporation of an additional predictive flag [IS_DISCARD] to determine the necessity of discarding the current filtering outcomes. When $\mathrm{IB}(\phi) < \mathrm{IB}(\tilde{x})$, which means candidate compression contains too little useful information to assist in model generation, we will set [IS_DISCARD] = True and vice versa.

3.3.3 Reinforcement Learning

Through merely supervised fine-tuning, the efficacy of the noise filtering is suboptimal, with its performance being constrained by the quality of compression candidates. Drawing parallels with the reinforcement learning from human feedback (RLHF, Ouyang et al. (2022)) that aligns large language models with human preference, we propose leveraging reinforcement learning to enhance the noise filter by incorporating guidance from the information bottleneck. Our approach involves utilizing direct preference optimization (DPO, Rafailov et al. (2023)), a recent iteration of RLHF that offers ease of implementation and robust stability. We define the preference probability as:

$$p^*(\tilde{x}_1 > \tilde{x}_2|q, x, y) = \sigma(\mathrm{IB}(\tilde{x}_2) - \mathrm{IB}(\tilde{x}_1)), \quad (12)$$

where our noise filter $p_\theta(\tilde{x}|x, q)$ is initially the policy $\pi_{\mathrm{ref}}(\tilde{x}|x, q)$ that needs to generate two compression samples $\tilde{x}_1$ and $\tilde{x}_2$ before. Then the offline dataset of preferences $\mathcal{D} = \{q, x, \tilde{x}_w, \tilde{x}_l\}$ is constructed, where samples ($\tilde{x}_w$ wins and $\tilde{x}_l$ losses) are automatically labeled with the information bottleneck preferences $p^*(\tilde{x}_1 > \tilde{x}_2|q, x, y)$. Hence, the objective for the noise filtering policy $\pi_\theta(\tilde{x}|x, q)$,

Method	NQ						TRIVIAQA					
	words	EM↑	TFR↓	FFR↑	F1↑	IB↓	words	EM↑	TFR↓	FFR↑	F1↑	IB↓
No Retrieval												
LLAMA2-13B	0	16.2	-	-	51.4	-4.46	0	49.9	-	-	76.7	-4.68
Retrieval without Noise Filtering												
Top 1 document	103.6	13.4	56.8	7.7	51.0	-4.29	102.9	46.5	28.3	21.4	75.7	-4.67
Top 5 documents	517.6	14.7	55.8	9.0	48.4	-4.21	514.6	40.7	39.6	21.1	70.9	-4.39
Retrieval with Filtering Method												
RANKGPT	103.6	16.5	51.0	10.3	53.7	-4.47	102.8	47.5	28.3	23.3	76.1	-4.70
LONGLLMLINGUA	141.1	14.7	47.3	7.4	49.9	-4.27	137.2	49.8	24.4	23.9	76.2	-4.61
LLAMA2-7B	37.3	18.3	43.7	**11.0**	52.2	-4.53	30.0	51.4	23.1	**26.0**	76.3	-4.76
Ours w/SFT	10.3	20.6	**17.6**	8.7	54.9	-4.60	11.6	50.3	14.6	15.2	77.4	-4.79
Ours w/SFT w/DPO	12.7	**21.5**	20.3	10.2	**55.9**	**-4.78**	13.3	**52.1**	**12.5**	16.8	**78.2**	**-4.88**

Table 1: Open-domain QA results with LLAMA2-13B as the generator. We report the **word** number of compressed retrieval evidence, which reflects the compression rate. Other evaluation metrics are in §4.1.

which is initialized with $\pi_{\text{ref}}(\tilde{x}|x,q)$, is as follows:

$$\mathcal{L}_{\text{DPO}} = - \mathbb{E}_{(q,x,\tilde{x}_w,\tilde{x}_l)\sim\mathcal{D}}\left[\log\sigma\left(\gamma\log\frac{\pi_\theta(\tilde{x}_w|x,q)}{\pi_{\text{ref}}(\tilde{x}_w|x,q)} - \gamma\log\frac{\pi_\theta(\tilde{x}_l|x,q)}{\pi_{\text{ref}}(\tilde{x}_l|x,q)}\right)\right], \tag{13}$$

where γ is a hyperparameter controlling the deviation from the base reference policy $\pi_{\text{ref}}(\tilde{x}|x,q)$. As shown in the right part of Figure 2C, our information bottleneck can also provide the reward function $\mathcal{R}_{\text{IB}}(\tilde{x}) = -\text{IB}(\tilde{x})$ of online policies.

4 Experiments

4.1 Experimental Settings

Datasets and Retrieval Corpus We conduct experiments on three question answering benchmarks: Natural Questions (NQ) (Kwiatkowski et al., 2019), TRIVIAQA (Joshi et al., 2017) and HOTPOTQA (Yang et al., 2018), with results reported on development sets. We utilize the adversarial Dense Passage Retriever[1] (DPR) (Karpukhin et al., 2020) to retrieve the top 5 passages from all Wikipedia passages for all datasets. The articles are truncated into non-overlapping documents of 100 words.

Implementation Details We use LLAMA2 (Touvron et al., 2023) as the backbone architecture of the large language model. We finetune the 7B model version with LORA (Hu et al., 2021) for noise filtering, where the optimizer is ADAMW with the learning rate of 5e-5 and the batch size of 32. The 13B version is also employed as the generator without any adjustments. For supervised finetuning, we set the coefficient of IB score $\alpha = 10$ to balance the compression.

Besides, filtering candidates are sampled from four different methods derived from traditional extractive summarization, with details in §5.1. For reinforcement learning, we utilize DPO and the hyperparameter $\gamma = 0.1$. The decoding strategy is top-p sampling with $p = 0.9$.

[1] Lucene index of Wikipedia with DPR 100-word splits

Dataset	Type	Examples	% of Recall Top 5	% of Recall Top 1
NQ	Train	106926	36.5	18.9
	Dev	2564	35.0	17.9
TRIVIAQA	Train	87622	68.3	49.1
	Dev	11313	68.5	49.4
HOTPOTQA	Train	90447	51.5	35.1
	Dev	7405	45.0	28.4

Table 2: Overview of the data quantities used for training and testing across three benchmark datasets and recall of the top 5 and top 1 DPR-retrieved passages.

Metrics Question-answering tasks typically employ Exact Match (**EM**) and **F1** as evaluation metrics, while each of them has its own limitations. For example, the EM metric requires an exact alignment between the generated content and the answer, exhibiting an over-strict criterion that lacks semantic compatibility. Besides, the F1 score, functioning as an uni-gram metric, is easy to be cheated by negative terms such as "no." Our **IB** score, on the contrary, is a comprehensive and versatile evaluation metric capable of assessing the *concise-*

ness and *correctness* of generated contents on a semantic level compared to ground answers, leveraging the capabilities of advanced large language models. In addition, to evaluate the performance variations resulting from retrieval augmentation in detail, we use the flip rate of EM to evaluate the extent to which generated responses are influenced by retrieved context. The True-Flip-Rate (**TFR**) and False-Flip-Rate (**FFR**) are defined as follows:

$$\text{TFR} = p(EM_{p_{\text{LM}}(y|[q,x])} = 0 | EM_{p_{\text{LM}}(y|q)} = 1)$$
$$\text{FFR} = p(EM_{p_{\text{LM}}(y|[q,x])} = 1 | EM_{p_{\text{LM}}(y|q)} = 0),$$

where TFR measures the degree of noise introduced by the retrieved information while FFR examines the amount of benefits from retrieval augmentation.

Baselines First we consider the results of the generator (LLAMA2-13B) without retrieval augmentation and with top-1 or top-5 retrieved passages. We next include two sets of filtering methods, reranking using RANKGPT (Sun et al., 2023) distilled from ChatGPT, and prompt compression with LONGLLMLINGUA (Jiang et al., 2023b). We also experiment with the base model LLAMA-7B doing summarization from top-5 passages.

4.2 Open-Domain Question Answering

Table 1 presents the experimental results on NQ and TRIVIAQA datasets. Initially, we demonstrate the performance of the generator (LLAMA2 13B) without retrieval, where a low EM score implies the limitation of large language models facing open-domain question answering task. When the generation is directly augmented with retrieved documents, the result is even worse with a drop of 9.2 EM score on TRIVIAQA at most, reflecting the presence of a large amount of noise in the retrieved content and the vulnerability of the generator to noise interference. Besides, the TFR reveals that noise in retrieved passages will cause over 50% of the generated answers wrong that could have been answered correctly on the NQ dataset.

Extractive noise filters including RANKGPT and LONGLLMLINGUA can compensate for the performance losses in EM, F1, and IB scores caused by retrieval, via noise filtering. However, they can hardly outperform the non-retrieval generator, because they enable the generator to answer more questions correctly with 10.3 of FFR, but also make more mistakes with 51.0 of TFR. Also as an extractive compression method, our approach significantly improves the EM scores on on NQ and TRIVIAQA by 5.3 and 2.2, compared with the non-retrieval generator. For F1 scores, the advancement are 4.5 and 1.5 respectively. Our information bottleneck objectives can achieve considerable FFR improvement while minimizing TFR performance degradation. For instance, on NQ dataset, we effectively minimize TFR (51.0 → 17.6) to alleviate the noise interference caused by retrieval, and maintain a comparable FFR (10.3 → 10.2) to minimize the loss of effective information as much as possible. It's worth noting that our compression rate achieves 2.5% and 2.6% for these datasets, which impressively reduces irrelevant noise and computational cost. Compared to LLAMA2-7B, a large language model with knowledge stored in parameters during pretraining stage where we consider it as a powerful abstractive summarization method, our method still outperforms it in EM and F1 scores.

In addition, our proposed information bottleneck metric IB assesses these models from a comprehensive perspective, which takes both conciseness and correctness of compression into account. The IB score of non-retrieval is the basic score for null compression $\tilde{X} = \phi$, which reflects the capability of the language model $p_{\text{LM}}(\cdot)$. The IB score generally aligns with the model's performance, where models exhibiting lower compression rates and reduced accuracy tend to yield lower scores. Our approach, with the best IB score, achieves the Pareto optimum between compression ratio and performance. We provide case study in Appendix C.

Method	HOTPOTQA					
	words	EM	TFR	FFR	F1	IB
LLAMA2-13B	0	18.5	-	-	53.6	-4.21
Top 1 document	102.9	23.2	31.1	12.8	57.3	-4.36
Top 5 documents	514.7	18.3	50.7	11.2	50.2	-4.10
RANKGPT	102.9	23.5	34.1	13.9	57.2	-4.39
LONGLLMLINGUA	137.7	23.9	32.7	14.0	56.4	-4.19
LLAMA2	27.7	25.9	31.1	**16.2**	57.8	-4.44
Ours	13.2	**26.1**	**22.1**	14.3	**58.3**	**-4.47**

Table 3: Results on the multi-hop HOTPOTQA dataset.

	words	EM↑	TFR↓	FFR↑	F1↑	IB↓
LLAMA2	-	16.2	-	-	51.4	-4.46
Top1	103.6	13.4	56.8	7.7	51.0	-4.29
Top5	517.6	14.7	55.8	9.0	48.4	-4.21
$I(\tilde{X};Y\|Q)$	13.1	19.2	24.9	8.5	54.6	-4.58
IB	**12.7**	**21.5**	**20.3**	**10.2**	**55.9**	**-4.78**

Table 4: Ablation study for *conciseness* on NQ.

Dataset	Filtering Candidates		HasAns	Em	F1	Words	IB	$I(\tilde{X}; X\|Y; Q)$
NQ	Exact	Paragraph-Level	31.4	21.2	53.6	78.1	-4.74	0.597
		Senetence-Level	33.5	23.8	55.4	28.4	-4.81	0.561
	Greedy	Query & Answer	26.6	19	52.1	26.2	-4.64	0.562
		Answer	34.2	24.3	56.8	18.2	-4.91	0.556
	IB Selection		**35.7**	**26.8**	**58.6**	31.6	-5.10	0.563
HotpotQA	Exact	Paragraph-Level	38.3	26.3	55.4	120.0	-4.55	0.679
		Senetence-Level	35.4	27.8	59.3	41.2	-4.63	0.619
	Greedy	Query & Supporting Facts & Answer	31.4	25.8	58.9	32.5	-4.51	0.614
		Supporting Facts & Answer	33.1	26.9	59.5	14.8	-4.63	0.604
	IB Selection		**38.3**	**30.9**	**61.9**	40.4	-4.88	0.619

Table 5: We validate the effectiveness of the information bottleneck in finding the oracle filtered data on the dev sets of NQ and TriviaQA. HasAns denotes the accuracy of filtered results having the answer.

4.3 Multi-Hop Question Answering

Filtering models encounter more significant hurdles when dealing with multi-hop problems, as solving these requires not only auxiliary information but multiple rounds of reasoning and analysis. As demonstrated in Table 3, the results obtained are marginally improved compared to those achieved through supervised training alone.

4.4 Ablation Study for *Conciseness*

We utilize ablation experiments to exhibit the significance of the *conciseness* term in the information bottleneck theory on NQ. As shown in Table 4, integrating the information bottleneck approach that combines both *conciseness* and *correctness* leads to superior outcomes compared to solely utilizing *correctness*. The filtered outcomes from the former approach are more concise and of superior quality.

5 Analysis

5.1 Silver Selection with IB

To visually demonstrate that applying the information bottleneck to select training data can enhance the upper bound, we conducted experiments on the validation sets of NQ and HotpotQA. Table 5 lists two basic filtering methods: exact search and greedy search. The application of IB selection is not limited to just two filtering methods and it can be generalized to other answer-guided silver mining strategies, such as the leave-one-out evidentiality mining strategy (Asai et al., 2022) and Cxmi (Wang et al., 2023). Our approach is not fixated on finding the optimal solution in the initial stages, but rather focuses on gradually approaching the optimal filter model through iterative training. Here we choose the two easiest methods, which can greatly reduce computational costs on the construction of training data

The goal of exact search is to find the paragraphs or sentences containing the ground answers. Greedy search (Nallapati et al., 2017) is one of the most popular heuristic method by far used in extractive summarization. This algorithm extracts oracle labels with the highest Rouge (Lin, 2004) scores compared to human-annotated abstracts. We considered two silver summaries, one that concatenates the query and answer, and the other that focuses solely on the answer itself. The former can cover more information, while the latter focuses more on the answer itself. Specially, the answer in intermediate state, supporting facts, are incorporated for multi-hop questions.

By using information bottleneck for selecting among the four filtering results, the obtained outcome is superior to any of them, which reveals that existing simple annotation methods only yield solutions that are far from optimal.

5.2 Length of Summary and *Conciseness*

Due to the adoption of extractive filtering method, the content retained after filtering is sourced from the original text. This leads to the hypothesis that there may be a correlation between the compression rate and the *conciseness* mutual information $I(\tilde{X}; X|Y; Q)$. We verify the relationship between compression rate and mutual information on NQ and HotpotQA with a toy experiment based on §5.1. For each query-answer pair with four filtering candidates, we calculate their corresponding length and *conciseness* $I(\tilde{X}; X|Y; Q)$. As the statistical information for different samples is independently and identically distributed, we convert the values of length and *conciseness* in each sam-

ple into their rankings among various compression methods. Then we calculate the Pearson correlation coefficient between length $Rank_L$ and *conciseness* $Rank_C$, which is 0.953, indicating a significant correlation at the 0.01 level (two-tailed).

6 Conclusion

We apply the information bottleneck principle to noise filters in retrieval-augmented generation, balancing the trade-off between the conciseness and correctness. Not only as an evaluation method for noise filtering, we also apply it to select supervised fine-tuning training data and provide reward for reinforcement learning. Our two-stage optimization gradually approaches the oracle filtering objective. Experimental results demonstrate that our filter significantly outperforms baseline methods and achieves an impressive compression rate.

Acknowledgements

Kun Zhu and Xiaocheng Feng contribute equally to to this work. Bing Qin is the corresponding author of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Key R&D Program of China via grant No. 2021ZD0112905, National Natural Science Foundation of China (NSFC) via grant 62276078 and U22B2059, the Key R&D Program of Heilongjiang via grant 2022ZX01A32, the International Cooperation Project of PCL, PCL2022D01 and the Fundamental Research Funds for the Central Universities (Grant No.HIT.OCEF.2023018).

Limitations

Although our method has exhibited effectiveness in enhancing the performance of noise filtering task on retrieval-augmented generation, it does have limitations such as performance reliance on the generator and trade-off between True-Flip-Rate (TFR) and False-Flip-Rate (FFR). In order to analyze the mutual information between filtered content $\tilde{X}$ and retrieval content X, as well as between $\tilde{X}$ and Y, it is crucial to utilize a white-box generator equipped with robust capabilities. Furthermore, by introducing the additional predictive flag to assess the necessity of discarding the current filtering outcomes, we successfully decrease TFR, while at a cost of potentially decreasing FFR. We mitigate it by engaging in training iterations, which inevitably leads to an escalation in training cost.

Ethics Statement

We are totally aware that text generation technology has a potential to be used maliciously to generate fake, toxic, or offensive content. If the retrieved content includes harmful or toxic information, it will influence the output of the generated content. Our approach is proposed to mitigate the influence of noise from retrieval, which includes toxic contents. However, there is no assurance that our approach will completely eliminate toxics.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jiaqi Bai, Hongcheng Guo, Jiaheng Liu, Jian Yang, Xinnian Liang, Zhao Yan, and Zhoujun Li. 2023. Griprank: Bridging the gap between retrieval and generation via the generative knowledge improved passage ranking. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 36–46.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Denvy Deng, and Qi Zhang. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation. *arXiv preprint arXiv:2303.08518*.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. 2020. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*.

Ian Fischer. 2020. The conditional entropy bottleneck. *Entropy*, 22(9):999.

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

Ron M Hecht, Elad Noor, and Naftali Tishby. 2009. Speaker recognition by gaussian information bottleneck. In *Tenth Annual Conference of the International Speech Communication Association*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

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*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. Llmlingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023b. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023c. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

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

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. 2023a. How does information bottleneck help deep learning? *arXiv preprint arXiv:2305.18887*.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. 2023b. How does information bottleneck help deep learning? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 16049–16096. PMLR.

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.

Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. 2021. Compressive visual representations. In *Advances in Neural Information Processing Systems*.

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.

Yucheng Li. 2023. Unlocking context constraints of llms: Enhancing context efficiency of llms with self-information-based content filtering. *arXiv preprint arXiv:2304.12102*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Junyi Liu, Liangzhi Li, Tong Xiang, Bowen Wang, and Yiming Qian. 2023. Tcra-llm: Token compression retrieval augmented large language model for inference cost reduction. *arXiv preprint arXiv:2310.15556*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. 2019. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020.

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.

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

Ravid Shwartz-Ziv and Yann LeCun. 2023. To compress or not to compress–self-supervised learning and information theory: A review. *arXiv preprint arXiv:2304.09355*.

Noam Slonim and Naftali Tishby. 2000. Document clustering using word clusters via the information bottleneck method. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 208–215.

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

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*.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE.

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*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. 2019. Bottlesum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. *arXiv preprint arXiv:1909.07405*.

Tailin Wu, Hongyu Ren, Pan Li, and Jure Leskovec. 2020. Graph information bottleneck. *Advances in Neural Information Processing Systems*, 33:20437–20448.

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

Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023a. Prca: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. *arXiv preprint arXiv:2310.18347*.

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

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

A Prompt

A.1 Filter Prompt

We show prompts used to train the filter in Table 6.

A.2 Generator Prompt

We show prompts used for inference in Table 7.

B Experimental Configuration

We use LLAMA2 (Touvron et al., 2023) as the backbone architecture of the large language model. We fine-tune the 7B model version with LoRA (Hu et al., 2021) for noise filtering, where the optimizer is ADAMW with the learning rate of 5e-5 and the batch size of 32. The 13B version is also employed as the generator without any adjustments.

For supervised fine-tuning, we set the coefficient of IB score $\alpha = 10$ to balance the compression. As the value of α decreases, there is a greater tendency

[INST]
≪SYS≫
You are now an intelligent assessment assistant. Your task is to read the context and then find coherent excerpts that can effectively answer the given question.After generating the answer, you need to determine whether the generated excerpt contributes to addressing the question.
≪/SYS≫
Question: {}
Context:
{}
[/INST]
Question: {}
Excerpt: {}
Contribution: [{}]

Table 6: Filter Prompt

to prioritize concise filtering results as the silver selection. Conversely, as the value of increases, there is a greater inclination to prioritize higher correctness. We balance the conciseness and correctness by the hyperparameter. In section 5.1, we conduct experiments on the validation sets of NQ and HotpotQA to visually demonstrate that applying the information bottleneck to select training data can enhance the upper bound. We also conduct experiments about the value of α here. As shown in the table 8 below, as long as the value of α is not zero, its value does not significantly impact the other performance metrics , such as HasAnswer, EM, and F1. $\alpha = 10$ is a good empirical value that balance the conciseness and correctness.

Besides, filtering candidates are sampled from four different methods derived from traditional extractive summarization, with details in §5.1. For reinforcement learning, we utilize DPO and the hyperparameter $\gamma = 0.1$. We use one NVIDIA-A100-40G to train 5 epochs given the question and 5 retrieval passages. For every additional 1W of data, the training time increases by 5.5 hours. We also use one NVIDIA-A100-80G to generate the answer given the question and compression results. We set a maximum length of 1024 tokens for all sequences during training and inference. The generator is configured to generate a maximum of 200 tokens and the decoding strategy is top-p sampling with $p = 0.9$.

C Case Study

In this section, we present exxamples from each of the three datasets within two distinct scenarios. The first scenario is when the retrieved content fails to directly address the question and the compressed content should be empty. Alternatively, the second scenario involves instances where the retrieved content effectively addresses the research question, and the compressed content should be as short as possible.

[INST]
≪SYS≫
You are a helpful, respectful and honest assistant. Your task is to predict the answer to the question based on the given context. If you don't know the answer to a question, please don't share false information. Answer the question as accurately as possible and put the answer in the form [answer].

Here is an example:
Question: Who was the first person killed in a car accident?
Answer: [Bridget Driscoll]

Question: Are both The New Pornographers and Kings of Leon American rock bands?
Answer: [no]

Question: What is the length of the track where the 2013 Liqui Moly Bathurst 12 Hour was staged?
Answer: [6.213 km long]

Question: Which was the first European country to abolish capital punishment?
Answer: [Norway]

(END OF EXAMPLE)
≪/SYS≫
Given the ['question', 'context'], predict the answer to the question.
Question: {}
Context:
{}
[/INST]
Answer: [{}]

Table 7: Generator Prompt

Dataset	α	HasAns↑	EM↑	F1↑	Words	IB↓	loss1↓	loss2↑
NQ	0	29.21	21.10	54.19	7.01	0.548	0.5481	0.5244
	1	35.69	26.79	58.67	16.46	-0.009	0.5541	0.5631
	2	36.08	27.11	58.74	19.74	-0.573	0.5561	0.5645
	3	36.19	27.11	58.81	22.05	-1.138	0.5579	0.5652
	4	36.00	26.95	58.63	24.97	-1.703	0.5595	0.5657
	5	35.88	26.99	58.68	26.43	-2.269	0.5604	0.5659
	6	35.8	26.95	58.66	27.53	-2.835	0.5612	0.5661
	7	35.69	26.87	58.62	28.68	-3.401	0.5618	0.5661
	8	35.61	26.79	58.53	29.70	-3.967	0.5623	0.5662
	9	35.73	26.83	58.59	30.38	-4.534	0.5627	0.5663
	10	**35.73**	**26.79**	**58.56**	**31.63**	**-5.100**	**0.5635**	**0.5663**
	11	35.73	26.79	58.53	31.87	-5.667	0.5637	0.5664
	12	35.65	26.87	58.57	32.73	-6.233	0.5642	0.5663
	13	35.65	26.87	58.57	33.95	-6.799	0.5649	0.5665
	14	35.73	26.95	58.63	34.31	-7.365	0.5652	0.5664
	15	35.69	26.91	58.61	34.49	-7.932	0.5653	0.5665
	16	35.73	26.95	58.63	34.93	-8.498	0.5657	0.5665
	17	35.76	26.95	58.64	35.23	-9.065	0.5659	0.5665
	18	35.76	26.95	58.64	35.33	-9.632	0.5660	0.5665
	19	35.80	26.95	58.63	35.90	-10.198	0.5664	0.5665
	20	35.80	26.95	58.61	35.94	-10.765	0.5664	0.5665
HotpotQA	0	31.74	25.39	57.77	9.46	0.600	0.6000	0.5132
	1	37.25	30.21	61.52	19.86	0.061	0.6058	0.5451
	2	37.61	30.49	61.70	24.91	-0.486	0.6089	0.5472
	3	37.84	30.70	61.91	28.46	-1.033	0.6112	0.5482
	4	37.97	30.80	61.96	31.30	-1.582	0.6130	0.5487
	5	38.11	30.93	62.00	33.37	-2.131	0.6144	0.549
	6	38.15	30.90	62.00	34.97	-2.680	0.6154	0.5492
	7	38.20	30.90	61.98	36.60	-3.229	0.6165	0.5494
	8	38.22	30.89	61.93	37.94	-3.779	0.6173	0.5485
	9	38.27	30.91	61.89	39.30	-4.328	0.6183	0.5496
	10	**38.31**	**30.91**	**61.90**	**40.43**	**-4.878**	**0.6190**	**0.5497**
	11	38.33	30.86	61.89	41.42	-5.428	0.6197	0.5498
	12	38.38	30.87	61.91	42.59	-5.977	0.6205	0.5498
	13	38.35	30.82	61.83	43.64	-6.527	0.6213	0.5499
	14	38.38	30.80	61.82	44.54	-7.077	0.6220	0.5499
	15	38.39	30.75	61.80	45.22	-7.627	0.6225	0.5500
	16	38.43	30.75	61.79	46.23	-8.177	0.6231	0.5500
	17	38.42	30.68	61.73	46.89	-8.727	0.6237	0.5500
	18	38.43	30.70	61.74	47.52	-9.277	0.6241	0.5500
	19	38.45	30.70	61.72	48.33	-9.827	0.6247	0.5501
	20	38.46	30.70	61.72	49.04	-10.377	0.6252	0.5501

Table 8: Experiments about the value of α.

Question: who won the 2017 sports personality of the year
Standard Answer: Mo Farah

Retrienval Passages

"Newsfirst Platinum Awards"

Angelo Mathews won the Most Popular Sports Person of The Year Award. The second Platinum ceremony launching started with January 2017, with the full support of the Ministry of Education. Nominations for Platinum Awards 2017 opened under 20 categories. The promotional campaign commenced at Thurstan College, Colombo on 30 January 2017. The campaign was held in every district in the island until the 28 February 2017. The Awards night was held on 31 March 2017 at Sirasa Stein Studios, Ratmalana. Five times world champion in Lamborghini motor racing, Dilantha Malagamuwa won the Most Popular Sports Person of The Year Award.

"BBC Sports Personality Team of the Year Award"

by Scottish teams; Celtic in 1967, after they became the first British football club to win the European Cup, and the 1990 Grand Slam winning Scotland rugby union squad. Football has had the highest representation among the winners, with 13 recipients. The most recent award was presented in 2017 to the England women's cricket team. This table lists the total number of awards won by nations that the teams have represented. This table lists the total number of awards won by the teams sporting discipline. BBC Sports Personality Team of the Year Award The BBC Sports Personality Team of the

"BBC Young Sports Personality of the Year"

In 2001, the award was replaced by the Young Sports Personality of the Year, and sprinter Amy Spencer was the first recipient of that award. Scottish tennis player Andy Murray, who won in 2004, is the only non-English recipient of the award. The only person to win the award more than once is diver Tom Daley, who won the award three times, in 2007, 2009, and 2010, and was nominated to the ten-person shortlist in five successive years (2007-2011). The most recent award was presented in 2017 to Manchester City midfielder Phil Foden. <nowiki>*</nowiki> Including a Newcomer of the Year

"Simone Biles"

Foundation. She was also one of the finalists for ""Time magazine""'s 2016 Person of the Year. Biles was also nominated for a 2016 ESPY award for Best Female Athlete along with Elena Delle Donne, Katie Ledecky, and Breanna Stewart; Stewart won the award. In July 2017, Biles won the ESPY Award for Best Female Athlete. She is the second gymnast to win this award after Nastia Liukin won it in 2009. In 2017, Simone won the Shorty Awards for the best in sports. At the 2017 Teen Choice Awards, Simone won favorite female athlete, and won Laureus World Sports Award

"BBC Sports Personality of the Year Award"

(2008) and Ennis-Hill (2017), received the BBC Sports Personality of the Year Lifetime Achievement Award. Princess Anne (1971) and her daughter Zara Phillips (2006) are the only award-winners to be members of the same family. The oldest recipient of the award is Dai Rees, who won in 1957 aged 44. Ian Black, who won the following year, aged 17, is the youngest winner. Torvill and Dean, who won in 1984, are the only non-individual winners of the award, so in the 61 years of the award there have been 62 recipients. Of these 13 have been female. 17 sporting disciplines

Without RAG answer: Mo Farah
Top-1-passage RAG answer: Greg Rusedski
Top-5-passage RAG answer: Simone Biles

Method	Summary	words	Answer
RankGPT	"Newsfirst Platinum Awards" Angelo Mathews won the Most Popular Sports Person of The Year Award. The second Platinum ceremony launching started with January 2017, with the full support of the Ministry of Education. Nominations for Platinum Awards 2017 opened under 20 categories. The promotional campaign commenced at Thurstan College, Colombo on 30 January 2017. The campaign was held in every district in the island until the 28 February 2017. The Awards night was held on 31 March 2017 at Sirasa Stein Studios, Ratmalana. Five times world champion in Lamborghini motor racing, Dilantha Malagamuwa won the Most Popular Sports Person of The Year Award.	103	Angelo Mathews
LONGLLMLINGUA	0] "BBC Sports Personality of the Award Document [3 "BBC Sports Person Team of the Angelo Matws won The Pl ceremony launch started with17 with the support of the Ministry of.ations01 2ionalenced Th3 The campaign in every district in the the7 The night117asa Stein Studiosana world in motor racing, D Malwa of The Year Award. Document [1] "BBC Young Sports Personality of the Year" Document [4] "Simone Biles" by Scottish teams; Celtic in 1967, after they became the first British football club to win the European Cup, and the 1990 Grand Slam winning Scotland rugby union squad. Football has had the highest representation among the winners, with 13 recipients. The most recent award was presented in 2017 to the England women's cricket team. This table lists the total number of awards won by nations that the teams have represented. This table lists the total number of awards won by the teams sporting discipline. BBC Sports Personality Team of the Year Award The BBC Sports Personality Team of the who won the 2017 sports personality of the year	177	Simone Biles
LLAMA2	* Winner of the 2017 BBC Sports Personality of the Year: Simone Biles * Winner of the 2017 BBC Sports Personality Team of the Year: England women's cricket team * Winner of the 2017 Platinum Awards Most Popular Sports Person of the Year: Angelo Mathe	45	Simone Biles
Ours	BBC Sports Personality Team of the Year Award The BBC Sports Personality Team of the "Simone Biles" Foundation. Contribution: [No]	0	Mo Farah

Table 9: An example of empty compressed content on NQ.

Question: who did the dominican republic gain its independence from
Standard Answer: Haiti

Retrienval Passages

"Dominican Day Parade"
leadership of General Gregorio Luperon, the war was ultimately won from Spain. In 1844, the Dominican Republic secured its independence from Haiti and became a sovereign state until 1861. Under the leadership of General Pedro Santana, segments of the Dominican population sought to annex the Republic back to Spain and did so during March 18, 1861. On August 16, 1863, the start of the war for the Restoration of the Dominican Republic under the command of General Luperon. The Dominican Republic originally declared its independence from Spain on December 1, 1821. Ultimately, the Dominican Republic was re-established, free from Spain,

"Colorism in the Caribbean"
into preference for lighter skin. Ritualistic skin bleaching to lighten one2̌019s skin, brown paper bag tests to verify one's skin tone, and degradation of darker-complected Haitians as ugly are contemporary manifestations of colorism in Haiti. After declaring its independence from Spanish rule in 1821, the Dominican Republic was overtaken by Haitian rule in 1822. The Dominican Republic did not achieve independence from Haiti until after their victory in the Dominican War of Independence in 1844. However, the country fell back under Spanish rule until it reclaimed its sovereignty after the Dominican War of Restoration of 1865. As the Dominican Republic

"La Trinitaria (Dominican Republic)"
in August 1843 as a result of his dissident activities. La Trinitaria's other members continued the fight in Duarte's absence. One of them was Francisco del Rosario S0̌0e1nchez, who corresponded with Duarte during the latter's exile in Venezuela, and Mat0̌0edas Ram0̌0f3n Mella, who along with Duarte and Sanchez became known as the founding fathers of the Dominican Republic. On January 16, 1844 La Trinitaria's manifesto in favor of independence was released, and the fight for independence began to gain the necessary momentum. On the heels of La Trinitaria's work, and after many battles and much bloodshed, the Dominican Republic was

"Dominican Navy"
Dominican Navy The Navy of the Dominican Republic or Armada de Republica Dominicana (ARD), is one of the three branches of the Military of the Dominican Republic, together with the Army and the Air Force. After the Dominican Republic gained its independence from Haiti on February 27, 1844, there was a need to create a naval fleet. Three schooners were commissioned for this, (flagship), ""Mar0̌0eda Chica"" and ""Leonor"". These were the original three Dominican vessels which were incorporated in the newly created Dominican Navy as authorized by the Junta Central Gubernativa with the Naval Act of 1844 on April 23,

"Dominican Navy"
Brigade, which was under Spanish command and operating in Southern Iraq. Raised in 2008 the MIC serves as the youngest arm of service in the Dominican Navy, which is tasked for amphibious and landing operations. One Marine Battalion is of active service as of the present. It also the youngest Marine unit in all of Latin America. Dominican Navy The Navy of the Dominican Republic or Armada de Republica Dominicana (ARD), is one of the three branches of the Military of the Dominican Republic, together with the Army and the Air Force. After the Dominican Republic gained its independence from

Without RAG answer: Spain
Top-1-passage RAG answer: Spain
Top-5-passage RAG answer: Haiti

Method	Summary	words	Answer
RankGPT	"Dominican Day Parade" leadership of General Gregorio Luperon, the war was ultimately won from Spain. In 1844, the Dominican Republic secured its independence from Haiti and became a sovereign state until 1861. Under the leadership of General Pedro Santana, segments of the Dominican population sought to annex the Republic back to Spain and did so during March 18, 1861. On August 16, 1863, the start of the war for the Restoration of the Dominican Republic under the command of General Luperon. The Dominican Republic originally declared its independence from Spain on December 1, 1821. Ultimately, the Dominican Republic was re-established, free from Spain,	103	Spain
LONGLLMLINGUA	0] "Domin Day Par [4 "La Trinitaria (Dominican Bade, was under Spanish command and operatingq. R0IC serves as theest arm of service in the Dominican, which task amphibious and operations. service of present also youngest in all of America.ican Navy The Navy of the Dominican Republic or Armada de), of Dominican Republic.ican Republic gained its [" ican Navy [Color the Caribbean" leadership of General Gregorio Luperon, the war was ultimately won from Spain. In 1844, the Dominican Republic secured its independence from Haiti and became a sovereign state until 1861. Under the leadership of General Pedro Santana, segments of the Dominican population sought to annex the Republic back to Spain and did so during March 18, 1861. On August 16, 1863, the start of the war for the Restoration of the Dominican Republic under the command of General Luperon. The Dominican Republic originally declared its independence from Spain on December 1, 1821. Ultimately, the Dominican Republic was re-established, free from Spain, who did the dominican republic gain its independence from	171	The Dominican Republic gained its independence from Spain.
LLAMA2	The Dominican Republic gained its independence from Haiti on February 27, 1844, after the Dominican War of Independence. Prior to that, the country had gained independence from Spain in 1821, but was overtaken by Haitian rule in 1822. The Dominican R	41	Haiti
Ours	In 1844, the Dominican Republic secured its independence from Haiti and became a sovereign state until 1861. Contribution: [Yes]	17	Haiti

Table 10: An example of the most concise compressed content on NQ.

Question: What was Walter Matthau's first movie?
Standard Answer: The Kentuckian

Retrienval Passages

"Pete 'n' Tillie"
Walter Matthau received a Golden Globe nomination for Best Actor Ž013 Motion Picture Musical or Comedy, and won the 1973 BAFTA Award for Best Actor in a Leading Role for his performance in this movie and for his performance in ""Charley Varrick"". Carol Burnett received a Golden Globe Award nomination for Best Actress - Motion Picture Musical or Comedy. Pete 'n' Tillie Pete 'n' Tillie is a 1972 American comedy-drama film directed by Martin Ritt and starring Walter Matthau and Carol Burnett. Its advertising tagline was: ""Honeymoon's over. It's time to get married."" Screenwriter Julius J. Epstein was nominated for

"Movers & Shakers"
Movers & Shakers Movers & Shakers is a 1985 American comedy film distributed by MGM, starring Walter Matthau and directed by William Asher. The story follows the head of production at a Hollywood studio who wants to make a movie to fulfill a promise made to a dying friend. The film was written by Charles Grodin, who also appears in the movie. The cast includes Tyne Daly, Gilda Radner, and Vincent Gardenia. Steve Martin makes a cameo appearance as Fabio Longio. Hollywood studio mogul Joe Mulholland (Matthau) vows to produce the pet project of a dying acquaintance, who has been

"The Fortune Cookie"
intention to sue the insurance company lawyers for invasion of privacy and report Purkey's racist remarks to various organizations. Hinkle drives to the stadium, where he finds Boom-Boom ready to leave the team and perhaps become a wrestler named ""The Dark Angel"". Hinkle manages to snap Boom-Boom out of his funk, and the two run down the fields passing and lateraling a football back and forth between them. This was the first film to feature the movie partnership of Jack Lemmon and Walter Matthau, who were to appear together in ten films: Ž020
 Matthau and Lemmon also both had roles

"The Secret Life of an American Wife"
a movie star who is known as an international sex symbol (Walter Matthau). Upon hearing that The Movie Star (the character is not given a name, and Matthau is credited as ""The Movie Star"" in the closing credits) indulges in the services of prostitutes, Victoria decides to pose surreptitiously as one in order to prove to herself that she is still sexually attractive. The Secret Life of an American Wife The Secret Life of an American Wife is a 1968 comedy film written and directed by George Axelrod. The film was released by 20th Century Fox in 1968, and was

"Charles Matthau"
Charles Matthau Charles ""Charlie"" Matthau (born December 10, 1962) is a film and television director and actor and the son of actor Walter Matthau and actress/author Carol Saroyan. Born in New York City, he appeared as a child actor alongside his father in such films as ""Charley Varrick"" (1973), ""The Bad News Bears"" (1976) and ""House Calls"" (1978). Among his directorial projects have been ""The Grass Harp"", from a novella by Truman Capote, and the made-for-TV movie ""The Marriage Fool"", both of which starred his father. He also directed ""Doin' Time on Planet Earth"" (1988), ""Her Minor Thing"" (2005), ""Baby-O""

Without RAG answer: The Kentuckian
Top-1-passage RAG answer: The Taking of Pelham One Two Three
Top-5-passage RAG answer: Pete 'n' Tillie

Method	Summary	words	Answer
RankGPT	"Pete 'n' Tillie" Walter Matthau received a Golden Globe nomination for Best Actor Ž013 Motion Picture Musical or Comedy, and won the 1973 BAFTA Award for Best Actor in a Leading Role for his performance in this movie and for his performance in ""Charley Varrick"". Carol Burnett received a Golden Globe Award nomination for Best Actress - Motion Picture Musical or Comedy. Pete 'n' Tillie Pete 'n' Tillie is a 1972 American comedy-drama film directed by Martin Ritt and starring Walter Matthau and Carol Burnett. Its advertising tagline was: ""Honeymoon's over. It's time to get married."" Screenwriter Julius J. Epstein was nominated for	103	Pete 'n' Tillie
LongLLMLingua	1] "les Matthau Document [Pete 'n' Till "The Fortune a movie star is as (W). that character nameauited Star) the, pose is. Theife 9 film Fox ention and report Purkey's racist remarks to various organizations. Hinkle drives to the stadium, where he finds Boom-Boom ready to leave the team and perhaps become a wrestler named ""The Dark Angel"". Hinkle manages to snap Boom-Boom out of his funk, and the two run down the fields passing and lateraling a football back and forth between them. This was the first film to feature the movie partnership of Jack Lemmon and Walter Matthau, who were to appear together in ten films: Ž020 Matthau and Lemmon also both had roles What was Walter Matthau's first movie?	122	The Fortune
LLAMA2	Walter Matthau's first movie was "The Secret Life of an American Wife" in 1968.	14	The Secret Life of an American Wife
Ours	"The Secret Life of an American Wife" a movie star who is known as an international sex symbol (Walter Matthau). Carol Burnett received a Golden Globe Award nomination for Best Actress - Motion Picture Musical or Comedy. This was the first film to feature the movie partnership of Jack Lemmon and Walter Matthau, who were to appear together in ten films: Ž020 Matthau and Lemmon also both had roles M̈overs & Shakers¨ Movers & Shakers Movers & Shakers is a 1985 American comedy film distributed by MGM, starring Walter Matthau and directed by William Asher. Contribution: [No]	0	The Kentuckian

Table 11: An example of empty compressed content on TriviaQA.

Question: Which was the only eastern bloc country to participate in the 1984 LA Olympics?
Standard Answer: Romania

Retrienval Passages

"1984 Summer Olympics boycott"
the majority of Soviet Bloc countries will not participate in the Games, Ceau0̆219escu's Romania is expected to attend. 1984 Summer Olympics boycott The boycott of the 1984 Summer Olympics in Los Angeles followed four years after the U.S.-led boycott of the 1980 Summer Olympics in Moscow. The boycott involved 14 Eastern Bloc countries and allies, led by the Soviet Union, which initiated the boycott on May 8, 1984. Boycotting countries organized another major event, called the Friendship Games, in July and August 1984. Although the boycott led by the Soviet Union affected a number of Olympic events that were normally

"Nicolae Ceaus0̆326escu"
visit of Egyptian president Anwar Sadat to Israel in 1977. Also Romania was the only country in the world to maintain normal diplomatic relations with both Israel and the PLO. In 1980, Romania participated in the 1980 Summer Olympics in Moscow with its other Soviet bloc allies, but in 1984 was one of the few Communist countries to participate in the 1984 Summer Olympics in Los Angeles when most of the Eastern Bloc's nations boycotted this event. In 1966, Ceau0̆219escu, in an attempt to boost the country's population, made abortion illegal and introduced Decree 770 to reverse the low birth

"Summer Olympic Games"
Eastern Bloc that did attend the 1984 Olympics. These games were perhaps the first games of a new era to make a profit. Although a boycott led by the Soviet Union depleted the field in certain sports, 140 National Olympic Committees took part, which was a record at the time. Again, without the participation of the Eastern European countries, the 1984 Games were dominated by their host country. The Games were also the first time mainland China (People's Republic) participated. According to British journalist Andrew Jennings, a KGB colonel stated that the agency's officers had posed as anti-doping authorities from

"Romania at the Olympics"
Romania at the Olympics Romania first participated at the Olympic Games in 1900, with a single participant. The National Olympic Committee for Romania is the Romanian Olympic and Sports Committee, and was created and recognized in 1914. The nation first sent a team to compete at the Games in 1924, and has only missed two editions each of the Summer Olympic Games and Winter Olympic Games since then. Notably, Romania was the lone Eastern Bloc nation to participate at the 1984 Summer Olympics, which the other nations boycotted. That was also Romania's most successful Olympic Games: they won 20 gold

"Craig Beardsley"
medal in the 200-meter butterfly at the Pan American Games in Caracas, Venezuela. Beardsley failed to qualify for the U.S Olympic team in 1984, by placing third by 0.36 of a second in the U.S. Olympic Trials, and he retired from competitive swimming afterward. Taking advantage of the hard feelings created by the American-led boycott of the 1980 Moscow Olympics, the Soviet Union and other Eastern Bloc countries had pushed for and received Olympic rules changes limiting each country to only two participants in each swimming event. The United States, as the dominant swimming power, was hurt the most. Beardsley

Without RAG answer: East Germany
Top-1-passage RAG answer: Romania
Top-5-passage RAG answer: Romania

Method	Summary	words	Answer
RankGPT	"1984 Summer Olympics boycott" the majority of Soviet Bloc countries will not participate in the Games, Ceau0̆219escu's Romania is expected to attend. 1984 Summer Olympics boycott The boycott of the 1984 Summer Olympics in Los Angeles followed four years after the U.S.-led boycott of the 1980 Summer Olympics in Moscow. The boycott involved 14 Eastern Bloc countries and allies, led by the Soviet Union, which initiated the boycott on May 8, 1984. Boycotting countries organized another major event, called the Friendship Games, in July and August 1984. Although the boycott led by the Soviet Union affected a number of Olympic events that were normally	104	Romania
LONGLLMLINGUA	Document [1] "1984 Summer Olympics boycott Document [] "Sum Olympic Games" ["Romania Eloc the4 These were first of a led the field was., the Eastern European4 by The also (Republic participated.GB thes had- from will not participate in the Games, Ceau0̆219escu's Romania is expected to attend. 1984 Summer Olympics boycott The boycott of the 1984 Summer Olympics in Los Angeles followed four years after the U.S.-led boycott of the 1980 Summer Olympics in Moscow. The boycott involved 14 Eastern Bloc countries and allies, led by the Soviet Union, which initiated the boycott on May 8, 1984. Boycotting countries organized another major event, called the Friendship Games, in July and August 1984. Although the boycott led by the Soviet Union affected a number of Olympic events that were normally Which was the only eastern bloc country to participate in the 1984 LA Olympics?	144	Romania
LLAMA2	Romania was the only Eastern Bloc country to participate in the 1984 Los Angeles Olympics, and it was also the country's most successful Olympic Games, winning 20 gold medals.	29	Romania
Ours	Notably, Romania was the lone Eastern Bloc nation to participate at the 1984 Summer Olympics, which the other nations boycotted. Contribution: [Yes]	20	Romania

Table 12: An example of the most concise compressed content on TriviaQA.

Question: What is the year of the event that occured first, Making Today a Perfect Day was produced, or Frozen was produced?
Standard Answer: 2013

Retrienval Passages

"Making Today a Perfect Day"
Making Today a Perfect Day ""Making Today a Perfect Day"" is a song from the 2015 Walt Disney Animation Studios computer-animated short film ""Frozen Fever"", with music and lyrics by Kristen Anderson-Lopez and Robert Lopez and performed throughout most of the short. It was released as a single in the United States on March 12, 2015. On September 2, 2014, during the ABC airing of """", Walt Disney Animation Studios' chief creative officer John Lasseter announced that a ""Frozen"" short film with a new song would be released in the future. On the same day, ""Variety"" announced that the short

"Making Today a Perfect Day"
the lyrics off-by-heart. Kat Brown of ""The Daily Telegraph"" referred to the short film as a ""musical video"", due to such a large proportion of it being taken up by this song. ""Us Weekly"" negatively compared its catchiness to ""Let It Go"", though described the ditty as ""fresh"", ""bright"", and ""fun"". In a negative review, ""Slate"" felt that ""the song itself, while hummable, is fatally damaged by its need to do too much."" Making Today a Perfect Day ""Making Today a Perfect Day"" is a song from the 2015 Walt Disney Animation Studios computer-animated short film ""Frozen Fever"", with music

"Making Today a Perfect Day"
would be released in early 2015 under the title ""Frozen Fever"", with Chris Buck and Jennifer Lee returning as co-directors, Peter Del Vecho returning as producer and a new song by Kristen Anderson-Lopez and Robert Lopez. In a mid-October interview, Idina Menzel revealed that the cast had already recorded their vocal tracks, stating ""We just worked on a short for ""Frozen""."" On December 3, 2014, it was announced that Aimee Scribner would be a co-producer and that ""Frozen Fever"" would debut in theaters alongside Walt Disney Pictures' ""Cinderella"" on March 13, 2015. In late December, the co-directors told the Associated

"Making Today a Perfect Day"
to start brainstorming possibilities. After early discussions about Olaf, head story artist Marc Smith pitched the idea of what might happen if Elsa had a cold, which became the basis for the short's plot. The directors began working on the short in June and by August were back in the recording studio with the cast to lay down vocal tracks. The short features the song ""Making Today a Perfect Day"", by Anderson-Lopez and Lopez. At the premiere of ""Cinderella"" and ""Frozen Fever"" at the El Capitan Theatre in Hollywood, California, on March 1, 2015, Josh Gad told ""USA Today"", ""I

"Making Today a Perfect Day"
want to apologize to parents everywhere for the fact that children are going to be singing a whole new ""Frozen"" song..."" Gad's wife noticed he was still humming it two days after he recorded his lines. In ""Making Today a Perfect Day"" there is a lyrical reference to ""Let It Go""2014Elsa notes to Anna that ""a cold never bothered me anyway"", this time re-purposing the line by referring to an actual cold. ""Billboard"" suggests that the songwriting duo included this Easter egg because they ""know exactly what the fanbase wanted"". The beginning of the song also includes a passage to

Without RAG answer: 2013
Top-1-passage RAG answer: 2015
Top-5-passage RAG answer: Frozen was produced first.

Method	Summary	words	Answer
RankGPT	"Making Today a Perfect Day" Making Today a Perfect Day ""Making Today a Perfect Day"" is a song from the 2015 Walt Disney Animation Studios computer-animated short film ""Frozen Fever"", with music and lyrics by Kristen Anderson-Lopez and Robert Lopez and performed throughout most of the short. It was released as a single in the United States on March 12, 2015. On September 2, 2014, during the ABC airing of """", Walt Disney Animation Studios' chief creative officer John Lasseter announced that a ""Frozen"" short film with a new song would be released in the future. On the same day, ""Variety"" announced that the short	105	2015
LONGLLMLINGUA	Making Today a Perfect ""aking Today a Perfect Day is a song from the 20 Walt Disney Animation Studios-animated short ""Fro F"", music and lyrics Kristen Anderson-Lopez and Robert Lopez throughout most of short It was single in United States on March 12, 2015. On September 2, 2014, during ABC airing of Walt Disney Animation chief creative officer John Lass announced aFrozen"" short film with a new would be released future same day,Vari announced that the [Making Today a Document [4M Day [fect Day" Frozen Fever"", with music What is the year of the event that occured first, Making Today a Perfect Day was produced, or Frozen was produced?	109	2015
LLAMA2	The event "Making Today a Perfect Day" was produced first, in 2015.	12	2015
Ours	The short features the song ""Making Today a Perfect Day"", by Anderson-Lopez and Lopez. Contribution: [No]	0	2013

Table 13: An example of empty compressed content on HotpotQA.

Question: Salisbury Woodland Gardens links a zoo with a park designed and built under the watchful eye of who?
Standard Answer: Thomas Mawson

Retrienval Passages

"Salisbury Woodland Gardens, Blackpool"
enable the local community to get more involved in the sites management and interpretation. Salisbury Woodland Gardens, Blackpool Salisbury Woodland Gardens is an open space located in the east of Blackpool, flanked by East Park Drive and Woodside Drive and linking Blackpool Zoo with Stanley Park. Known simply as the 'Woodland Gardens' to local people, the site was acquired in 1924 by Blackpool Corporation and was originally developed as a shelter belt for the adjacent Stanley Park Golf Course. The gardens were later developed in the 1940s as an arboretum and public open space for all to enjoy. It was

"Salisbury Woodland Gardens, Blackpool"
Salisbury Woodland Gardens, Blackpool Salisbury Woodland Gardens is an open space located in the east of Blackpool, flanked by East Park Drive and Woodside Drive and linking Blackpool Zoo with Stanley Park. Known simply as the 'Woodland Gardens' to local people, the site was acquired in 1924 by Blackpool Corporation and was originally developed as a shelter belt for the adjacent Stanley Park Golf Course. The gardens were later developed in the 1940s as an arboretum and public open space for all to enjoy. It was renovated in 1967 by Peter Perry and his 'Flying Squad (see below). Popular once

"Salisbury Woodland Gardens, Blackpool"
as a wedding photograph location, the site went into decline during the 1990s. The Council's Ranger Service manage and protect the gardens which they took over in September 2006 and have been funding and undertaking the restoration of the woodland. In 1967, Parks Director Norman Leach appointed gardener Pete Perry and his Flying Squad of gardeners to plant up the gardens. All plants, (primulas, meconopsis, etc.) were grown from seed in the greenhouses at Stanley Park, and planted ""en masse"". Extra shrubs, including azalea were also planted. The neighbouring Blackpool Zoo site was formerly Blackpool's municipal airport. In 1927 the

"Stanley Park, Blackpool"
Stanley Park, Blackpool Stanley Park is a public park in the town of Blackpool on the Fylde coast in Lancashire, England. It is the town's primary park and covers an area of approximately . The park was designed to include significant sporting provisions, along with formal gardens, a boating lake and woodland area. It was designed and built in the 1920s, under the eye of Thomas Mawson. It is located in the Great Marton and Layton areas of the town. It is Grade II* listed and is on the Register of Historic Parks and Gardens of special historic interest in

"ZooTampa at Lowry Park"
Larry Killmar, the zoo's Director of Collections who had authorized many of Salisbury's questionable animal transfers. Under Killmar, the zoo reorganized its internal policies over several months, and on March 27, 2009, the AZA reinstated the membership of both Lowry Park Zoo and its director of collections. The saga came to a close in August 2009 when Salisbury and the Lowry Park Zoo board agreed to a settlement in which Salisbury paid $2,200 and agreed to return all the structures, fencing, and equipment that the zoo had built at Safari Wild but did not admit to any wrongdoing. ZooTampa at

Without RAG answer: Capability Brown
Top-1-passage RAG answer: Blackpool Corporation
Top-5-passage RAG answer: Peter Perry.

Method	Summary	words	Answer
RankGPT	"Salisbury Woodland Gardens, Blackpool" enable the local community to get more involved in the sites management and interpretation. Salisbury Woodland Gardens, Blackpool Salisbury Woodland Gardens is an open space located in the east of Blackpool, flanked by East Park Drive and Woodside Drive and linking Blackpool Zoo with Stanley Park. Known simply as the 'Woodland Gardens' to local people, the site was acquired in 1924 by Blackpool Corporation and was originally developed as a shelter belt for the adjacent Stanley Park Golf Course. The gardens were later developed in the 1940s as an arboretum and public open space for all to enjoy. It was	104	Blackpool Corporation
LONGLLMLINGUA	Document [] "Salisbury Woodland Gard Blackpool [1 "Salisbury Woodland Gard, Blackpool"]Sbury, Sburypool located eastpool by East Drive Drive and linkingpooloo withown asood local site acquired4pool Corporationt for Theens were0um and for. It97 by andFlying Squad (see below). Popular once enable the local community to get more involved in the sites management and interpretation. Salisbury Woodland Gardens, Blackpool Salisbury Woodland Gardens is an open space located in the east of Blackpool, flanked by East Park Drive and Woodside Drive and linking Blackpool Zoo with Stanley Park. Known simply as the 'Woodland Gardens' to local people, the site was acquired in 1924 by Blackpool Corporation and was originally developed as a shelter belt for the adjacent Stanley Park Golf Course. The gardens were later developed in the 1940s as an arboretum and public open space for all to enjoy. It was Salisbury Woodland Gardens links a zoo with a park designed and built under the watchful eye of who?	158	Flying Squad
LLAMA2	Salisbury Woodland Gardens in Blackpool was designed and built under the watchful eye of Thomas Mawson in the 1920s, as part of Stanley Park. The gardens were later renovated in 1967 by Pete Perry and his 'Flying Squad'. The site is now managed by th	45	Thomas Mawson
Ours	It was designed and built in the 1920s, under the eye of Thomas Mawson. It was designed and built in the 1920s, under the eye of Thomas Contribution: [Yes]	27	Thomas Mawson

Table 14: An example of the most concise compressed content on HotpotQA.