Familiarity-Aware Evidence Compression for Retrieval-Augmented Generation

Dongwon Jung[1] **Qin Liu**[1] **Tenghao Huang**[2] **Ben Zhou**[3] **Muhao Chen**[1]
[1]University of California, Davis, [2]University of Southern California, [3]Arizona State University
{dwojung,qinli,muhchen}@ucdavis.edu tenghaoh@usc.edu benzhou@asu.edu

Abstract

Retrieval-augmented generation (RAG) improves large language models (LMs) by incorporating non-parametric knowledge through evidence retrieved from external sources. However, it often struggles to cope with inconsistent and irrelevant information that can distract the LM from its tasks, especially when multiple evidence pieces are required. While compressing the retrieved evidence with a compression model aims to address this issue, the compressed evidence may still be unfamiliar to the target model used for downstream tasks, potentially failing to utilize the evidence effectively. We propose FAVICOMP (**FA**miliarity-aware E**VI**dence **COMP**ression), a novel inference-time evidence compression technique that makes retrieved evidence more familiar to the target model, while seamlessly integrating parametric knowledge from the model. Experimental results show that FAVICOMP consistently outperforms the most recent evidence compression baselines across multiple open-domain QA datasets, improving accuracy by up to 28.1% while achieving high compression rates. Additionally, we demonstrate the effective integration of both parametric and non-parametric knowledge during evidence compression. [1]

1 Introduction

Retrieval-augmented generation (RAG) has become a common paradigm for large language models (LMs) to leverage external knowledge beyond their inherent knowledge boundaries to perform better in knowledge-intensive tasks such as open-domain question answering (QA) (Lewis et al., 2020; Izacard and Grave, 2021; Guu et al., 2020) and fact-checking (Pan et al., 2023; Li et al., 2024c). In particular, incorporating multiple evidence pieces is crucial in solving complicated tasks such as multi-hop and complex reasoning (Trivedi et al., 2023; Jiang et al., 2023b; Li et al., 2024b; Lu et al., 2023), which require various sources of information to solve the questions.

Nevertheless, RAG often struggles to cope with inconsistent and irrelevant information from the multiple evidence pieces, which can interfere with downstream tasks (Shi et al., 2023). This highlights the need for evidence compression to identify and retain only the essential information for LMs to utilize effectively. Traditionally, evidence compression has focused on reranking documents or sentences by relevance and then incorporating a top-ranked subset (Nogueira et al., 2020; Zhuang et al., 2023; Wang et al., 2023c) or compressing the documents into a compact form that retains only essential context (Jiang et al., 2023a; Xu et al., 2024; Yoon et al., 2024). However, the compressed evidence might be unfamiliar to the LM employed for the downstream task (referred to as the target model), particularly due to discrepancies in the internal knowledge and prompt preferences between the compression model and the target model (Gonen et al., 2023; Lee et al., 2024; Li et al., 2024a; Mallen et al., 2023). When LMs encounter unfamiliar contextual information, they often fail in balancing parametric and non-parametric knowledge, either by overly relying on their parametric knowledge (Longpre et al., 2021; Wang et al., 2023a; Zhou et al., 2023) or by utilizing retrieved evidence without considering its relevance to the input (Wu et al., 2024).

To address these challenges, we propose **FA**miliarity-aware E**VI**dence **COMP**ression (FAVICOMP), an inference-time evidence compression method that consolidates multiple evidence into an abstractive summary that is more familiar to the target model, while seamlessly integrating parametric knowledge from the model. Inspired by the prior findings that an LM's familiarity with a prompt is generally reflected by low perplexity

[1]Code and data are available at https://github.com/luka-group/FaviComp

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 16181–16196
November 4-9, 2025

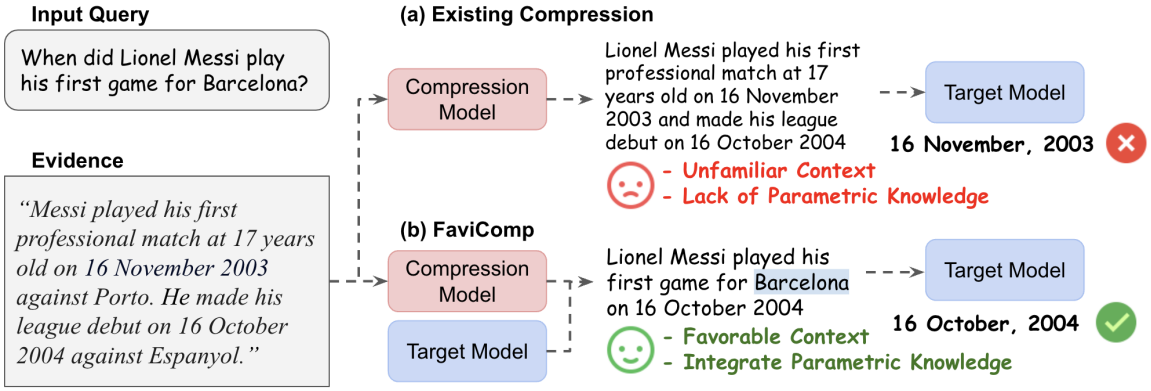

Figure 1: An overview of FAVICOMP. Instead of relying solely on compressed evidence from the compression model (upper), FAVICOMP familiarizes the compressed evidence to the target model while integrating parametric knowledge through ensemble decoding, resulting in improved downstream performance (lower).

(Liu et al., 2024; Gonen et al., 2023; Wang et al., 2023b), FAVICOMP proactively composes the compressed evidence in a way to lower the perplexity of the target model. Specifically, instead of directly selecting the highest probability token from the compression model at each decoding step, FAVICOMP selects the token from the ensemble of the token probabilities from both the compression and target models. This ensemble decoding therefore constrains the token search space of the compression model to those with lower perplexity for the target model, making the context more familiar to the target model (Liu et al., 2024).

Furthermore, FAVICOMP potentially synergizes the retrieved knowledge with the target model's parametric knowledge introduced during ensemble decoding. It can effectively discern when to leverage internal or external knowledge, which is particularly beneficial in the presence of noisy contextual evidence in complex tasks such as multi-document or multi-hop QA (Wang et al., 2024).

Our experiments show that FAVICOMP outperforms most recent evidence compression baselines in five open-domain QA datasets, improving accuracy by up to 28.1% while maintaining high compression rates. Additionally, we conduct ablation studies by varying the degree of decoding ensemble and analyzing its impact on performance and context perplexity. Moreover, we investigate how FAVICOMP effectively integrates parametric and non-parametric knowledge during evidence compression.

2 Method

We present FAVICOMP, a inference-time evidence compression method that familiarizes retrieved evidence with the target model while synergizing them with the model's parametric knowledge. We first illustrate the motivation for FAVICOMP in §2.1 and provide the preliminaries of evidence compression in RAG §2.2, followed by a detailed definition of our proposed framework in §2.3.

2.1 Motivation and Method Overview

Figure 1 illustrates the overview of FAVICOMP. Existing evidence compression methods employ the compression model to filter out irrelevant information from the retrieved documents. However, since the compression model and the target model are different, the target model might not be familiar to the compressed evidence due to the difference in internal knowledge and prompt preferences between the two models (Gonen et al., 2023; Lee et al., 2024; Mallen et al., 2023). In addition, the compressed evidence cannot be supplemented with the rich parametric knowledge from the target model. In the example, even though the compression model successfully summarizes the essential information, the target model produces an inaccurate answer due to the unfamiliarity with the target model and the lack of integration of the parametric knowledge. On the other hand, FAVICOMP compresses the given evidence more favorable to the target model by using a novel ensemble decoding technique and leverages its parametric knowledge to supplement the missing evidence (*"Lionel Messi made his league debut in Barcelona"*), effectively combining evidential and parametric knowledge.

2.2 RAG with Evidence Compression

Given a set of k retrieved evidence snippets $D = \{d_1, d_2, \ldots, d_k\}$ and a textual input sequence x, RAG aims to generate an output sequence y, con-

ditioned on both D and x. However, RAG directly utilizes D which often contains irrelevant information to x, potentially confusing the target model in downstream tasks (Shi et al., 2023). Thus, we use an additional compression model to condense D into a concise and input-relevant context c, which is then used in place of D during the downstream generation process. Thus, the RAG with evidence compression is formalized as:

$$y^* = \arg\max_y P_{\text{tar}}(y \mid x, \hat{c}),$$
$$\hat{c} = P_{\text{comp}}(c \mid x, [d_1, d_2, \ldots, d_k]),$$

where y^* is the final output sequence, $[\cdot, \cdot]$ denotes concatenation, and P_{tar} and P_{comp} represent the probability distributions of the target and compression models, respectively. In this work, we consider any natural language prompting tasks, such as open-domain QA tasks, where x represents the input prompt (also known as the query in QA tasks) and y^* denotes the output sequence.

The compression model's objective is to produce a concise yet informative summary c of the evidential documents D that captures the essential information relevant to the input query x. We use an unsupervised approach, where the model is instructed to generate a query-relevant summary of D in a zero-shot manner using an evidence compression instruction prompt, denoted as I_{comp}, such as the one below:

Evidence Compression Instruction

Given a question and multiple document snippets, generate one summarized context that is helpful to answer the question.

Specifically, the evidence compression is done in an auto-regressive way formalized as,

$$P_{\text{comp}}(c \mid \mathcal{C}_{\text{comp}}) = \prod_{i=1}^{|c|} P_{\text{comp}}(c_i \mid \mathcal{C}_{\text{comp}}, c_{<i}),$$

where $\mathcal{C}$ denotes the input prompt, constructed by stringifying $\{I_{\text{comp}}, x, D\}$ using a predefined prompt template and $|c|$ is the length of the summary c.

2.3 Ensemble Decoding for FAVICOMP

Simple compression techniques might lead to subpar performance in downstream tasks because the compressed evidence may not be familiar to the target model. To better align the context to the target model, FAVICOMP proactively composes it to lower the target model's perplexity by introducing a constraint in decoding space from the target model during the evidence compression. FAVICOMP achieves this goal through ensemble decoding, which involves a multiplicative ensemble of two LMs—compression model and target model—at each decoding step.

Specifically, the target model is instructed to generate a context c that would be helpful in answering the question x without referencing the evidence set. This is also done in zero-shot using a context generation instruction prompt I_{gen} such as:

Context Generation Instruction

Given a question, generate a context that is helpful to answer the question.

The context generation is also performed in an auto-regressive fashion, represented as:

$$P_{\text{tar}}(c \mid \mathcal{C}_{\text{gen}}) = \prod_{i=1}^{|c|} P_{\text{tar}}(c_i | \mathcal{C}_{\text{gen}}, c_{<i}),$$

where $\mathcal{C}_{\text{gen}}$ denotes the input prompt constructed using $\{I_{\text{gen}}, x\}$[2] and $|c|$ denotes the length of the generated context c.

Once the compression model and the target model generate their respective probability distributions for the next token, the subsequent token is chosen by maximizing the weighted sum of the log probabilities from both models. The selected token is the continuation of the previously generated text aligned with their objectives. This process is formalized as follows:

$$c_i = \arg\max_{c_i', c_i'' \in V} (\alpha \cdot \log P_{\text{tar}}(c_i' \mid \mathcal{C}_{\text{gen}}, c_{<i}) + (1 - \alpha) \cdot \log P_{\text{comp}}(c_i'' \mid \mathcal{C}_{\text{comp}}, c_{<i})),$$

where c_i is the subsequent token, and α is the ensemble coefficient that weighs between the two probability distributions. We demonstrate how the coefficient α impacts both the perplexity and the downstream performance in §4.2.

Ensemble decoding proactively shifts the token search space in evidence compression by upweighting those tokens with lower perplexity from the target model's perspective, resulting in a compressed evidence that is more familiar to the target model. Note that since both objectives ultimately share the

[2]We provide the prompt templates for evidence compression and context generation in Table 10.

goal of generating context relevant to the question, combining the logits ensures alignment with this ultimate goal.

In addition, ensemble decoding enables FAVICOMP to seamlessly integrate both retrieval knowledge from the external evidence set and the target model's parametric knowledge. Specifically, FAVICOMP selects the $\arg\max$ token from the target model only when the token's probability is higher than that of the compression model, demonstrating that FAVICOMP draws on parametric knowledge only when necessary—potentially when the compression model is uncertain about the next token. This is particularly beneficial for complex tasks like multi-document QA, where the evidence set may not include all the necessary information (Mallen et al., 2023). In such cases, the missing information in compressed evidence can be supplemented by tokens generated from context generation by the target model, which is entirely based on parametric knowledge. We demonstrate in §4.3 and §5 that FAVICOMP can incorporate knowledge from both sources effectively, leading to a performance boost compared to compression methods that solely focus on distilling knowledge from the evidence set.

3 Experimental Settings

We assess the effectiveness of FAVICOMP on knowledge-intensive QA tasks. In this section, we delve into the details of the experimental settings.

3.1 Datasets

We evaluate FAVICOMP on five open-domain QA datasets, including two single-document QA datasets, Natural Questions (NQ; Kwiatkowski et al. 2019) and TriviaQA (TQA; Joshi et al. 2017), and three multi-document QA datasets, HotpotQA (HQA; Yang et al. 2018), 2WikiMultiHopQA (Wiki; Ho et al. 2020), and MuSiQue (MQ; Trivedi et al. 2022). Following prior studies (Asai et al., 2023; Xu et al., 2024), we evaluate the performance on the development set of each dataset using two evaluation metrics, Accuracy (Acc) and token-level F1.

3.2 Implementation Details

For all the comparison methods, we utilize `Llama3-8B-Instruct` and `Mixtral-8x7B-Instruct` as the target model to tackle downstream QA tasks with RAG. For FAVICOMP and Zero-shot Summarization, we employ two compression models, one for each target model: **`Llama3.2-3B-Instruct`** for `Llama3-8B-Instruct` target model and **`Mistral-7B-Instruct`** for `Mixtral-8x7B-Instruct` target model. For each question, we retrieve five documents from 2018 Wikipedia corpus (Karpukhin et al., 2020) using Contriever-MSMARCO (Izacard et al., 2021), so as to be consistent with previous studies (Xu et al., 2024; Yoon et al., 2024). We set ensemble coefficient α of FAVICOMP to 0.5 by default, for which more analyses are given in §4.2. The prompts used in the experiment are presented in Appx. §C.

3.3 Baselines

We consider the following categories of baselines. (1) **No Context**: RAG without any context. (2) **Gold Compression**: RAG using directly relevant evidence from the retrieved documents if they exist. (3) **Raw Document**: RAG with raw documents that have not undergone any compression. (4) **Generated Context** (Yu et al., 2023): RAG with context generated by the same LM as the target model. This is equivalent to FAVICOMP with $\alpha = 1$, as we rely solely on the target model to generate context when $\alpha = 1$. (5) **Reranking-based Methods**: We rerank sentences in the evidence set and choose top-ranked sentences as the context. We utilize two rerankers—Sentence-BERT (Reimers and Gurevych, 2020) and RECOMP-extractive (Xu et al., 2024). (6) **Compression-based Methods**: We employ four compressors—LongLLMLingua (Jiang et al., 2023a), RECOMP-abstractive (Xu et al., 2024), CompAct (Yoon et al., 2024), and Zero-shot Summarization. For Zero-shot Summarization, we use the same evidence compression instruction prompt of FAVICOMP to summarize multiple evidence using the same LM as the target model. This is equivalent to FAVICOMP with $\alpha = 0$, as we depend entirely on the compression model without any intervention from the target model.[3]

4 Experimental Results

In this section, we compare the overall performance of FAVICOMP with other baselines across the five datasets (§4.1), explore the impact of ensemble coefficient α on performance and perplexity (§4.2),

[3] A more detailed explanation of the implementation of the baselines is provided in Appx. §A.

Methods	Size	NQ		TQA		HQA		Wiki		MQ	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Llama3-8B-Instruct											
Gold Compression	-	-	-	-	-	42.3	51.3	35.7	40.0	10.2	17.7
No Context	-	26.9	31.9	57.2	61.2	19.1	25.5	20.5	25.0	5.4	13.0
Raw Document	-	42.6	**47.1**	67.6	70.8	30.3	38.7	22.0	26.8	8.2	15.0
Generated Context	-	32.3	36.6	59.7	62.4	22.7	29.7	24.8	28.7	7.6	14.8
Sentence-BERT	110M	30.3	35.4	59.2	62.9	22.4	29.6	18.1	22.9	7.7	14.8
RECOMP-extractive	110M†	33.7	38.1	59.4	62.8	22.5	29.8	18.0	22.4	8.1	15.5
LongLLMLingua	7B†	35.4	40.9	64.8	67.6	25.9	34.7	19.2	24.2	7.7	14.4
RECOMP-abstractive	775M†	39.3	43.3	62.9	66.1	27.0	34.8	20.5	25.0	7.3	14.8
CompAct	7B†	42.3	46.1	67.0	69.7	29.8	37.5	21.4	26.6	9.2	16.9
Zero-shot Summarization	3B	39.4	43.2	64.2	67.1	30.1	38.5	25.7	31.1	7.7	15.3
FAVICOMP	3B	**42.8**	46.8	**68.0**	**70.9**	**33.0**	**41.6**	**29.6**	**35.2**	**10.8**	**19.9**
Mixtral-8x7B-Instruct											
Gold Compression	-	-	-	-	-	48.2	55.1	49.9	51.9	12.9	18.6
No Context	-	36.7	38.4	68.9	72.0	25.1	31.6	32.5	35.9	6.4	11.8
Raw Document	-	**46.3**	42.1	72.1	71.1	34.0	39.0	32.9	36.3	10.1	15.6
Generated Context	-	33.6	33.9	61.4	62.9	26.5	32.9	30.2	34.3	7.2	13.4
Sentence-BERT	110M	36.8	36.8	67.0	68.7	28.3	34.5	32.5	36.2	9.9	15.2
RECOMP-extractive	110M†	38.0	37.9	66.7	68.0	28.7	34.3	31.8	34.9	9.4	15.6
LongLLMLingua	7B†	40.1	39.4	70.5	71.0	32.0	38.3	31.9	36.1	9.7	15.9
RECOMP-abstractive	775M†	42.1	41.3	68.4	69.4	32.3	38.5	32.2	36.2	7.9	13.6
CompAct	7B†	44.1	43.4	70.3	71.4	35.2	41.6	35.9	39.5	11.2	16.9
Zero-shot Summarization	7B	42.1	40.6	65.9	67.0	31.4	38.1	28.5	32.8	8.4	13.8
FAVICOMP	7B	43.6	**44.5**	**72.6**	**73.9**	**36.3**	**44.4**	**40.5**	**45.2**	**13.4**	**19.9**

Table 1: Experimental results on five open-domain QA datasets. **Size** column represents the size of the compression model used for each method. † indicates a fully-supervised compression model, where the compressor is trained.

investigate how effectively FAVICOMP incorporate parametric and non-parametric knowledge (§4.3), and compare the compression rates with other baselines (§4.4).

4.1 Main Results

The overall performance of FAVICOMP and the baselines across the five datasets are presented in Table 1.[4] To start with, the compression-based methods consistently outperform the reranking-based methods, due to the fact that the reranking-based methods are prone to losing more question-relevant information by discarding lower-ranked sentences.

Next, FAVICOMP outperforms all other baselines across all the datasets, except for the Gold Compression which is regarded as the upper bound of the performance. It is noteworthy that FAVICOMP, as a training-free strategy, outperforms all the supervised compression-based baselines that use similar or larger compression models[5]. This result suggests that knowledge distillation from a larger teacher LM to a smaller compression model may not generalize well, as the context preferences and prior knowledge of the target model and the teacher model are likely to differ. In contrast, the superior performance of FAVICOMP is attributed to its ability to familiarize evidence with the target model and its effective incorporation of parametric knowledge from ensemble decoding. Moreover, for the MQ dataset, FAVICOMP even outperforms Gold Compression baseline which can be viewed as a perfect compressor. This demonstrates that explicitly incorporating parametric knowledge from the target model can significantly enhance performance in multi-document QA, even when the context is imperfect.

Finally, given that Zero-shot Summarization corresponds to FAVICOMP with $\alpha = 0$ and Generated Context corresponds to FAVICOMP with $\alpha = 1$,

[4]We present additional experimental results using other combinations of compression and target model at Appx. §B.1.

[5]We conduct a fair comparison with RECOMP-abstractive by using the same base compression model in Appx. §B.2.

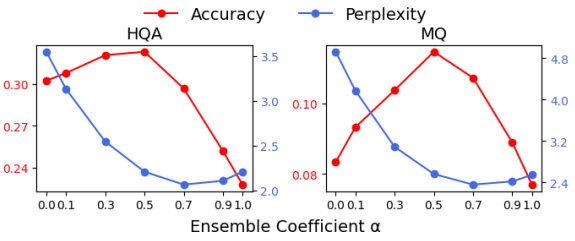

Figure 2: Impact of coefficient α on performance and perplexity when using `Llama3.2-3B-Instruct` and `Llama3-8B-Instruct` compression-target pairs.

Methods	NQ	TQA	HQA	Wiki	MQ
Generated Context	36.6	62.4	29.7	28.7	14.8
Zero-shot Summarization	43.2	67.1	38.5	31.1	15.3
Concatenation	42.5	66.7	36.5	29.0	15.6
FAVICOMP	**46.8**	**70.9**	**41.6**	**35.2**	**19.9**

Table 2: Performance (F1) comparison against concatenation of parametric and non-parametric knowledge.

the fact that FAVICOMP outperforms both baselines highlights its ability to effectively incorporate tokens from both sources—evidence summary and generated context. This results in superior performance compared to relying on one source alone.

4.2 Impact of Ensemble Coefficient on Performance and Perplexity

Figure 2 illustrates how performance and perplexity change as the ensemble coefficient α is varied across the values when using `Llama3.2-3B-Instruct` and `Llama3-8B-Instruct` compression-target pairs on HQA and MQ datasets[6]. We calculate the perplexity of the compressed evidence conditioned on the preceding inputs, i.e. instruction, demonstrations, and the question. For all the datasets, performance is the highest when $\alpha = 0.5$, indicating that proactively lowering perplexity by equally weighting both input sources yields the best results. When α is below 0.5, performance improves as the perplexity of compressed evidence decreases, which aligns with the previous works (Liu et al., 2024; Gonen et al., 2023). However, when α exceeds 0.5, performance declines as perplexity decreases due to the lack of evidential knowledge during evidence compression. Additionally, when α reaches 0.9 or 1.0, there is a slight rise in the perplexity due to LM's increased uncertainty with limited evidential knowledge.

[6]Results for other datasets are included in Figure 6.

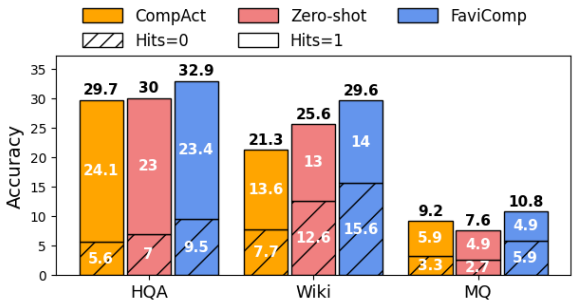

Figure 3: Accuracy of baselines methods on Hits $= 0$ and Hits $= 1$ subset of multi-document QA datasets.

4.3 Integration of Parametric and Non-parametric Knowledge

The effective integration of parametric and non-parametric knowledge is crucial for complex tasks such as multi-document QA, where the evidence set may not contain all the necessary information. To this end, we evaluate how effectively FAVICOMP incorporates parametric knowledge from the target model and non-parametric knowledge from the compression model on the multi-document QA datasets. We begin by dividing the test samples of each dataset into evidence-relevant and evidence-irrelevant subsets, using the Hits metric. The Hits metric is set to 1 (evidence-relevant) if the retrieved evidence set contains the correct answer, and 0 (evidence-irrelevant) if it does not. We then assess the downstream performance of each subset. The underlying intuition is that if a method performs better on the evidence-relevant subset, it suggests that the method is more effectively utilizing the provided evidential knowledge. Conversely, if a method excels on the evidence-irrelevant subset, it indicates that the method is more effectively leveraging parametric knowledge without relying on potentially irrelevant evidence.

As shown in Figure 3, we compare the accuracy of FAVICOMP with `Llama3.2-3B-Instruct` and `Llama3-8B-Instruct` compression-target pairs on Hits $= 0$ and Hits $= 1$ subsets with the top-performing baselines, Zero-shot Summarization and CompAct[7]. FAVICOMP outperforms other baselines in the Hits $= 0$ subset while performing comparably with others in the Hits $= 1$ subset. This proves that FAVICOMP effectively relies on parametric knowledge rather than evidential knowledge when faced with irrelevant evidence, while maintaining similar effectiveness in utilizing

[7]We provide results of FAVICOMP on various alpha values in Appx. §B.3

evidential knowledge when relevant evidence is present.

In addition, we conduct another experiment to demonstrate FAVICOMP's superior ability to synergize two sources of knowledge. We compare it against a straightforward approach that concatenates parametric and non-parametric knowledge as context for downstream generation. Specifically, we concatenate the compressed evidence from the Zero-shot Summarization with the generated context from the Generated Context and use this concatenated context for evaluation. The results, shown in Table 2, reveal that simple concatenation underperforms compared to the Zero-shot Summarization baseline. This suggests that naively merging non-parametric and parametric knowledge in-context can be less effective than relying solely on non-parametric knowledge. In contrast, FAVICOMP effectively integrates both knowledge sources during compression, leveraging their synergy to achieve superior performance.

4.4 Compression Rate Comparisons

Since one of the functionalities of evidence compression in RAG is to reduce the number of tokens from the evidence set, we report the compression rate of FAVICOMP with `Llama3.2-3B-Instruct` and `Llama3-8B-Instruct` compression-target pairs in Table 3. We compute the compression rate as $\frac{\textit{\# of tokens in retrieved documents}}{\textit{\# of tokens in compressed documents}}$. Overall, RECOMP-abstractive and FAVICOMP consistently score the highest compression rates. RECOMP-abstractive exhibits high compression rates because the compression model is trained to output an empty string when no relevant evidence is found, which is often the case in multi-document QA datasets. FAVICOMP compresses the evidence to make it familiar to the target model by lowering its perplexity at each decoding step, typically resulting in a shorter context. Notably, when compared to Zero-shot Summarization, which is equivalent to FAVICOMP with $\alpha = 0$, FAVICOMP consistently achieves higher compression rates. This demonstrates that the ensemble decoding strategy, combining token logits from both evidence compression and context generation, leads to greater compression efficiency.

5 Case Study

Table 4 presents two examples from HQA to illustrate how FAVICOMP effectively familiarizes evidence while seamlessly integrating both parametric and non-parametric knowledge during evidence compression. We compare its output with Raw Document, which does not apply any compression, and Zero-shot Summarization.

Methods	NQ	TQA	HQA	Wiki	MQ
LongLLMlingua	1.87	1.84	1.83	1.83	1.83
RECOMP-abstractive	**17.96**	17.79	19.72	**32.06**	**32.05**
CompAct	8.85	8.92	9.45	10.71	8.96
Zero-shot Summarization	14.12	17.12	18.75	21.39	16.19
FAVICOMP	16.43	**22.40**	**22.55**	23.10	18.95

Table 3: Compression rates of the baselines and FAVICOMP.

In both examples, Raw Document fails to produce the correct answer, even though the evidence contains the necessary information, highlighting the need for effective evidence compression. In the first example, while the difference between the compressed evidence from Zero-shot Summarization and FAVICOMP appears subtle, FAVICOMP delivers the correct answer with a lower perplexity in compression, underscoring the significance of evidence familiarization. The second example highlights the importance of parametric knowledge when the retrieved evidence set lacks complete information. Since the evidence set does not mention "Skeptic", Zero-shot Summarization introduces irrelevant information ("Philanthropy magazine"), ultimately leading to an incorrect answer. In contrast, FAVICOMP integrates parametric knowledge about "Skeptic" and incorporates it into the evidence compression. Notably, FAVICOMP selects the $\arg\max$ token from the target model only when the token's probability is higher than that of the compression model, demonstrating that FAVICOMP draws on parametric knowledge only when necessary—potentially when the compression model is uncertain about the next token.

6 Related Works

Evidence Compression for RAG. Recent efforts on evidence compression seek to compress retrieved evidence pieces to filter out unnecessary information and retain only the essential context (Wang et al., 2023c; Li et al., 2024d; Ke et al., 2024; Xu et al., 2024; Yoon et al., 2024). Most recently, Xu et al. (2024) and Yoon et al. (2024) train a compression model to generate an abstractive summary of the documents by distilling knowledge from larger language models.

While these methods are successful to some ex-

Question: This film is an adaption of a Jacques Offenbach's opera that was written by a Hungarian British screenwriter?			
Methods	**(Compressed) Evidence**	**Prediction**	**Perplexity**
Raw Document	The Tales of Hoffmann is a 1951 British Technicolor film adaptation of Jacques Offenbach's opera "The Tales of Hoffmann", written, produced and directed by the team of Michael Powell and Emeric Pressburger ...	Emeric Pressburger ✗	12.429
Zero-shot Summarization	The 1951 film "The Tales of Hoffmann" is an adaptation of Jacques Offenbach's opera, written, produced, and directed by Michael Powell and Emeric Pressburger.	Emeric Pressburger ✗	2.298
FAVICOMP	The 1951 film "The Tales of Hoffmann" is an adaptation of Jacques Offenbach's opera, written by Emeric Pressburger, a Hungarian-British screenwriter, and directed by Michael Powell and Emeric Pressburger.	The Tales of Hoffmann ✓	1.959
Question: Which magazine was first published earlier, The Chronicle of Philanthropy or Skeptic?			
Raw Document	The Chronicle of Philanthropy is a magazine that covers the nonprofit world. ... It was founded in 1988 by editor Phil ... First published as a newsletter in 1987, "Philanthropy" became a glossy magazine in 1996.	Philanthropy ✗	4.856
Zero-shot Summarization	The Chronicle of Philanthropy was founded in 1988, while Philanthropy magazine was first published as a newsletter in 1987 and became a glossy magazine in 1996.	Philanthropy magazine ✗	3.196
FAVICOMP	The Chronicle of Philanthropy was first published in 1988, while Skeptic was first published in 1992.	The Chronicle of Philanthropy ✓	1.345

Table 4: Case study of evidence compression: FAVICOMP vs. Raw Document and Zero-shot Summarization. For FAVICOMP, the colors red and blue highlight tokens that are the $\arg\max$ of the compression model and the target model, respectively. Purple indicates a token that is the $\arg\max$ of neither model. Tokens with no coloring represent those that are the $\arg\max$ of both models.

tent, they often achieve suboptimal performance because of the discrepancy between the compression model and the target model, leading unfamiliarity of the context. In contrast, FAVICOMP proactively compresses the evidence pieces in a way to lower the target model's perplexity using an ensemble decoding technique without any training, thereby improving the downstream performance.

Parametric and Non-parametric Knowledge in RAG. There has been a lack of research focused on effectively combining both sources. A few of these efforts introduce counterfactual augmentation (Longpre et al., 2021; Fang et al., 2024; Zhang et al., 2024) and causal intervention (Zhou et al., 2023; Wang et al., 2023a) to mitigate knowledge conflict, which, however, requires explicitly knowing the features of the input that causes such conflict. Zhang et al. (2023) seek to address this issue by incorporating LM-generated context into the LM's input along with the retrieved documents, thereby integrating both sources of knowledge. However, merely concatenating both contexts is a suboptimal solution, as LMs may still show bias toward one source over the other when generating responses (Longpre et al., 2021; Wu et al., 2024). To address this, FAVICOMP employs ensemble decoding during the evidence compression, ensuring that both types of knowledge are seamlessly fused together to create a consistent context.

Constrained Decoding. Constrained decoding has been previously proposed in text generation tasks for various purposes, including optimizing prompts (Liu et al., 2024), enhancing plausibility (Li et al., 2023) or controllability (Meng et al., 2022; Huang et al., 2023), and reducing hallucination (Shi et al., 2024). Our work is closely connected with the method by Liu et al. (2024) which employs ensemble decoding to paraphrase prompts to enhance zero-shot LM prompting and generalization. Their approach focuses on the robustness and generalizability of instruction prompts for tasks without retrieval augmentation. In contrast, our approach compresses externally retrieved evidence while integrating parametric knowledge during compression, specifically targeting knowledge-intensive tasks that require balancing both evidential and parametric knowledge.

7 Conclusion

In this study, we introduce FAVICOMP, a training-free, inference-time evidence compression method designed to enhance RAG performance by consolidating retrieved evidence set to be more familiar to the target model, while seamlessly integrating parametric knowledge. Our extensive experiments validate the effectiveness of FAVICOMP on open-domain QA tasks, showing significant improvements over recent evidence compression baselines

in multiple datasets. Additionally, FAVICOMP's model-agnostic nature allows it to be incorporated into various RAG workflows at inference time, making it a versatile tool for enhancing LMs in complex tasks.

Acknowledgment

We appreciate the reviewers for their insightful comments and suggestions. This work was partly supported by the Amazon Nova Trusted AI Prize, the NSF of the United States Grants ITE 2333736 and OAC 2531126, and the DARPA FoundSci Grant HR00112490370.

Limitations

Although FAVICOMP exhibits superior performance in RAG compared to the recent evidence compression baselines, it has some limitations. (1) FAVICOMP consumes approximately twice as much computation compared to methods that only use a compression model since it needs two inferences (compression and target model) during the ensemble decoding. However, it is a training-free strategy that can be easily plugged into any RAG application. We provide insights on the tradeoff between latency and performance in Appx. §B.4. (2) Ensemble decoding requires the compression and target model to share the same vocabulary and tokenizer, which can limit the range of compatible models. Nonetheless, recent studies, such as Gu et al. (2024), have introduced techniques to enable model-agnostic ensemble decoding. This implies that there will be a potential direction of incorporating model-agnostic ensemble decoding with our framework to enable more flexible integration of various models, which we leave as future work.

Ethics Statement

This work follows the ACL Code of Ethics. We believe no potential risk is directly associated with the presented work.

References

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*.

Tianqing Fang, Zhaowei Wang, Wenxuan Zhou, Hongming Zhang, Yangqiu Song, and Muhao Chen. 2024. Getting sick after seeing a doctor? diagnosing and mitigating knowledge conflicts in event temporal reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3846–3868.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148.

Kevin Gu, Eva Tuecke, Dmitriy Katz, Raya Horesh, David Alvarez-Melis, and Mikhail Yurochkin. 2024. Chared: Character-wise ensemble decoding for large language models. *arXiv preprint arXiv:2407.11009*.

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.

Tenghao Huang, Ehsan Qasemi, Bangzheng Li, He Wang, Faeze Brahman, Muhao Chen, and Snigdha Chaturvedi. 2023. Affective and dynamic beam search for story generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11792–11806.

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 Edouard 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. Association for Computational Linguistics.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. 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. 2023b. 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. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

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

Zixuan Ke, Weize Kong, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Bridging the preference gap between retrievers and llms. *arXiv preprint arXiv:2401.06954*.

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*, 7:452–466.

Yoonsang Lee, Pranav Atreya, Xi Ye, and Eunsol Choi. 2024. Crafting in-context examples according to lms' parametric knowledge. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2069–2085.

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.

Bangzheng Li, Ben Zhou, Xingyu Fu, Fei Wang, Dan Roth, and Muhao Chen. 2024a. Famicom: Further demystifying prompts for language models with task-agnostic performance estimation. *arXiv preprint arXiv:2406.11243*.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2024b. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7668–7681.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2024c. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312.

Zhonghao Li, Xuming Hu, Aiwei Liu, Kening Zheng, Sirui Huang, and Hui Xiong. 2024d. Refiner: Restructure retrieval content efficiently to advance question-answering capabilities. *arXiv preprint arXiv:2406.11357*.

Qin Liu, Fei Wang, Nan Xu, Tianyi Yan, Tao Meng, and Muhao Chen. 2024. Monotonic paraphrasing improves generalization of language model prompting. *arXiv preprint arXiv:2403.16038*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. 2023. Multi-hop evidence retrieval for cross-document relation extraction. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

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

Tao Meng, Sidi Lu, Nanyun Peng, and Kai-Wei Chang. 2022. Controllable text generation with neurally-decomposed oracle. *Advances in Neural Information Processing Systems*, 35:28125–28139.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

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.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. 2023a. A causal view of entity bias in (large) language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15173–15184.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19206–19214.

Zezhong Wang, Luyao Ye, Hongru Wang, Wai Chung Kwan, David Ho, and Kam-Fai Wong. 2023b. Readprompt: A readable prompting method for reliable knowledge probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7468–7479.

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

Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Preprint*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *The Twelfth 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.

Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. 2024. Compact: Compressing retrieved documents actively for question answering. *arXiv preprint arXiv:2407.09014*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, S Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *International Conference on Learning Representations*.

Hao Zhang, Yuyang Zhang, Xiaoguang Li, Wenxuan Shi, Haonan Xu, Huanshuo Liu, Yasheng Wang, Lifeng Shang, Qun Liu, Yong Liu, et al. 2024. Evaluating the external and parametric knowledge fusion of large language models. *arXiv preprint arXiv:2405.19010*.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Merging generated and retrieved knowledge for open-domain qa. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4710–4728.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2308–2313.

A Implementation Details

A.1 Generation Configuration

For all the baselines and FAVICOMP, we use default temperature and top-p values of the compression model during evidence compression and fix the temperature of the target model to 1.0 during evaluation.

A.2 Dataset Statistics

We provide the statistics of the evaluation dataset utilized in our experiments in Table 6.

A.3 Implementation Details of Baselines

(1) **Gold Compression**: We implement the Gold Compression baseline following the approach outlined by (Yoon et al., 2024). We evaluate only on HQA, Wiki, and MQ, as these datasets contain gold documents. We first identify the presence of any gold documents in the retrieved documents. If found, we use the documents as the context. If none of the retrieved documents are identified as gold, we utilize the entire set of retrieved documents as the context for the evaluation. To identify the gold documents within the retrieved documents, we compare each gold document with the retrieved ones. If 50% or more of the content matches, we classify it as a gold document. This approach is necessary because the documents are chunked, and

Methods	Train	Compression Model	NQ		TQA		HQA	
			Acc	F1	Acc	F1	Acc	F1
RECOMP-abstractive	O	T5-large	38.0	37.8	62.1	65.0	27.4	34.3
RECOMP-abstractive	O	Mistral-7B-Instruct-v0.3	38.3	38.2	63.0	65.4	29.5	36.6
FaviComp	X	Mistral-7B-Instruct-v0.3	40.3	40.4	65.9	68.9	32.0	40.5

Table 5: Head-to-head comparison results with RECOMP

Dataset	NQ	TQA	HQA	Wiki	MQ
# of Samples	3610	11313	7405	12576	4834

Table 6: Number of samples in each dataset.

the retrieved documents may not exactly match the gold documents.

(2) **Generated Context**: We use the context generation prompt in Table 10 to generate the context.

(3) **Zero-shot Summarization**: We use the evidence compression prompt in Table 10 to compress the retrieved documents.

(4) **RECOMP-extractive**: We utilize the same Contriever models trained by the authors for each dataset, to encode both the question and the sentences in the evidence set. For Wiki and MQ, since there are no fine-tuned models available, we use the Contriever fine-tuned on HQA. Following the original paper, we select one sentence as the context for NQ and TQA, whereas for the other datasets, we utilize two sentences.

(5) **RECOMP-abstractive**: Similar to RECOMP-extractive, we use the same T5-large models trained by the authors for each dataset to compress the retrieved evidence. For the Wiki and MQ, we employ the T5-large model fine-tuned on HQA.

(6) **LongLLMLingua**: We use `Llama2-7B`[8] trained by the authors as the prompt compressor model. We use the default hyperparameters in the original paper, where the dynamic context compression rate is set to 0.3, and the maximum compression rate is set to 0.5.

(7) **CompAct**: We use the same `Mistral-7B-Instruct`[9] model instruction-tuned by the authors for evidence compression. The number of documents per segment is set to 5 with 1 iteration.

[8]https://huggingface.co/NousResearch/Llama-2-7b-hf

[9]https://huggingface.co/cwyoon99/CompAct-7b

B Additional Experiment Results

B.1 Other Compression and Target Models

We conduct an experiment where we use `Llama3-8B-Instruct` and `Mistral-7B-Instruct` for both compression and target models. The result in Table 8 demonstrates that FAVICOMP outperforms all other baselines, supplementing the effectiveness shown in §4.1.

B.2 Head-to-Head Comparison with RECOMP-abstractive

Since the lower performance of RECOMP-abstractive might possibly be due to the use of smaller base model for compression (`T5-large`), we conduct a head-to-head experiment on FAVICOMP and RECOMP-abstractive by using the same base compression model. We construct training data on NQ, TQA, and HQA according to Xu et al. (2024) and finetune `Mistral-7B-Instruct` on each of the training data. We train for 7 epochs using LoRA with Adam optimizer with a learning rate of 2e-6 and a batch size of 64. We present the evaluation results in Table 5. Even though using larger base model for compression enhances the performance of RECOMP-abstractive to some extent, it still underperforms compared to training-free FAVICOMP. This underscores that the familiarization during evidence compression and integration of parametric and non-parametric knowledge are more helpful to the downstream generation than relying on a trained model for evidence compression.

B.3 Performance of $\text{Hits} = 0$ and $\text{Hits} = 1$ on Varying Alpha Values

We evaluate FAVICOMP's performance on evidence-relevant ($\text{Hits} = 1$) and evidence-irrelevant ($\text{Hits} = 0$) subsets by varying α values. Figure 4 shows that $\alpha = 0.5$ or $\alpha = 0.7$ performs the best on the $\text{Hits} = 0$ subset, while performance declines as α deviates further from the value. This pattern in the $\text{Hits} = 0$ subset mirrors the overall performance trend, suggesting that appropriately

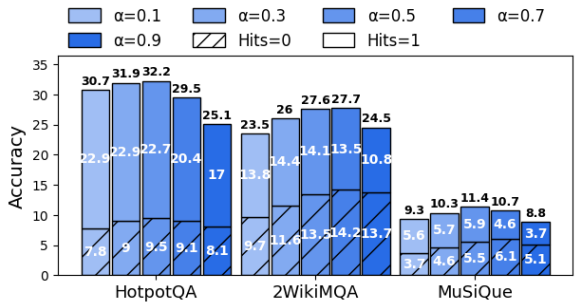

Figure 4: Accuracy of FAVICOMP with various α values on Hits $= 0$ and Hits $= 1$ subset of multi-document QA datasets.

utilizing parametric knowledge when the evidence is irrelevant is crucial to the overall performance. In the Hits $= 1$ subset, performance remains consistent for α values up to 0.5 but decreases significantly when α exceeds 0.5 due to the diminished utilization of the relevant evidential context.

B.4 Latency Ablation Study

Table 7 shows the latency of our method along with other major baselines to provide insights on the trade-offs between accuracy and latency. We used Llama-3-8B-Instruct as the target model and tested on NQ dataset for the experiment. Although there are trade-offs between latency and accuracy across all methods, training-free FaviComp demonstrates lower latency while achieving higher accuracy than CompAct, which is the supervised baseline that previously achieved SOTA performance.

C Prompt Templates

Evaluation Prompt Template

```
{System Prompt}
{Demonstrations}
Question: {Question}
Context: {Context}
Answer:
```

Figure 5: Evaluation Prompt Template.

C.1 Evaluation

The evaluation prompt template is shown in Figure 5. For all the evaluations throughout the experiment, we switch the positions of the Question and Context if doing so results in better performance. System prompts and demonstrations used in the evaluations are presented in Table 9 and Table 11, respectively.

C.2 FAVICOMP

The prompt templates for evidence compression and context generation of FAVICOMP are presented in Table 10.

D Licenses

We include the licenses of datasets and models we used in this work.
Dataset Licenses:

- NQ: Apache-2.0
- TQA: Apache-2.0
- HQA: CC BY-SA 4.0
- Wiki: Apache-2.0
- MQ: CC-BY-4.0

Model Licenses:

- Llama3: Custom License https://www.llama.com/llama3/license/
- Mistral & Mixtral: Apache-2.0

Methods	Compression Model	Avg latency per sample (s)	Performance	
			Acc	F1
RECOMP-abstractive	T5-large	0.22	39.3	43.3
CompAct	Mistral-7B-Instruct-v0.2	8.72	42.3	46.1
Zero-shot Summarization	Llama-3.2-3B-Instruct	3.99	39.4	43.2
FAVICOMP	Llama-3.2-3B-Instruct	6.43	42.8	46.8

Table 7: Latency and of the baselines and FAVICOMP

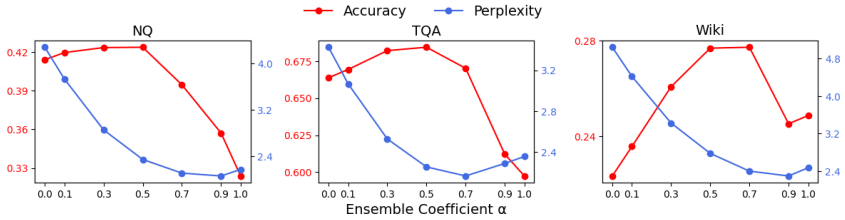

Figure 6: Impact of coefficient α on performance and perplexity for NQ, TQA and Wiki.

Methods	NQ		TQA		HQA		Wiki		MQ	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Llama3-8B-Instruct										
Gold Compression	-	-	-	-	42.3	51.3	35.7	40.0	10.2	17.7
No Context	26.9	31.9	57.2	61.2	19.1	25.5	20.5	25.0	5.4	13.0
Raw Document	42.6	**47.1**	67.6	70.8	30.3	38.7	22.0	26.8	8.2	15.0
Generated Context	32.3	36.6	59.7	62.4	22.7	29.7	24.8	28.7	7.6	14.8
Sentence-BERT	30.3	35.4	59.2	62.9	22.4	29.6	18.1	22.9	7.7	14.8
RECOMP-extractive†	33.7	38.1	59.4	62.8	22.5	29.8	18.0	22.4	8.1	15.5
LongLLMLingua†	35.4	40.9	64.8	67.6	25.9	34.7	19.2	24.2	7.7	14.4
RECOMP-abstractive†	39.3	43.3	62.9	66.1	27.0	34.8	20.5	25.0	7.3	14.8
CompAct†	42.3	46.1	67.0	69.7	29.8	37.5	21.4	26.6	9.2	16.9
Zero-shot Summarization	41.3	45.1	66.3	69.5	30.2	38.6	22.3	28.1	8.3	16.3
FAVICOMP	42.3	46.6	**68.4**	**71.5**	32.3	41.0	27.6	33.6	**11.4**	**20.1**
Mistral-7B-Instruct										
Gold Document	-	-	-	-	41.0	50.5	38.1	40.3	9.6	15.2
No Context	28.1	27.5	58.8	60.9	19.7	24.8	21.9	22.8	5.2	9.7
Raw Document	40.2	39.3	**66.2**	68.6	30.3	37.2	26.6	28.5	7.5	13.1
Generated Context	30.1	31.7	57.3	60.7	23.7	30.6	25.1	29.5	7.1	12.8
Sentence-BERT	29.8	30.1	57.8	60.7	23.8	30.3	22.9	24.7	7.5	12.3
RECOMP-extractive†	31.7	32.2	57.2	60.0	24.1	30.2	23.2	24.4	7.4	12.5
LongLLMLingua†	34.3	36.4	63.8	66.9	27.0	34.7	25.5	28.0	7.1	13.0
RECOMP-abstractive†	38.0	37.8	62.1	65.0	27.4	34.3	25.1	27.4	6.4	12.0
CompAct†	38.8	38.9	65.1	67.1	30.2	37.1	24.9	27.6	8.2	13.6
Zero-shot Summarization	38.4	38.2	62.3	64.8	28.2	35.2	23.2	27.1	6.8	11.8
FAVICOMP	**40.3**	**40.4**	65.9	**68.9**	**32.0**	**40.5**	**29.7**	**35.1**	**9.2**	**15.2**

Table 8: Additional experimental results. `Llama3-8B-Instruct` and `Mistral-7B-Instruct` are used for both compression and target models.

Target Models	System Prompt
Llama-3-8B-Instruct	You are an expert in Question Answering. Your job is to answer questions in 1 to 5 words based on the given context.
Mixtral-8x7B-Instruct	You are an expert in Question Answering. Your job is to answer questions in 1 to 5 words based on the given context. Just output the answer as concisely as possible, no other words
Mistral-7B-Instruct	You are an expert in Question Answering. Your job is to answer questions in 1 to 5 words based on the given context. Just output the answer as concisely as possible, no other words

Table 9: System prompts used in evaluation

Instruction	Prompt Template
Evidence Compression	You are an expert in summarization. Given a question and multiple document snippets, generate one summarized context that is helpful to answer the question. Just summarize, no other words. Question: `{Question}` Documents: `{Evidence}` Summarized Context:
Context Generation	You are an expert in context generation. Given a question, generate a context that is helpful to answer the question. Just generate the context, no other words. Question: `{Question}` Context:

Table 10: Prompt Templates for FAVICOMP

Dataset	Demonstrations
NQ	Question: who sings i've got to be me Answer: Sammy Davis, Jr Question: who wrote i will follow you into the dark Answer: Ben Gibbard Question: who won season 2 of total drama island Answer: Owen (Scott McCord) Question: what part of the mammary gland produces milk Answer: cuboidal cells Question: when did the golden compass book come out Answer: 1995
TQA	Question: Who sang the theme for the James Bond film 'Thunderball'? Answer: Tom Jones Question: A hendecagon has how many sides? Answer: Eleven Question: In the 1968 feature film Chitty Chitty Bang Bang, of what country is Baron Bomburst the tyrant ruler? Answer: Vulgaria Question: Artists Chuck Close, Henri-Edmond Cross, John Roy, Georges-Pierre Seurat, Paul Signac, Maximilien Luce and Vincent van Gogh painted in what style? Answer: Pointillism Question: What is the study of the relation between the motion of a body and the forces acting on it? Answer: Dynamics
HQA	Question: Which magazine was started first Arthur's Magazine or First for Women? Answer: Arthur's Magazine Question: The Oberoi family is part of a hotel company that has a head office in what city? Answer: Delhi Question: Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who? Answer: President Richard Nixon Question: Are Jane and First for Women both women's magazines? Answer: Yes Question: Were Pavel Urysohn and Leonid Levin known for the same type of work? Answer: No
Wiki	Question: Where was the place of death of Marie Thérèse Of France (1667–1672)'s father? Answer: Palace of Versailles Question: Who is the paternal grandmother of Przemysław Potocki? Answer: Ludwika Lubomirska Question: Who lived longer, Herbert Findeisen or Léonie Humbert-Vignot? Answer: Léonie Humbert-Vignot Question: Are Alison Skipper and Diane Gilliam Fisher from the same country? Answer: Yes Question: Are director of film Move (1970 Film) and director of film Méditerranée (1963 Film) from the same country? Answer: No
MQ	Question: Who is the child of the director and star of Awwal Number? Answer: Suneil Anand Question: What county shares a border with the county where Black Hawk Township is located? Answer: Dodge County Question: Who is the sibling of the person credited with the reinvention and popularization of oil paints? Answer: Hubert Van Eyck Question: Who heads the Catholic Church, in the country that a harp is associated with, as a lion is associated with the country that Queen Margaret and her son traveled to? Answer: Eamon Martin

Table 11: Demonstrations used in evaluation for each dataset