K-COMP: Retrieval-Augmented Question Answering With Knowledge-Injected Compressor

Anonymous ACL submission

Abstract

Retrieval-augmented question answering (QA) integrates external information, and thereby increases the QA accuracy of reader models that lack domain knowledge. However, documents retrieved for closed domains require high expertise, so the reader model may have difficulty fully comprehending the text. Moreover, the retrieved documents contain thousands of tokens, some unrelated to the question. As a result, the documents include some inaccurate information, which may lead the reader model to mistrust the passages and could result in hallucinations. To solve these problems, we propose **K-COMP** (**K**nowledge-injected **COMP**ressor) which provides the knowledge required to answer the question correctly. The compressor automatically generates the prior knowledge needed to answer before compressing the retrieved passages, and then compresses passages autoregressively, injecting the knowledge into the compression process. This process ensures alignment between the question intent and the compressed context. By augmenting this prior knowledge and concise context, the reader models are guided toward relevant answers and trust the context.

1 Introduction

Retrieval-augmented question answering (QA) is a task where passages related to a question are appended into the prompt, such that a reader model can reference them and infer correct answer (Ahmad et al., 2019; Guo et al., 2021). Towards this, many studies like Jiang et al. (2023b); Yu et al. (2023a); Lin et al. (2024); Shi et al. (2024b) utilize retrieval augmentation techniques to significantly reduce the occurrence of hallucinations and enhance overall answer reliability without necessitating additional parameter updates for the reader model. This approach significantly increases QA accuracy in both open and closed domains (Wang et al., 2024b; Louis et al., 2024; Frisoni et al., 2024).

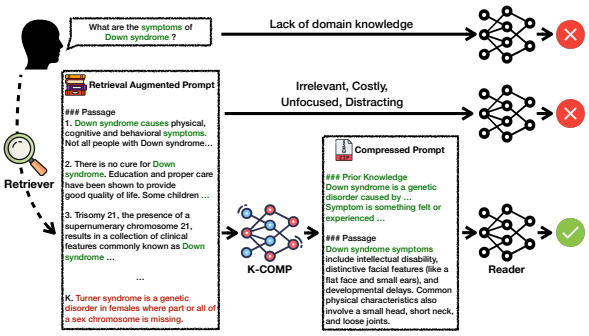

Figure 1: K-COMP helps the reader model infer accurate responses by using domain knowledge and compressed context aligned with the question.

However, several limitations impede use of retrieval-augmented approaches in closed domains with large language models (LLMs). First, the documents retrieved for closed domains require domain expertise, so the reader may not trust the whole text. When faced with unfamiliar input, the model exhibits an availability bias towards commonly known knowledge, making it more willing to believe in information they can easily recall (Jin et al., 2024). Also, retrieved passages contain thousands of tokens and are sometimes unrelated to the question, so they include inaccurate information, which can cause the language model to distrust the passages and perceive them as irrelevant noise, and generate answers that do not consider them. These problems lead to hallucinations (Ji et al., 2023a), which cause the model to generate inaccurate answers or infer plausible but false responses. Lastly, LLMs are sensitive to the order of retrieved documents and the prompting method. Specifically, LLMs can have difficulty finding the necessary information within lengthy input prompts, especially when key information or correct answer clues are located in the middle of the prompt (Liu et al., 2024; Xu et al., 2024b).

To tackle these issues, we propose K-COMP

(Knowledge-injected COMPressor). We aim to use an autoregressive LLM as a compressor with the domain knowledge needed to answer the question, and increase the alignment of the retrieved passages with the question intent. Additionally, when the compressor is trained domain-related terms and knowledge, it becomes able to recognize the entities that occur in the question, and provide descriptions for them. This process is significant for closed domains that require substantial prior knowledge. For retrieval augmentation, we use a large amount of text from domain-specific sources including Wikipedia. We exploit the advantages of domain relevance by efficiently reusing it when annotating prior knowledge, not just for retrieval. Furthermore, we use a causal masking objective (Aghajanyan et al., 2022) during the training phase to inject domain knowledge into the compressor.

More specifically, we focus on medical domain. Our proposed process for generating knowledge-infused summaries learns the correlation between medical entities and the summary, allowing the summary to accurately incorporate the intent of the question. We evaluate the relevance of the summary and demonstrate that our approach increases answer accuracy by using prompts that include prior knowledge.

In summary, our contributions are as follows:

- We propose a novel approach to generate knowledge-injected summaries adapted for the medical domain. We incorporate causal masking to inject knowledge into the compressor without modifying its structure. This approach aligns the summary with the question.
- Even without domain knowledge in the reader model, K-COMP generates prior knowledge to answer the question, thereby enabling LLMs with diverse backgrounds to handle medical domain questions more accurately.
- We efficiently annotate entity-knowledge pairs by using title-text pairs from a retrieval corpus, and thereby avoid the need for additional data. Furthermore, after K-COMP is fine-tuned, it autonomously generates entity-knowledge pairs without referencing the complex corpus.

2 Related Work

Text Infilling. Models such as BERT (Devlin et al., 2019), SpanBERT (Joshi et al., 2020), T5 (Raffel et al., 2020), and BART (Lewis et al., 2020), are pre-trained using masked language modeling within a bidirectional encoder architecture. They have shown strong performance in infilling short and contiguous masked token spans. However, the bidirectional attention mechanism typically restricts the fillable span length to dimensions significantly shorter than a sentence.

In contrast, decoder-only models such as GLM (Du et al., 2022), CM3 (Aghajanyan et al., 2022), and InCoder (Fried et al., 2023) operate by left-to-right generation. They can accommodate variable infill span lengths. Causal masking (Aghajanyan et al., 2022) or fill-in-the-middle (Bavarian et al., 2022) methods predict masked spans from the posterior context. These methods have their generative capabilities, which increase the length of infill spans. They can also exploit the advantages of considering contextual relationships that surround the masked span. Not only is the proposed K-COMP able to fill the span by considering bidirectional context, but also is able to align the generated summary with the question by regressively encoding the infilled span.

Retrieval-Augmented Generation (RAG). Efforts to mitigate hallucination by augmenting snippets with relevant information retrieved from external knowledge repositories have proven effective in enhancing the performance of natural language processing tasks (Izacard and Grave, 2021; Yu et al., 2023c; Luo et al., 2023; Shi et al., 2024a; Anantha and Vodianik, 2024; Xu et al., 2024c). RAG uses reader LLMs that have been trained for general purposes, then provides the LLMs with external information for closed domain tasks. This method enables the LLM to adapt to various domains without requiring architectural changes or fine-tuning (Khandelwal et al., 2020; Jiang et al., 2023b). However, the reader model can still give inaccurate answers if it becomes overly dependent on retrieved documents that contain noise. To solve this problem, a document-validation step is essential (Nan et al., 2021; Yu et al., 2023a). The overall quality and reliability of the generated content are increased by considering the suitability of documents (Asai et al., 2024a) or adding a step to verify further the factual accuracy and relevance of the documents (Yu et al., 2023b).

Prompt Compression. Several studies have demonstrated that prompt augmentations effectively enhance the performance of LLMs across various tasks (Liu et al., 2023a; Ram et al., 2023; Ryu et al., 2023; Wang et al., 2024c; Long et al., 2023; Yagnik et al., 2024). Yet, the relevance and reliability of the augmented passages are significant challenges in prompt augmentations. To address this, recent studies have attempted to directly extract contents from ambiguous and lengthy passages. Kim et al. (2024) eliminates irrelevant information while maximizing the extraction of accurate information, whereas Yang et al. (2023) leverages the black-box LLMs by applying a reward-based method during compressor training to generate summaries. RECOMP (Xu et al., 2024a) selects and augments the summary with the highest end-task performance by using prompts in which non-essential summaries are set to empty strings if necessary. LLMLingua (Jiang et al., 2023a) dynamically assigns different compression rates to various components within the prompt, and thereby maintains the original meaning while achieving maximum compression. In contrast, K-COMP focuses on the keywords needed to answer the question, emphasizing the alignment between the compressed context and the question.

3 Causal Knowledge Injection

Causal models that have been trained using autoregressive language modeling depend exclusively on the context to the left of the generated tokens to predict subsequent tokens (Brown et al., 2020). This attribute confers an ***advantage in causally generating entire documents***, such as text generation. However, these models show limited proficiency in tasks that require understanding of post-positional relationships for span infilling. Conversely, masked language models excel at predicting masked spans by referencing attention scores from tokens located ***both anteriorly and posteriorly***. Nonetheless, Their training objective is limited to decoding only short segments of the passages (Devlin et al., 2019; Joshi et al., 2020).

We adopt a causal masking (Aghajanyan et al., 2022) to combine the advantages of both objectives. We focus on the masked medical entities within the ***question (prior context)*** and aim to predict them by considering the ***retrieved snippets (subsequent context)***. Afterward, by ***auto-regressively compressing the retrieved snippets***, we can leverage both advantages.

4 Methods

In this section, we report our proposed approach for knowledge-injected compression and retrieval augmentation. To retrieve passages similar to a question, we construct a retrieval pipeline composed of a large corpus (§4.1). Next, we explain the data processing steps for training, with details about identification of entities within the question and matching descriptions from the retrieval corpus with the knowledge (§4.2). Finally, we detail the training scheme for K-COMP with the proposed objective and explain the inference phase for retrieval augmentation (§4.3). Figure 1 shows an overview of the prompts that K-comp consists of.

4.1 Retrieval Framework

Corpora. Closed domain tasks have not been as thoroughly explored as open domain tasks, which have achieved notable performance enhancements using Wikipedia as a retrieval corpus (Karpukhin et al., 2020). In contrast to open domains, the challenge in closed domains is that unified corpora have not been established. Research endeavors, such as Xiong et al. (2024); Wang et al. (2024b), are currently underway to address this gap. To ensure coverage of both general and domain knowledge, we adopt the MedCorp corpus (Xiong et al., 2024) as our retrieval corpus. It combines Wikipedia, PubMed[1], StatPearls[2], and textbooks (Jin et al., 2021).

Retriever. We employ a lexical-based sparse retriever to emphasize the entities present in the question. Simultaneously, to mitigate bottlenecks and efficiently execute similarity searches on our large-scale corpus comprising four distinct text corpora, we use an embedding-based k-NN search (Johnson et al., 2019). To build an integrated retrieval system, we employ BM25 (Robertson et al., 2009) and Contriever (Izacard et al., 2022). Specifically, we encode a large-scale corpus $\mathbb{C}$ in a data-parallel manner by using dense retrieval, then store each embedding offline in advance. Given a question q, Contriever retrieves multiple relevant passages $\mathbb{P} \subseteq \mathbb{C}$ based on vector similarity, then re-ranks them lexically by using BM25 to select only the top-k passages $\mathcal{P} = \{p_i | p_i \in \mathbb{P}\}$, $|\mathcal{P}| = k$. This approach semantically selects a bundle of passages

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://www.statpearls.com/

	MedQuAD			MASH-QA			BioASQ		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
Original	13,127	1,640	1,640	27,728	3,587	3,493	3,209	803	707
After filtering	9,077	1,098	1,562	20,546	2,665	3,264	2,288	566	651
% Filtered	30.9	33	4.8	25.9	25.7	6.6	28.7	29.5	7.9

Table 1: Dataset sizes before and after filtering in the entity recognition step. For test data, filtering is applied exclusively to questions lacking any entities. For other datasets, filtering is additionally conducted for the absence of corresponding descriptions for the recognized entities.

from an extensive range of text chunks and refines the final retrieved passages to be word-centric and relevant to the question.

4.2 Data processing

Entity Recognition. We rely on off-the-shelf tools to perform named-entity recognition[3], which identifies biomedical entities $\mathcal{E} = \{e_i\}$ in each question for causal masking. $\mathbb{C}$ consists of title and text pairs, with the first sentence of each text assumed to be a short description of the title (Xu et al., 2023). We then match these pairs of titles and short descriptions with corresponding entities and their corresponding knowledge d_i. Given our assumption that each question contains at least one medical entity, all entities discerned within the question can be aligned with corresponding titles and short descriptions available in the retrieval corpus. If a question does not have an entity, its data are excluded. Any instances that does not have a corresponding titles in the retrieval corpus is also filtered for the training dataset (Table 1).

In the test data, even if corresponding titles for the entities are absent in the retrieval corpus, K-COMP unveils a novel contribution by automatically generating domain-specific entity descriptions during inference. Thus, K-COMP provides these descriptions without needing costly and unnecessary tasks, such as searching for medical terms or finding definitions within the corpus.

Ground-Truth Summary. To synthesize gold summaries, GPT-3.5[4] compresses the passages by considering $\{\mathcal{P}, \mathcal{E}\}$ input pairs, and the number of passages used for summary synthesis is set to 5, i.e., $|\mathcal{P}| = 5$. Notably, we explicitly prohibit the inclusion of the question in the summary synthesis process. This is because incorporating the question into the input prompt for generating the summary risks shifting the focus from crafting keyword-focused summaries to formulating a summary that is aimed at answering the question. Detailed instructions for the summary synthesis are provided in Table 6 of the Appendix A.

4.3 K-COMP

Training. $q = [q^1, q^2, ..., q^N]$, where q^N represents the N-th token in the q. We use the special token <ent> to mask each medical entity spans within q, $q_m = [q^1, ..., \texttt{<ent>}, ..., q^{N-l}]$. Also, a special <eom> token is appended at the end of the description of the corresponding entity, $d_i = [d_i^1, ..., d_i^M, \texttt{<eom>}]$. An example is provided as follow:

q_m = What are the <ent> of <ent>?

d_1 = symptom: {description}<eom>

d_2 = Down syndrome: {description}<eom>

We define the dataset for the compressor as $\{\mathcal{P}, \mathcal{E}, \mathcal{D}, s, q_m\}$, where $\mathcal{D} = \{d_i\}$ and s is a gold summary. By encoding q_m and $\mathcal{P}$, we fill the <ent> tokens and generate short descriptions for the masked words:

$$P_\theta(\mathcal{E}, \mathcal{D} | \mathcal{P}, q_m) = \prod_i \left(\prod_{\alpha,\beta} P_\theta(e_i^\alpha, d_i^\beta | e_i^{<\alpha}, d_i^{<\beta}, \mathcal{P}, q_m) \right)$$

where θ represents the parameters of K-COMP.

This approach facilitates the incorporation of descriptions into the prompt for the reader model and ensures that the generated entities and their descriptions are regressively encoded. Consequently, a summary is generated causally, focusing on the entities present in the question and their related content, thereby composing a summary centered on these domain entities.

$$P_\theta(s | \mathcal{E}, \mathcal{D}, \mathcal{P}, q_m) = \prod_\gamma P_\theta(s^\gamma | s^{<\gamma}, \mathcal{E}, \mathcal{D}, \mathcal{P}, q_m)$$

[3]We use ScispaCy (Neumann et al., 2019) package.

[4]We use gpt-3.5-turbo-0125 (https://openai.com/index/chatgpt/).

	General-purpose LLMs								Medical-purpose LLMs			
	Llama-2-13B		Llama-2-70B		Mixtral-8x7B		GPT-3.5-turbo[4]		MedAlpaca-13B		Meditron-70B	
	BertScore	UniEval	BertScore	UniEval	BertScore	UniEval	BertScore	UniEval	BertScore	UniEval	BertScore	UniEval
						MedQuAD						
Without compressor												
Top-1 passage	66.62	61.39	78.57	54.51	62.11	44.88	85.14	58.48	73.86	36.06	77.03	53.27
Top-5 passages	72.91	65.85	78.46	56.31	63.35	46.17	84.39	65.14	14.91	8.17	73.65	51.09
With compressor												
RECOMP	69.82	65.03	79.11	55.91	58.94	45.15	85.78	58.85	77.95	34.99	74.98	53.42
LLMLingua	61.22	51.42	61.11	45.24	61.43	45.92	85.46	57.5	77.91	42.78	74.45	53.44
FT	71.83	68.13	82.29	61.23	71.7	58.58	85.91	63.02	79.79	41.1	76.34	56.01
K-COMP	**74.08**	**69.21**	**85.21**	**64.15**	**78.6**	**60.97**	**86.12**	**65.65**	**83.8**	**45.58**	**78.27**	**58.18**
						MASH-QA						
Without compressor												
Top-1 passage	59.68	44.49	81.58	58.37	73	50.81	84.4	59.16	77.42	44.92	81.68	55.54
Top-5 passages	63.81	46.77	79.53	58.23	73.79	52.93	84.87	64.11	29.34	17.1	79.85	56.97
With compressor												
RECOMP	58.21	44.13	81.92	59.17	73.89	51.31	**85.21**	61.29	78.52	36.89	81.17	55.68
LLMLingua	53.31	40.54	68.7	51.54	77.17	58.76	84.83	58.96	80.77	48	81.07	58.1
FT	62.25	48.4	83.17	62.57	79.52	59.93	84.91	63.39	80.23	44.38	82.12	59.1
K-COMP	**71.48**	**55.89**	**84.07**	**68.97**	**82.71**	**63.48**	85.2	**64.99**	**82.93**	**51.12**	**84.07**	**61.07**
						BioASQ						
Without compressor												
Top-1 passage	68.27	57.38	84.89	61.9	83.72	59.85	88.08	53.62	75.15	38.28	85.74	58
Top-5 passages	71.34	60.61	83.9	64.51	83.84	64.3	88.56	**62.61**	19.2	11.15	83.6	60.63
With compressor												
RECOMP	63.92	47.23	85.33	63.45	82.81	60.72	**88.82**	57.71	79.11	33.03	85.6	58.24
LLMLingua	65.08	50.09	79.46	58.71	81.87	60.37	88.36	55.33	82.03	42.66	82.62	59.14
FT	67.32	58.6	86.89	62.43	86.79	61.88	88.46	58.01	81.56	38.13	85.47	59.03
K-COMP	**72.43**	**66.16**	**87.28**	**65.05**	**86.93**	**64.61**	88.73	59.44	**84.62**	**44.96**	**86.56**	**61.4**

Table 2: Main results. We report automatic evaluation for retrieval-augmented QA with and without compressors.

We train the compressor using the standard next token objective $J(\theta)$:

$$P_\theta(\mathcal{E}, \mathcal{D}, s|\mathcal{P}, q_m) = P_\theta(\mathcal{E}, \mathcal{D}|\mathcal{P}, q_m) \times P_\theta(s|\mathcal{E}, \mathcal{D}, \mathcal{P}, q_m)$$

$$\therefore\ J(\theta) = \max_\theta \mathbb{E}(\log P_\theta(\mathcal{E}, \mathcal{D}, s \mid \mathcal{P}, q_m))$$

Inference. At inference time, documents are retrieved in advance to construct the compressor input batch $\{q, \mathcal{P}\}$. Unlike the training phase, which relied on the NER library[3] to pre-identify masking spans, K-COMP can generate knowledge based on the encoded passages even in the absence of masked spans in the question. This enables the sequential autoregressive generation of the entities and descriptions from the question until the <eom> token is produced. Considering the overall context, including entities and descriptions, a summary that aligns more closely with the question is then generated. This process ultimately constructs the input prompt for the reader model, ensuring a reliable response to the question. The prompt for the reader model can be found in Table 7 of the Appendix A.

5 Experiments

In this section, we evaluate K-COMP trained by causal knowledge injection and the retrieval-augmented QA task. We report the datasets and settings used in the experiments (§5.1) and discuss the main results (§5.2) and analyze the results from various perspectives (§5.3).

5.1 Settings

Datasets. To reduce potential biases from fine-tuned medical LLMs (Han et al., 2023; Chen et al., 2023), we conduct experiments using the medical QA datasets MedQuAD (Ben Abacha and Demner-Fushman, 2019), MASH-QA (Zhu et al., 2020), and BioASQ (Krithara et al., 2023), which were not directly used for training both models. MedQuAD encompasses a wide range of question types related to biomedicine, such as diseases, drugs, and medical tests. MASH-QA is a dataset from the consumer health domain where answers need to be extracted from multiple, non-consecutive parts of a long document. BioASQ is a biomedical dataset derived from PubMed, designed to support a range of tasks, including question-answering, information retrieval, and summarization. Although MASH-QA and BioASQ provide gold passages containing answers, our experiments do not utilize these gold passages. Instead, we rely on passages retrieved by our retrieval framework.

Evaluation Metrics. Since all datasets consist of long-form answers, we use the trained model to evaluate answers. We quantify the relevance of answers by using BertScore (Zhang* et al., 2020), which evaluates the similarity between two sen-

	Llama-2-13B		Llama-2-70B		MedAlpaca-13B		Meditron-70B	
	BertScore	UniEval	BertScore	UniEval	BertScore	UniEval	BertScore	UniEval
				MedQuAD				
K-COMP	**74.08**	69.21	**85.21**	**64.15**	**83.8**	**45.58**	**84.07**	**61.07**
$-Prior$	72.22	**69.28**	82.77	61.14	80.94	40.82	76.43	56.67
FT	71.83	68.13	82.29	61.23	79.79	41.1	76.34	56.01
				MASH-QA				
K-COMP	**71.48**	**55.89**	**84.07**	**68.97**	**82.93**	**51.12**	**84.07**	**61.07**
$-Prior$	61.63	48.11	83.32	62.72	80.84	43.97	82.19	61.07
FT	62.25	48.4	83.17	62.57	80.23	44.38	82.12	59.1
				BioASQ				
K-COMP	**72.43**	**66.16**	**87.28**	**65.05**	**84.62**	**44.96**	**86.56**	**61.4**
$-Prior$	67.12	58.83	87.23	62.37	81.78	39.04	86.41	59.75
FT	67.32	58.6	86.89	62.43	81.56	38.13	85.47	59.03

Table 3: Ablation studies. $-Prior$ denotes the scenario where K-comp does not provide prior knowledge to the reader LLMs.

tences by exploiting the contextual embeddings of the encoder. We also use UniEval (Zhong et al., 2022), which is a multi-dimensional evaluation metric that has high correlation and similarity with human judgment. We explicitly assess the factual consistency between generated and gold answers.

Implementation Details. We fine-tuned Gemma-2B (Team et al., 2024) with our knowledge injection objective as the compressor. K-COMP was trained for 3 epochs with a batch size of 8, using the AdamW (Loshchilov and Hutter, 2019) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. We set the peak learning rate to 1×10^{-4} with 3% warm-up ratio and linear decay. For compressors and reader models, we employ top-p sampling (Holtzman et al., 2020) with p=1.0 and a temperature of 0.01. Both training and inference were run on 1-2 NVIDIA A100 GPUs with 80GB memory. We use vLLM (Kwon et al., 2023) to accelerate inference. To evaluate K-COMP, we use various models with differing parameters and purposes (Touvron et al., 2023; Han et al., 2023; Chen et al., 2023; Jiang et al., 2024) within the constraints of the available hardware.

5.2 Results

Baselines. We compare K-COMP with standard RAG approach with top-1 and top-5 retrieved passages without applying prompt compression. We also compare with previous state-of-the-art prompt compression methods, including RECOMP (Xu et al., 2024a) and LLMLingua (Jiang et al., 2023a). Specifically, for implementing RECOMP, we use an abstractive compressor fine-tuned on the Natural Questions dataset (Kwiatkowski et al., 2019), and for LLMLingua, we use Llama-2-7B (Touvron et al., 2023) for compression. Furthermore, we evaluate against a model fine-tuned (FT) using only the standard language modeling objective for summarization, without causal knowledge injection, to verify the importance of automatically generating prior knowledge.

Overall Performance. Table 2 shows the main results of K-COMP compared to the baselines across various reader LLMs. For MedAlpaca, which has the smallest context window size of 2048 among the reader models, answer accuracy declines significantly with Top-5 passages input due to the limited window size. Overall, compressing the context and providing it to the reader model is effective. Chunking snippets for retrieval is inherently imperfect, making the Top-1 and Top-5 passages suboptimal. Consequently, a reprocessing stage, such as compression, is required to improve the quality of chunked text and enable the reader model to reference it appropriately. Among baselines, although RECOMP is trained in an open domain, it performs relatively better than other baselines when applied to the medical domain. However, for the BioASQ dataset constructed from PubMed, directly providing the retrieved passages to the reader model without compression proves exceptionally effective. As a result, some baselines perform better without the compression process than models fine-tuned (FT) on each dataset with compressed context. Nonetheless, K-COMP directly provides focused and concise compressed context and supplies domain knowledge, and is therefore

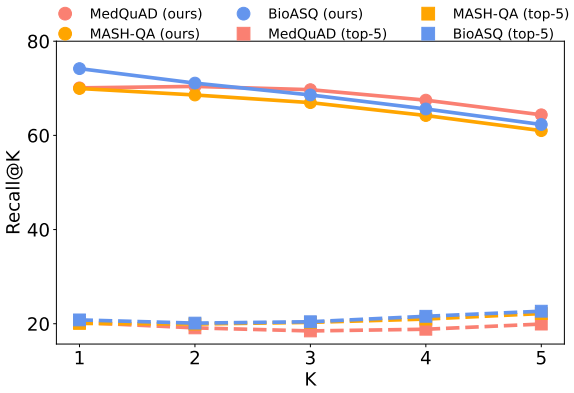

Figure 2: Percentage of Recall@K according to the variation of K for the retrieved passages and our compressed contexts, where top-5 denotes the five passages with the highest similarity scores among the 15 passages retrieved by the retriever.

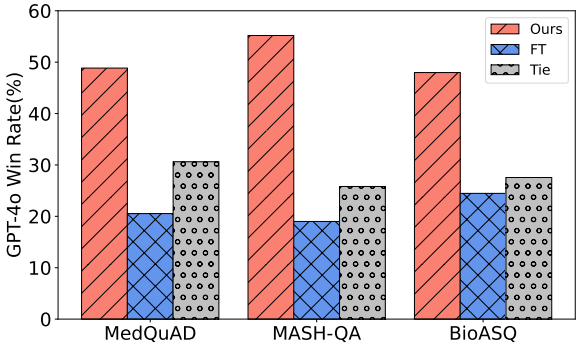

Figure 3: GPT-4o evaluation. Comparison between K-COMP's summaries and FT's summaries.

suitable for reader models with diverse parameters and backgrounds.

Table 3 highlights the importance of automatically generating prior knowledge by comparing it with prompts that do not provide knowledge ($-Prior$). Even $-Prior$ is comparable to the baseline fine-tuned for summarization tasks. However, it is clear that providing prior knowledge to the reader model significantly improves the accuracy of the final answers compared to FT. Additionally, for the BioASQ data, although the FT is relatively inferior across several metrics, the injection of prior knowledge offers a potential solution. This analysis is confined to the QA accuracy of reader LLMs as they are influenced by changes in the components that form the prompt. The following sections will discuss the relevance and alignment of the summary.

5.3 Analyses

Reranking Preference. In addition to QA task performance, it is essential to ensure that summaries are generated to be relevant to the question. Although human evaluation is valuable, it demands significant resources and domain expertise, which are not readily available in our case. Instead, we propose employing a state-of-the-art sentence embedding model[5] (Li and Li, 2024) as a reranker to measure the relevance between the context and the question. For each question q, we execute the compressor to produce five contexts using a high-temperature setting (temperature=1) based on $\{q, \mathcal{P}\}$ pairs, in a manner similar to the inference. Next, we retrieve the top-15 passages related to q. Thus, we gather a total of 20 passages to be fed to the reranker. By applying Recall@K to these 20 passages, we observe the K passages that are most similar to q, and quantify the proportion of K-COMP varied as K varied.

Figure 2 illustrates Recall@K across different values of K. Specifically, we achieved Recall@1 scores of 70%, 70%, and 74% on MedQuAD, MASH-QA, and BioASQ, whereas the top-5 retrieved passages achieved 20%, 20%, and 21%. This comparison demonstrates that the reranker strongly prefers our compressed contexts across all three benchmarks. Additionally, for the BioASQ, where the performance of the top-5 passages in the QA task is relatively high, we estimate that this result is influenced by the overlap between our retrieval corpora and PubMed used to create the dataset.

Qualitative Evaluation with GPT-4. We further explore the additional alignment of the context. Given that GPT-4 has been demonstrated to correlate highly with human judgments (Liu et al., 2023b), we used GPT-4o[6] to perform a comparative evaluation of summaries generated by K-comp and FT. As can be seen in Figure 3, GPT-4o preferred summaries generated by K-COMP. Specifically, FT's summaries have an average preference of 21.3%, whereas ours have a preference of 50.7%. Additionally, the average lengths of the summaries generated by the proposed methodology (532 characters) are comparable with FT (571 characters). This indicates that effect of GPT's potential bias towards longer responses (Wang et al., 2024a) on

[5]Following the MTEB Leaderboard (Muennighoff et al., 2023), we use WhereIsAI/UAE-Large-V1.

[6]We use gpt-4o-2024-05-13 (https://openai.com/index/hello-gpt-4o/).

	Question: How common are **ear infections in adults** ?	
	Augmented Passage	Answer
FT	### Passage Ear pain is a common presentation in acute care settings, with infections of the ear being a frequent occurrence. Acute otitis media is a common infection affecting children, often accompanied by cold symptoms ...	Ear infections are less common in adults than in children, but they can still occur. Adults are more likely to experience ear pain due to other causes.
K-COMP	### Prior Knowledge **Infection**:{description} **Adult**: {description} ### Passage **Ear infections in adults**, such as chronic otitis media, are common and can be caused by various bacteria and fungi ...	**Ear infections in adults** are common and can be caused by various bacteria and fungi. The most common causative bacteria include Staphylococcus aureus, Pseudomonas aeruginosa, Proteus mirabilis, and Prevotella melaninogenica. In some cases, ear infections in adults may be misdiagnosed as other conditions, highlighting the importance of proper diagnosis and treatment.

Table 4: Case study. We show the passages used to augment the reader prompt and the answers. **Red texts** highlight the medical entities within the question.

the preferences is minor. Detailed prompts for the GPT-4o are provided Table 8 in the Appendix A.

Inference Speed. In Table 5, we report the number of tokens in the augmented documents and the inference time as metrics for evaluating efficiency. Specifically, we used Llama-2-70B as the reader model and measured the GPU runtime on the BioASQ test set. Both the compressor and reader model are executed on a single A100 GPU. Even when considering the time needed for the compressor inference, our method was able to triple the throughput compared to prepending the top-5 passages, making it more efficient. Moreover, we note that inference speed is dependent on the implementation and size of the reader model. For instance, models with more parameters will suffer increased latency by increasing the number of input tokens. This phenomenon amplifies the speed advantage of K-COMP.

Case Study with K-COMP and FT. In Table 4, we evaluate how K-COMP generates a summary when aligned with the question and the prior knowledge required to answer it. Here, K-COMP is able to address the incidence of ear infections in adults, and provided comprehensive information on common characteristics and the types of bacteria frequently responsible for them. In contrast, the context generated by FT offers information on ear pain and the incidence of ear infections in children, but fails to provide a focused context on the prevalence of ear infections in adults. FT merely summarizes the passages retrieved based on semantic and overall lexical similarities, including keyword matches, to the question without considering the queried intent. Consequently, the reader model does not fully trust the augmented passages; instead, it perceives them as irrelevant noise and generates answers not based on the passages. This result can lead to inaccuracies and potential hallucinations.

Settings	**Top-1**	**Top-5**	**K-COMP**
Input tokens	321	1450	203
Inference time	1,486s	3,926s	1,043s
Compression time	-	-	248s
Total time	1,486s	3,926s	1,291s

Table 5: Inference speed of Llama-2-70B on BioASQ.

6 Conclusion

In this paper, we have proposed a novel method to improve retrieval augmented QA by compressing retrieved documents into text summaries focused on questions. We design a comprehensive scheme that begins with identifying medical entities and annotating data to automatically generate prior knowledge, then extend training and inference methods that enable the autoregressive generation of summaries that incorporate domain knowledge while considering the context causally. Our experiments demonstrate that the prior knowledge and summaries generated by K-COMP positively impact the reader model's ability to answer and increase the performance of retrieval-augmented generation in the medical domain.

Limitations

We rely on an off-the-shelf NER library to work in scenarios where medical entities exist in the question. However, our methodology is ambiguous for QA where the NER tool does not automatically detect keywords or entities absent in the questions. To mitigate these issues, expanding the retrieval corpus with additional text chunks can inject more knowledge into the compressor and learn domain-relevant entities, but this will drastically increase the cost of annotating the data and require enormous resources for retrieval to perform nearest-neighbor searches. Therefore, we consider the problem of extending these retrieval datastores as an important task in retrieval augmentation, and this method can be extended in future work.

Also, our study mainly focuses on English medical QA, which limits generalization to other languages and domains. Additional approaches are required to investigate potential language and domain adaptation tasks. Addressing these aspects will enable the proposed methodology to be applied in other settings, which will provide a more extensive understanding and application of the approach in diverse linguistic and multi-domain environments.

Ethical Statement

We utilized public datasets such as MedQuAD (CC-BY-4.0 License), MASH-QA (Apache License), and BioASQ (CC-BY-2.5 License) in our research. When synthesizing ground-truth summaries, we ensure that no personally identifiable information is used and that all data are anonymized. Our methodology is still in its early stages and is not yet ready for direct practical use in medical domains, where reliability and accuracy are paramount. In particular, hallucinations can have a critical impact on patient care and clinical decision-making. Therefore, our methodology is considered to mitigate hallucination by emphasizing the domain knowledge in healthcare QA research rather than substituting professional medical judgment and by highlighting the alignment of summaries with questions, thus posing no risk of harm.

References

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. 2022. Cm3: A causal masked multimodal model of the internet. *Preprint*, arXiv:2201.07520.

Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. ReQA: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 137–146, Hong Kong, China. Association for Computational Linguistics.

Raviteja Anantha and Danil Vodianik. 2024. Context tuning for retrieval augmented generation. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 15–22, St Julians, Malta. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024a. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen tau Yih. 2024b. Reliable, adaptable, and attributable language models with retrieval. *Preprint*, arXiv:2403.03187.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022. Efficient training of language models to fill in the middle. *Preprint*, arXiv:2207.14255.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Scott Yih, Luke Zettlemoyer, and Mike Lewis. 2023. Incoder: A generative model for code infilling and synthesis. In *The Eleventh International Conference on Learning Representations*.

Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9878–9919, Bangkok, Thailand. Association for Computational Linguistics.

Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2021. MultiReQA: A cross-domain evaluation forRetrieval question answering models. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 94–104, Kyiv, Ukraine. Association for Computational Linguistics.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

Gautier Izacard and Edouard Grave. 2021. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*.

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

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376, Singapore. Association for Computational Linguistics.

Zhengbao Jiang, Frank 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, Singapore. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Li Qiuxia, and Jun Zhao. 2024. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16867–16878, Torino, Italia. ELRA and ICCL.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2024. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. In *Proceedings*

of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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.

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

Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024. Sure: Summarizing retrievals using answer candidates for open-domain QA of LLMs. In *The Twelfth International Conference on Learning Representations*.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xianming Li and Jing Li. 2024. AoE: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1839, Bangkok, Thailand. Association for Computational Linguistics.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2024. RA-DIT: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023a. RECAP: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8404–8419, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Quanyu Long, Wenya Wang, and Sinno Pan. 2023. Adapt in contexts: Retrieval-augmented domain adaptation via in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6525–6542, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22266–22275.

Hongyin Luo, Tianhua Zhang, Yung-Sung Chuang, Yuan Gong, Yoon Kim, Xixin Wu, Helen M. Meng, and James R. Glass. 2023. Search augmented instruction learning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of*

the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2727–2733, Online. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Houxing Ren, Mingjie Zhan, Zhongyuan Wu, and Hongsheng Li. 2024. Empowering character-level text infilling by eliminating sub-tokens. *Preprint*, arXiv:2405.17103.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Cheol Ryu, Seolhwa Lee, Subeen Pang, Chanyeol Choi, Hojun Choi, Myeonggee Min, and Jy-Yong Sohn. 2023. Retrieval-based evaluation for LLMs: A case study in Korean legal QA. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 132–137, Singapore. Association for Computational Linguistics.

Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Geunbae Lee, and Jungseul Ok. 2024. Key-element-informed sllm tuning for document summarization. *Preprint*, arXiv:2406.04625.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Wen tau Yih, and Mike Lewis. 2024a. In-context pretraining: Language modeling beyond document boundaries. In *The Twelfth International Conference on Learning Representations*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024b. REPLUG: Retrieval-augmented black-box language models. 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 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2024a. How far can camels go? exploring the state of instruction tuning on open resources. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2024b. Augmenting black-box llms with medical textbooks for clinical question answering. *Preprint*, arXiv:2309.02233.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024c. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *Preprint*, arXiv:2403.05313.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024a. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024b. Retrieval meets long context large language models. *Preprint*, arXiv:2310.03025.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024c. Unsupervised information refinement training of large language models for retrieval-augmented generation. *Preprint*, arXiv:2402.18150.

Yan Xu, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu, and Dilek Hakkani-Tür. 2023. Kilm: Knowledge injection into encoder-decoder language models. *arXiv preprint*.

Niraj Yagnik, Jay Jhaveri, Vivek Sharma, and Gabriel Pila. 2024. Medlm: Exploring language models for medical question answering systems. *Preprint*, arXiv:2401.11389.

Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023. 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.

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

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023b. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *Preprint*, arXiv:2311.09210.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023c. Augmentation-adapted retriever improves generalization of language models as generic plug-in. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2421–2436, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3840–3849, Online. Association for Computational Linguistics.

A Appendix

Questions: What are the treatments for Complex Regional Pain Syndrome ?
(Question is not included in the prompt.)

Instruction

Please extract the content about the entity in fewer than four sentences.

Passage

Complex regional pain syndrome: a review of evidence-supported treatment options.

Complex regional pain syndrome consists of pain and other symptoms that are unexpectedly severe or protracted after an injury. In type II complex regional pain syndrome, major nerve injury, often with motor involvement, is the cause; in complex regional pain syndrome I, the culprit is a more occult lesion, often a lesser injury that predominantly affects unmyelinated axons.

... (skip)

Other treatments with encouraging published results (eg, neural stimulators) are not used often enough. We hope to encourage clinicians to rely more on evidence-supported treatments for complex regional pain syndrome.

Physical modalities for complex regional pain syndrome.

Hand therapy is the backbone of a treatment program for complex regional pain syndrome. Various treatment techniques and physical modalities are described in the framework of a clear set of treatment goals. Hand therapy is often the only treatment necessary for simple cases. Adjunct treatments, such as injections or other pharmacologic agents, may be needed when pain control is problematic.

[Spinal cord stimulation for complex regional pain syndrome: report of 2 cases.]

Two adolescents with complex regional pain syndrome (CRPS) were treated safely and effectively by spinal cord stimulation (SCS). They complained of intractable pain resistant to conservative therapies. Whereas continuous epidural anesthesia temporarily reduced pain, SCS was more effective in alleviating chronic severe pain and improving the quality of life. With careful selection of patients, SCS therapy might be recommended even in young cases.

Complex Regional Pain Syndrome – Treatment / Management – Pharmacotherapy

Multiple pharmacotherapeutic agents are used in the management of CRPS. The commonly used therapeutic options in this category include anti-inflammatory medications, anticonvulsants, antidepressants, transdermal lidocaine, opioids, NMDA antagonists, and bisphosphonates. Using a multimodal pharmacologic regimen that combines several different classes may lead to superior outcomes.

[Complex regional pain syndrome-An interdisciplinary view from the surgical consultation.]

Chronic pain disorders are common and have a substantial impact on the patients' daily life. The specific syndrome of complex regional pain syndrome (CRPS, Sudeck's disease) is comparatively rare and characterized by additional sensorimotor, vascular and trophic dysfunctions.

... (skip)

Bisphosphonates, steroids and antiepileptic drugs are well-established as medicinal treatment but should always be used in combination with functional therapy. Interventional treatment options are reserved for patients with complicated and enduring symptoms and should be carried out in specialized centers. The course of the disease is highly individual and frequently requires a long-term interdisciplinary treatment.

Entity

treatment, Complex Regional Pain Syndrome

Table 6: Prompt for summary synthesis.

Passage

Psoriasis in the mouth is rare, with lesions appearing as white or grey-yellow plaques. Fissured tongue is a common finding in those with oral psoriasis, occurring in 6.5-20% of people with psoriasis affecting the skin. Psoriasis in the mouth may be asymptomatic or present as white or grey-yellow plaques in the mouth

Prior Knowledge

psoriasis: Skin disease

mouth: First portion of the alimentary canal that receives food

Questions

What does psoriasis on your lips look like?

Passage

Psoriasis

Seborrheic-like psoriasis Seborrheic-like psoriasis is a common form of psoriasis with clinical aspects of psoriasis and seborrheic dermatitis, and it may be difficult to distinguish from the latter. This form of psoriasis typically manifests as red plaques with greasy scales in areas of higher sebum production such as the scalp, forehead, skin folds next to the nose, the skin surrounding the mouth, skin on the chest above the sternum, and in skin folds.

Clinical presentation of psoriasis.

Psoriasis is a chronic, inflammatory disease affecting 1-3% of the world's population. Joints can be affected in up to 30% of patients. About one third of patients have either severe or moderate (involving more than 10% of body surface area) disease.

... (skip)

Nail psoriasis shows various features: nail pits; oil spots; subungual hyperkeratosis; onycholysis. Rare forms include psoriasis circinata, lip psoriasis and oral psoriasis. Differential diagnosis includes many other dermatological conditions.

Psoriasis

Mouth Psoriasis in the mouth is very rare, in contrast to lichen planus, another common papulosquamous disorder that commonly involves both the skin and mouth.

... (skip)

The microscopic appearance of oral mucosa affected by geographic tongue (migratory stomatitis) is very similar to the appearance of psoriasis. However, modern studies have failed to demonstrate any link between the two conditions.

Oral changes in patients with psoriasis.

Psoriasis is one of the most frequent skin diseases. The cause of psoriasis is not fully expained as there are many factors (infectious, traumatic, hormonal, and chemical) that may play a role in the manifestation of its symptoms.

... (skip)

The psoriasis arthritis changes can also affect temporomandibular joint and impair the function of stomatognathic system. Because of these reports, cooperation of dermatologists and dentists in psoriasis care seems to be necessary.

Psoriasis – History and Physical

Erythrodermic psoriasis presents with widespread inflammation in the form of erythema and exfoliation of the skin covering more than 90% of the body area. It is associated with severe itching, swelling, and pain.

... (skip)

Fissured tongue is the most common finding of oral psoriasis and has been reported to occur in 6.5% to 20% of people with psoriasis affecting the skin.

Questions

What does psoriasis on your lips look like?

Table 7: Prompt for reader LLMs. (Above: K-COMP, Below: Top-5 passages)

Instruction

Select which summary (Summary 1 or Summary 2 or Tie) is more relevant and plausible as a rationale to answer a given question. Choice: [Summary 1, Summary 2, Tie], do not offer any opinions other than the choice.

Summary 1

X-chromosome inactivation (XCI) is a process that silences one of the two X chromosomes in female cells, leaving one X active and one inactive. Some genes escape XCI, allowing them to remain active in some somatic cells. This escape is important for genes like TLR7, which are essential for innate immunity and autoimmune diseases. Additionally, some genes can be expressed from both active and inactive X chromosomes, indicating the presence of double dosage in females. This double dosage can lead to differences in gene expression between males and females, with some genes being more active in females compared to males.

Summary 2

Escape from X inactivation is a process that allows some genes on the X chromosome to escape silencing and be expressed in somatic cells. This process is crucial for maintaining X chromosome inactivation in female cells, as some genes may escape silencing and be expressed in somatic cells. Escape from X inactivation is a phenomenon that has been studied in various organisms, including humans, and has implications for immune responses and autoimmune diseases.

Question

In which cells does TLR7 escape X-chromosome inactivation?

Table 8: Prompt for GPT-4o evaluation. (Summary 1: K-COMP, Summary 2: FT)