Say Less, Mean More:
Leveraging Pragmatics in Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

We propose a simple, unsupervised method that injects pragmatic principles in retrieval-augmented generation (RAG) frameworks such as Dense Passage Retrieval (Karpukhin et al., 2020) to enhance the utility of retrieved contexts. Our approach first identifies which sentences in a pool of documents retrieved by RAG are most relevant to the question at hand, cover all the topics addressed in the input question and no more, and then highlights these sentences within their context, before they are provided to the LLM, without truncating or altering the context in any other way. We show that this simple idea brings consistent improvements in experiments on three question answering tasks (ARC-Challenge, PubHealth and PopQA) using five different LLMs. It notably enhances relative accuracy by up to 19.7% on PubHealth and 10% on ARC-Challenge compared to a conventional RAG system.

1 Introduction

Retrieval-augmented generation (RAG) (Lewis et al., 2020) has emerged as a solution to the limited knowledge horizon of large language models (LLMs). RAG combines "pre-trained parametric and non-parametric memory for language generation," (Lewis et al., 2020) with the non-parametric memory typically retrieved from large collections of documents. RAG has been shown to dramatically improve the performance of LLMs on various question-answering and reasoning tasks (see section 2). However, we argue that RAG often overwhelms the LLM with too much information, only some of which may be relevant to the task at hand. This contradicts Grice's four maxims of effective communication (Grice, 1975), which state that the information provided should be "as much as needed, and no more" and that it should be "as clear, as brief" as possible. The four maxims are enumerated as follows: (1) *Maxim of Quantity*: Provide as much information as needed, but no more; (2) *Maxim of Quality*: Be truthful; avoid giving information that is false or unsupported; (3) *Maxim of Relation*: Be relevant, sharing only information pertinent to the discussion; (4) *Maxim of Manner*: Be clear, brief, and orderly; avoid obscurity and ambiguity. While these maxims were originally formulated in the context of human communication, we argue that they are also applicable in a RAG setting.

We propose a simple, unsupervised method that injects pragmatics in any RAG framework. In particular, our method: (a) identifies which sentences in a pool of documents retrieved by RAG are most relevant to the question at hand (maxim of relation), and cover all the topics addressed in the input question and no more (maxim of quantity and manner);[1] and (b) highlights these sentences within their original contexts before they are provided to the LLM. Table 1 shows an example of our method in action.

The contributions of our paper are:

(1) We introduce a strategy to introduce pragmatics into any RAG method such as Dense Passage Retrieval (Karpukhin et al., 2020). To our knowledge, we are the first to investigate the impact of pragmatics for RAG.

(2) We evaluate the contributions of pragmatics in RAG on three datasets: ARC-Challenge (Clark et al., 2018), PubHealth (Kotonya and Toni, 2020) and PopQA (Mallen et al., 2022) and with five different LLMs ranging from 1B to 7B parameters: Mistral-7B-Instruct-v0.1 (Jiang et al., 2023a), Alpaca-7B (Taori et al., 2023), Llama2-7B-chat (Touvron et al., 2023), Qwen2.5-3B (Team, 2024) and AMD-OLMo-1B-SFT (Liu et al., 2024). Our results indicate that pragmatics helps the most when the QA task primarily involves single-hop or multi-hop logical deduction where the highlighted evidence comprises factual statements that can be

[1] We envision that the maxim of quality could be considered too by identifying factual statements (Rudinger et al., 2018). We leave this for future work.

sequentially chained to derive the answer. Our post-hoc analysis further shows that this approach fares especially well for queries that benefit from analogical reasoning; with highlighted evidence sentences resembling in-context learning exemplars, proving especially useful for smaller language models with limited reasoning capabilities such as AMD-OLMo-1B-SFT, enabling a 10% relative improvement on ARC-Challenge for this model.

(3) We find that pragmatics is less effective when the QA task requires arithmetic manipulation, or involves subtleties such as *double negation*. Furthermore, we find that for factoid QA tasks, if a set of ambiguous contexts are first retrieved by DPR for a given query where the query lacks disambiguating information and multiple plausible answers could be derived, our method struggles to identify the appropriate evidence sentences for highlighting. In such cases, incorrect evidence highlighting can yield a slight degradation in LLM performance.

(4) Our empirical evidence suggests that our method is complementary when paired with a strong retriever like DPR; in favorable cases it can improve performance by up to 20%, while exhibiting minimal degradation (approximately 1%) in less optimal scenarios. Thus, we present it as a low risk and low overhead default augmentation to standard DPR implementations.

2 Related Work

Since it was first proposed (Lewis et al., 2020), RAG has become an essential arrow in the quiver of LLM tools. However, many of the proposed RAG approaches rely on supervised learning to jointly optimize the retrieval component and the LLM (Lewis et al., 2020; Guu et al., 2020; Xu et al., 2024; Kim and Lee, 2024, inter alia) or to decide "when to retrieve" (Asai et al., 2024). Instead, our approach is training free: it uses a set of unsupervised heuristics that approximate Grice's maxims (refer to Section 1). Part of our method is similar to Active-RAG, which also reformulates the input query (Jiang et al., 2023b). However, unlike Active-RAG, we use pragmatics to reformulate the input query and retrieve evidence for it, instead of relying on LLM probabilities. Our work is also similar to (Xu et al., 2024) and (Sarthi et al., 2024), which also touch on pragmatics by reducing the quantity of text presented to the LLM through summarization. However, the method used in (Xu et al., 2024) is supervised. Furthermore, both of these methods

Highlighted evidence	[...] Bats are famous for using echolocation to hunt down their prey, using sonar sounds to capture them in the dark. Another reason for nocturnality is avoiding the heat of the day. **<evidence>This is especially true in arid biomes like deserts, where nocturnal behavior prevents creatures from losing precious water during the hot, dry daytime.</evidence>** This is an adaptation that enhances osmoregulation. One of the reasons that (cathemeral) lions prefer to hunt at night is to conserve water.
MCQ	Question: Many desert animals are only active at night. How does being active only at night most help them survive in a hot desert climate? Choices: A. They can see insects that light up at night. B. Their bodies lose less water in the cool night air. C. They are able to find more plant food by moonlight. D. Their bodies absorb sunlight in the daytime while they sleep.

Table 1: Example of a multiple-choice question (MCQ) from the ARC-C dataset (Clark et al., 2018) together with a fragment of a supporting document retrieved, in which the relevant evidence is highlighted with "<evidence>" tokens by our pragmatics-inspired algorithm. This evidence highlighting allows the downstream LLM to identify the correct answer (option B).

exhibit considerably higher overhead compared to our proposed approach, which relies on simple yet robust heuristics.

Our method adopts a *pre-retrieval* reasoning approach that is complementary to post-retrieval reasoning approaches such as (Trivedi et al., 2023; Kim et al., 2023), which reason after document retrieval. Further, we do not focus on reasoning about whether the retrieval was useful or not (Islam et al., 2024). Further, we do not focus on reasoning about whether the retrieval was useful or not (Islam et al., 2024). For example, current approaches that incorporate reasoning into the QA task, such as rStar (Qi et al., 2024), use an LLM to guide MCTS, where each intermediate step in the tree is verified by another LLM. (Jiang et al., 2024) demonstrate that, rather than relying solely on the LLM's parametric knowledge, retrieved contexts can also enhance tree search. Another reasoning-based approach, STaR (Zelikman et al., 2022), employs an LLM to iteratively generate and refine a training set of rationales. The LLM is then fine-tuned on these rationales, generates a new set of rationales, and repeats the process. In contrast, our method integrates reasoning directly into retrieval in a more efficient manner; specifically, we first reason about the task and then retrieve using the simple technique described in (Zheng et al., 2024).

Lastly, our work focuses on improving the utility of retrieved documents, somewhat similar to CRAG (Yan et al., 2024). However, we do not improve utility by retrieving more documents (e.g., from a web search) but rather by highlighting useful information already present in the current set of documents through pragmatics. All previous methods, especially those based on summarization (Xu et al., 2024) reduce the text by chopping it. Ours does not. The key idea of our work is to extract more utility *while keeping the full text*.

3 Approach: Combining Step-Back Reasoning With Pragmatic Retrieval

Conceptually, our approach is a simple plug-and-play extension that emphasizes important information in any standard RAG setup (as shown in Figure 1). In this paper, we apply our extension to a collection of documents retrieved by a dense passage retriever (DPR) (Izacard et al., 2021).[2] We adapt the unsupervised iterative sentence retriever proposed by (Yadav et al., 2020) to identify important sentences in the documents retrieved by RAG with DPR, as follows: **(1)** Given a query and associated passages retrieved by DPR, the query is first conjoined with a more abstract *step-back* version of itself created by a *step-back LLM* (Zheng et al., 2024). **(2)** In the first sentence retrieval iteration, this conjoined query is used to retrieve a set of relevant evidence sentences from the corresponding passages (see Eqs. 1 and 2). **(3)** In the next iteration(s), the query is reformulated to focus on *missing information*, i.e., query keywords not covered by the current set of retrieved evidence sentences (see Eq. 3) and the process repeats until all question phrases are covered. As such, this strategy implements Grice's maxims of relation (because the evidence sentences are relevant to the question), quantity, and manner (because we identify as many sentences as needed to cover the question and no more). By aggregating sets of retrieved evidence sentences across iterations, this retrieval strategy allows constructing *chains* of evidence sentences for a given query, which can extend dynamically until a parameter-free termination criteria is reached. Further, by varying the first evidence sentence in the top N^3 retrieved evidences, we can trivially extend this retriever to extract *parallel evidence chains*, each of varying lengths, to create a more diverse set of evidence sentences that support the query.

Lastly, we condition the generation of the Question Answering (QA) LLMs on the retrieved evidences, highlighted with special *evidence tokens*, embedded in their original DPR contexts, in order (see Table 1 for an example). We describe each of these stages in more detail below.

3.1 Step-Back Query Expansion

In this work, we employ *Step-Back Prompting* (Zheng et al., 2024), a simple technique to integrate LLM driven reasoning into the retrieval process. A step-back prompt elicits from the LLM an abstract, higher-level question derived from the original query, encouraging higher-level reasoning about the problem. For example, a step-back version of the query: "As bank president, Alex Sink eliminated thousands of Florida jobs while taking over \$8 million in salary and bonuses. True or False?" could be: "What were the actions taken by Alex Sink as bank president?". We hypothesize that step-back queries, representing a more generalized query formulation, when utilized as initialization seeds for the iterative retrieval (refer to Figure 1), will generate a more diverse yet still relevant set of candidate evidence sentences. For multiple-choice questions (MCQs), we generate step-back answer choices for each option, combining them with the step-back query to guide retrieval. This approach introduces an additional dimension of parallelism in constructing evidence chains for MCQs. The stepback prompts used for multi-hop reasoning are adapted from (Zheng et al., 2024) (refer to appendix C for prompts and Table 8 for examples of stepback questions).

3.2 Parallel Iterative Evidence Retrieval

Computing an alignment score between queries and documents is a critical step in any retrieval system. Keeping in mind the Gricean maxim's of *quality* and *relation* (Section 1), which emphasize relevance and factual grounding, we leverage a principle similar to "late interaction" (Khattab and Zaharia, 2020) & (Santhanam et al., 2022), where evidences are selected based on token-level similarities between queries and KB passages. We align

[2] We use the same KB collection of documents as Self-RAG (Asai et al., 2024) and CRAG (Yan et al., 2024).

[3] In our experiments, we set $N = 3$.

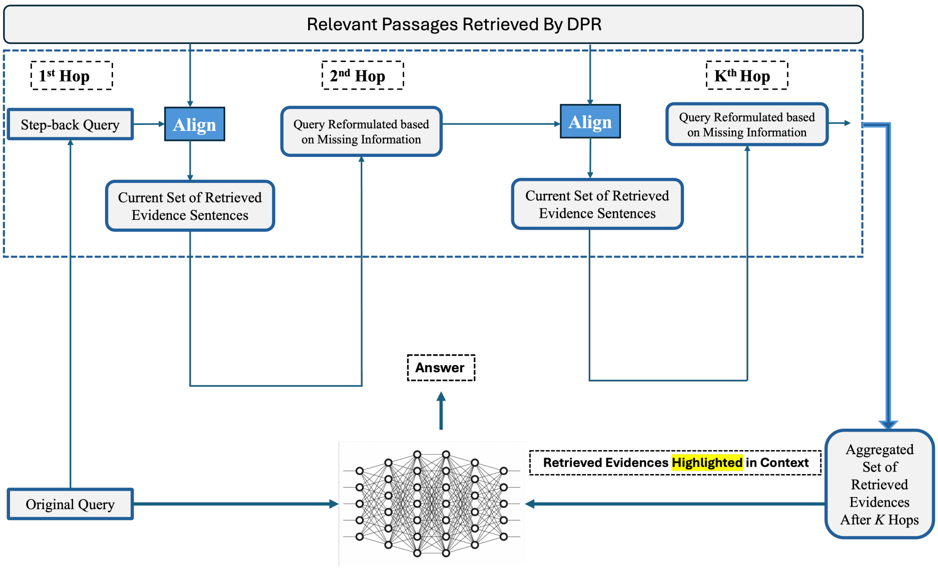

Figure 1: Our proposed method. Each query is concatenated with a more abstract *Step-back* version of itself synthesized by a *Step-back* LLM. This new query is used initiate multi-hop retrieval where in each hop the query is aligned with passages retrieved by DPR to select one evidence sentence. These sentences are aggregated across hops with alignment at each hop driven by query reformulation based on *missing information* (maxim of relation) between the current set of selected evidence sentences and current query. After all query keywords are covered by the retrieved evidences (maxim of quantity), our method highlights them within their original contexts and provides them to the LLM.

query tokens with tokens from each sentence in the KB passages to construct evidence sentences, by selecting the most maximally similar token from the KB passage based on cosine similarity scores over dense embeddings[4] (Equation 1).

$$s(Q, P_j) = \sum_{i=1}^{|Q|} align(q_i, P_j) \tag{1}$$

$$align(q_i, P_j) = \max_{k=1}^{|P_j|} cosSim(q_i, p_k) \tag{2}$$

where q_i and p_k are the i^{th} and k^{th} terms of the query (Q) and evidence sentence (P_j) respectively.

Query reformulation is driven by remainder terms, defined as the set of query terms which have not yet been covered by the set of evidence sentences which were retrieved in the first i iterations of the multi-hop retriever (Equation 3):

$$Q_r(i) = t(Q) - \bigcup_{s_k \in S_i} t(s_k) \tag{3}$$

where $t(Q)$ represents the unique set of query terms, $t(s_k)$ represents the unique terms of the k^{th} evidence sentence in set S_i, which is the set of evidences retrieved in the i^{th} iteration of the retrieval process.

The notion of coverage here is based on soft matching alignment: a query term is considered to be included in the set of evidence terms if its cosine similarity with a evidence term is greater than M[5]. Note that the goal of query reformulation is to maximize the coverage of the query keywords by the retrieved chain of evidences, which aligns with the notion of the maxim of *quantity* (Section 1).

Ambiguous queries are mitigated by dynamically expanding the current query with terms from all previously retrieved evidence sentences if the number of uncovered terms in the query falls below T,[6] which also satisfies the last of Grice's maxims (maxim of *manner*).

4 Results

Evaluation & Datasets: We evaluate our method on the test sets of ARC-Challenge (a *MCQ* reasoning dataset), PubHealth (a fact *verification* dataset about public health) & PopQA (open-domain question-answering). For closed-tasks (ARC-Challenge, PubHealth), we evaluate Accuracy. For the short-form generation task (PopQA),

[4]While (Yadav et al., 2020) align tokens based on similarity over GloVe embeddings, we use sentence transformer embeddings: `https://huggingface.co/jinaai/jina-embeddings-v2-base-en`

[5]In this work, we set $M = 0.98$.

[6]In this work, we set $T = 4$.

Settings	ARC-C	PubHealth	PopQA
No Retrieval			
Mistral-7B-Instruct	62.39 (+6.72%)	74.82 (+0.96%)	32.52 (-49.73%)
Alpaca-7B	34.02 (-17.43%)	43.25 (-7.78%)	30.24 (-53.04%)
Llama2-7B	40.94 (-9.78%)	68.02 (+10.57%)	23.73 (-64.07%)
Qwen-2.5-3B	**78.12** (+7.28%)	65.89 (-7.15%)	26.88 (-62.39%)
AMD-OLMo-1B-SFT	25.81 (-0.17%)	60.81 (+0.00%)	33.38 (-44.14%)
DPR (No Evidence Highlighting)			
Mistral-7B-Instruct	58.46	74.11	64.69
Alpaca-7B	41.20	46.90	64.40
Llama2-7B-chat	45.38	61.52	66.05
Qwen-2.5-3B	72.82	70.96	71.48
AMD-OLMo-1B-SFT	25.64	60.81	59.76
DPR + Evidence Highlighting + No Step-back			
Mistral-7B-Instruct	59.23 (+1.32%)	76.04 (+2.60%)	63.90 (-1.22%)
Alpaca-7B	41.28 (+0.19%)	50.56 (+7.80%)	63.83 (-0.89%)
Llama2-7B-chat	47.44 (+4.54%)	62.64 (+1.82%)	65.98 (-0.10%)
Qwen-2.5-3B	73.42 (+0.82%)	71.17 (0.3%)	**73.05** (+2.2%)
AMD-OLMo-1B-SFT	28.21 (+10.02%)	61.02 (+0.35%)	60.54 (+1.31%)
DPR + Evidence Highlighting + Step-back			
Mistral-7B-Instruct	59.57 (+1.90%)	**76.14** (+2.74%)	64.19 (-0.77%)
Alpaca-7B	41.37 (+0.41%)	56.14 (+19.70%)	64.05 (-0.54%)
Llama2-7B-chat	47.95 (+5.66%)	66.40 (+7.94%)	65.76 (-0.43%)
Qwen-2.5-3B	74.19 (+1.88%)	70.15 (-1.14%)	72.91 (+2.0%)
AMD-OLMo-1B-SFT	28.21 (+10.02%)	62.03 (+2.01%)	60.47 (+1.19%)

Table 2: Our pragmatics driven RAG versus a Standard DPR RAG setup. **Bold** numbers indicate the best performance among all methods and LLMs for a specific dataset. Percentage changes relative to the *DPR without Evidence Highlighting* setting are shown in parentheses. Positive changes are highlighted in green, negative in red. In the *No Retrieval* setting, we do not retrieve any documents and test the LLM's parametric knowledge. *DPR (No Evidence Highlighting)* refers to the setting where we provide the top-K passages for each query to the LLM without highlighting any evidence sentences within those passages. In the *DPR + Evidence Highlighting + No Step-back* setting, we provide DPR passages annotated with highlighted evidences using "<evidence>" tokens. The *DPR + Evidence Highlighting + Step-back* setting extends the previous setting by introducing reformulated queries and answer choices using Step-back prompting.

the metrics indicate performance based on whether gold answers are included in the model generations instead of strictly requiring exact matching (Appendix C). Table 2 shows that integrating pragmatic hints into RAG can enhance performance over DPR. For example, on ARC-Challenge, combining evidence highlighting with step-back reasoning improves Llama-2-7B by up to 5.66% and AMD-OLMo-1B by up to 10% (relative, compared to using just the DPR passages without evidence highlighting). On PubHealth, our method improves Alpaca-7B by up to 19.7% and Llama-2-7B by up to 7.94%. For both PubHealth and ARC-Challenge, the "*DPR + Evidence Highlighting + Step-back reasoning*" setting consistently outperforms the "*Dense Passage Retrieval (DPR) (No Evidence Highlighting)*" setting and the "*DPR + Evidence Highlighting + No Step-back reasoning*" setting.

Choice of LLMs We primarily utilize older language models to mitigate data contamination risks (Sainz et al., 2023). For instance, we excluded DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI et al., 2025) after observing its 90% accuracy on ARC-Challenge under *No Retrieval Setting*—a clear indication of data leakage. While our selected models may still exhibit some contamination (evidenced by strong performance in *No Retrieval* settings), our method demonstrates improvements over these models even when paired with Dense Passage Retrieval, establishing a comparative baseline. Please refer to Appendix D for details of the prompts used and other experimental details.

Dataset and Setting	Llama-2–7B-chat	Alpaca-7B	Mistral-7B-Instruct
ARC-C *(Evidences w/ Context)*	47.95	41.37	59.57
ARC-C *(Evidences w/o Context)*	47.69 (-0.54%)	38.03 (-8.07%)	58.29 (-2.14%)
PubHealth *(Evidences w/ Context)*	66.40	56.14	76.14
PubHealth *(Evidences w/o Context)*	54.82 (-17.44%)	49.34 (-12.11%)	62.23 (-18.27%)

Table 3: Performance of various models on ARC-C and PubHealth datasets when using highlighted evidences within their original context versus using highlighted evidences while discarding surrounding context. Percentage changes (decreases) are shown in parentheses relative to the full context setting. Using highlighted evidence without its surrounding context can significantly degrade the LLMs QA performance.

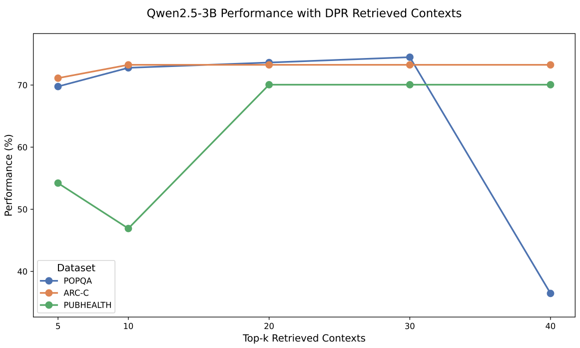

Figure 2: Performance of Qwen2.5-3B with *DPR + Evidence Highlighting + Step-back Reasoning* & varying top-k where k is the number of DPR contexts retrieved

.

5 Analysis

When Does Pragmatics Help? Our error analysis indicates that leveraging pragmatics is effective when answering the query requires connecting facts along a causal path to deduce the answer (as shown in the example of Good Evidence in Table 6, appendix A). We also observe that highlighted evidence often functions as implicit few-shot exemplars, facilitating analogical reasoning. For instance, given the question "In the design process, what is an example of a trade-off?", our method highlights two analogous scenarios: a career decision ("$50,000 salary worker sacrificing income to pursue medical training with the goal of increasing their future income after becoming a doctor") and a biological principle ("beneficial trait changes linked to detrimental ones"). We hypothesize that such examples stimulate the model's in-context learning capabilities, possibly explaining the observed 10% relative improvement in OLMo-1B's performance on ARC-C. However, our method exhibits a few limitations in specific scenarios (refer to Table 7, appendix B). First, it fails to highlight relevant evidences for queries which require arithmetic manipulation or comparison of physical quantities, as these tasks depend more on mathematical reasoning than factual knowledge. Second, it struggles with complex linguistic phenomena, particularly negation patterns. For example, consider the question: "Which human activities would have a positive effect on the natural environment?" Most retrieved passages focus on negative environmental impacts, reflecting their prevalence in real world corpora. The task here requires identifying contrary evidence from the long tail of the distribution, but our unsupervised retrieval heuristics do not account for such semantic inversions.
Lastly, we find that for factoid QA tasks like PopQA, evidence highlighting can slightly degrade performance compared to DPR, likely because these tasks rely more on the model's parametric knowledge. For instance, PopQA queries like "What is Antonio Álvarez Alonso's occupation??" often retrieve ambiguous contexts with multiple roles (e.g., Spanish retired footballer, Spanish paracanoeist, Spanish pianist and composer), offering insufficient signals for disambiguation. In such scenarios, our method may either highlight all potential evidences or arbitrarily select one, confusing the model and potentially leading to incorrect answers.

Time Complexity of Retrieval The computational complexity of our retrieval method can be decomposed into two main components: First, for every query, we make one call to a step-back LLM for query expansion (i.e., creating an abstract step-back version of the query, refer to section 3.1). Second, for evidence selection and highlighting (Yadav et al., 2020), given S sentences retrieved by DPR, we select a subset of K evidence sentences from S passage sentences. In each hop of the iterative retriever, one evidence sentence is chosen from S. The number of hops is upper bounded by the hyperparameter K (where we set $K \leq 6$). Thus the cost of this step is $O(K \times S)$ (constant). Since we allow the retriever to extract N parallel evidence chains by varying the top-scoring evidence (see section 3), the total cost of parallel evidence

Settings	ARC-C	PubHealth	PopQA
No Retrieval			
Mistral-7B-Instruct	62.39 (+9.11%)	**74.82** (+34.23%)	32.52 (+18.17%)
Alpaca-7B	34.02 (-16.02%)	43.25 (+17.05%)	30.24 (-22.66%)
Llama2-7B	40.94 (+0.22%)	68.02 (+0.15%)	23.73 (-0.29%)
Qwen-2.5-3B	78.12 (-0.98%)	65.89 (+51.30%)	26.88 (+0.83%)
AMD-OLMo-1B-SFT	25.81 (+0.00%)	60.81 (+0.00%)	33.38 (+4.25%)
BM25 (No Evidence Highlighting)			
Mistral-7B-Instruct	57.18	55.74	27.52
Alpaca-7B	40.51	36.95	**39.10**
Llama2-7B	40.85	67.92	23.80
Qwen-2.5-3B	**78.89**	43.55	26.66
AMD-OLMo-1B-SFT	25.81	60.81	32.02
BM25 + Evidence Highlighting + No Step-back			
Mistral-7B-Instruct	58.38 (+2.10%)	62.23 (+11.64%)	29.16 (+5.96%)
Alpaca-7B	40.17 (-0.84%)	53.91 (+45.90%)	37.81 (-3.30%)
Llama2-7B	47.69 (+16.74%)	62.23 (-8.38%)	33.88 (+42.35%)
Qwen-2.5-3B	75.13 (-4.77%)	42.84 (-1.63%)	37.03 (+38.90%)
AMD-OLMo-1B-SFT	25.13 (-2.63%)	59.39 (-2.33%)	33.10 (+3.37%)
BM25 + Evidence Highlighting + Step-back			
Mistral-7B-Instruct	58.72 (+2.69%)	62.64 (+12.38%)	29.24 (+6.25%)
Alpaca-7B	40.00 (-1.26%)	45.69 (+23.65%)	38.46 (-1.64%)
Llama2-7B	47.61 (+16.55%)	61.93 (-8.82%)	34.31 (+44.16%)
Qwen-2.5-3B	74.62 (-5.41%)	43.05 (-1.15%)	36.45 (+36.72%)
AMD-OLMo-1B-SFT	25.38 (-1.67%)	60.61 (-0.33%)	33.02 (+3.12%)

Table 4: Our pragmatics driven RAG versus a BM25 RAG setup. **Bold** numbers indicate the best performance among all methods and LLMs for a specific dataset. Percentage changes relative to the BM25 *without Evidence Highlighting* setting are shown in parentheses. Positive changes are highlighted in green, negative in red. In the *No Retrieval* setting, we do not retrieve any documents and test the LLM's parametric knowledge. *BM25 (No Evidence Highlighting)* refers to the setting where we provide the top-K passages for each query to the LLM without highlighting any evidence sentences within those passages. In the *BM25 + Evidence Highlighting + No Step-back setting*, we provide BM25 passages annotated with highlighted evidences using "<evidence>" tokens. The *BM25 + Evidence Highlighting + Step-back* setting extends the previous setting by introducing reformulated queries and answer choices using Step-back prompting.

retrieval is $O(N \times K \times S)$ (constant). Evidence highlighting requires a linear scan of the S passage sentences with complexity $O(S)$ (constant). Therefore, the total computational complexity is: $\text{Cost}_{\text{total}} = \text{Cost}(\text{LLM}_{\text{stepback}}) + O(n)$, where n represents the number of tokens in the retrieved passages S. We note two important considerations: (a) the base retrieval cost is inherent to any RAG system and thus unavoidable, and (b) our method introduces minimal computational overhead compared to alternative reasoning-enhanced QA approaches such as STaR (Zelikman et al., 2022).

Is Full DPR Context necessary? We conduct an experiment where we compare how dropping the context surrounding the highlighted evidence sentences versus keeping it affects QA performance. As shown in Table 3, on both ARC-C and PubHealth with three different LLMs, we find that just providing the highlighted evidence sentences without context can significantly degrade QA performance relative to the scenario where we highlight evidence while keeping the full, surrounding context.

How does the quality of the retrieved passages impact our method? To assess the relationship between initial retrieval quality and our method's effectiveness, we conduct comparative experiments using the sparse retrieval method BM25 (Robertson and Zaragoza, 2009) in place of DPR. For each query, we retrieve the top-20 passages using BM25, then apply our iterative retrieval approach with step-back reasoning (Section 3) to identify and highlight key evidence sentences within these contexts. As shown in Table 4, retrieval quality

Category	Frequency (ARC-Challenge)	Frequency (PubHealth)
Bad (0)	6	8
Medium (0.5)	10	4
Good (1)	4	8

Table 5: Highlighted Evidence Quality Scores for 20 randomly sampled queries from the ARC-Challenge and PubHealth datasets. The frequencies represent the number of instances falling into each quality category for the highlighted evidence in both datasets.

significantly influences our method's performance. We observe substantial improvements across multiple models and datasets: Llama-2-7B achieves a 16.74% gain on ARC-Challenge, Alpaca-7B shows up to a 45.90% improvement on PubHealth, while Llama-7B and Qwen2.5-3B demonstrate gains of up to 44.16% and 38.90% on PopQA, respectively, relative to their baseline BM25 performance. However, the efficacy of our method when applied to BM25-retrieved passages is inconsistent, with several models also demonstrating performance deterioration compared to both baseline BM25 and the "*No Retrieval*" setting. We hypothesize that this is because of two reasons: (a) BM25's lexical overlap-based retrieval mechanism yields passages containing necessary but insufficient information for query resolution. For instance, on ARC-Challenge (refer to Table 4), Alpaca-7B improves by 16% when using BM25-retrieved passages as context, but subsequent evidence highlighting on top of these passages diminishes this gain. (b) Evidence highlighting more effectively grounds the LLM in the retrieved context, potentially overriding useful parametric knowledge. This effect is particularly pronounced with qwen-2.5 3B, where the model significantly degrades by 51.3% when provided with BM25 retrieved passages as contexts relative to "*No Retrieval*", and the application of evidence highlighting over these contexts further reduces performance by 1.6%. This suggests that while evidence highlighting effectively directs model attention in high-quality passages, it creates a bias that may be counterproductive when retrieved passages are of lower quality.[7] In such instances, our method may constrain the model to prioritize highlighted information over potentially superior parametric knowledge (which the model acquired through test data appearing in its pre-training corpus). These results suggest that our approach is more complementary to DPR and similar neural retrieval methods than to lexical matching approaches like BM25.

[7]We do not imply that BM25-retrieved passages are always lower quality than those retrieved by DPR; rather, in this specific case, the DPR *Contriever* has been finetuned on web-domain data (Bajaj et al., 2018) similar to our evaluation datasets, making it a more effective retrieval method. We acknowledge that BM25 can be more robust than DPR out-of-domain.

Evaluating Quality of Highlighted Evidence: We conduct a human evaluation of the quality of evidence highlighting on 40 questions (split evenly between ARC-Challenge and PubHealth), rating each question's set of highlighted evidences for a sample of 40 questions, 20 of which are sampled from ARC-Challenge and 20 of which are sampled from the PubHealth dataset. We score each highlighted evidence according to the following scale: **0 (bad)**, **0.5 (medium)** and **1 (good)**. Overall, 60% to 70% of highlighted evidences were rated at least "medium" by the human evaluator across both datasets. See Appendix A for the evaluation criteria used and examples of 'good', 'medium' and 'bad' evidence sentences.

Understanding the Impact of Top-*k* Retrieval on our approach We analyze the effect of varying DPR's top-k retrieved contexts on Qwen2.5-3B's performance with evidence highlighting and step-back reasoning. Our results (figure 2) indicate a "Goldilocks zone" for k: while larger k values generally improve performance on ARC-C and PubHealth by increasing the likelihood of retrieving relevant information, excessive context ($k > 30$) proves detrimental for PopQA, where additional contexts introduce more ambiguity that degrades LLM performance.

6 Conclusions

We present an unsupervised method that enhances retrieval-augmented generation (RAG) by highlighting key sentences in retrieved documents. We find that this approach can improve QA performance across 3 different datasets and 5 different LLMs.

Limitations

This study investigates the effectiveness of pragmatics in enhancing Retrieval Augmented Generation (RAG) systems. Our evaluation, however, is limited to a comparison against standard Dense Passage Retriever (DPR) and BM25 baselines. The proposed method has potential for integration with more sophisticated RAG systems, such as those developed by (Asai et al., 2024), (Xu et al., 2024), (Sarthi et al., 2024). Our assessment encompasses three datasets, but a more comprehensive evaluation would involve a broader range of single-hop and multi-hop tasks. Moreover, there are several scenarios which our approach does not cover, such as handling linguistic phenomena like negation, mathematical reasoning tasks and reconciling retrieved contexts that are ambiguous. Our current approach is also limited by the fact that it is unsupervised and query reformulation is mostly driven by a bag-of-words. One could trivially improve query reformulation by using an LLM, or using a weakly supervised strategy that fine-tunes an LLM to retrieve pragmatic evidence (using supervision from the current retriever) via a joint loss that learns to retrieve evidence sentences while simultaneously answering the query correctly (motivated by the relevance estimator and answer marginalization losses proposed by (Kim and Lee, 2024)). We leave the exploration of supervised pragmatic RAG methods as future work. While we hypothesize that our retrieved & highlighted justifications constitute "shallow chains of thought" which are faithfully utilized by the Large Language Model in its generations, this assertion remains to be formally validated through rigorous analysis.

References

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

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Herbert Paul Grice. 1975. Logic and conversation. *Syntax and semantics*, 3:43–58.

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

Shayekh Bin Islam, Md Asib Rahman, KSM Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-RAG: Enhanced retrieval augmented reasoning with open-source large language models. In *The 2024 Conference on Empirical Methods in Natural Language Processing*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. 2024. Rag-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. *Preprint*, arXiv:2412.12881.

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. *Preprint*, arXiv:2305.06983.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Preprint*, arXiv:2004.12832.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. *Preprint*, arXiv:2310.14696.

Kiseung Kim and Jay-Yoon Lee. 2024. Re-rag: Improving open-domain qa performance and interpretability with relevance estimator in retrieval-augmented generation. *Preprint*, arXiv:2406.05794.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

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.

Jiang Liu, Jialian Wu, Prakamya Mishra, Zicheng Liu, Sudhanshu Ranjan, Pratik Prabhanjan Brahma, Yusheng Su, Gowtham Ramesh, Peng Sun, Zhe Li, Dong Li, Lu Tian, and Emad Barsoum. 2024. Amd-olmo: A series of 1b language models trained from scratch by amd on amd instinct™ mi250 gpus.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint*.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *Preprint*, arXiv:2408.06195.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *Preprint*, arXiv:2112.01488.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *Proceedings of the International Conference on Machine Learning*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Preprint*, arXiv:2212.10509.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2024. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. In *Proceedings of the International Conference on Machine Learning*.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering. *Preprint*, arXiv:2005.01218.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Preprint*, arXiv:2203.14465.

Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V Le, and Denny Zhou. 2024. Take a step back: Evoking reasoning via abstraction in large language models. *Preprint*, arXiv:2310.06117.

A Human Evaluation of Evidence Quality

A.1 Evaluation Criteria

We categorize highlighted evidence as "bad" (score: 0) when it includes completely irrelevant sentences or sentences within contexts that are somewhat related to the query but fail to provide any meaningful support in addressing it. In the case of fact-checking datasets like PubHealth, we also classify highlighted evidence as "bad" if it appears to support a claim but overlooks negations in the surrounding context that would ultimately refute the claim.

Highlighted evidence is categorized as "medium" (score: 0.5) when it consists of sentences situated in relevant contexts that may allow the correct answer to be inferred indirectly in some instances but lack the direct or explicit support needed to effectively answer the query.

Highlighted evidence is categorized as "good" (score: 1) when it includes a sufficient number of sentences that directly address the query while ensuring no confounding factors (e.g., negations in the surrounding context) are overlooked.

Table 6 shows an example of good, medium and bad quality evidences as assessed by a human evaluator. The example of `Good Evidence` shown is rated as such because connecting the evidence sentences together allows the reader to deduce the answer to the query "What is the atomic mass of the atom?" even without extensive prior knowledge of chemistry.

B Low Quality Evidence

In Table 7, we include some examples of retrieved evidences from the ARC-C dataset that do not help the model to deal with specific tasks, especially those which requiring modeling negation and arithmetic reasoning.

C Step-Back Reasoning Examples

Please refer to Table 8 for examples of original queries and the more abstract *Step-back* questions elicited from those queries.

C.1 Step-back Prompt for Query Expansion

```
You are an expert at world
    knowledge. Your task is to
    step back and paraphrase a
    question to a more generic
    step-back question, which is
    easier to answer. Here are a
    few examples:

Original Question: Which position
     did Knox Cunningham hold from
     May 1955 to Apr 1956?
Stepback Question: Which
    positions have Knox Cunningham
     held in his career?

Original Question: who has scored
     most runs in t20 matches as
    of 2017
```

Category	Examples of Evidences
Good Evidence	**Question:** A certain atom has 20 electrons, 21 neutrons, and 20 protons. What is the atomic mass of the atom? **Highlighted Evidence**: – "Mass number (symbol 'A', from German 'Atomgewicht') is the total number of protons and neutrons (nucleons) in a nucleus." – "Atomic mass is approximately the mass number times an atomic mass unit (approximate mass of a proton, neutron, or hydrogen-1 atom)."
Medium Evidence	**Question:** A law in Japan makes it illegal for citizens of that country to be fat. **Highlighted Evidence**: – "Japan implemented the 'metabo' law in 2008 to combat rising obesity rates." – "The New York Times reported that the law aims to shrink the overweight population by 10% over 4 years and 25% over 7 years via financial penalties." – "In 2008, Japan passed the "Metabo Law," addressing metabolic syndrome—a cluster of conditions increasing the risk of heart disease, stroke, and diabetes." – "The law requires models to have a minimum BMI and warns against photoshopped images."
Bad Evidence	**Question:** Ted Cruz Says Democrats are embracing abortion up until (and even after) birth. **Highlighted Evidence**: – "In January 2016, Cruz announced his "Pro-Lifers for Cruz" coalition, with statements about executing abortion doctors to expunge bloodguilt." – "Kamala Harris refuted Republican claims about Democrats' abortion views." – "In the mid-1990s, Moynihan supported banning the procedure known as partial-birth abortion."

Table 6: Examples of Good, Medium, and Bad Highlighted Evidences

```
Stepback Question: What are the
    runs of players in t20 matches
     as of 2017

Original Question: When was the
    abolishment of the studio that
     distributed The Game?
Stepback Question: which studio
    distributed The Game?

Original Question: What city is
    the person who broadened the
    doctrine of philosophy of
    language from?
Stepback Question: who broadened
    the doctrine of philosophy of
    language

Original Question: Would a
    Monoamine Oxidase candy bar
    cheer up a depressed friend?
Stepback Question: What are the
    effects of Monoamine Oxidase?

What is the Stepback Question for
     this?: {
    original_question_text}
Answer with only the Stepback
    Question and no extra text.
```

C.2 Step-back Prompt for MCQ Answer Choices

```
You are an expert at world
    knowledge. You are given a
    statement. Your task is to
    extract the concepts and
    principles underlying the
    statement. Answer only with
    the concepts and principles
    without any extra text.
If there are multiple concepts
    and principles, list them
    separated by commas.
Original Statement: {answer_text}
Answer:
```

D Experimental Details

Our experimental results for Mistral-7B-Instruct v0.1, Alpaca-7B & Llama-2-7B differ from those reported by other works such as Self-RAG (Asai et al., 2024) & CRAG (Yan et al., 2024), and Speculative RAG due to the following methodological variations:

1. **Evaluation Function:** We employ a different evaluation criteria for assessing accuracy between Large Language Model (LLM) generations and gold labels in tasks such as ARC-Challenge, PopQA, and PubQA. Our

ARC-Challenge	**Question:** Scott filled a tray with juice and put it in a freezer. The next day, Scott opened the freezer. How did the juice most likely change? **Evidence:** - Most recently, Scott produced the documentary film "Apple Pushers" with Joe Cross (filmmaker) juicer and a generator. - However, in March 1996, 70,000 Juice Tiger juicers (9% of its models) were recalled after 14 injury incidents were reported.
ARC-Challenge	**Question:** A physicist wants to determine the speed a car must reach to jump over a ramp. The physicist conducts three trials. In trials two and three, the speed of the car is increased by 20 miles per hour. What is the physicist investigating when he changes the speed? **Evidence:** - Objects in motion often have variations in speed (a car might travel at 50 km/h, slow to 0 km/h, then reach 30 km/h). - Preparing an object for g-tolerance (avoiding damage when subjected to high speeds). - Hence, the round-trip time on traveler clocks will be $\Delta\tau = 4\left(\frac{c}{\alpha}\right)\cosh(\gamma)$.
ARC-Challenge	**Question:** Human activities affect the natural environment in many ways. Which action would have a positive effect on the natural environment? **Evidence:** - This environment encompasses the interaction of all living species, climate, weather, and natural resources affecting human survival and economic activity. - For instance, actions by the U.S. Army Corps of Engineers that threatened ecosystems in Florida's Oklawaha River valley and issues in preserving Pacific Coast Redwood communities are cited as case studies. - Humans have contributed to the extinction of many plants and animals.
Pop-QA	**Question:** What is Antonio Álvarez Alonso's occupation? **Evidence:** - Antonio De Diego Antonio de Diego Álvarez is a Spanish paracanoeist and member of the National Spanish Canoeist Team, Paracanoe class A (maximum level of disability). - Antonio Álvarez Alonso Antonio Álvarez Alonso (11 March 1867 - 22 June 1903) was a Spanish pianist and composer.

Table 7: Examples of low-quality evidences retrieved for various types of queries from ARC-Challenge & Pop-QA

approach considers an LLM generation correct based on the principle of "inclusion," i.e., if the generation includes the correct answer as a substring, post-normalization.

2. **Number of retrieved passages in DPR and BM25 (top-K):** In both BM25 and DPR retrieval, we set $K = 11$ for models which have a 4096 token limit context (e.g., Llama-2-7B), where 10 passages are from the Wikipedia KB mixed with a web search result from CRAG. For Alpaca=7B and AMD-OlMo-1B-SFT, owing to their small context window size of 2048, we keep just the top-9 documents ($K = 9$). For Alpaca and OlMo, we observe significant degradation if we use 10 or more documents causing the DPR setting to perform worse than even the *No-Retrieval model*. For models with larger context windows e.g., Mistral-7B and Qwen2.5-3B we use all DPR and BM25 retrieved passages.

3. **Prompt Engineering:** Our prompts differ slightly from those used in Self-RAG and C-RAG. We have engineered our prompts to adhere more closely to the recommended Instruction Tuning format, particularly for Alpaca-7B (Taori et al., 2023) and Llama-2-7B-chat (Touvron et al., 2023).

4. **Stepback-LLM:** In all experiments, we use Mistral-7B-Instruct v0.1 as the step-back LLM.

These methodological distinctions should be considered when comparing our results with those of previous studies.

E Example Prompts

Examples of the task specific prompts utilized in our study are as follows:

- **ARC-Challenge**
  - Mistral-7B-Instruct:

```
Refer to the following documents
    , follow the instruction and
     answer the question.

Documents: {highlighted_passages
    }

Question: {question}
```

Dataset	Original Question and Step-back Question
ARC-Challenge	**Original Question:** An astronomer observes that a planet rotates faster after a meteorite impact. Which is the most likely effect of this increase in rotation? **Step-back Question:** What effects do meteorite impacts on planets have?
ARC-Challenge	**Original Question:** A group of engineers wanted to know how different building designs would respond during an earthquake. They made several models of buildings and tested each for its ability to withstand earthquake conditions. Which will most likely result from testing different building designs? **Step-back Question:** What are the testing methods used by the engineers to determine the earthquake resilience of the different building models?
PopQA	**Original Question:** What is Henry Feilden's occupation? **Step-back Question:** What are the important aspects of Henry Feilden's academic work?
PubHealth	**Original Question:** A mother revealed to her child in a letter after her death that she had just one eye because she had donated the other to him. **Step-back Question:** What are the circumstances surrounding the donation of the mother's second eye to her child after her death?

Table 8: Examples of Step-back questions created from original questions in the three datasets.

```
Instruction: Given four answer
    candidates, A, B, C and D,
    choose the best answer
    choice.
Please answer with the
    capitalized alphabet only,
    without adding any extra
    phrase or period.
```

– Alpaca-7B:

```
Below is an instruction that
    describes a task. Write a
    response that appropriately
    completes
the request.

### Instruction: Given four
    answer candidates, A, B, C
    and D, choose the best
    answer choice.
Please answer with the
    capitalized alphabet only,
    without adding any extra
    phrase or period.

### Input:
Documents: {highlighted_passages
    }
Question: {question}
Choices: {choices_str}

### Response:
```

– Llama-2-7B-chat:

```
Below is an instruction that
    describes a task. Write a
    response that appropriately
    completes
the request.

### Instruction: Given four
    answer candidates, A, B, C
    and D, choose the best
    answer choice.
Please answer with the
    capitalized alphabet only,
    without adding any extra
    phrase or period.

### Input:
Documents: {highlighted_passages
    }
Question: {question}
Choices: {choices_str}

### Response:
```

- **PopQA**

  – Mistral-7B-Instruct:

```
Refer to the following documents
    , follow the instruction and
     answer the question.

### Input:
Documents: {highlighted_passages
    }

### Instruction: Answer the
    question: {question}
### Response:
```

  – Alpaca-7B:

```
Below is an instruction that
    describes a task. Write a
    response that appropriately
    completes
the request.

### Instruction: Refer to the
    following documents and
    answer the question.
### Input:
Documents: {highlighted_passages
    }

Question: {question}
### Response:
```

  – Llama-2-7B:

```
<s>[INST] <<SYS>>
    You are a helpful,
        respectful and honest
        assistant. Always answer
         as helpfully as
        possible,
```

```
        while being safe. Your
            answers should not
            include any harmful,
            unethical, racist,
            sexist,
        toxic, dangerous, or illegal
             content. Please ensure
            that your responses are
            socially unbiased
        and positive in nature.

        If a question does not make
            any sense, or is not
            factually coherent,
            explain why instead of
        answering something not
            correct. If you don't
            know the answer to a
            question, please don't
        share false information.
<</SYS>>

Below is an instruction that
    describes a task. Write a
    response that appropriately
    completes
the request.

Instruction: Refer to the
    following documents and
    answer the question.

Documents: {highlighted_passages
    }

Question: {question}
### Response: [/INST]
```

- **PubHealth**
  - Mistral-7B-Instruct:

```
Read the documents and answer
    the question: Is the
    following statement correct
    or not?
Only say true if the statement
    is true; otherwise say false
    . Don't capitalize or add
    periods,
just say ``true'' or ``false''.

Documents: {highlighted_passages
    }

Statement: {question}
### Response:
```

  - Alpaca-7B:

```
Below is an instruction that
    describes a task. Write a
    response that appropriately
    completes
the request.

### Instruction: Read the
    documents and answer the
    question: Is the following
    statement correct
or not? Only say true if the
    statement is true; otherwise
     say false. Don't capitalize
     or add
periods, just say ``true'' or ``
    false''.

### Input:
Documents: {highlighted_passages
    }

Statement: {question}
### Response:
```

  - Llama-2-7B:

```
<s>[INST] <<SYS>>
    You are a helpful,
        respectful and honest
        assistant. Always answer
         as helpfully as
        possible,
    while being safe. Your
        answers should not
        include any harmful,
        unethical, racist,
        sexist,
    toxic, dangerous, or illegal
         content. Please ensure
        that your responses are
        socially unbiased
    and positive in nature.

    If a question does not make
        any sense, or is not
        factually coherent,
        explain why instead of
    answering something not
        correct. If you don't
        know the answer to a
        question, please don't
    share false information.
<</SYS>>

Below is an instruction that
    describes a task. Write a
    response that appropriately
    completes
the request.

### Instruction: Read the
    documents and answer the
    question: Is the following
    statement correct or not?
    Only say true if the
    statement is true; otherwise
     say false. Don't capitalize
     or add
periods, just say ``true'' or ``
    false''.

### Input:
Documents: {highlighted_passages
    }

Statement: {question}
### Response: [/INST]
```