ASRank: Zero-Shot Re-Ranking with Answer Scent for Document Retrieval

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) models have drawn considerable attention in modern open-domain question answering. The effectiveness of RAG depends on the quality of the top retrieved documents. However, conventional retrieval methods sometimes fail to rank the most relevant documents at the top. In this paper, we introduce ASRANK, a new re-ranking method based on scoring retrieved documents using zero-shot answer scent which relies on a pretrained large language model to compute the likelihood of the document-derived answers aligning with the answer scent. Our approach demonstrates marked improvements across several datasets, including NQ, TriviaQA, WebQA, ArchivalQA, HotpotQA, and Entity Questions. Notably, ASRANK increases Top-1 retrieval accuracy on NQ from 19.2% to 46.5% for MSS and from 22.1% to 47.3% for BM25. Finally, ASRANK shows strong retrieval performance on several datasets compared to state-of-the-art methods 47.3 Top-1 by ASRANK vs 35.4 by UPR (Sachan et al., 2022) by BM25[1].

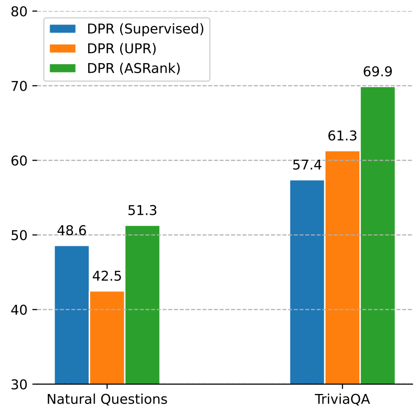

Figure 1: After re-ranking the top 1,000 passages retrieved by DPR (Karpukhin et al., 2020a) with ASRANK, our method surpasses the performance of strong unsupervised models like UPR (Sachan et al., 2022) on the Natural Questions and TriviaQA datasets.

1 Introduction

Document retrieval is a core sub-task in many NLP problems, including open-domain question answering (ODQA), where a document is retrieved and then read to answer an input query. This process tries to find the most relevant documents or passages given the query. The Retrieval-Augmented Generation (RAG) model has achieved a significant improvement in the field of open-domain question answering (ODQA) (Lewis et al., 2020). RAG models combine retrieved documents and advanced pre-trained large language models (LLMs) generating responses based on the retrieved information (Lewis et al., 2020; Lála et al., 2023). However, the performance of RAG models depends on the top retrieved documents, especially on the first document (Setty et al., 2024; Zhang et al., 2024). The RAG model usually uses the first retrieved document, which is the primary source for generating the response. In RAG, queries and documents are embedded in a shared representation space to enable efficient search before using a task-specific model to perform a deeper, token-level document analysis.

The answer scent is a concept analogous to the way in which animals track the scent of their prey (Maxwell and Azzopardi, 2018). Cognitive psychologists (Winerman, 2012) have found that people search for information online in much the same way as animals hunt for food, leading to the establishment of the concept of *Information scent* in the Information Retrieval field. It refers to the trail of relevant information that leads a user to the correct answer. Our proposal is built upon a similar concept of tracing the answer scent.

[1]because the anonymous code and the dataset will be available after reviewing the paper.

In this paper, we introduce ASRANK, a simple, effective, and fast re-ranking method that leverages the concept of answer scent. Our approach first utilizes larger LLMs like GPT-3.5 or Llama 7-70B to generate an answer scent. This is done just once, hence is computationally efficient. Subsequently, a smaller model such as T5 is employed to re-rank the documents based on the received answer scent. This two-tiered approach allows leveraging the generative capabilities of a larger LLM to boost the re-ranking capabilities of smaller models thanks to improved contextual understanding. Our method scores retrieved documents using a zero-shot answer scent, which relies on a pre-trained LLM to compute the likelihood of the document-derived answers aligning with the answer scent. This approach allows to rank documents not just based on their initial retrieval scores but also on the likelihood of containing an answer (via answer scent) and the degree to which they contain information that aligns with the expected answer. By applying a cross-attention mechanism to every token in both the question and the passage, ASRANK tracks the answer scent within the document corpus. Our approach successfully addresses the challenge of ensuring that the most relevant document is ranked at the top, a significant problem in open-domain question answering and RAG systems (Figure 1).

2 Method

In this section, we detail the methodology of ASRANK, starting with retrieving documents based on either sparse or dense techniques. Subsequently, we introduce our concept of generating an Answer Scent using a large language model (Section 2.2), followed by an efficient re-ranking process that employs a smaller model (Section 2.3), which enhances the alignment and relevance of the retrieved documents to the query in our RAG system. Figure 2 shows an overview of the ASRANK framework.

2.1 Retriever

Let $\mathcal{M} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M\}$ represent a collection of evidence documents. Given a query $\boldsymbol{q}$, the retriever's task is to select a subset of relevant documents $\mathcal{D} \subset \mathcal{M}$, aiming to include those that likely contain the answer to $\boldsymbol{q}$. Our framework is designed to operate on documents retrieved by arbitrary methods, hence ones that can either utilize sparse or dense representations. **Sparse representation** methods such as BM25 (Robertson et al., 2009), a non-neural approach, rely on term frequency and inverse document frequency to rank documents. This method is effective for scenarios where lexical matching is crucial, providing a strong baseline due to its simplicity and proven efficiency in various information retrieval tasks. **Dense representation** methods like Dense Passage Retrieval (DPR) (Karpukhin et al., 2020b) employ neural network architectures to encode queries and documents into dense vector spaces. The relevance of documents is assessed based on the similarity of these vectors, allowing to capture semantic relationships that go beyond keyword matching. Regardless of the retrieval technique employed, the retrieval system identifies the top-K most relevant documents, denoted as $\mathcal{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_K\}$.

2.2 Answer Scent Generation

Large language models (LLMs) such as GPT-3 (Brown et al., 2020) and LLaMA (Touvron et al., 2023) define probability distributions over sequences of tokens. Given a sequence $x_1, ..., x_n$, these models typically predict the sequence's probability using an autoregressive approach $p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i|x_{<i})$, where $x_{<i} := x_1, ..., x_{i-1}$ represents the sequence of tokens preceding x_i, also referred to as its *prefix*. This modeling is implemented via a transformer network parameterized by θ_1, typically employing a causal self-attention mask $p(x_1, ..., x_n) = \prod_{i=1}^{n} p_{\theta_1}(x_i|x_{<i})$, which effectively models the conditional probabilities of each token.

In our approach, we incorporate the concept of *answer scent*, which guides the model in generating answers that are contextually appropriate for the query, inspired by the success of In-Context Learning (Brown et al., 2020; Ram et al., 2023; Dong et al., 2022). This context is derived using a zero-shot approach, where the model infers the scent without explicit prior training on such task: $p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i|x_{<i}, \mathcal{S}(x_{<i}))$, where $\mathcal{S}(x_{<i})$ denotes the inferred answer scent in the form of a generated text content. The objective of Scent is to encode the essence of what the answer should represent, enriching the input to the LLM reranker.

2.3 ASRANK Re-Ranking

ASRANK introduces an unsupervised re-ranking utilizing LLM to evaluate the relevance of documents based on $\mathcal{S}(q)$, which serves as the guiding context corresponding to the target query q.

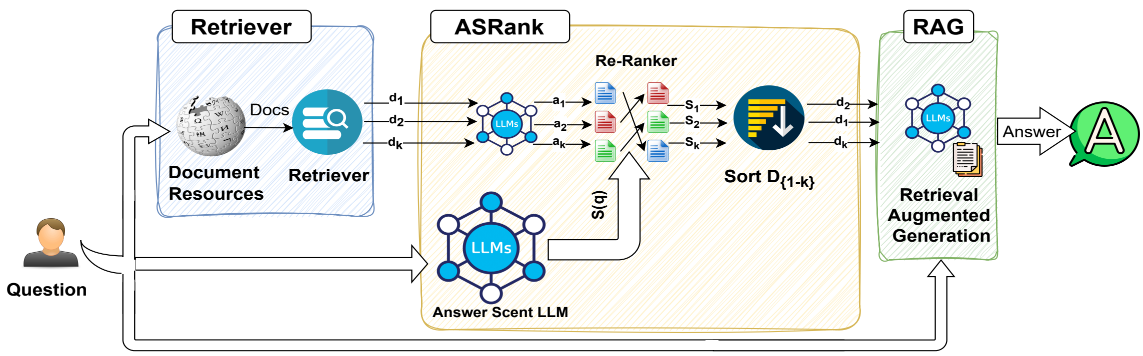

Figure 2: Our ASRank framework, starting with document retrieval, re-ranking using the answer scent from LLMs, and finally passing top-k in the RAG system.

The core of ASRank's method is the calculation of a relevance score for each document, leveraging both the content of the document and its alignment with the inferred answer scent. The score is formulated as:

$$s(\boldsymbol{d}_i) = \sum_{t=1}^{|\boldsymbol{a}|} - \log p(a_t \mid \boldsymbol{a}_{<t}, \boldsymbol{d}_i, \boldsymbol{q}, \mathcal{S}(\boldsymbol{q}); \Theta_2),$$

where $\boldsymbol{d}_i$ represents an individual document within the set of retrieved documents $\mathcal{D}$, $|\boldsymbol{a}|$ denotes the length of the $\boldsymbol{a}$ - an answer generated based on $\boldsymbol{d}_i$ by the rank model, and $\mathcal{S}(\boldsymbol{q})$ represents the answer scent derived from the query $\boldsymbol{q}$. The term $\log p(a_t \mid \boldsymbol{a}_{<t}, \boldsymbol{d}_i, \boldsymbol{q}, \mathcal{S}(\boldsymbol{q}); \Theta_2)$ is the log probability of each token a_t of the answer conditional on the prior tokens $\boldsymbol{a}_{<t}$, the document $\boldsymbol{d}_i$, the query $\boldsymbol{q}$, and the answer scent, parameterized by the model's parameters Θ_2.

To elaborate, the computation of the conditional probabilities can be decomposed as follows:

$$\log p(\boldsymbol{a} \mid \boldsymbol{d}_i, \boldsymbol{q}, \mathcal{S}(\boldsymbol{q})) = \sum_{t=1}^{|\boldsymbol{a}|} \log p(a_t \mid \boldsymbol{a}_{<t}, \boldsymbol{d}_i, \boldsymbol{q}, \mathcal{S}(\boldsymbol{q}); \Theta_2),$$

where $\log p(\boldsymbol{a} \mid \boldsymbol{d}_i, \boldsymbol{q}, \mathcal{S}(\boldsymbol{q}))$ is the log probability of generating the entire answer sequence from the document, the query, and the answer scent. The relevance score, denoted by $s(\boldsymbol{d}_i)$, is reformulated using Bayes' Theorem. This score is represented as follows:

$$s(\boldsymbol{d}_i) \propto \log p(\boldsymbol{a} \mid \boldsymbol{d}_i, \boldsymbol{q}, \mathcal{S}(\boldsymbol{q})) + \log p(\boldsymbol{d}_i \mid \boldsymbol{q}) - \log p(\boldsymbol{a} \mid \boldsymbol{q}),$$

where $\log p(\boldsymbol{a} \mid \boldsymbol{d}_i, \boldsymbol{q}, \mathcal{S}(\boldsymbol{q}))$ represents the log probability of generating the answer $\boldsymbol{a}$ given the document $\boldsymbol{d}_i$, the query $\boldsymbol{q}$, and the inferred answer scent $\mathcal{S}(\boldsymbol{q})$. $\log p(\boldsymbol{d}_i \mid \boldsymbol{q})$ is the log probability that the document $\boldsymbol{d}_i$ is relevant to the query $\boldsymbol{q}$, based on the initial retrieval. $-\log p(\boldsymbol{a} \mid \boldsymbol{q})$ normalization term adjusts for the base likelihood of the answer $\boldsymbol{a}$ being related to the query $\boldsymbol{q}$ across all documents.

The decision to select the most relevant document employs a maximization approach $\hat{i} = \arg\max_{i\in[1,K]} s(\boldsymbol{d}_i)$, indicating the selection of the document $\boldsymbol{d}_i$ that maximizes the relevance score, thereby enhancing the likelihood that the document contains the information necessary to answer the query effectively, aligned with the derived answer scent.

3 Experiment Settings

3.1 Datasets

We utilize several common datasets for our experiments, whose detailed statistics are provided in Appendix A:

Open Domain QA datasets: TriviaQA (Joshi et al., 2017) is a collection of trivia questions sourced from trivia and quiz-league websites. Natural Questions (NQ) (Kwiatkowski et al., 2019) is a question-answering dataset containing 79,168 training examples, 8,757 development examples, and 3,610 test question-answer pairs.WebQuestions (Berant et al., 2013) is a question-answering dataset that was created using Freebase as a knowledge base and which contains 5,810 question-answer pairs.

Entity-centric Questions: Entity Questions (Sciavolino et al., 2021) contains 22K short questions about named entities based on facts from Wikipedia.

Temporal Questions: ArchivalQA (Wang et al., 2022) is a large-scale question answer collection designed specifically for temporal news QA, containing 532,444 question-answer pairs, often on detailed or minor aspects. These pairs are derived from the New York Times Annotated Corpus, which spans news articles published between January 1, 1987, and June 19, 2007. We follow prior work (Wallat et al., 2024) and evaluate ASRANK on the subset of ArchivalQA dataset, which comprises 7,500 questions.

Multi-hop Questions: HotpotQA (Yang et al., 2018) contains 113K crowd-sourced questions constructed in a way that the introduction paragraphs of two Wikipedia articles are required to answer questions (i.e., two hops). We focus on the full-wiki setting, in which two Wikipedia passages are required to answer the questions. We follow prior work (Khalifa et al., 2022) and evaluate ASRANK on the development set, which has 7,405 questions.

3.2 Retrievers

In our re-ranking experiments, we retrieve passages using both unsupervised and supervised retrievers, as detailed below.

Unsupervised Retrievers: BM25 (Robertson and Zaragoza, 2009) is a ranking function used by search engines to estimate the relevance of documents to a given search query. Masked Salient Spans (MSS) (Sachan et al., 2021a) is a dense retriever trained by predicting masked salient spans like named entities with the help of a reader network. Contriever (Izacard et al., 2022) is a framework for pre-training and fine-tuning models for information retrieval using contrastive learning.

Supervised Retrievers: Dense Passage Retrieval (DPR) (Karpukhin et al., 2020b) uses annotated question-context paragraphs and hard negative examples to train a supervised dense retriever. MSS-DPR (Sachan et al., 2021a) further improves DPR performance by first pre-training the dense retriever using MSS followed by DPR-style supervised fine-tuning. A detailed explanation of Unsupervised/Supervised retrievers is given in Appendix D.

3.3 LLM Models

This section overviews the large language models (LLMs) utilized in our experiments. These models are essential for generating the "answer scent" and re-ranking documents based on their inferred relevance to the query.

Answer Scent Models: We leverage a variety of Large Language Models (LLMs), each bringing unique strengths to our re-ranking methodology. The **Llama** models, developed by Meta, are known for their robust performance in dialogue applications, having undergone extensive pre-training and fine-tuning (Touvron et al., 2023). **Mistral** and **Mixtral**, from Mistral AI, push the boundaries of efficiency and computational optimization, employing instruction fine-tuning and a sparse mixture of experts approach respectively (Jiang et al., 2023, 2024). **Gemma**, a product of Google, offers both base and instruction-tuned versions in different sizes, designed for adaptability across various hardware platforms (Team et al., 2024). **GPT**, from OpenAI, is renowned for its general-purpose capabilities, pre-trained on vast data pools to generate semantically rich responses (Brown et al., 2020). Lastly, **Qwen**, by Alibaba Cloud, encapsulates a broad pre-training regime across multiple languages and domains, optimized for long-context interactions, highlighting its scalability and depth in handling complex linguistic tasks (Bai et al., 2023). A detailed explanation of the LLM models is in Appendix C.

Rank Model: In our experiments, we specifically utilize the T5-Base model, a variation of the original T5 architecture (Raffel et al., 2020) adapted for language modelling tasks. This model, part of the T5 series, features encoder and decoder transformers pre-trained to improve their ability to handle input text sequences.

3.4 Experimental Setup

All re-ranking experiments were conducted on a high-performance computing cluster using NVIDIA A100 48GB GPUs, with specific experiments outlined in Section §6 run on NVIDIA A40 GPUs. We evaluated our method on five retrievers (BM5, MSS, MSS-DPR, DPR, and Contriver) for retrieving 1,000 passages, same as in (Sachan et al., 2022), while temporal questions (ArchivalQA) were evaluated with two additional retrievers, Ance (Xiong et al., 2020) and Rocket (Qu et al., 2020). Additionally, HotspotQA question scenarios followed the dataset and retrieval configurations as described in (Khalifa et al., 2022), ensuring a comprehensive assessment across various question types and retrieval technologies. To assess the performance of ASRANK, we use top-K retrieval (Sachan et al., 2022) accuracy. For RAG, we evaluate how accurately and completely the model answers questions using exact match, recall, and F1 scores. More details about the used framework implementation and metrics are in Appendix B and E.

Retriever	NQ				TriviaQA				WebQ			
	Top-1	Top-5	Top-10	Avg	Top-1	Top-5	Top-10	Avg	Top-1	Top-5	Top-10	Avg
Unsupervised Retrievers												
MSS	19.2	41.2	51.2	37.2	30.7	52.6	60.5	47.9	11.6	29.0	39.1	26.6
MSS + UPR	38.7	64.8	72.2	58.6	57.2	75.5	78.9	70.5	29.9	57.4	65.0	50.7
MSS + ASRANK †	45.2	**64.7**	**70.6**	60.1	65.3	77.2	79.8	74.1	42.5	61.3	67.7	57.1
MSS + ASRANK ‡	**46.5**	64.4	69.8	**60.2**	**66.3**	**77.6**	**80.1**	**74.6**	**45.0**	**63.6**	**68.8**	**59.1**
BM25	22.1	43.7	54.4	40.1	46.3	66.2	71.7	61.4	18.8	41.8	52.1	37.6
BM25 + UPR	35.4	63.4	70.2	56.3	55.7	76.5	80.2	70.8	30.1	57.3	66.5	51.3
BM25 + ASRANK †	46.2	65.3	**72.3**	61.2	67.2	**77.9**	**80.7**	75.2	44.8	63.7	68.7	**59.0**
BM25 + ASRANK ‡	**47.3**	**65.6**	71.4	**61.4**	**67.3**	**77.9**	**80.7**	**75.3**	**45.4**	**62.9**	**68.9**	**59.0**
Contriever	22.1	47.2	58.7	42.7	34.1	59.4	68.0	53.8	19.9	43.4	56.3	39.9
Contriever + UPR	36.4	64.6	72.4	57.8	56.7	77.0	80.2	71.3	30.0	58.5	68.2	52.2
Contriever + ASRANK §	41.5	61.3	68.4	57.0	57.9	72.8	76.8	69.1	42.9	62.7	69.8	58.4
Contriever + ASRANK ‡	**48.0**	**66.6**	**72.5**	**62.3**	**66.8**	**78.9**	**81.4**	**76.0**	**46.8**	**64.8**	**70.8**	**60.8**
Supervised Retrievers												
DPR	48.6	68.7	74.5	63.9	57.4	72.4	76.5	68.7	44.8	65.0	70.6	60.1
DPR + UPR	42.5	70.6	78.1	63.8	61.3	78.7	81.9	74.0	34.9	63.6	71.7	56.7
DPR + ASRANK †	50.2	69.9	**76.1**	65.3	68.8	79.8	**82.4**	77.0	48.2	**68.1**	73.2	63.1
DPR + ASRANK ‡	**51.3**	**70.6**	76.0	**65.9**	**69.9**	**79.8**	82.1	**77.3**	**49.3**	67.3	**73.4**	**63.3**
MSS-DPR	50.1	**71.8**	**77.4**	**66.5**	61.6	75.2	79.1	71.9	44.2	65.0	71.6	60.3
MSS-DPR + UPR	41.4	69.8	77.9	63.0	60.5	78.9	82.5	74.0	31.8	61.6	70.3	54.5
MSS-DPR + ASRANK †	48.8	69.3	76.1	64.7	69.4	80.4	82.9	77.5	47.7	**67.0**	**73.0**	62.5
MSS-DPR + ASRANK ‡	**50.6**	69.3	75.2	65.0	**69.9**	**80.5**	**82.9**	**77.7**	**49.7**	66.6	72.6	**62.9**

Table 1: Top-1, 5, 10 retrieval accuracy of re-ranking methods including ASRANK and baseline models on the NQ, TriviaQA and WebQ Datasets. † refers to Llama 70B, ‡ refers to GPT175B. For a comparison between LLama 7b vs UPR see table 9 in Appendix G.

4 Experiment Results

In this section, we evaluate ASRANK on a variety of question-answering tasks, leveraging several datasets to assess its performance. The datasets employed cover different QA challenges, ranging from open domain to entity-centric, temporal, and multi-hop questions. The primary objective is to evaluate ASRANK's capability to rank the Top-1, 5, 10 retrieved passages. For this purpose, an initial retrieval of 1,000 passages per question is conducted for reranking.

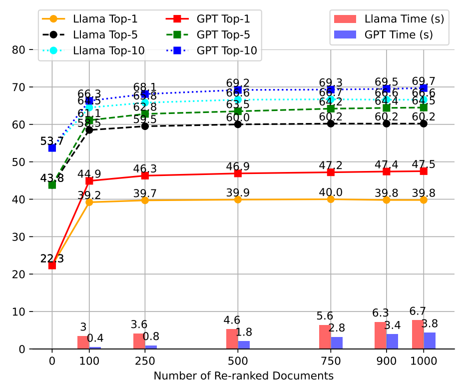

Figure 3: Effect of the number of passage candidates on the accuracy of Top-1, 5, 10 results, and latency when re-ranked with LLama 8B and GPT 175. The results were computed on the NQ development set using BM25 retrieved passages.

4.1 ODQA Re-ranking

In this section, we focused on evaluating ASRANK across several ODQA datasets (NQ, TriviaQA, and WebQ). The results, as summarized in Table 1, show improvements in retrieval Top-K accuracy. ASRANK enhances the retrieval of Top-K results across various settings, often outperforming the UPR model. For instance, when combined with the MSS retriever on the NQ dataset, ASRANK † (Llama 70B) increases the relevance of Top-1 result to 45.2%, a notable improvement over UPR's increase to 38.7%. Similarly, the TriviaQA dataset, ASRANK with BM25 achieves a Top-1 accuracy of 47.3%, surpassing BM25 + UPR's performance of 35.4%. Also, the combination of ASRANK with the MSS retriever results in a remarkable uplift in Top-1 accuracy for NQ, from an initial 19.2% to 46.5%. Similarly, on TriviaQA, the integration of ASRANK with the BM25 retriever increases the accuracy of Top-1 from 22.1% to 47.3%.

The performance improvement of the ASRANK is focused on using zero-shot answer scent generation with a cross-attention mechanism within its re-ranking framework. ASRANK uses the advanced capabilities of LLMs to interpret and generate answer scents. The answer scent is not static but dynamically interacts with the passage tokens through a cross-attention mechanism employed in the model's architecture. Each token of the generated answer scent considers every token in the passage, enabling a deeper and more contextual understanding before determining the relevance of each passage. By focusing on the semantic and contextual alignment between the question and the document, ASRANK improves the retrieval and ensures that the top-ranked documents are relevant to the information needs.

4.2 Impact of Answer Scent on Re-Ranking and Latency Implications

Answer scents improve the alignment of retrieved documents with the question. The process is captured through the computation of log-likelihood of each document given the question $\log p(\boldsymbol{d} \mid \boldsymbol{q}; \Theta) = \frac{1}{|\boldsymbol{d}|} \sum_t \log p(d_t \mid \boldsymbol{d}_{<t})$, where $\boldsymbol{d}$ denotes the document tokens, $\boldsymbol{q}$ the question, Θ the LLM parameters, and $|\boldsymbol{d}|$ the number of tokens in the document.

Re-ranking with Answer Scent has shown improvements in Top-K on the NQ development set. The Top-1 accuracy increases from 22.3% at 100 documents to around 39.8% at 750 documents (see Fig. 3). The ASRANK significantly reduces latency challenges, especially as the number of re-ranked documents grows. For example, re-ranking 1,000 documents takes up to 6.7 seconds with Llama models and 3.8 seconds with GPT models, compared to 11.6 seconds with UPR. This means ASRANK cuts latency by nearly 42% compared to UPR, as shown in Figure 3, and Table 8 in Appendix F.

5 Ablation Studies

5.1 Evaluation on NQ development

In this section, we compare our approach with UPR using different model sizes (T0-3B, T0-11B) (Sanh et al., 2021) to assess the efficacy in the context of the NQ development set. This comparison highlights the significant advantages offered by the ASRANK, across different retrievers like BM25, MSS, and DPR. The results are shown in Table 2. ASRANK enhances retrieval performance across all Top-1, 5, 10, 20 results. Notably, after re-ranking with ASRANK using the Llama 70B configuration, the accuracy of Top-1 for the MSS-DPR combination reaches 48.1%, which is an improvement over its performance with UPR, where the Top-1 achieves 39.7%. Similarly, the Top-1 for DPR alone ascends to 50.4% with GPT3.5, surpassing the 41.1% recorded with UPR.

5.2 Evaluation on Diverse Question Answering Datasets

The ablation studies were conducted across three distinct datasets—Entity Questions, HotpotQA, and ArchivalQA. As summarized in Tables 3, 5,

Retriever	NQ (dev)				
	Top-1	Top-5	Top-10	Top-20	Avg
BM25	22.3	43.8	53.6	62.3	45.5
MSS	17.7	38.6	48.7	57.4	40.6
MSS+BM25	17.6	38.7	48.8	57.8	40.7
Contriever	19.6	45.4	55.8	64.9	46.4
DPR	47.8	67.3	73.0	77.4	6.4
MSS-DPR	**48.9**	**69.9**	**75.7**	**80.4**	**68.7**
After Re-ranking with UPR					
BM25+MSS+T5-lm-adapt (3B)	29.7	59.9	-	76.9	55.5
BM25+MSS+T5-lm-adapt (11B)	32.1	62.3	-	78.5	57.6
BM25+MSS+ T0-3B	36.7	64.9	-	79.1	60.2
BM25+MSS+ T0-11B	37.4	64.9	-	79.1	60.5
DPR +T0-3B	**41.1**	**69.5**	**77.0**	81.9	**67.4**
MSS+T0-3B	36.6	62.9	70.8	75.7	61.5
MSS+DPR+T0-3B	39.7	68.6	76.5	**82.0**	66.7
After Re-ranking with GPT3.5					
MSS	46.2	63.5	69.1	73.2	63.0
BM25	47.5	64.5	69.7	74.3	64.0
Contriever	47.7	65.5	71.2	76.2	65.2
BM25+MSS	47.9	65.5	71.2	76.4	65.3
MSS-DPR	50.1	68.9	74.8	**79.8**	68.4
DPR	**50.4**	**68.9**	**74.9**	79.4	**68.4**
After Re-ranking with LLama 70B					
MSS	44.9	63.7	69.4	73.9	62.9
BM25	44.8	64.1	69.9	75.0	63.5
Contriever	45.7	65.4	71.5	76.2	64.7
BM25+MSS	45.4	65.4	71.0	76.6	64.6
MSS-DPR	48.1	**68.6**	**74.6**	**79.8**	**67.8**
DPR	**48.2**	67.9	73.8	78.6	67.1

Table 2: Performance comparison of different retrievers on the NQ development set, illustrating the significant improvement provided by ASRANK over methods like UPR.

and 4, ASRANK enhances Top-1, 5, and 10 retrieval accuracies, across different retrievers.

Retriever	Entity Questions			
	Top-1	Top-5	Top-10	Avg
Baselines				
MSS	19.3	35.9	43.1	32.8
DPR	25.3	39.5	45.3	36.7
MSS-DPR	**30.3**	47.7	54.1	**44.0**
Contriever	27.1	**48.0**	**55.7**	43.6
After Re-ranking with Llama 70b				
MSS†	44.5	58.3	62.7	55.2
DPR†	41.7	53.8	58.2	51.2
MSS-DPR†	46.4	60.1	64.5	57.0
Contriever†	**46.6**	**61.1**	**65.9**	**57.9**
After Re-ranking with GPT3.5				
MSS	46.6	60.5	64.5	57.2
DPR	43.6	55.6	59.4	52.9
MSS-DPR	48.4	62.1	66.2	58.9
Contriever	**48.9**	**63.2**	**67.5**	**59.8**

Table 3: Top-1, 5, 10 retrieval accuracy for the Entity Questions dataset, comparing baseline retrievers with results after re-ranking using Llama 70b and GPT3.5 models.

The Entity Questions dataset, when used Llama 70B and GPT3.5 boosts performance, achieving Top-1 retrieval accuracy up to 48.9%, which indicates an increase of over 25% compared to baselines. For the HotpotQA dataset, which requires reasoning over multiple documents, ASRANK achieves substantial enhancements in

Retriever	ArchivalQA Questions			
	Top-1	Top-5	Top-10	Avg
Baselines				
Contriever	1	3.2	5.0	3.0
BM25	**18.2**	**32.3**	**38.6**	**29.7**
DPR	17.0	30.1	36.8	27.9
Rocket	15.7	29.3	35.6	26.9
ANCE	18.0	31.8	37.7	29.2
After Re-ranking with Llama 70b				
Contriever	3.9	8.1	10.4	7.4
BM25	26.2	37.3	42.4	35.3
DPR	**27.5**	**38.2**	**43.3**	**36.3**
Rocket	26.2	37.4	42.4	35.3
ANCE	27.3	**38.2**	**43.3**	**36.3**
After Re-ranking with GPT3.5				
Contriever	4.2	8.7	10.9	7.9
BM25	27.6	37.7	42.4	35.9
DPR	27.7	38.5	43.5	36.6
Rocket	26.5	37.9	42.7	35.7
ANCE	**28.1**	**38.1**	**42.9**	**36.3**

Table 4: Top-1, 5, 10 retrieval accuracy for the ArchivalQA dataset, comparing baseline retrievers with results after re-ranking using Llama 70b and GPT3.5 models.

Top-1, surpassing fully-supervised baselines like DPR (Karpukhin et al., 2020a) and DrKit (Dhingra et al., 2020), MDR (Xiong et al., 2021), PathRetriever (Asai et al., 2020) when combined with TF-IDF. This shows ASRANK's strength in multi-hop question answering, supporting complex inference tasks across linked data points. Notably, ASRANK combined with DPR achieves a Top-1 accuracy of 42.6%, which not only surpasses the fully-supervised baselines such as DPR at 18.5% and DrKit at 38.3%, but also outperforms unsupervised models like PromptRank-GPT2-XL and PromptRank-T5-XL (Khalifa et al., 2022), which score 36.6% and 42.8%, respectively. On the ArchivalQA dataset, which contains temporal questions, ASRANK shows also good improvements. After re-ranking with Llama 70B and GPT3.5, the model significantly boosts Top-1 accuracies across different retrievers, demonstrating its effectiveness in extracting temporally relevant information. Specifically, after re-ranking with Llama 70B, BM25 improves from 18.2% to 26.2% in Top-1, DPR from 17.0% to 27.5%, and ANCE from 18.0% to 27.3%. When using GPT3.5, BM25 improves further to 27.6% in Top-1, DPR reaches 27.7%, and ANCE advances to 28.1%.

5.3 Impact of Answer Scent LLM

In this section, we evaluate the impact of different LLMs on Top-1, 5, 10, 20 by re-ranking the Top-1000 passages from the NQ development set. The performance of these LLMs on the NQ development set is detailed in Table 6. The baseline retrieval using BM25 achieves a Top-1 of 22.3%. However, with LLMs like Llama-2 and GPT3.5, there's an increase in all Top-K. For instance, Llama-2 70B improves the Top-1 accuracy to 45.3%, and GPT3.5 pushes it further to 46.3%. As the model size increases from 7B to over 70B, there's a performance improvement. The Mixtral model achieves a Top-1 of 42.5%. In Appendix I, we show random examples from NQ dev and WebQA after and before re-ranking.

6 RAG for Open-Domain Question Answering

Method In the Retrieval-Augmented Generation (RAG) framework, we employ a large language model (LLM), leveraging its capacity to utilize retrieved documents dynamically for generating responses. The RAG method combines the robust retrieval capabilities of DPR with the generative models, thereby enabling understanding and response generation based on the context provided by the retrieved documents. The RAG model is formulated as $p(a \mid q, D) = \sum_{d \in D} p(d \mid q) \cdot p(a \mid q, d)$, where a represents the answer, q the query, and D is the set of retrieved documents relevant to q. The term $p(d \mid q)$ denotes the document's retrieval probability, and $p(a \mid q, d)$ represents the probability of generating answer a given the query q and document d.

Results We evaluated the RAG method on NQ, TriviaQA, and WebQA revealing significant performance gains as shown in Figure 4. For instance, before applying our ASRANK re-ranking strategy, the BM25+LLama7B achieves a baseline EM of 16.0% on NQ. After re-ranking with ASRANK, the EM increased to 24.8%. We show a detailed comparison between the baselines (BM25, DPR, MSS, Contriever, MSS-DPR), UPR, and ASRANK in Table 10 (Appendix H).

7 Related Work

Recent developments in the field of information retrieval have increasingly focused on the integration of LLMs for enhancing retrieval and reranking mechanisms. LLMs have demonstrated a substantial impact in retrieval tasks, largely due to their deep generative capabilities. Innovative approaches like InPars (Bonifacio et al., 2022; Jeronymo et al.,

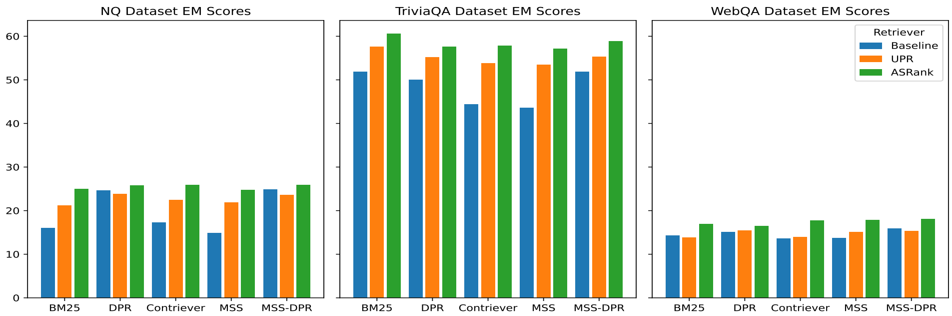

Figure 4: Comparison of Exact Match (EM) scores across three datasets (NQ, TriviaQA, and WebQA) for various retrieval models.

Retriever	HotpotQA			
	# Ex.	top-2	top-10	top-20
Fully-supervised Baselines				
DPR	-	18.5	37.2	47.1
DPR+ASRank	-	42.6	68.8	79.2
DrKit	~90K	38.3	67.2	71.0
MDR	~90K	65.9	77.5	**80.2**
PathRetriever	~90K	**66.4**	**77.8**	78.7
Unsupervised Baselines				
TF-IDF	–	9.9	27.6	35.0
TF-IDF + BM25	–	19.1	54.7	61.8
PromptRank-GPT2-XL	-	36.6	60.5	65.9
PromptRank-T5-XL	-	42.8	68.9	74.1
TF-IDF+ASRank	-	36.9	61.1	72.5
TF-IDF+ASRank †	-	**45.1**	**69.1**	**78.9**

Table 5: Top-2, 10, 20 retrieval performance on HotpotQA comparing ASRank to baselines.

Retriever	#Parameters	NQ (dev)				
		Top-1	Top-5	Top-10	Top-20	Avg
BM25	-	22.3	43.8	53.7	62.3	
Gemma	7B	21.2	37.7	45.9	54.2	39.8
Mistral	7B	27.9	46.3	54.8	62.3	47.8
Qwen1.5	7B	30.3	50.4	58.6	66.2	51.4
Llama-2	7B	39.2	58.6	65.8	71.4	58.8
Llama3	8B	39.8	60.2	66.6	71.9	59.6
Qwen1.5	14B	34.9	54.4	62.7	69.1	55.3
Qwen1.5	32B	39.9	60.3	67.2	72.9	60.1
Mixtral	8x7B	42.5	61.9	68.2	73.0	61.4
Llama3	70B	44.8	64.1	69.9	**75.0**	**63.5**
Qwen1.5	72B	43.2	62.6	68.9	73.9	62.2
Llama-2	70B	45.3	64.0	**69.9**	74.4	63.4
Qwen1.5	110B	44.0	63.3	69.8	74.4	62.9
GPT3.5	175B	**46.3**	**63.6**	69.1	73.8	63.2

Table 6: Performance metrics of different LLMs utilizing the answer scent concept for document retrieval across Top-1, 5, 10, and 20 rankings on the NQ (dev) dataset.

2023) and Promptagator (Dai et al., 2022) have explored the generation of synthetic datasets to improve domain-specific retrieval performance. Concurrently, models like SGPT (Muennighoff, 2022) and UPR (Sachan et al., 2022) have showcased the direct utility of GPT-based and T5 models as effective rankers in bi-encoder architectures, with UPR utilizing query likelihood for scoring. Notably, PRP (Qin et al., 2023) and Ma et al. (2023a) have demonstrated that fine-tuning LLMs like LLaMA enhances retrieval performance beyond smaller models, positioning LLMs as powerful tools for reranking tasks. The integration of unsupervised and supervised retrieval techniques such as BM25 (Robertson et al., 2009), MSS (Sachan et al., 2021a), Contriever (Izacard et al., 2022), and DPR (Karpukhin et al., 2020b) has been pivotal. These methods, including enhancements like MSS-DPR (Sachan et al., 2021b), leverage dense and sparse retrieval techniques to enhance the initial retrieval stages, subsequently improved through reranking. Moreover, newer supervised methods like ColBERT (Khattab and Zaharia, 2020) and SPLADE (Formal et al., 2021) further refine retrieval accuracy. A growing body of work has investigated the role of LLMs in reranking by prompting them to reorder documents based on relevance, with methods like RankVicuna (Pradeep et al., 2023) and LRL (Ma et al., 2023b) demonstrating significant advancements. These studies illustrate that LLMs with prompts can handle reranking tasks efficiently.

8 Conclusion

In this paper, we introduced ASRank, a novel zero-shot re-ranking method that uses the concept of answer scent to enhance document retrieval for open-domain question answering. Our experiments across diverse datasets demonstrate that the ASRank outperforms both unsupervised and supervised baselines. ASRank not only enhances the top-1 accuracy but also shows substantial gains in Top-5, and 10 retrieval metrics, which makes it a valuable tool for improving the efficiency and efficacy of question-answering systems.

Limitations

While ASRank demonstrates significant improvements in document re-ranking with the incorporation of answer scent, there are several limitations that warrant discussion:

1. The computational cost associated with ASRank increases with the number of documents due to the need to compute the answer scent with the answer generated from each document.

2. The effectiveness and consistency of ASRank are contingent upon the specific pre-trained language models used for generating the answer scent. Variations in these models, due to different training data or updates, can introduce biases and affect the stability of the re-ranking outcomes.

3. ASRank's performance heavily depends on the quality of the initial retrieval phase.

Ethical Considerations and Licensing

Our research utilizes the GPT models, which is available under the OpenAI License and Apache-2.0 license, and the Llama model, distributed under the Llama 2 Community License Agreement provided by Meta. We ensure all use cases are compliant with these licenses. Additionally, the datasets employed are sourced from repositories permitting academic use. We are releasing the artifacts developed during our study under the MIT license to facilitate ease of use and modification by the research community. We have ensured that all data handling, model training, and dissemination of results are conducted in accordance with ethical guidelines and legal stipulations associated with each used artifact.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2387–2392.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

Bhuwan Dhingra, Manzil Zaheer, Vidhisha Balachandran, Graham Neubig, Ruslan Salakhutdinov, and William W. Cohen. 2020. Differentiable reasoning over a virtual knowledge base. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *arXiv preprint arXiv:2301.01820*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020a. 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.

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

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2022. Few-shot reranking for multi-hop qa via language model prompting. *arXiv preprint arXiv:2205.12650*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

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.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023a. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023b. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

David Maxwell and Leif Azzopardi. 2018. Information scent, searching and stopping: Modelling serp level stopping behaviour. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 210–222. Springer.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, et al. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv preprint arXiv:2306.17563*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*.

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

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. *arXiv preprint arXiv:2204.07496*.

Devendra Singh Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L Hamilton, and Bryan Catanzaro. 2021a. End-to-end training of neural retrievers for open-domain question answering. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Devendra Singh Sachan, Siva Reddy, William L. Hamilton, Chris Dyer, and Dani Yogatama. 2021b. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Advances in Neural Information Processing Systems*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *Preprint*, arXiv:2110.08207.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Spurthi Setty, Katherine Jijo, Eden Chung, and Natan Vidra. 2024. Improving retrieval for rag based question answering models on financial documents. *arXiv preprint arXiv:2404.07221*.

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*.

Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. Temporal blind spots in large language models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 683–692.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022. Archivalqa: a large-scale benchmark dataset for open-domain question answering over historical news collections. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3025–3035.

L Winerman. 2012. Tracking the scent of information. *Monitor on psychology*, 43(3):44–47.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick S. H. Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021. Answering complex open-domain questions with multi-hop dense retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.

A Datasets' Details

In this section, we present a table that details the statistics of the datasets utilized in our study. These tables include comprehensive data such as sample sizes, feature counts, and other relevant metrics, providing a clear overview of the datasets' composition and scope. This information is crucial for understanding the context and the robustness of our analysis, enabling a deeper insight into the methodologies applied and the validity of our results. We utilize several well-known datasets for our experiments:

Open Domain QA datasets: TriviaQA (Joshi et al., 2017) is a collection of trivia questions sourced from trivia and quiz-league websites. The dataset includes 78,785 examples in the training set, 8,837 examples in the development set, and 11,313 examples in the test set.

Natural Questions (NQ) (Kwiatkowski et al., 2019) is a question-answering dataset containing 79,168 training examples, 8,757 development examples, and 3,610 test question-answer pairs. Each example is comprised of a question obtained from a Google query and a corresponding Wikipedia page.

WebQuestions (Berant et al., 2013) is a question-answering dataset that was created using Freebase as a knowledge base and that contains 5,810 question-answer pairs. It was constructed by crawling questions through the Google Suggest API and then obtaining corresponding answers using Amazon Mechanical Turk.

Entity-centric Questions: Entity Questions (Sciavolino et al., 2021) contains 22K short questions about named entities based on facts from Wikipedia. Previous work using this dataset has shown that dense retrievers struggle to retrieve relevant passages, while sparse approaches like BM25 tend to be more successful.

Temporal Questions: ArchivalQA (Wang et al., 2022) is a large-scale question answer collection designed specifically for temporal news QA, containing 532,444 question-answer pairs, often on detailed and minor aspects. These pairs are derived from the New York Times Annotated Corpus, which spans news articles published between January 1, 1987, and June 19, 2007. The dataset-constructing framework with automatic question generation and filtering steps ensures high-quality and non-ambiguous questions.

Multi-hop Questions: HotpotQA (Yang et al., 2018) is a question answering dataset created from the English Wikipedia. It contains about 113K crowd-sourced questions that are constructed to require the introduction paragraphs of two Wikipedia articles to answer. We focus on the full wiki setting, in which two Wikipedia passages are required to answer the questions. Since the gold passages for the test set are not available, we follow prior work and evaluate ASRANK on the development set, which has 7,405 questions.

B Evaluation Metrics

To assess the performance of the ASRANK, we use top-K retrieval accuracy and several other metrics for the RAG. Top-K retrieval accuracy measures whether the correct answer appears within the top-K retrieved passages, calculated as:

$$\text{TOP@}k = \frac{\sum(\text{any}(\text{Correct@}k))}{\text{Total Data}}$$

Given that LLMs tend to generate verbose answers, many standard QA metrics are not well suited to evaluate the answer quality; the Exact Match will always be less given the occurrence of other non-ground-truth tokens, and the F1 score will be penalized by other, potentially helpful tokens. Therefore, we utilize a set of model-agnostic metrics (i.e., token recall and answer string containment).

Table 7: Statistics of TriviaQA, NQ and WebQ datasets.

Dataset	Train	Dev	Test
TriviaQA	78,785	8,837	11,313
NQ	79,168	8,757	3,610
WebQA	3,417	361	2,032
Entity Questions	-	-	22000
HotpotQA	90,564	7,405	-
ArchivalQA	-	-	7500

C Answer Scent Models

Llama (Touvron et al., 2023) is a part of the Llama 2 family of Large Language Models (LLMs) developed by Meta. These models are pre-trained and fine-tuned generative text models, specifically optimized for dialogue use cases.

Mistral (Jiang et al., 2023), a model released by Mistral AI, is renowned for its power and efficiency. It surpasses the Llama 2 13B on all benchmarks. The model leverages instruction fine-tuning, where the prompt should be surrounded by [INST] and [/INST] tokens.

Mixtral (Jiang et al., 2024), another innovation from Mistral AI, is a trained generative Sparse Mixture of Experts that outperforms the Llama 2 70B model on most benchmarks. The model leverages up to 45B parameters but only uses about 12B during inference, leading to better inference throughput at the cost of more vRAM.

Gemma (Team et al., 2024) is a family of Large Language Models (LLMs) developed by Google. It is based on the Gemini family of LLMs. Gemma comes in two sizes: 2B and 7B parameters, each with base instruction-tuned versions. All the variants can be run on various types of consumer hardware, even without quantization, and have a context length of 8K tokens.

GPT (Brown et al., 2020) is a series of Generative Pre-trained Transformers developed by OpenAI. These models are pre-trained on massive amounts of data, such as books and web pages, to generate contextually relevant and semantically coherent language12. The most recent of these, GPT-4. GPT models are general-purpose language prediction models.

Qwen (Bai et al., 2023) is a comprehensive language model series developed by Alibaba Cloud. It includes Qwen, the base pretrained language models, and Qwen-Chat, the chat models fine-tuned with human alignment techniques. These models have been pre-trained with data from a wide range of domains and languages, supporting the context length of 32768 tokens.

D Retrievers

In our re-ranking experiments, we retrieve passages using both unsupervised and supervised retrievers, as detailed below.

Unsupervised Retrievers BM25 (Robertson and Zaragoza, 2009) is a ranking function used by search engines to estimate the relevance of documents to a given search query. It is based on the probabilistic retrieval framework and uses term-frequency (TF) and inverse document frequency (IDF) of the keywords present in the question and passage. Masked Salient Spans (MSS) (Sachan et al., 2021a) is a dense retriever trained by predicting masked salient spans like named entities with the help of a reader network. The objective function for training the MSS retriever can be represented as:

$$\mathcal{L}_{MSS} = -\mathbb{E}_{(q,d^+,d^-)\sim D}[\log p(d^+|q) + \log(1 - p(d^-|q))]$$

where D is the dataset, (q, d^+, d^-) is a triplet of the question, positive document, and negative document, and $p(d|q)$ is the probability of a document d being relevant to a question q.

Contriever is a framework for pre-training and fine-tuning models for information retrieval using contrastive learning. The objective function for training the Contriever model is:

$$\mathcal{L}_{Contriever} = -\mathbb{E}_{(q,d^+,d^-)\sim D}[\log \sigma(s(q, d^+)) + \log(1 - \sigma(s(q, d^-)))]$$

where $s(q, d)$ is the similarity score between question q and document d, and σ is the sigmoid function (Izacard et al., 2022).

Supervised Retrievers Dense Passage Retrieval (DPR) (Karpukhin et al., 2020b) uses annotated question-context paragraphs and hard negative examples to train a supervised dense retriever. The objective function for training the DPR model is:

$$\mathcal{L}_{DPR} = -\mathbb{E}_{(q,d^+,d^-)\sim D}[\log \sigma(s(q, d^+)) + \log(1 - \sigma(s(q, d^-)))]$$

where $s(q, d)$ is the similarity score between question q and document d, and σ is the sigmoid function.

MSS-DPR (Sachan et al., 2021a) is an approach that further improves DPR performance by first pre-training the dense retriever using MSS followed by DPR-style supervised fine-tuning. The objective function for training the MSS-DPR model is:

$$\mathcal{L}_{MSS-DPR} = \alpha\mathcal{L}_{MSS} + (1 - \alpha)\mathcal{L}_{DPR}$$

where α is a hyperparameter that controls the trade-off between the MSS and DPR losses.

E Implementation Framework

Our implementation of ASRANK utilizes the PyTorch (Paszke et al., 2019) framework alongside the transformers (Wolf et al., 2019) library from Hugging Face to handle the computational demands of our document re-ranking tasks.

F Impact of Passage Number on Retrieval Accuracy and Latency

In this section, we analyze the relationship between the number of passages re-ranked and both retrieval accuracy and latency. This study highlights how the ASRANK performs as we increase the number of passage candidates, focusing on Top-K retrieval accuracy and the time taken per query. We conducted experiments using the NQ development set to evaluate the performance of ASRANK with different quantities of retrieved passages. The passages were

Retriever / Re-Ranker	#Document	NQ (dev)					Time/Question
		Top-1	Top-5	Top-10	Top-20	Top-100	
BM25	-	22.3	43.8	53.7	62.3	76.0	-
Llama3 8B	100	39.2	58.5	64.5	69.8	76.0	3s
Llama3 8B	200	39.6	59.4	65.7	70.9	78.5	3.4s
Llama3 8B	250	39.7	59.5	65.8	71.2	79.1	3.6s
Llama3 8B	500	39.9	60.0	66.6	71.9	80.2	4.6s
Llama3 8B	750	40.0	60.2	66.7	71.9	80.74	5.6s
Llama3 8B	900	39.8	60.2	66.6	72.0	80.9	6.3s
Llama3 8B	1000	39.8	60.2	66.6	71.9	80.9	6.7
GPT 175B	100	44.9	61.1	66.3	70.8	76.0	0.4s
GPT 175B	250	46.3	62.8	68.1	72.4	79.4	0.8s
GPT 175B	500	46.9	63.5	69.2	73.6	80.8	1.8s
GPT 175B	750	47.2	64.2	69.3	74.2	81.6	2.8s
GPT 175B	900	47.4	64.4	69.5	74.3	81.8	3.4s
GPT 175B	1000	47.5	64.5	69.7	74.3	81.9	3.8s

Table 8: Impact of the Number of Passage Candidates on Top-1, Top-5, Top-10 Retrieval Accuracy, and Latency per Query.

retrieved using BM25 and re-ranked using LLaMA (8B) and GPT (175B) models. We varied the number of passages from 100 to 1000 to observe the impact on Top-K accuracy and latency. The results of these experiments are presented in Table 8. The table illustrates how increasing the number of re-ranked passages affects the Top-1, 5, 10 retrieval metrics, and the latency per query.

Retrieval Accuracy: The Top-1 accuracy significantly improves as the number of re-ranked passages increases. For example, using Llama3 8B, Top-1 accuracy increases from 39.2% with 100 passages to 40.0% with 750 passages. Similarly, GPT 175B shows an increase in Top-1 accuracy from 44.9% with 100 passages to 47.5% with 1000 passages.

Latency: As expected, the latency per query increases with the number of passages. With Llama3 8B, the latency grows from 3 seconds for 100 passages to 6.7 seconds for 1000 passages. GPT 175B, while providing better accuracy, also shows an increase in latency, from 0.4 seconds for 100 passages to 3.8 seconds for 1000 passages.

G Comparative Analysis of LLama 7B and UPR for Document Re-Ranking

In this section, we present a comparison between the performance of ASRANK utilizing the LLama 7B model and the UPR method. This analysis is aimed at understanding how ASRANK, enhanced with the capabilities of LLama 7B, measures up against UPR in terms of improving retrieval accuracy across various question-answering datasets.

We evaluated both LLama 7B with ASRANK and UPR across three major datasets: NQ, TriviaQA, and WebQ. The goal was to assess the improvements in retrieval accuracy, specifically focusing on Top-1, Top-5, and Top-10 metrics. The retrieval setups included unsupervised and supervised retrievers. The detailed results are summarized in Table 9. The analysis highlights the performance of the two methods under different retrievers, providing insights into their effectiveness across varying retrieval conditions.

Performance across Datasets: Both methods improve retrieval accuracy across all datasets. However, ASRANK with LLama 7B consistently achieves a higher Top-1 metric compared to UPR, suggesting that the inclusion of the answer scent concept might be more effective at distinguishing the most relevant documents at the top of the retrieval list.

Influence of Retrieval Method: When combined with MSS, ASRANK with LLama 7B surpasses UPR in Top-1 retrieval accuracy by a notable margin (e.g., 41.3% vs. 38.7% on NQ). This indicates that ASRANK's approach to utilizing deep contextual embeddings effectively captures nuances that improve the alignment between the query and retrieved documents.

H RAG

In the realm of Retrieval-Augmented Generation (RAG), our study delves into the effects of utilizing LLaMA 7B and LLaMA 13B models, along with varying the number of documents considered in the re-ranking process. Our examination reveals differences in performance across two scenarios: using either one or two documents during the re-ranking phase.

Starting with the LLaMA 7B model, we observed that increasing the number of documents from one to two generally improves the recall and contextual understanding of the model, which is critical in generating accurate responses. For instance, when using the MSS-DPR retriever with LLaMA 7B, the exact match (EM) score sees a slight improvement from 24.3% with one document to 24.9% with two documents. This pattern is consistent across other retrievers like BM25 and Contriever, suggesting that the additional context from a second document helps the model refine its answers.

Switching to the LLaMA 13B model, which offers more capacity and potentially finer understanding due to its larger size. For example, when using the BM25 retriever with LLaMA 13B, the

Retriever	NQ				TriviaQA				WebQ			
	Top-1	Top-5	Top-10	Avg	Top-1	Top-5	Top-10	Avg	Top-1	Top-5	Top-10	Avg
Unsupervised Retrievers												
MSS	19.2	41.2	51.2	37.2	30.7	52.6	60.5	47.9	11.6	29.0	39.1	26.6
MSS + UPR	38.7	**64.8**	**72.2**	**58.6**	57.2	**75.5**	**78.9**	**70.5**	29.9	57.4	65.0	50.7
MSS + ASRank §	**41.3**	60.3	67.2	56.2	**58.5**	71.8	75.6	68.6	**40.1**	**59.9**	**66.6**	**55.5**
BM25	22.1	43.7	54.4	40.1	46.3	66.2	71.7	61.4	18.8	41.8	52.1	37.6
BM25 + UPR	35.4	**63.4**	**70.2**	56.3	55.7	**76.5**	**80.2**	**70.8**	30.1	57.3	66.5	51.3
BM25 + ASRank §	**42.1**	61.1	67.4	**56.8**	**58.2**	71.1	74.7	68.0	**40.9**	**61.1**	**68.1**	**56.7**
Contriever	22.1	47.2	58.7	42.7	34.1	59.4	68.0	53.8	19.9	43.4	56.3	39.9
Contriever + UPR	36.4	**64.6**	**72.4**	**57.8**	56.7	**77.0**	**80.2**	**71.3**	30.0	58.5	68.2	52.2
Contriever + ASRank §	**41.5**	61.3	68.4	57.0	**57.9**	72.8	76.8	69.1	**42.9**	**62.7**	**69.8**	**58.4**
Supervised Retrievers												
DPR	48.6	68.7	74.5	63.9	57.4	72.4	76.5	68.7	44.8	65.0	70.6	60.1
DPR + UPR	42.5	**70.6**	**78.1**	**63.8**	61.3	**78.7**	**81.9**	**74.0**	34.9	63.6	71.7	56.7
DPR + ASRank §	**43.5**	64.9	72.2	60.2	**61.8**	74.6	78.3	71.5	**45.9**	**66.7**	**72.4**	**61.6**
MSS-DPR	**50.1**	**71.8**	**77.4**	**66.5**	61.6	75.2	79.1	71.9	44.2	65.0	71.6	60.3
MSS-DPR + UPR	41.4	69.8	77.9	63.0	60.5	**78.9**	**82.5**	**74.0**	31.8	61.6	70.3	54.5
MSS-DPR + ASRank §	43.5	65.1	72.5	60.3	**61.7**	74.8	78.6	71.7	**44.6**	**65.4**	**72.2**	**60.7**

Table 9: Top-1, 5, 10 retrieval accuracy on the test set of datasets before and after re-ranking the top 1000 retrieved passages. § refers to Llama 7B

EM score increases from 18.5% to 28.8% with two documents. This suggests that the larger model can leverage the extra information more effectively, leading to better overall performance.

I Case Study

In this section, we present a detailed case study to illustrate the effectiveness of ASRank in re-ranking documents retrieved by different retrieval systems. Tables 11, 12, and 13 showcase examples from the NQ dev dataset, WebQA, and TriviaQA, respectively. Each table lists the document IDs retrieved before and after applying ASRank, indicating whether each document contains the answer ('has_answer: True' or 'False'). These case studies demonstrate how ASRank enhances the precision of document retrieval across varied contexts and query types by leveraging the answer scent generated from advanced language models.

Model	top-K	NQ			TriviaQA			WebQA		
		EM	Recall	Con	EM	Recall	Con	EM	Recall	Con
LLama 7B+Baselines										
BM25	1	16.0	29.3	21.7	51.9	63.5	57.2	14.3	35.7	25.6
MSS	1	14.9	27.4	20.8	43.6	55.4	49.3	13.7	37.1	26.9
Contriever	1	17.3	31.1	23.9	44.4	56.5	50.2	13.6	38.5	23.8
DPR	1	24.6	**40.5**	**32.1**	50.0	62.6	56.6	15.1	**40.3**	**29.3**
MSS-DPR	1	**24.9**	40.4	32.0	**51.9**	**64.7**	**58.4**	**15.9**	40.1	29.0
LLama 13B+Baselines										
BM25	1	18.5	30.8	23.9	54.6	65.8	59.4	14.6	35.1	25.5
MSS	1	17.9	29.5	22.7	47.8	58.5	52.6	14.9	35.6	25.5
Contriever	1	20.3	32.4	25.2	49.0	59.5	53.6	17.5	38.6	28.3
DPR	1	27.9	43.2	33.9	53.6	65.3	58.9	18.2	**41.2**	**30.9**
MSS-DPR	1	**28.9**	**43.8**	**34.5**	**54.7**	**66.7**	**60.5**	**19.6**	40.7	30.7
LLama 7B+Baselines										
MSS	2	15.1	28.2	21.2	44.6	57.3	51.3	14.3	37.8	27.2
BM25	2	16.3	29.9	22.8	52.2	65.0	58.5	13.1	36.9	26.1
Contriever	2	16.9	31.3	23.9	44.8	58.2	51.9	13.9	39.3	28.4
DPR	2	23.9	39.3	31.5	49.8	63.3	57.0	14.8	**40.3**	**28.7**
MSS-DPR	2	**24.3**	**40.3**	**32.2**	**50.8**	**64.5**	**58.1**	**15.1**	39.7	29.0
LLama 7B+UPR										
MSS	1	21.9	37.3	29.2	53.5	66.5	60.1	15.1	39.9	**29.0**
BM25	1	21.2	36.2	28.4	**57.6**	**70.3**	**63.6**	13.9	37.0	25.6
Contriever	1	22.5	38.5	30.5	53.8	67.5	61.3	14.0	38.9	27.6
DPR	1	**23.8**	**39.8**	**31.3**	55.2	68.5	61.8	**15.5**	**40.1**	28.7
MSS-DPR	1	23.6	39.4	30.8	55.3	68.5	62.2	15.3	39.8	28.4
LLama 13B+UPR										
MSS	1	25.2	39.4	31.2	56.4	68.4	62.1	16.9	39.3	28.0
BM25	1	25.1	39.3	30.7	57.3	68.8	**63.5**	16.8	36.8	26.7
Contriever	1	26.0	40.4	31.9	56.5	68.0	62.7	17.4	38.3	28.4
DPR	1	**27.4**	**42.2**	**33.0**	**57.3**	**69.6**	63.2	**17.5**	**40.6**	**29.8**
MSS-DPR	1	26.3	41.3	32.7	57.2	69.2	62.9	17.1	37.9	27.2
LLama 7B+UPR										
MSS	2	21.6	37.3	29.9	54.1	67.9	61.3	15.2	39.1	28.1
BM25	2	22.0	37.8	30.2	**58.2**	**71.4**	**64.8**	14.7	39.7	28.1
Contriever	2	22.3	38.4	30.5	54.9	68.2	65.0	14.5	38.8	27.2
DPR	2	23.2	38.9	31.3	55.1	69.3	62.9	**15.7**	**40.4**	**28.8**
MSS-DPR	2	**24.1**	**40.0**	**32.0**	54.9	69.2	62.7	14.4	39.6	27.5
LLama 7B+ASRANK										
MSS	1	24.8	40.6	32.7	57.1	70.5	64.1	17.9	42.3	31.2
BM25	1	25.0	40.4	32.3	**60.6**	**73.2**	**66.8**	16.9	42.5	31.6
Contriever	1	25.9	41.9	33.6	57.8	71.0	64.9	17.7	43.9	33.1
DPR	1	25.8	42.2	33.7	57.6	71.1	64.6	16.5	43.7	31.2
MSS-DPR	1	**25.9**	**42.6**	**34.2**	58.9	71.8	65.4	**18.1**	**43.9**	**32.8**
LLama 13B+ASRANK										
MSS	1	28.5	43.4	34.6	60.1	72.4	65.9	20.5	43.9	33.6
BM25	1	28.8	44.2	35.4	**63.3**	**74.9**	**68.5**	19.3	43.0	31.5
Contriever	1	29.7	45.1	36.3	60.1	72.4	66.1	20.6	44.0	33.2
DPR	1	28.9	44.9	35.5	60.8	72.9	66.7	19.9	43.0	32.2
MSS-DPR	1	**30.2**	**45.5**	**36.6**	60.9	73.3	66.9	**20.6**	**44.7**	**33.7**
LLama 7B+ASRANK										
MSS	2	25.5	41.3	33.2	56.5	70.8	64.2	17.9	43.8	32.9
BM25	2	24.9	40.8	33.4	**59.8**	**73.3**	**67.1**	**16.6**	**41.7**	**30.4**
Contriever	2	25.5	41.6	33.8	56.5	71.4	64.9	17.0	43.4	31.8
DPR	2	25.9	41.9	33.8	57.3	71.2	64.7	17.6	43.6	32.3
MSS-DPR	2	**26.0**	**42.3**	**34.4**	57.7	72.1	65.5	17.6	43.7	32.5

Table 10: Exact match scores for the open-domain QA task.

Answer Scent Prompt: Generate a brief, insightful answer scent to the following question: q

Question: who sang i just called to say i love you?
Answer Scent: The song "I Just Called to Say I Love You" was sung by Stevie Wonder. It was released in 1984 and was featured in the romantic comedy film "The Woman in Red". The song was a huge commercial success, reaching number one on the Billboard Hot 100 chart and winning several awards, including the Academy Award for Best Original Song.
True Answer: Stevie Wonder
DPR:
`6069735, has_answer: True`
`6069737, has_answer: True`
`8421731, has_answer: False`
`8421729, has_answer: False`
`8421736, has_answer: False`
DPR+ASRANK:
`6069735, has_answer: True`
`6069737, has_answer: True`
`12952376, has_answer: True`
`5693954, has_answer: False`
`6069736, has_answer: False`

Question: who is the coach of arizona state men's basketball?
Answer Scent: the head coach of the Arizona State University men's basketball team is Bobby Hurley.
True Answer: Bobby Hurley
DPR:
`12295065, has_answer: True`
`20172309, has_answer: False`
`10047172, has_answer: False`
`10047129, has_answer: False`
`17545880, has_answer: False`
DPR+ASRANK:
`4833080, has_answer: True`
`4833089, has_answer: True`
`20047916, has_answer: True`
`20776179, has_answer: True`
`20776180, has_answer: True`

Question: who played mason on wizards of waverly place?
Answer Scent: The character Mason Greyback, a werewolf and love interest of Alex Russo, was played by actor Gregg Sulkin on the Disney Channel show "Wizards of Waverly Place".
True Answer: Sulkin
DPR:
`4309105, has_answer: False`
`15243289, has_answer: False`
`19878866, has_answer: False`
`17449330, has_answer: False`
`5763564, has_answer: False`
DPR+ASRANK:
`10341922, has_answer: True`
`9651553, has_answer: True`
`17449330, has_answer: False`
`17449350, has_answer: False`
`17449363, has_answer: False`

Question: who plays the queen on game of thrones?
Answer Scent: The role of Cersei Lannister, the Queen Regent of the Seven Kingdoms, on the hit HBO series Game of Thrones is played by the talented actress Lena Headey.
True Answer: Headey
DPR:
`15742943, has_answer: False`
`16420109, has_answer: False`
`13408263, has_answer: False`
`15179481, has_answer: False`
`13408259, has_answer: False`
DPR+ASRANK:
`4456301, has_answer: True`
`5396490, has_answer: True`
`4456267, has_answer: True`
`5396478, has_answer: True`
`12800938, has_answer: True`

Table 11: Case study of the retrieved passage from DPR retriever before/after reranking for NQ dev dataset.

Answer Scent Prompt: Generate a brief, insightful answer scent to the following question: q

Question: where is jamarcus russell from?
Answer Scent: JaMarcus Russell, the former NFL quarterback, is from Mobile, Alabama.
True Answer: Mobile
MSS:
```
7564117, has_answer: False
9565056, has_answer: False
7564095, has_answer: False
7564124, has_answer: True
8298815, has_answer: False
```
MSS+ASRANK:
```
7564126, has_answer: True
7564096, has_answer: True
7564097, has_answer: True
7564125, has_answer: True
7564127, has_answer: False
```

Question: where is the fukushima daiichi nuclear plant located?
Answer Scent: The Fukushima Daiichi Nuclear Power Plant is located in Okuma, Fukushima Prefecture, Japan. It is situated on the Pacific coast of Honshu, the largest island of Japan, approximately 150 miles (240 km) north of Tokyo.
True Answer: "Japan", "Okum"
MSS:
```
20927197, has_answer: False
3072697, has_answer: False
20927199, has_answer: False
10806688, has_answer: False
18499449, has_answer: True
```
MSS+ASRANK:
```
15436005, has_answer: False
11148375, has_answer: True
10483298, has_answer: True
3541428, has_answer: False
1682872, has_answer: True
```

Question: what does jamaican people speak?
Answer Scent: In Jamaica, the official language is English, which is used in government, education, business, and formal settings. However, the most widely spoken language in Jamaica is Jamaican Patois, also known as Jamaican Creole or Patwa.
True Answer: Jamaican English
MSS:
```
5665719, has_answer: False
9912963, has_answer: False
5665720, has_answer: False
11838832, has_answer: False
20587290, has_answer: False
```
MSS+ASRANK:
```
4423284, has_answer: False
1353789, has_answer: False
8404038, has_answer: False
4423299, has_answer: False
4423301, has_answer: False
```

Question: what is the best sandals resort in st lucia?
Answer Scent: St. Lucia is a beautiful island with several amazing Sandals Resorts to choose from. Each resort has its unique features, amenities, and atmosphere, so the "best" one ultimately depends on your personal preferences and priorities.
True Answer: "Micoud Quarter", "Choiseul Quarter", "Praslin Quarter", ..
MSS:
```
18392196, has_answer: False
18461202, has_answer: False
11371584, has_answer: False
16577459, has_answer: False
3764126, has_answer: False
```
MSS+ASRANK:
```
5476353, has_answer: False
18392196, has_answer: False
3401309, has_answer: False
3401311, has_answer: True
6134966, has_answer: False
```

Table 12: Case study of the retrieved passage from MSS retriever Before/after Reranking for WebQA.

Answer Scent Prompt: Generate a brief, insightful answer scent to the following question: q

Question: which 70s show was based on the british show till death us do part?
Answer Scent: The 1970s show based on the British show "Till Death Us Do Part" is "All in the Family".
True Answer: "All In The Family", "Justice For All (TV pilot)", "Stretch Cunningham", ...
Contriever:
`9539720, has_answer: False`
`6899634, has_answer: False`
`475319, has_answer: False`
`9549805, has_answer: False`
`475315, has_answer: False`
Contriever+ASRANK:
`9607452, has_answer: True`
`1413988, has_answer: True`
`1834891, has_answer: True`
`5285410, has_answer: True`
`1941863, has_answer: True`

Question: what is the name of terence and shirley conran's dress designer son?
Answer Scent: Jasper Conran!
True Answer: "Jaspis", "Bruneau jasper", "Egyptian jasper"
Contriever:
`4935862, has_answer: False`
`4935861, has_answer: True`
`7176709, has_answer: False`
`14139592, has_answer: False`
`5848573, has_answer: True`
Contriever+ASRANK:
`5848571, has_answer: True`
`5848575, has_answer: False`
`5848577, has_answer: True`
`5848576, has_answer: False`
`5848573, has_answer: True`

Question: in which country is the sky train rail bridge?
Answer Scent: The SkyTrain Rail Bridge is located in Vancouver, British Columbia, Canada.
True Answer: "Canada", "Kenadian", "Canadialand", "Xanada", "Dominion of Canada", "Canadaa"
Contriever:
`11617523, has_answer: False`
`11617522, has_answer: False`
`7697355, has_answer: False`
`3375880, has_answer: False`
`4904611, has_answer: True`
Contriever+ASRANK:
`8509738, has_answer: True`
`1145807, has_answer: True`
`1145854, has_answer: True`
`1145799, has_answer: True`
`8509740, has_answer: True`

Question: bandar seri begawan international airport is in which country?
Answer Scent: Bandar Seri Begawan International Airport is located in Brunei.
True Answer: "Abode of Peace", "BRUNEI", "Health in Brunei", ...
Contriever:
`2693267, has_answer: False`
`6595413, has_answer: False`
`10932719, has_answer: False`
`670520, has_answer: True`
`10932726, has_answer: True`
Contriever+ASRANK:
`670503, has_answer: True`
`670496, has_answer: True`
`10893158, has_answer: True`
`5225731, has_answer: True`
`11964123, has_answer: True`

Table 13: Case study of the retrieved passage from Contriever retriever Before/after Reranking for TriviaQA.