Entity Retrieval for Answering Entity-Centric Questions

Hassan S. Shavarani
School of Computing Science
Simon Fraser University
BC, Canada
sshavara@sfu.ca

Anoop Sarkar
School of Computing Science
Simon Fraser University
BC, Canada
anoop@sfu.ca

Abstract

The similarity between the question and indexed documents is a crucial factor in document retrieval for retrieval-augmented question answering. Although this is typically the only method for obtaining the relevant documents, it is not the sole approach when dealing with entity-centric questions. In this study, we propose *Entity Retrieval*, a novel retrieval method which rather than relying on question-document similarity, depends on the salient entities within the question to identify the retrieval documents. We conduct an in-depth analysis of the performance of both dense and sparse retrieval methods in comparison to *Entity Retrieval*. Our findings reveal that our method not only leads to more accurate answers to entity-centric questions but also operates more efficiently.

https://github.com/shavarani/EntityRetrieval

1 Introduction

Information retrieval has significantly enhanced the factual reliability of large language model (LLM) generated responses (Shuster et al., 2021) in question answering (Zhu et al., 2021; Zhang et al., 2023). This improvement is particularly evident in Retrieval-Augmented Generation (RAG; Lewis et al., 2020b; Izacard and Grave, 2021b; Singh et al., 2021), which typically employs the Retriever-Reader architecture (Chen et al., 2017). RAG retrievers can be sparse (Peng et al., 2023), dense (Karpukhin et al., 2020), or hybrid (Glass et al., 2022), while the readers are usually generative language models[1] such as BART (Lewis et al., 2020a), T5 (Raffel et al., 2020), or GPT-4 (OpenAI, 2023) that generate answers based on the documents identified by the retriever. Recent RAG methodologies leverage the in-context learning capabilities of LLMs to incorporate retrieved documents into the prompt (Shi et al., 2023; Peng et al., 2023; Yu et al., 2023).

Entity-centric questions seek concise factual answers about the real world, typically in the form of single words or short phrases. These answers often reference or directly stem from a knowledge base entity (Ranjan and Balabantaray, 2016), and Retrieval-Augmentation enhances LLM performance in answering such questions, particularly for rare entities that appear infrequently in LLM training and fine-tuning data (Kandpal et al., 2023).

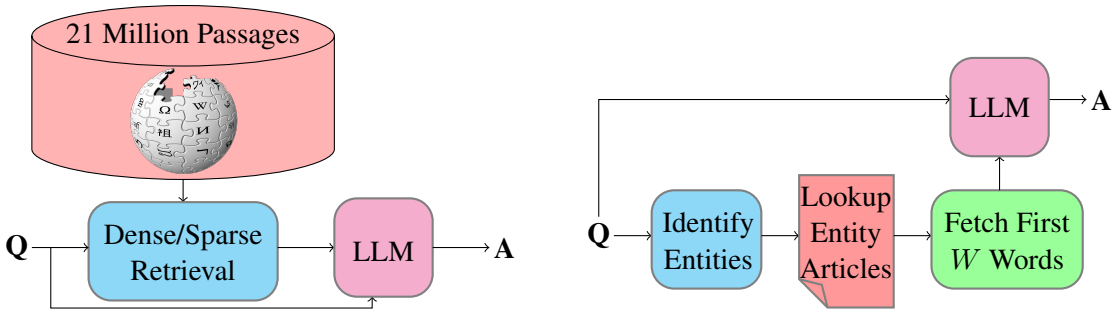
But is there a correlation between the quality of the retrieved documents and the generated response quality? Sciavolino et al. (2021) found that dense retrievers retrieve less relevant documents for answering entity-centric questions than simpler sparse retrievers. Additionally, Cuconasu et al. (2024) show that the presence of irrelevant documents leads to worse answers. These findings underscore the crucial role of the retrieval module, particularly for entity-centric questions.

In this paper, we propose *Entity Retrieval* (Figure 1b), which uses salient entities in the question to lookup knowledge base (e.g., Wikipedia) articles that correspond to each entity. Each article is truncated to the first W words to form a document set that augments the question passed to the LLM.

Our contributions are as follows: (1) we propose *Entity Retrieval*, a novel method of acquiring augmentation documents using salient entities in the questions, (2) we compare the retrieval performance quality of several retrieval techniques (both dense and sparse) to *Entity Retrieval* for questions within two entity-centric question answering datasets, (3) we study the Retrieval-Augmentation quality of the compared techniques and *Entity Retrieval*, using salient entity annotations of the questions, and (4) we examine the application of a recent state-of-the-art entity linking method for *Entity Retrieval* in the absence of entity annotations in entity-centric questions.

[1]The readers in the original architecture were designed to extract answer spans rather than generate answers.

Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing (KnowledgeNLP'25), pages 1–17



(a) Retrieval-Augmented QA with Dense/Sparse Retrieval

(b) Retrieval-Augmented QA with *Entity Retrieval*

Figure 1: *Entity Retrieval* simplifies the process of obtaining augmentation documents by replacing the need to search through large indexed passages with a straightforward lookup. For **Q**: `What is the capital of Seine-Saint-Denis?` *Entity Retrieval* considers the first few sentences of `Seine-Saint-Denis` Wikipedia article which states "`Its prefecture is` **`Bobigny`**." and returns **A = Bobigny** where the other retrieval methods return **A = Saint-Denis** or **A = Paris**.

2 Retrieval for Retrieval-Augmentation

Retrieval-Augmentation (Lewis et al., 2020b) can be employed as a method of converting Closed-book question answering[2] (Roberts et al., 2020) into extractive question answering (Abney et al., 2000; Rajpurkar et al., 2016), where the answers can be directly extracted from the retrieved documents. Despite the abundance of effective retrieval techniques for Retrieval-Augmented Question Answering in existing literature (Zhan et al., 2020a,b; Yamada et al., 2021; Chen et al., 2022; Izacard et al., 2022; Santhanam et al., 2022; Ni et al., 2022, *inter alia.*), this section will concentrate on a select few methods[3] utilized to study answering entity-centric questions in this paper.

BM25 (Robertson et al., 1994, 2009) is a probabilistic retrieval method that ranks documents based on the frequency of query terms appearing in each document, adjusted by the length of the document and overall term frequency in the collection. It operates in the sparse vector space, relying on precomputed term frequencies and inverse document frequencies to retrieve documents based on keyword matching.

DPR (Dense Passage Retrieval; Karpukhin et al., 2020) leverages a bi-encoder architecture, wherein the initial encoder processes the question and the subsequent encoder handles the passages to be retrieved. The similarity scores between the two encoded representations are computed using a dot product. Typically, the encoded representations of the second encoder are fixed and indexed in FAISS (Johnson et al., 2019), while the first encoder is optimized to maximize the dot-product scores based on positive and negative examples.

ANCE (Xiong et al., 2021) is another dense retrieval technique similar to DPR[4]. It employs one encoder to transform both the questions and passages into dense representations. The key distinction from DPR is that ANCE uses hard negatives generated by periodically updating the passage embeddings during training, which helps the model learn more discriminative features, thereby enhancing retrieval performance over time.

3 Entity Retrieval for Question Answering

While quite powerful, most Retrieval-Augmented systems are notably time and resource-intensive, necessitating the storage of extensive lookup indices and the need to attend to all retrieved documents to generate the response (see Section 4.7). This attribute renders such methods less desirable, particularly given the drive to run LLMs locally and on mobile phones (Alizadeh et al., 2023).

Entity recognition has been an integral component of statistical question answering systems (Aghaebrahimian and Jurčíček, 2016, *inter alia*). Additionally, the extensively studied field of Knowledge Base Question Answering (Cui et al., 2017, *inter alia*) has underscored the significance of entity information from knowledge bases in question answering (Salnikov et al., 2023). A traditional neural question answering pipeline may

[2]Closed-book QA focuses on answering questions without additional context during inference.

[3]We selected the methods supported by `pyserini.io` for the similarity between the underlying modules, minimizing discrepancies across different implementations.

[4]We have also implemented DKRR (Izacard and Grave, 2021a), however, due to its significantly poorer performance compared to other methods, we exclude it from our analysis.

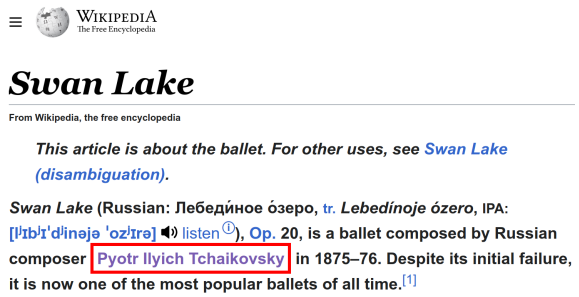


Figure 2: The first paragraph of the Wikipedia article typically provides an informative summary for the entity. For example, the first paragraph of `Swan Lake` Wikipedia article contains the answer to "`Who is the composer of The Swan Lake ballet?`"

contain entity detection, entity linking, relation prediction, and evidence integration (Mohammed et al., 2018; Lukovnikov et al., 2019), where entity detection can employ LSTM-based (Hochreiter and Schmidhuber, 1997) or BERT-based (Devlin et al., 2019) encoders. Inspired by this body of work, we investigate the relevance of retrieval based on entity information as an alternative strategy to the proposed retrieval methods of Section 2, especially for answering entity-centric questions with LLMs.

Our proposed method, *Entity Retrieval*, leverages the salient entities within the questions to identify and retrieve their corresponding knowledge base articles. We will then truncate these articles to the first W words[5] to form the list of the documents augmenting entity-centric questions when prompting LLMs. Figure 1 presents a schematic comparison between *Entity Retrieval* and other retrieval methods in identifying retrieval documents to enhance question answering with LLMs. Figure 2 provides an intuitive example to motivate the effectiveness of *Entity Retrieval*.

4 Experiments and Analysis

4.1 Setup

We focus on Wikipedia as the knowledge base and utilize the pre-existing BM25, DPR, and ANCE retrieval indexes in Pyserini (Lin et al., 2021). These indexes, follow established practices (Chen et al., 2017; Karpukhin et al., 2020) and segments the articles into non-overlapping text blocks of 100 words, resulting in 21,015,300 passages. For dense retrievers, the passages are processed with a pre-trained context encoder, generating fixed embedding vectors stored in a FAISS index (Douze et al., 2024). Our experimental entity-centric questions are encoded using the question encoder, and the top k relevant passages to the encoded question are retrieved from the FAISS index. For BM25 sparse retriever, the passages are stored in a Lucene index and the questions are keyword-matched to this index.

As outlined in Section 3, the document retrieval process will require loading the entire index (as well as the question encoder for dense retrieval) into memory which entails significant time and memory consumption. To address this challenge, following Ram et al. (2023), we treat document retrieval as a pre-processing step, caching the most relevant passages for each question before conducting the question answering experiments.

For *Entity Retrieval*, similar to BM25, DPR, and ANCE, we maintain document lengths at 100 words. However, our approach diverges in sourcing documents: rather than drawing from a large index of 21 million passages, we employ the salient entities within the question and retrieve their corresponding Wikipedia articles, which we then truncate to the initial 100 words.

We conduct our Retrieval-Augmented Question Answering experiments using LLaMA 3 model[6], and in all such experiments[7], we prevent it from generating sequences longer than 10 subwords.

We do not use any instructional question-answer pairs in the prompts of our models[8]. In the Closed-book setting, the prompt includes only the question, along with a simple instruction to answer it. In Retrieval-Augmented settings using BM25, DPR, and ANCE, the prompt incorporates pre-fetched retrieved documents from the corresponding retrieval index alongside the question and the instruction. Similarly, in the *Entity Retrieval* settings, the prompt consists of the first W words of the Wikipedia articles corresponding to the salient entities in the question. We follow Ram et al. (2023) for question normalization and prompt formulation. Appendix A provides the prompts, and example retrieved documents for each setting.

[5]The first sentences of Wikipedia articles have been proven informative for document classification (Shavarani and Sekine, 2020) as well as question answering (Choi et al., 2018).

[6]`https://llama.meta.com/llama3/`.

[7]We run our experiments on one server containing 2 RTX A6000s with 49GB GPU memory each.

[8]Further exploration into few-shot experimental setups involving additional (context, question, answer) in-context examples is left for future investigation.

4.2 Data

We use the following datasets in our experiments[9]:

EntityQuestions (Sciavolino et al., 2021) is created by collecting 24 common relations (e.g., 'author of' and 'located in') and transforming fact triples (subject, relation, object) that contain these relations, into natural language questions using predefined templates. The dataset comprises 176,560 train, 22,068 dev, and 22,075 test question-answer pairs. To expedite our analytical experiments in this paper, given the extensive size of the dev and test sets, we constrain the question-answer pairs in these subsets to those featuring salient entities within the top 500K most linked Wikipedia pages, as suggested by Shavarani and Sarkar (2023). Thus, the dev and test subsets of EntityQuestions considered in our experiments consist of 4,710 and 4,741 questions, respectively.

FactoidQA (Smith et al., 2008) contains 2,203 hand crafted question-answer pairs derived from Wikipedia articles, with each pair accompanied by its corresponding Wikipedia source article included in the dataset.

StrategyQA (Geva et al., 2021) is a complex boolean question answering dataset, constructed by presenting individual terms from Wikipedia to annotators. Its questions contain references to more than one Wikipedia entity, and necessitate implicit reasoning for binary (Yes/No) responses. The dataset comprises 5,111 answered questions initially intended for training question answering systems, with the system later tested on test set questions with unreleased answers. This training set is split into two subsets resulting in `train` and `train_filtered` subsets containing 2,290 and 2,821 questions, respectively.

4.3 Evaluation

We evaluate the performance of the retrieval methods using the following metrics; in each of which a document is considered *relevant* if it contains a normalized form of the expected answer to the question:

- nDCG@k (normalized Discounted Cumulative Gain at rank k; Järvelin and Kekäläinen, 2002) evaluates the quality of a ranking system by considering both the relevance and the position of documents in the top k results. Mathematically, it is represented as

$$\text{nDCG@}k = \frac{\sum_{i=1}^{k} \frac{2^{r_i}-1}{\log_2(i+1)}}{\sum_{i=1}^{|REL_k|} \frac{2^{r_i}-1}{\log_2(i+1)}}$$

  Where, r_i denotes the relevance score of a document at the i^{th} position for a question, with relevance score $r_i = 1$ if the document is *relevant*, and $r_i = 0$, otherwise. REL_k refers to the *relevant* subset of the retrieved documents. nDCG@k scores range between 0 and 1, where a score of 1 signifies an optimal ranking with the most *relevant* documents positioned at the top.

- MRR (Mean Reciprocal Rank; Voorhees and Harman, 1999) is the average of the reciprocal ranks of the first *relevant* document for each question. Mathematically, it is represented as

$$\text{MRR} = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{r_j}$$

  where $|Q|$ represents the total number of questions and r_j denotes the rank of the first *relevant* document for the j-th question.

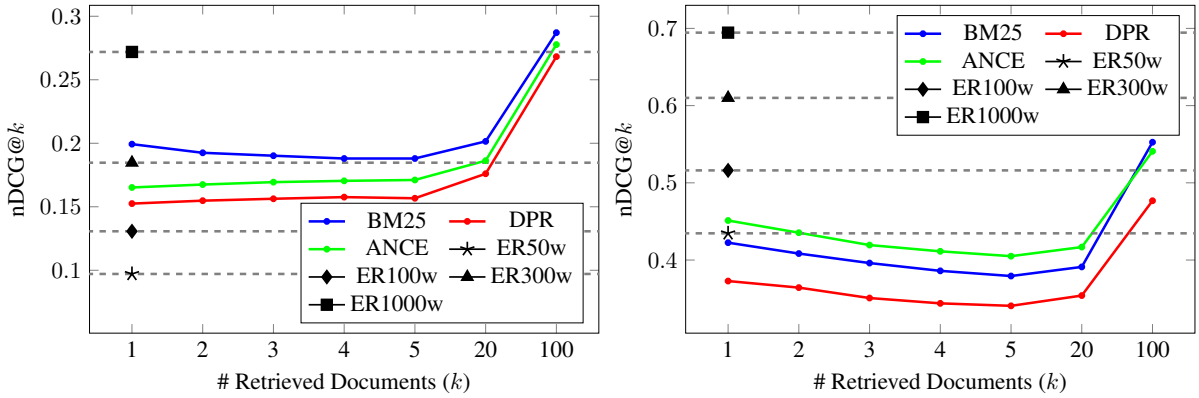
- Top-k Retrieval Accuracy, as reported by Sciavolino et al. (2021), is calculated as the number of questions with at least one *relevant* document in the top k retrieved documents divided by the total number of questions in the dataset.

We evaluate the performance of the Retrieval-Augmented Question Answering models with each retrieval method as follows:

- For FactoidQA and EntityQuestions datasets, we use `OpenQA-eval` (Kamalloo et al., 2023) scripts to evaluate model performance, and report exact match (EM) and F1 scores by comparing expected answers to normalized model responses.

- For StrategyQA, we present accuracy scores by comparing model responses to the expected boolean answers in the dataset. As well, to assess model comprehension of the task, we count the number of answers that deviate from Yes or No and report this count in a distinct column labeled "`Inv #`" for each experiment.

[9]Please note that since *Entity Retrieval* does not involve training, all mentioned dataset subsets (e.g., train, dev, or test) will be used for evaluation regardless of their names.



(a) FactoidQA

(b) EntityQuestions (dev)

Figure 3: nDCG@k scores evaluate the quality of BM25, DPR, ANCE, and *Entity Retrieval* by considering both the relevance and the position of documents in the top k retrieved passages for each question. Note that *Entity Retrieval* typically results in k=1 document since the datasets under study often have one salient entity. The horizontal lines aid in visually comparing the performance of *Entity Retrieval*, which averages one document, to other methods retrieving k>1 documents.

4.4 Entity Retrieval Performance using Question Entity Annotations

We begin our analysis by comparing *Entity Retrieval* performance to BM25, DPR, and ANCE. For this experiment, we calculate nDCG with various retrieved document sets of size $k = 1, 2, 3, 4, 5, 20$, and 100. We use the entity annotations provided with the questions from FactoidQA and the dev set of EntityQuestions to fetch their corresponding Wikipedia articles, excluding StrategyQA from our analysis as it does not include entity annotations. On average, FactoidQA and EntityQuestions datasets contain one salient entity per question.

Apart from a few questions, the majority of FactoidQA questions, and all questions in the EntityQuestions dataset, contain only one entity annotation (leading to one augmentation document). This puts *Entity Retrieval* at a disadvantage. To address this, we consider truncating the *Entity Retrieval* documents to varying lengths. We compare *Entity Retrieval* using the first 100 words (equivalent to the size of documents returned by BM25, DPR, and ANCE, noted as *ER100w*) and also consider the first 50, 300, and 1000 words of the retrieved Wikipedia articles (noted as *ER50w*, *ER300w*, and *ER1000w*). A 300-word *Entity Retrieval* document matches the word count of three documents returned by BM25 or DPR.

Figure 3 presents the computed nDCG@k scores across varying document sizes, highlighting the superior performance of *Entity Retrieval* over other retrieval methods in the context of the entity-centric datasets under study. Notably, *ER1000w*, which corresponds to ten BM25 retrieved passages in terms of word count, exhibits a retrieval performance on par with 100 retrieved documents in FactoidQA and surpasses BM25, the top-performing retriever on EntityQuestions, by 25%. This impressive performance by *Entity Retrieval* can be attributed to its ability to retrieve fewer, yet more relevant, documents. This observation aligns with the conclusion drawn by Cuconasu et al. (2024), which emphasizes that the retrieval of irrelevant documents can negatively impact performance. *Entity Retrieval* effectively minimizes the retrieval of such documents. Further insights can be gleaned from the comparison of nDCG scores along the x-axis of the plots in Figure 3. As the number of retrieved documents increases, the likelihood of retrieving irrelevant documents also rises, leading to a decline in retrieval performance when moving from 1 to 5 retrieved documents.

Table 1 showcases the calculated MRR scores, emphasizing the quicker attainment of relevant retrieval documents in *Entity Retrieval* compared to other retrieval methods. Concurrently, Figure 4 illustrates the impact of incrementing the number of retrieved documents on the expansion of the expected answers' coverage for the EntityQuestions dev subset.

While it may be appealing to consider 100 or more documents to simultaneously enhance both nDCG and Retrieval Accuracy, it is important to note that 100 retrieved documents would comprise 10,000 words. This could potentially overwhelm the model with excessive noise (irrelevant documents), and as well, could make it extremely costly

	FactoidQA	EntityQuestions (dev)
BM25	0.245	0.522
DPR	0.209	0.456
ANCE	0.222	0.536
ER50w	0.097	0.435
ER100w	0.131	0.516
ER300w	0.185	0.610
ER1000w	**0.272**	**0.695**

Table 1: MRR scores comparing the retrieval quality of BM25, DPR, ANCE, and *Entity Retrieval* through the average of the reciprocal ranks of the first relevant document for each question.

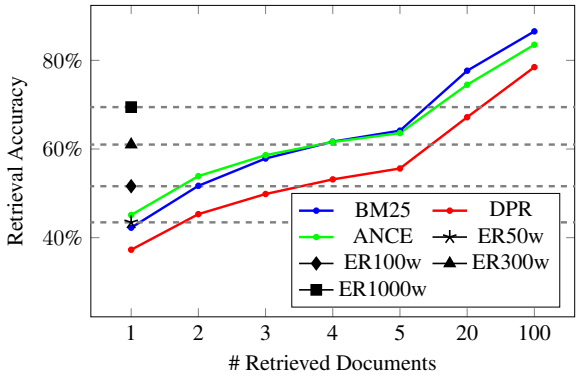


Figure 4: Retrieval Accuracy scores showcasing the correlation between the number of retrieved documents and the expected answers' coverage in EntityQuestions (dev) subset.

to execute Retrieval-Augmented Question Answering, especially when the cost of API calls is calculated per token. We would need at least 10,000 tokens (optimistically, assuming each word equates to only one token) in addition to the tokens in the question. These factors suggest that retrieving a few documents for each question is more beneficial.

Taking these considerations into account, along with the nDCG@k, MRR, and Retrieval Accuracy results from this section, we gain a comprehensive understanding of the trade-off between the quality of the retrieved documents, which diminishes as we consider more documents, and the answer coverage, which increases as the model has a higher chance of encountering the right document with the correct hint for the answer. Consequently, we opt for $k = 4$ as a default, and we will always retrieve the top-4 documents in our Retrieval-Augmented Question Answering experiments.

4.5 Retrieval-Augmented Question Answering

Next, we examine the effectiveness of our proposed *Entity Retrieval* method compared to other retrieval methods in improving the quality of responses to entity-centric questions. We explore three settings: Closed-book, Retrieval-Augmented, and *Entity Retrieval* with question entity annotations (Section 4.1). The primary purpose of using question entity annotations is to demonstrate their ability to accurately identify relevant augmentation documents. These experiments establish an expected performance ceiling for *Entity Retrieval* and can inspire future research to meet or exceed this threshold.

The initial eight rows of Table 2 present the results of our experiments using LLaMA 3 (8B) model. Upon examining these results, it is evident that *ER100w*, the most analogous *Entity Retrieval* setting to other retrieval methods, outperforms in terms of both EM and F1 scores. This setting, like the other retrieval methods, returns 100-word documents. However, as we noted earlier, *Entity Retrieval* generally retrieves fewer documents overall, making it both more accurate and more efficient.

Our dense retrieval results align with the observations of Sciavolino et al. (2021), asserting that entity-centric questions indeed challenge dense retrievers. Although the BM25 method proves successful in enhancing the results compared to the Closed-book setting, it is noteworthy that even *Entity Retrieval* with the initial 50 words of the articles corresponding to the salient entities within questions yields superior results. This is particularly significant when compared to other retrieval methods which necessitate indexing the entire knowledge base on disk and loading the index into memory; a process required in inference time where caching is not an option.

4.6 Entity Retrieval in absence of Question Entity Annotations

Section 4.5 establishes *Entity Retrieval* as a viable augmentation method for entity-centric questions. Next, we aim to reach the established performance ceiling in the absence of question entity annotations. Here, we examine the potential of entity linking as an automated method to provide these annotations. Our primary research question is: how effectively can current entity linking methods help *Entity Retrieval* achieve optimal performance?

Ideally, we would like to evaluate all recent entity linking methods to identify the most effective one. However, due to time and budget limitations,

LLaMA3 (8B)	FactoidQA		EntityQuestions			
			dev		test	
	EM	**F1**	**EM**	**F1**	**EM**	**F1**
Closed-book	30.5±0.4	39.3±0.0	22.9±0.5	37.9±0.7	22.9±0.2	38.3±0.5
Retrieval-Augmented QA						
BM25	32.4±0.8	42.6±0.3	23.7±0.3	38.5±0.6	23.4±0.2	38.7±0.3
DPR	29.8±1.0	38.9±1.1	21.9±0.3	36.2±0.2	20.7±0.6	35.4±0.4
ANCE	30.4±0.4	39.9±0.3	23.1±0.5	37.9±0.4	22.7±0.5	37.9±0.6
Entity Retrieval w/ Question Entity Annotations						
ER50w	34.4±0.5	43.7±0.5	24.9±0.1	41.2±0.1	24.1±0.6	41.1±0.3
ER100w	33.6±0.3	42.9±0.4	**26.3**±0.2	**42.8**±0.1	**25.7**±0.1	**42.4**±0.0
ER300w	33.7±0.9	43.0±1.1	26.2±0.3	42.7±0.1	25.5±0.7	**42.4**±0.8
ER1000w	**35.0**±0.3	**44.9**±0.5	25.1±0.4	41.9±0.4	24.2±0.9	41.1±0.6
Entity Retrieval w/ SpEL Entity Annotations						
ERSp50w	29.6±0.3	38.6±0.5	24.1±0.5	39.1±0.2	23.6±0.8	39.4±0.5
ERSp100w	28.7±0.9	37.7±1.0	24.8±0.5	40.0±0.2	24.4±0.3	39.9±0.2
ERSp300w	26.9±0.4	35.6±0.5	24.5±0.3	39.9±0.4	24.4±0.5	40.2±0.3
ERSp1000w	21.7±0.7	30.8±1.0	24.2±0.2	39.6±0.3	22.9±0.5	39.0±0.7

Table 2: Question answering efficacy comparison between Closed-book and Retrieval-Augmentation using BM25, DPR, ANCE, and *Entity Retrieval*. EM refers to the exact match between predicted and expected answers, disregarding punctuation and articles (`a`, `an`, `the`).
* Results represent the average of three runs, accompanied by a margin of error based on a 99% confidence interval.

we depend on the recent benchmarking studies by Ong et al. (2024) to choose a method. They examine the latest entity linking methods in terms of performance against unseen data and endorse SpEL (Shavarani and Sarkar, 2023) as the top performer. Consequently, we investigate *Entity Retrieval* using entities identified with SpEL, while reserving the examination of other entity linking techniques for *Entity Retrieval* for future research.

We maintain the *Entity Retrieval* settings as before, defining *ERSp50w*, *ERSp100w*, *ERSp300w*, and *ERSp1000w* for performing entity linking with SpEL, then retrieving the Wikipedia articles corresponding to the SpEL identified entities, and using the first 50, 100, 300, and 1000 words of these articles as documents to augment the question when prompting the LLM. Table 3 presents the aggregated entity identification statistics of SpEL across various subsets of each dataset under study.

The final four rows of Table 2 showcase the comparative results of utilizing entities identified by SpEL for *Entity Retrieval*. Given that one-third of EntityQuestions and approximately half of FactoidQA lack identified annotations, the exact match

	Max.	Avg.	Linked %
FactoidQA	8	0.8	56.5%
EntityQuestions	3	0.7	65.6%
StrategyQA	4	1.1	74.9%

Table 3: Maximum and Average SpEL identified entity count as well as the total percentage of questions with at least one identified entity in each dataset. SpEL successfully identifies and links entities in 1,244 FactoidQA, 3,108 EntityQuestions (dev), 3,095 EntityQuestions (test), 1,735 StrategyQA (train), and 2,094 StrategyQA (train_filtered) questions. For the remaining questions in each dataset where no entities are identified, they will be introduced to the LLM without any augmented documents in the *Entity Retrieval* settings.

scores reveal that *Entity Retrieval* performs robustly and surpasses BM25, the top-performing competitor, for EntityQuestions while approaching DPR's performance for FactoidQA. This underscores the potential of *Entity Retrieval* within this paradigm. In addition, the disparity between the results with and without question entity annotations strongly indicates the necessity for further research in Entity Linking, which could enhance

Question	Who performed Alexis Colby?	What is the capital of Seine-Saint-Denis?
Answer	Joan Collins	Bobigny
Closed-Book	Diana Ross	Paris
BM25	Linda Evans	Saint-Denis
DPR	Alexis Cohen	Saint-Denis
ANCE	Nicollette Sheridan performed Alexis Colby.	Saint-Denis
ERSp100w	Joan Collins	Bobigny
Question	Where did John Snetzler die?	Where was Brigita Bukovec born?
Answer	Schaffhausen	Ljubljana
Closed-Book	He died in London, England, in 178	Brigita Bukovec was born in Slovenia
BM25	John Snetzler died in London.	Slovenia
DPR	John Snetzler died in London	in Slovakia
ANCE	in England	Rîbnița
ERSp100w	Schaffhausen	Ljubljana

Table 4: Example questions from EntityQuestions (dev) to demonstrate the performance of *Entity Retrieval*.

LLaMA3 (8B)	train		train_filtered	
	Acc.	Inv #	Acc.	Inv #
BM25	43.5±0.6	608±14	48.9±0.7	673±12
ANCE	46.6±1.3	552±11	51.8±0.7	647±35
ERSp50w	50.1±1.1	370±28	**56.3**±0.9	417±21
ERSp100w	**50.3**±1.4	**369**±15	56.2±0.8	**384**±9
ERSp300w	46.2±1.3	504±17	53.5±1.5	546±20
ERSp1000w	39.5±1.4	775±6	43.4±0.5	919±14

Table 5: Comparison of *Entity Retrieval* using SpEL identified entities to the best-performing dense and sparse retrieval methods of Table 2 on the StrategyQA dataset. Given the expected boolean results for StrategyQA questions, we restricted LLaMA 3 to generate only one token. *Acc.* indicates the fraction of answers that correctly match the expected Yes or No responses in the dataset, while *Inv #* represents the count of labels that are neither Yes nor No, but another invalid answer.

⋆ Results represent the average of three runs, accompanied by a margin of error based on a 99% confidence interval.

entity-centric question answering as a downstream task. Table 4 provides some example questions where *Entity Retrieval* has led to better answers.

Table 5 compares of the performance of *Entity Retrieval* using SpEL identified entities against other retrieval methods on the StrategyQA dataset. The results clearly demonstrate the superior performance of *Entity Retrieval* over the top-performing retrieval methods of Table 2. It is important to note that the 100-word setting (*ERSp100w*) is the most analogous to other retrieval methods. Interestingly, the results from the 1000-word setting suggest that longer documents do not necessarily enhance the model's recall. In fact, beyond a certain length, the model may become overwhelmed by the sheer volume of noise, leading to confusion. Lastly, the invalid count values suggest that *Entity Retrieval* is more effective in assisting the model to comprehend the boolean nature of expected responses, eliminating the need to rely on retrieval from millions of passages.

4.7 Real-time Efficiency Analysis

Our analysis thus far has primarily focused on the retrieval performance, without consideration for the time and memory efficiency; crucial factors in retrieval method selection. In this section, we shift our focus to these aspects.

We begin by replacing our pre-built retrieval cache document sets with the original retrieval modules that were used in creating the cached sets. We load the indexes and the necessary models for fetching the retrieval documents. We then record the peak main memory requirement of each method during the experiment. It is important to note that all retrieval methods primarily rely on main memory, with minimal differences in GPU memory requirements. Therefore, we report an average GPU memory requirement of 35GB for LLaMA 3 (8B) and exclude it from our results ta-

	Total Time	Disk Storage	Main Memory
BM25	45min	11GB	2.3GB
ANCE	960min	61.5GB	64.2GB
ERSp100w	34min	9.4GB	6.3GB

Table 6: Comparison of the required resources for each retrieval method in real-time execution. The reported total time values exclude the time taken to load the indexes and models, focusing solely on the time used to answer the questions.

ble. We then feed all 2,203 FactoidQA questions into the BM25, ANCE, and *Entity Retrieval* (using SpEL identified entities) to fetch the top-4 documents. We report the total time taken to generate answers to all the questions, which includes the time for querying the BM25 or ANCE indexes in the Retrieval-Augmented settings, or the time for performing on-the-fly entity linking and fetching the Wikipedia articles from disk in the *Entity Retrieval* setting. Additionally, we keep track of all the pre-built models and indexes that each method requires for download and storage. We report the total size of all downloaded files to disk.

Table 6 presents our findings on time and memory requirements. It is evident that ANCE requires significantly more time to fetch and provide documents, six times more disk space to store its indexes, and over ten times higher main memory demands to load its dense representations[10]. In contrast, BM25 and *Entity Retrieval* are more resource-friendly. Notably, *Entity Retrieval* is 25% faster than BM25 in response generation while demanding the total memory and disk space of a standard personal computer. Future research can be directed towards reducing the memory requirements of *Entity Retrieval*; a direction which we find quit promising.

5 Related Work

Similar to our studies, Kandpal et al. (2023) investigate the impact of salient entities on question answering, and propose constructing oracle retrieval documents as the 300-word segment surrounding the ground-truth answer from the Wikipedia page that contains the answer (entity name). Our approach leverages salient entities from questions without directly involving answers. Additionally, they primarily use entities to classify questions into those concerning frequent knowledge base entries versus those about rare entries on the long-tail, whereas our approach assigns a more substantial role to entities, treating them as pointers guiding the retrieval of relevant documents to augment questions.

Sciavolino et al. (2021) compare DPR and BM25 retrievers for entity-centric questions, and demonstrate that DPR greatly underperforms BM25. They attribute this to dense retrievers' difficulty with infrequent entities, which are less represented in training data. In contrast, BM25's frequency-based retrieval is not sensitive to entity frequency. We take a parallel approach and propose a simple yet effective method that leverages salient entities in the question for identifying augmentation documents.

Similar to our studies, Dhingra et al. (2020); Asai et al. (2020) focus on answering questions with minimal lexical overlap between the retrieved documents and the question text. However, they emphasize multi-hop question answering, using entity linking to extract entities from the question and leveraging knowledge base articles to guide the multi-hop process. In contrast, we utilize entity links to directly identify augmentation documents. Sun et al. (2018) employ entity linking to identify entities in the question, generating a set of seed entities, which are then expanded using the PPR algorithm to create a subgraph of the knowledge base containing relevant entities. A graph propagation algorithm subsequently learns representations for each node in the subgraph, and each representation is binary classified to determine if it answers the question. Our approach differs as we focus on using LLMs, employing entity linking in a Retrieval-Augmented setting without relying on graph propagation.

6 Conclusion

In this study, we focused on Retrieval-Augmented Question Answering, and explored various retrieval methods that rely on the similarity between the question and the content of the passages to be retrieved. We introduced a novel approach, *Entity Retrieval*, which deviates from the conventional textual similarity-based mechanism. Instead, it capitalizes on the salient entities within the question to identify retrieval documents. Our findings indicate that our proposed method is not only more accurate but also faster in the context of entity-centric question answering.

[10] Our empirical results demonstrate that DPR follows the same trend.

Limitations and Ethical Considerations

Our proposed *Entity Retrieval* method is specifically tailored for answering entity-centric questions, with its performance heavily reliant on the presence of question entities. In scenarios where entity annotations are absent, the method's effectiveness is directly tied to the performance of external entity linking methods. We acknowledge that our exploration of potential entity linking methods has not been exhaustive, and further investigation may yield insights that could enhance the *Entity Retrieval* method, even in the absence of question entity annotations.

Furthermore, we recognize that entity linking can occasionally result in ambiguous entities. Our research has not delved into the impact of such ambiguities on the *Entity Retrieval* method, and we propose that future studies should focus on ensuring the selection of the most contextually appropriate entities for retrieval.

Our research is primarily centered on Wikipedia as the knowledge base, a choice heavily influenced by previous studies for the sake of comparability. However, we acknowledge the importance of exploring other knowledge bases and ontologies, particularly in different domains, such as UMLS (Bodenreider, 2004) in the medical field.

In terms of benchmarking, we have compared the *Entity Retrieval* method against a limited selection of existing retrieval methods, guided by our judgement, experience, and considerations of implementation availability. We concede that our comparison has not been exhaustive, and this reasoning extends to our comparison using different LLMs and their available sizes.

Our research is on English only, and we acknowledge that entity-centric question answering in other languages is also relevant and important. We hope to extend our work to cover multiple languages in the future. We inherit the biases that exist in the data used in this project, and we do not explicitly de-bias the data. We are providing our code to the research community and we trust that those who use the model will do so ethically and responsibly.

References

Steven Abney, Michael Collins, and Amit Singhal. 2000. Answer extraction. In *Sixth Applied Natural Language Processing Conference*, pages 296–301, Seattle, Washington, USA. Association for Computational Linguistics.

Ahmad Aghaebrahimian and Filip Jurčíček. 2016. Open-domain factoid question answering via knowledge graph search. In *Proceedings of the Workshop on Human-Computer Question Answering*, pages 22–28, San Diego, California. Association for Computational Linguistics.

Keivan Alizadeh, Iman Mirzadeh, Dmitry Belenko, Karen Khatamifard, Minsik Cho, Carlo C Del Mundo, Mohammad Rastegari, and Mehrdad Farajtabar. 2023. Llm in a flash: Efficient large language model inference with limited memory. *arXiv preprint arXiv:2312.11514*.

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Xilun Chen, Kushal Lakhotia, Barlas Oguz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2022. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 250–262, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. *arXiv preprint arXiv:2401.14887*.

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2017. Kbqa: Learning question answering over qa corpora and knowledge bases. *Proceedings of the VLDB Endowment*, 10(5).

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

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

Bhuwan Dhingra, Manzil Zaheer, Vidhisha Balachandran, Graham Neubig, Ruslan Salakhutdinov, and William W. Cohen. 2020. Differentiable reasoning over a virtual knowledge base. In *International Conference on Learning Representations*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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*.

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

Gautier Izacard and Edouard Grave. 2021b. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

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

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

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

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

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Denis Lukovnikov, Asja Fischer, and Jens Lehmann. 2019. Pretrained transformers for simple question answering over knowledge graphs. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18*, pages 470–486. Springer.

Salman Mohammed, Peng Shi, and Jimmy Lin. 2018. Strong baselines for simple question answering over knowledge graphs with and without neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 291–296, New Orleans, Louisiana. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nicolas Ong, Hassan S. Shavarani, and Anoop Sarkar. 2024. Unified examination of entity linking in absence of candidate sets. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Mexico. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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.

Prakash Ranjan and Rakesh Chandra Balabantaray. 2016. Question answering system for factoid based question. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 221–224. IEEE.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

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

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Text Retrieval Conference*.

Mikhail Salnikov, Hai Le, Prateek Rajput, Irina Nikishina, Pavel Braslavski, Valentin Malykh, and Alexander Panchenko. 2023. Large language models meet knowledge graphs to answer factoid questions. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 635–644, Hong Kong, China. 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. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hassan S. Shavarani and Anoop Sarkar. 2023. SpEL: Structured prediction for entity linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11123–11137, Singapore. Association for Computational Linguistics.

Hassan S. Shavarani and Satoshi Sekine. 2020. Multiclass multilingual classification of Wikipedia articles using extended named entity tag set. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1197–1201, Marseille, France. European Language Resources Association.

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

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

Noah A Smith, Michael Heilman, and Rebecca Hwa. 2008. Question generation as a competitive undergraduate course project. In *Proceedings of the NSF Workshop on the Question Generation Shared Task and Evaluation Challenge*, volume 9.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4231–4242, Brussels, Belgium. Association for Computational Linguistics.

Ellen M Voorhees and Donna Harman. 1999. Overview of the eighth text retrieval conference (trec-8). In *Proceedings of the Eighth Text REtrieval Conference (TREC-8), NIST Special Publication*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Ikuya Yamada, Akari Asai, and Hannaneh Hajishirzi. 2021. Efficient passage retrieval with hashing for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 979–986, Online. Association for Computational Linguistics.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. *arXiv preprint arXiv:2305.14002*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020a. Learning to retrieve: How to train a dense retrieval model effectively and efficiently. *arXiv preprint arXiv:2010.10469*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020b. Repbert: Contextualized text embeddings for first-stage retrieval. *arXiv preprint arXiv:2006.15498*.

Qin Zhang, Shangsi Chen, Dongkuan Xu, Qingqing Cao, Xiaojun Chen, Trevor Cohn, and Meng Fang. 2023. A survey for efficient open domain question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14447–14465, Toronto, Canada. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Example Prompts for Different Experimental Settings

In this section, we present the prompts used in our experimental settings. For each setting, we provide the prompt template, and explain the processes needed to obtain the augmentation documents if a Retrieval-Augmented setting is being discussed.

A.1 Closed-book Setting

In this setting, we do not have any augmentation documents, so the prompt contains the instruction, followed by the question:

```
Answer this question:
Q: {question}
A:
```

Here is an example prompt with the question mentioned in Figure 1 and Table 4:

```
Answer this question:
Q: What is the capital of Seine-Saint
  -Denis?
A:
```

A.2 Retrieval-Augmented Settings

In this setting, we examine two variations of prompts based on the number of available augmented documents. For a single document, the prompt is as follows:

```
{document}

Based on this text, answer this
question:
Q: {question}
A:
```

When multiple documents are available, they are presented sequentially, followed by the instruction and question:

```
{document1}

{document2}

...

{documentN}

Based on these texts, answer this
question:
Q: {question}
A:
```

Doc#	Content
1	Pierrefitte-sur-Seine<newline>Pierrefitte-sur-Seine Pierrefitte-sur-Seine is a commune in the Seine-Saint-Denis department and Ile-de-France region of France. Today forming part of the northern suburbs of Paris, Pierrefitte lies from the centre of the French capital. The town is served by Pierrefitte - Stains railway station on line D of the RER regional suburban rail network. The south of the commune, where the National Archives of France relocated in 2013, is also served by Saint-Denis - Universite station on Paris Metro Line 13. This station lies on the border between the communes of Pierrefitte-sur-Seine and Saint-Denis. Primary and secondary schools in the commune include:
2	"Saint-Ouen, Seine-Saint-Denis"<newline>Saint-Ouen, Seine-Saint-Denis Saint-Ouen () is a commune in the Seine-Saint-Denis department. It is located in the northern suburbs of Paris, France, from the centre of Paris. The communes neighbouring Saint-Ouen are Paris, to the south, Clichy, to the west, Asnieres-sur-Seine and L'Ile-Saint-Denis, to the north, and Saint-Denis to the east. The commune of Saint-Ouen is part of the canton of Saint-Ouen, which also includes L'Ile-Saint-Denis and part of Epinay-sur-Seine. Saint-Ouen also includes the Cimetiere de Saint-Ouen. On 1 January 1860, the city of Paris was enlarged by annexing neighbouring communes. On that occasion, a part of the commune of Saint-Ouen
3	"Ile-de-France"<newline>of France. The population of immigrants is more widely distributed throughout the region than it was in the early 2000s, though the concentrations remain high in certain areas, particularly Paris and the department of Seine-Saint-Denis. The proportion of residents born outside of Metropolitan France has dropped since the 1999 census (19.7 percent) and the 2010 census (23 percent). . The Petite Couronne (Little Crown, i.e. ""Inner Ring"") is formed by the 3 departments of Ile-de-France bordering with the French capital and forming a geographical ""crown"" around it. The departments, until 1968 part of the disbanded Seine department, are Hauts-de-Seine, Seine-Saint-Denis
4	"Saint-Denis, Seine-Saint-Denis"<newline>Saint-Denis, Seine-Saint-Denis Saint-Denis () is a commune in the northern suburbs of Paris, France. It is located from the centre of Paris. Saint-Denis is a subprefecture () of the department of Seine-Saint-Denis, being the seat of the arrondissement of Saint-Denis. Saint-Denis is home to the royal necropolis of the Basilica of Saint Denis and was also the location of the associated abbey. It is also home to France's national football and rugby stadium, the Stade de France, built for the 1998 FIFA World Cup. Saint-Denis is a formerly industrial suburb currently changing its economic base. Inhabitants of Saint-Denis are called

Table 7: Top 4 documents retrieved from the BM25 Lucene index for the question `What is the capital of Seine-Saint-Denis?` from the EntityQuestions (dev) dataset.

Next, we examine the various Retrieval-Augmentation techniques studied in this paper: BM25, DPR, and ANCE, showcasing their top four retrieved documents for `What is the capital of Seine-Saint-Denis?`. Tables 7, 8, and 9 present these retrieved documents. The finalized prompt template will include the four retrieved documents alongside the question, as previously discussed.

In analyzing the retrieved documents, you can verify the originating Wikipedia articles mentioned in the beginning of each passage. Notably, passages are drawn from three or four different articles, and given the entity-centric nature of the question, relying on multiple sources could mislead the LLM, as suggested by Cuconasu et al. (2024). Additionally, these methods primarily focus on lexical similarity, particularly the presence of `capital`, `Seine`, `Saint`, and `Denis`. However, this focus has not consistently led to retrieval of passages containing the correct answer: `Bobigny`.

A.3 *Entity Retrieval* Settings

For *Entity Retrieval*, we utilize an entity linker to identify entities within the question. In this

Doc#	Content
1	"L'Ile-Saint-Denis"<newline>L'Ile-Saint-Denis L'Ile-Saint-Denis (the island of Saint Denis) is a commune in the northern suburbs of Paris, France. It is located from the center of Paris. The commune is entirely contained on an island of the Seine River, hence its name. Several transit connections are located nearby. The closest station to L'Ile-Saint-Denis is Saint-Denis station, which is an interchange station on Paris RER line D and on the Transilien Paris - Nord suburban rail line. This station is located in the neighboring commune of Saint-Denis, from the town center of L'Ile-Saint-Denis. Tram T1 stops near Ile-Saint-Denis's town hall. Bus route 237
2	"15th arrondissement of Paris"<newline>15th arrondissement of Paris The 15th arrondissement of Paris (""XV arrondissement"") is one of the 20 arrondissements of the capital city of France. In spoken French, this arrondissement is referred to as ""quinzieme"". The arrondissement, called Vaugirard, is situated on the left bank of the River Seine. Sharing the Montparnasse district with the 6th and 14th arrondissements, it is the city's most populous arrondissement. The ""Tour Montparnasse"" - the tallest skyscraper in Paris - and the neighbouring Gare Montparnasse are both located in the 15th arrondissement, at its border with the 14th. It is also home to the convention center
3	"L'Ile-Saint-Denis"<newline>few of the students were White. There are three primary schools in the commune: Ecole Samira Bellil, Ecole Paul Langevin, and Ecole Jean Lurcat. College Alfred Sisley, a junior high school, is on the island. L'Ile-Saint-Denis L'Ile-Saint-Denis (the island of Saint Denis) is a commune in the northern suburbs of Paris, France. It is located from the center of Paris. The commune is entirely contained on an island of the Seine River, hence its name. Several transit connections are located nearby. The closest station to L'Ile-Saint-Denis is Saint-Denis station, which is an interchange station on Paris RER line D and
4	"2nd arrondissement of Paris"<newline>2nd arrondissement of Paris The 2nd arrondissement of Paris (""II arrondissement"") is one of the 20 arrondissements of the capital city of France. In spoken French, this arrondissement is colloquially referred to as ""deuxieme"" (second/the second). Also known as Bourse, this arrondissement is located on the right bank of the River Seine. The 2nd arrondissement, together with the adjacent 8th and 9th arrondissements, hosts an important business district, centred on the Paris Opera, which houses the city's most dense concentration of business activities. The arrondissement contains the former Paris Bourse (stock exchange) and a large number of banking headquarters, as

Table 8: Top 4 documents retrieved from the DPR index for the question `What is the capital of Seine-Saint-Denis?` from the EntityQuestions (dev) dataset.

study, we employed SpEL, though any suitable entity linking method can be used. The primary requirement is that the linker accepts a string (the question) as input and returns a list containing (`begin character`, `end character`, `identified entity`) tuples. The `begin character` and `end character` values help determine the order of entity annotations in the text, ensuring proper sequence if the returned list is unordered.

The `identified entity` values are then used to search the Wikipedia dump on disk, fetching articles corresponding to the identified entities. Efficient implementation of this lookup process is crucial for the method's performance. Our approach involves a two-step indexing and lookup process. First, we index the file bytes marking the beginning of each Wikipedia article in the dump file. When an article is needed, we use this index to locate the byte number and employ the `seek` method to navigate to the correct position in the file and read the article.

After gathering the relevant articles, we truncate each one to the first W words (suffixed with the Wikipedia identifier, as per convention) to create a list of augmentation documents to accompany the question when querying the LLM.

To prompt the LLM, we use the same prompts previously mentioned. If no entities are found in

Doc#	Content
1	"L'Ile-Saint-Denis"<newline>L'Ile-Saint-Denis L'Ile-Saint-Denis (the island of Saint Denis) is a commune in the northern suburbs of Paris, France. It is located from the center of Paris. The commune is entirely contained on an island of the Seine River, hence its name. Several transit connections are located nearby. The closest station to L'Ile-Saint-Denis is Saint-Denis station, which is an interchange station on Paris RER line D and on the Transilien Paris - Nord suburban rail line. This station is located in the neighboring commune of Saint-Denis, from the town center of L'Ile-Saint-Denis. Tram T1 stops near Ile-Saint-Denis's town hall. Bus route 237
2	"L'Ile-Saint-Denis"<newline>few of the students were White. There are three primary schools in the commune: Ecole Samira Bellil, Ecole Paul Langevin, and Ecole Jean Lurcat. College Alfred Sisley, a junior high school, is on the island. L'Ile-Saint-Denis L'Ile-Saint-Denis (the island of Saint Denis) is a commune in the northern suburbs of Paris, France. It is located from the center of Paris. The commune is entirely contained on an island of the Seine River, hence its name. Several transit connections are located nearby. The closest station to L'Ile-Saint-Denis is Saint-Denis station, which is an interchange station on Paris RER line D and
3	"Saint-Denis, Seine-Saint-Denis"<newline>one private elementary, middle, and high school (""Ensemble Scolaire Jean-Baptiste de la Salle-Notre Dame de la Compassion"") and one private middle and high school (""College et lycee Saint-Vincent-de-Paul""). Saint-Denis is twinned with: Saint-Denis, Seine-Saint-Denis Saint-Denis () is a commune in the northern suburbs of Paris, France. It is located from the centre of Paris. Saint-Denis is a subprefecture () of the department of Seine-Saint-Denis, being the seat of the arrondissement of Saint-Denis. Saint-Denis is home to the royal necropolis of the Basilica of Saint Denis and was also the location of the associated abbey. It is also home to France's
4	"Saint-Ouen, Seine-Saint-Denis"<newline>Saint-Ouen, Seine-Saint-Denis Saint-Ouen () is a commune in the Seine-Saint-Denis department. It is located in the northern suburbs of Paris, France, from the centre of Paris. The communes neighbouring Saint-Ouen are Paris, to the south, Clichy, to the west, Asnieres-sur-Seine and L'Ile-Saint-Denis, to the north, and Saint-Denis to the east. The commune of Saint-Ouen is part of the canton of Saint-Ouen, which also includes L'Ile-Saint-Denis and part of Epinay-sur-Seine. Saint-Ouen also includes the Cimetiere de Saint-Ouen. On 1 January 1860, the city of Paris was enlarged by annexing neighbouring communes. On that occasion, a part of the commune of Saint-Ouen

Table 9: Top 4 documents retrieved from the ANCE index for the question `What is the capital of Seine-Saint-Denis?` from the EntityQuestions (dev) dataset.

the question, we refer to the prompt in Appendix A.1. If one entity is recognized, resulting in one augmentation document, we use the first prompt from Appendix A.2. If multiple entities are identified, we use the second prompt from the same appendix section. In rare cases where the number of identified entities exceeds k (the expected number of documents to retrieve), we simply consider the first k unique entities to form the list of augmentation documents.

Table 10 presents the single document retrieved for `What is the capital of Seine-Saint-Denis?`, which contains the answer: `Bobigny`. Examining the lexical distribution in this document, we observe that unlike the BM25 method, *Entity Retrieval* treats the salient entity `Seine-Saint-Denis` as an atomic term rather than emphasizing each word in the question. This focused approach, coupled with the retrieval of fewer documents, allows the model to concentrate on the relevant information, reducing noise and potential confusion.

However, the effectiveness of *Entity Retrieval* in real-world scenarios, where question entity annotations are not available, largely depends on the quality of the entity linker used to identify salient entities in the question. Therefore, further research into developing more accurate entity linking mod-

Doc#	Content
1	Seine-Saint-Denis<newline>Seine-Saint-Denis In 2019, it had a population of 1,644,903 across 40 communes. In French, the learned but rarely used demonym for the inhabitants of Seine-Saint-Denis is ; more common is . The department is surrounded by the departments of Hauts-de-Seine, Val-de-Marne, Paris, Val-d'Oise, and Seine-et-Marne. It is thus the only one of the five French departments surrounded entirely by other departments of the same region. Image:Petite couronne.png The most populous commune is Saint-Denis; the prefecture Bobigny is the eleventh-most populous. As of 2019, there are 5 communes with more than 70,000 inhabitants: is made up of three departmental and 40

Table 10: The only document retrieved by *Entity Retrieval* using SpEL annotations for the question `What is the capital of Seine-Saint-Denis?` from the EntityQuestions (dev) dataset. SpEL identifies only one entity in the question: `Seine-Saint-Denis` and returns the first 100 words (considering W=100) of its Wikipedia article as the retrieved document. The answer to the question: `Bobigny` is highlighted for ease of verification.

els could enhance *Entity Retrieval* performance.