Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering

Shamane Siriwardhana[*◇]**, Rivindu Weerasekera**[◇]**, Elliott Wen**[◇]**,**
Tharindu Kaluarachchi[◇]**, Rajib Rana**[†]**, and Suranga Nanayakkara**[△◇]

[◇] Augmented Human Lab, Auckland Bioengineering Institute,
The University of Auckland, New Zealand
`Shamane@ahlab.org`
[△] Department of Information Systems & Analytics, National University of Singapore, Singapore
[†] University of Southern Queensland, Australia
`Rajib.Rana@usq.edu.au`

Abstract

Retrieval Augment Generation (RAG) is a recent advancement in Open-Domain Question Answering (ODQA). RAG has only been trained and explored with a Wikipedia-based external knowledge base and is not optimized for use in other specialized domains such as healthcare and news. In this paper, we evaluate the impact of joint training of the retriever and generator components of RAG for the task of domain adaptation in ODQA. We propose *RAG-end2end*, an extension to RAG that can adapt to a domain-specific knowledge base by updating all components of the external knowledge base during training. In addition, we introduce an auxiliary training signal to inject more domain-specific knowledge. This auxiliary signal forces *RAG-end2end* to reconstruct a given sentence by accessing the relevant information from the external knowledge base. Our novel contribution is that, unlike RAG, RAG-end2end does joint training of the retriever and generator for the end QA task and domain adaptation. We evaluate our approach with datasets from three domains: COVID-19, News, and Conversations, and achieve significant performance improvements compared to the original RAG model. Our work has been open-sourced through the HuggingFace Transformers library, attesting to our work's credibility and technical consistency.

1 Introduction

Open Domain Question Answering (ODQA) (Lee et al., 2019; Lewis et al., 2020c) is an important task in natural language understanding. ODQA methods generally feature a two-stage pipeline: a retriever that selects passages relevant to a given question and a reader that generates the answers from selected passages. Conventionally, these two components are trained separately using ground truth context passages relevant to question-answer (QA) pairs. However, for many real-world scenarios, it is hard to find explicitly annotated context-question-answer triplets (Lee et al., 2019; Lewis et al., 2020b; Guu et al., 2020).

Recently, Retrieval Augmented Models (RAGs) have drawn considerable attention from researchers. RAG consists of a state-of-the-art neural retriever called Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) and BART seq2seq language model (Lewis et al., 2020a). Compared to the conventional two-staged ODQA pipelines, RAG merges the retriever and reader stages into one architecture. Moreover, unlike expensive language models with billions of parameters (e.g., GPT-3 [Brown et al., 2020] and Megatrone-LM [Narayanan et al., 2021]) where the model's parametric memory represents the complete knowledge, RAG can also extract knowledge from an external knowledge base. Using both parametric and non-parametric memory generally leads to reduced hallucinations and higher interpretability in tasks like question answering and summarization (Xu et al., 2021; Komeili et al., 2021; Guu et al., 2020; Lewis et al., 2020b).

In this work, we focus on exploring retrieval augmented architectures for the task of domain-specific ODQA. Although there are several similar retrieval augmented architectures, such as REALM (Guu et al., 2020) and RETRO (Borgeaud et al., 2021), we used RAG in our

---

[*]The corresponding author - `shamane@ahlab.org`.

experiments due to its excellent open-source documentation and availability.

When the RAG model is finetuned for downstream QA tasks, the original implementation keeps the encoding of passages and the external knowledge base fixed. This is because re-encoding the external knowledge base is computationally expensive and relies on a sophisticated implementation. Despite not finetuning the passage encodings, the RAG model performs well for datasets with Wikipedia-like knowledge bases because the DPR retriever components have already been trained on Wikipedia-based datasets (Kwiatkowski et al., 2019; Joshi et al., 2017). However, the feasibility of adapting RAG to specific ODQA domains such as research papers and news is not well understood. This is a critical research gap to address, as improved domain adaptation can further improve the ODQA performance of RAG.

This paper explores the feasibility of using RAG in specialized domains for ODQA. In particular, we propose two modifications to the original RAG to improve its domain adaptability. Motivated by recent end2end retrieval augmented mechanisms (Guu et al., 2020; Sachan et al., 2021; Singh et al., 2021), we first propose a method to finetune the RAG model with its neural retriever and update its knowledge encodings asynchronously during training. We refer to this as *RAG-end2end* since it allows us to update all RAG components during training, including the external knowledge base, the DPR model, and the BART model. Secondly, we propose an auxiliary training signal to help our model learn more domain-specific knowledge. This took the form of generating a concise and factual statement about a document using a self-retrieved set of passages from the provided domain-specific knowledge base. These two modifications offer a unique feature to RAG-end2end over RAG: joint training of the retriever and generator for the end QA task and domain adaptation. Although asynchronous updates to the knowledge encoder have been proposed before in the REALM, previous work has not evaluated the effects of joint training of the RAG's retriever and the generator for the domain adaptation in ODQA.

We evaluate our proposed approach on three different datasets from three domains: COVID-19 research (Wang et al., 2020), Conversations (Wu et al., 2021b), and News (Trischler et al., 2016).

The major finding of our work is that the adaptation of the retriever component plays a critical role in overall domain adaptation performance in RAG-like architectures. Updating only the question encoder without updating the knowledge base encoding could degrade performance. Instead of finetuning the DPR retriever separately, our experiments show that finetuning it as a part of the RAG-end2end mechanism gives better overall results. Our results also show that using the auxiliary signal improves both the retriever component and the overall accuracy.

In addition, we open-source the implementation of *RAG-end2end* with the HuggingFace Transformers (Wolf et al., 2019) Library,[1] providing the opportunity for the scientific community to use/test/build on our work.

2 Background and Related Work

Open-domain QA systems (Yang et al., 2015; Kwiatkowski et al., 2019) generally have a two-stage pipeline: passage retrieval (i.e., finding relevant text chunks related to an input question from a knowledge base) and machine comprehension (i.e., generating an answer from a set of selected documents). Traditionally sparse vector methods such as TF-IDF and BM25 are used for document retrieval (Robertson and Zaragoza, 2009). Researchers have recently moved to use dense text representations, which allows modeling textual similarity more semantic level. A recent example is the DPR (Karpukhin et al., 2020), which generates embeddings for questions and text passages using two BERT (Devlin et al., 2018) models. The dot product of the embeddings is used as a similarity score between a question and a passage. DPR has demonstrated that higher retrieval precision results in a higher end-to-end QA accuracy. For the answer generation component of QA systems, recent studies have used either extractive language models like BERT or generative language models like BART/GPT-2 (Min et al., 2021; Lewis et al., 2021).

2.1 Retrieval Augmented Architecture

Recently, Retrieval Augmented Architectures (Lewis et al., 2020b; Guu et al., 2020) have drawn considerable attention due to their explainable,

---

[1]Huggingface Transformers implementation.

scalable, and adaptable nature. Unlike other open-domain QA architectures, RAG (Lewis et al., 2020b) combines the information retrieval stage and answer generation stage in a differentiable manner. It uses a combination of parametric and non-parametric memory, where the parametric memory consists of a pre-trained seq2seq BART (Lewis et al., 2019) generator, and the non-parametric memory consists of dense vector representations of Wikipedia articles indexed with the FAISS library (Johnson et al., 2017). RAG first encodes a question into a dense representation, retrieves the relevant passages from an indexed Wikipedia knowledge base, and then feeds them into the generator. The loss function can finetune both the generator and the question encoder at the same time. Lewis et al. (Lewis et al., 2020b) highlight RAG's ability to perform well in Wikipedia-based general question-answering datasets like Natural Questions (Kwiatkowski et al., 2019). Other recent work also highlights how the outputs generated from RAG models are much more factual due to RAG being conditioned on the retrieved documents, possibly providing an answer to the hallucination problem of generative language models. Shuster, Kurt, et al. (Shuster et al., 2021) also highlight how RAG reduces hallucinations in knowledge-grounded conversational tasks, where the task is to generate responses to dialogues based on a large Wikipedia knowledge base. Xu et al. (2021) illustrate the effectiveness of RAG in chatbot frameworks and highlight how RAG models are able to recall and summarize conversations compared to standard seq2seq models with only parametric memory. This paper aims to understand how RAG could be extended to an end2end model and adapted to specific domains. To the best of our knowledge, this is the first time RAG is being investigated on domain adaptation for the task of ODQA systems.

2.2 REALM-like end2end Retrieval Augment Architectures

REALM (Guu et al., 2020) is a similar Retrieval Augmented model to RAG. REALM introduced a novel masked language pre-training step that involves an end-to-end trainable retriever. In the REALM work, the authors first train the entire model on the masked language prediction task and then finetune it on question-answering tasks (keeping the retriever frozen). In comparison

to REALM, the original RAG model uses an already-trained DPR retriever and conducts partial end-to-end training with a BART reader model. Compared to REALM, RAG is less computationally expensive, and its code is available open-source. We explore and extend the original RAG architecture for domain adaptation in our work. We adapted some concepts of our RAG-end2end extension from REALM. REALM only updates its retriever during the pre-training process that uses the masked language modeling (Devlin et al., 2018) task. Then during the downstream finetuning task, REALM keeps its retriever fixed. However, the REALM end-to-end training code is not open-sourced, possibly due to its computational complexity. Compared to REALM, RAG is a combination of already pre-trained language models where the users do not need to go through a heavy pre-training stage. Due to these engineering-friendly features and high availability, we conducted our experiments with RAG and extended RAG into an end-to-end trainable retrieval augmentation model. It is also important to highlight that none of the prior work has explored the domain adaptation of retrieval augment models for question answering; instead, most focus on general question answering with Wikipedia-based knowledge bases.

Similar to REALM's end2end architecture, recent work (Sachan et al., 2021) extended RAG and highlighted that the retriever training could improve the overall performance in question-answering datasets like Natural Questions. Compared to our work, the authors did not focus on the domain adaptation of retrieval augment models. The authors mainly explore the ability to train neural retrievers in an end-to-end way using retrieval augment models. Similarly, another related work (Singh et al., 2021) extended retrieval augmented architectures to an end-to-end model and illustrated that it could improve the question answering accuracy. Singh et al. (2021) mainly focused on improving the document reading ability and answer generation rather than domain adaptation.

3 Model Architecture and Training Procedure

In this work, we extend RAG to finetune all components, including the DPR retriever, and dynamically update the external knowledge base
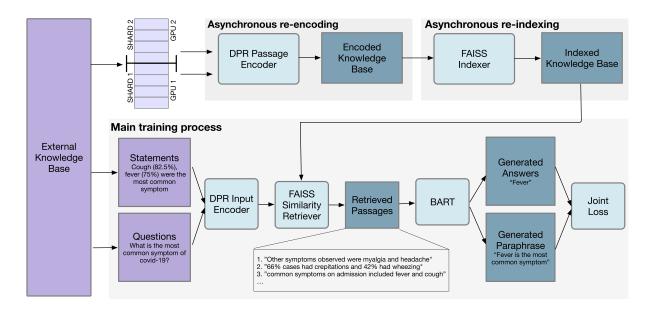
Figure 1: System Overview. Our RAG-end2end training architecture uses asynchronous processes to dynamically re-encode and re-index the knowledge base while optimizing a joint QA and paraphrasing signal loss. The training dataset consists of both reconstruction signals and QA pairs. The network learns to generate answers to questions and useful statements jointly. The input to the BART reader is illustrated in Equation 3, where the model can differentiate the answer generation task and statement reconstruction task with the use of a control token. During the training, embeddings and the knowledge base index get updated asynchronously.

during training. We hypothesize that the use of asynchronous updates helps with domain adaptation. Figure 1 demonstrates the main workflow of our model. In the following sections, we describe our extensions and training signals.

3.1 RAG Retriever and Generator

The retriever is a DPR (Karpukhin et al., 2020) model pre-trained on Wikipedia-based question-answering datasets (Kwiatkowski et al., 2019; Joshi et al., 2017). It consists of two tower BERT-based networks: the Question Encoder (E_Q) and the Passage Encoder (E_P). We use their CLS token embeddings as representations for questions and passages. The similarity between a question (q) and a passage (p) is calculated by taking the dot product of the two embeddings, as shown in Equation 1.

$$sim(p, q) \propto E_Q(q)^T E_P(p). \qquad (1)$$

RAG's generator consists of a pre-trained BART (Lewis et al., 2019) seq2seq language model. To train these retriever and generator components, RAG enhances the traditional sequence-to-sequence cross-entropy loss function by setting the retrieved passages as a latent variable (Z) (Guu et al., 2020; Lewis et al.,

2020b). The loss value of generating each token is marginalized on the probability of selecting documents given a context X (i.e., Document Score $p(Z|X)$). The formula (RAG-Token-Loss) can be written as illustrated in Equation 2.

$$P_{RAG-Token-Loss}(y|x) =$$
$$\prod_i^n \sum_{z \varepsilon_{top-k} P(.|x)} P_\eta(z|x) P_\theta(y_i|x, z, y_{1:i-1}). \quad (2)$$

3.2 Indexing of the External Knowledge Base

Before the training phase, we need to encode all passages in the external knowledge base using E_P. Then we need to retrieve similar passages from the external knowledge base given the output from E_Q. This process mainly involves dot product calculation between input question embeddings and encoded passages. The retrieval process will likely result in a performance bottleneck during the training since there are usually millions of passages in the knowledge base. To address this issue, RAG adopts the FAISS indexing approach proposed in (Johnson et al., 2017). With the help of the indexes, we can skip a considerable amount of repeated computation and significantly accelerate the retrieval process.

4

3.3 End-to-End Retriever Training

Although the DPR module makes use of two BERT models (E_P, E_q), the original RAG architecture only finetunes the question encoder E_Q in the retriever. The passage encoder E_P and the external knowledge base's encoding are fixed during the training phase. In other words, the pre-trained passage encoder of DPR is only used once to encode the external knowledge base. The RAG authors suggest that such a design performs well for Wikipedia-based ODQA datasets (Kwiatkowski et al., 2019; Joshi et al., 2017). Such settings work because the DPR model was also pre-trained with Wikipedia-based datasets, and their experiment uses an external knowledge base consisting of Wikipedia articles.

However, it may be beneficial to finetune all the DPR components during RAG training for domain adaptation since the model needs access to different domain-specific external knowledge bases. In this work, we introduce *RAG-end2end*, where we augment RAG to be fully end-to-end trainable. We finetune the passage encoder and question encoder and then update the index of the external knowledge base during the training process.

It is straightforward to propagate gradients to both the passage and question encoders with RAG's loss function because this loss function employs the passage selection probability known as doc-score ($p_\eta(z|x)$ term illustrated in Equation 2). However, for it to have a true effect on the overall model training process, we have to iteratively update the embeddings with the updated context encoder and then update the index of the external knowledge base. In other words, we need to re-encode and re-index the knowledge base using the updated passage encoder. When the external knowledge base possesses tens of millions of passages, the re-encoding and re-indexing steps can be very time-consuming. Re-encoding can take several GPU hours, and re-indexing with FAISS can take several CPU hours, depending on the size of the knowledge base. Therefore, it is inefficient to stall the training loop while the re-encoding re-indexing steps are being carried out.

To have an efficient training mechanism, we designed our training framework into three main processes: (1) The main training loop, which updates the gradients, (2) Re-encoding processes with several GPUs that update the knowledge-base encoding with the updated DPR's context encoder, and (3) A Re-indexing process that uses FAISS to build an index with the updated encoding. Figure 1 illustrates these three processes. Our implementation uses two asynchronous processes to re-encode and re-index the external knowledge base that runs independently to the main training loop. We first distribute the external knowledge base to a set of GPUs that are not used in the main training loop. Then we encode the passages with an updated passage encoder, which we call the re-encoding process. Once the re-encoding process has finished, we re-index the knowledge base in another parallel process that uses FAISS (re-indexing process). Inside the main training loop, we ensure that the re-indexing process always starts after finishing the re-encoding process. Then as soon as the new index of the external knowledge base is created, we load that to the main training loop. Once the new index loading is completed again, we start the re-encoding process, which repeats the entire embedding updating process. It is important to note that the first re-encoding process should be finished, and new embeddings should get saved to the hard disk, before the start of the FAISS indexing process. If the knowledge base is not entirely updated with the new embeddings, the re-indexing process fails. We use Python multiprocessing handles to keep the order, and re-indexing and re-encoding processes are only asynchronous with respect to the main training loop process. The number of steps between each re-encoding process depends on the size of the dataset. To test the number of steps between the knowledge-base updates, we experimented with a knowledge base consisting of 250,000 passages and used four dedicated GPUs for the re-encoding process with a batch size of 32 each. Our computation machine consists of 96 CPU cores. We found that it takes an average of 750 updates. However, the computation time can be easily improved when using more GPUs for encoding and using a machine with a higher number of CPU cores (FAISS indexing process depends on the number of CPU cores). These steps are repeated throughout the training loop. Since the training and knowledge base's index update processes are running asynchronously, it may result in stale gradients. This, however, does not significantly degrade the model performance according to previous research (Guu et al., 2020).

5

3.4 Statement Reconstruction

We explore the incorporation of statement reconstruction as an auxiliary signal assuming that it forces the model to gain more domain-specific knowledge. As illustrated in Figure 1, we first encode input statements using the input/question encoder (E_Q). Then the retriever retrieves the most similar set of passages from the indexed external knowledge base by conducting a similarity search. Afterward, the final output generator attempts to re-construct the input statements using only the selected support set of documents. We ensure that the external knowledge base does not contain the input statement to prevent the model from over-fitting on just the lexical content. To differentiate the paraphrasing signal from the QA signal, we prepend a special token $<p>$ (represents passages) in front of the reconstruction statements, which acts as a control token in the seq2seq language modeling (Raffel et al., 2019; Keskar et al., 2019). Concretely, when training the RAG architecture on QA pairs, the questions are prepended to the retrieved passages before being fed to the BART generator. As illustrated in Equation 3, for the input reconstruction signal, we only prepend the $<p>$ token to the retrieved passages before feeding them to the BART generator.

$$
\begin{aligned}
\text{QA Input:} \quad &\langle \text{Question} \rangle + \langle \text{Retrieved Passages} \rangle \\
\text{Reconstruction Input:} \quad &\langle \text{<p>} \rangle + \langle \text{Retrieved Passages} \rangle
\end{aligned} \quad (3)
$$

4 Experiments and Results

4.1 Domain-specific Dataset Setup

In this work, our main intention is to explore the adaptation of domain-specific retrieval augmentation with regard to ODQA. As mentioned in recent work (Lewis et al., 2020b), most ODQA datasets like Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), Web-Questions (Berant et al., 2013), and CuratedTrec (Baudiš and Šedivỳ, 2015) are answered with Wikipedia-based knowledge-bases. Since neural retrievers like DPR are already trained with Wikipedia-based datasets, it is hard for us to explore the domain adaptation of RAG fairly in this setting. Therefore, we selected three domain-specific datasets for our experiment: COVID-19 QA, News QA, and Conversation QA. Since the availability of domain-specific ODQA datasets is minimal, in our work, we open-source all domain-

specific knowledge-bases and question-answer pairs to support future research.[2]

COVID-19 QA Domain

Knowledge Base Generation: To create the external knowledge base, we use 5,000 full-text scientific articles extracted from the CORD-19 (Wang et al., 2020) dataset. The external knowledge base is created with 250,000 100-word passages. Each passage is pre-pended with the title of the research paper.

Reconstruction Statement Generation: We use sentences from the abstract section of research articles for the reconstruction signal. We first extract the abstract sections in 10K papers and split them into sentences using the NLTK library (Loper and Bird, 2002). We filter out the sentences that are too short (less than 15 words) or too long (more than 35 words). In this process, approximately 50,000 abstract statements are generated. It is important to note that when generating the knowledge base, we exclude the abstract sections.

Synthetic QA Generation: In this domain, we only use synthetic data for training and validation. Following prior work (Shakeri et al., 2020), we use a BART seq2seq model trained on the SQuAD dataset (Rajpurkar et al., 2016) to generate synthetic QA pairs given a passage. We used the Squad dataset's passages as the input and corresponding question-answer pairs as the expected output. We trained a BART-large checkpoint for two epochs. Then, we followed round-trip consistency (Alberti et al., 2019) to filter synthetic QA pairs. Our final synthesized QA dataset consisted of 225,000 QA pairs. We use 90% of these QA pairs as training data and 10% as validation data. As the test data, we use 2000 human-labeled question-answer pairs from the COVID-QA dataset (Moller et al., 2020).

News QA Domain

Knowledge Base Generation: We extract 85,000 100-word passages as the knowledge base using 10,000 news articles from the NewsQA dataset (Trischler et al., 2016).

Reconstruction Statement Generation: We extract corresponding news summary sentences

---

[2]Domain-specific datasets.

from the CNN/DM dataset (Hermann et al., 2015) for the reconstruction signal. Every article consists of more than one summary sentence. However, we use the first sentence as the title of the article, which we used in knowledge base generation and the rest of the statements as reconstruction statements. Our final dataset contains 35,000 summary statements.

QA Generation: The NewsQA dataset (Trischler et al., 2016) consists of 100,000 human annotated QA pairs from 10,000 news articles from the CNN/DM dataset (Hermann et al., 2015). We use the train (90,000), valid (5,000), and test (5,000) splits given in the dataset to train and evaluate our model. All questions in the NewsQA dataset focus on the high-level content of articles. So, to answer these questions, the model must access a large span of passages to conduct the reasoning process.

Conversation QA Domain

Knowledge Base Generation: We create the external knowledge base of 110,000 passages by splitting the 10,000 conversations given in the QAConv dataset (Wu et al., 2021b) into passages, each with at most 100 words. We prepend the identifier of each conversation (found in the original dataset) as the title of the passages. We also appended the speaker's name, followed by the '':'' symbol, to the starting position of each dialogue to keep each conversation connected to its speakers.

Reconstruction Statement Generation: We use the state-of-the-art abstractive conversation summarization model[3] (Wu et al., 2021a) to generate one-sentence (TLDR) summary (approximately 45 words per conversation). We then use this as the auxiliary signal. We only generate summaries of conversations with more than 45 words. By doing this, we collect 35,000 synthetic summary/ reconstruction statements.

QA Generation: We use the QAConv dataset (Wu et al., 2021b), which contains 35,000 QA pairs generated from 10,000 conversations that involved two or more parties. We use the train (25,000), valid (5,000), and test (5,000) splits given in the dataset to train and evaluate our model.

---

[3]Salseforce checkpoint.

4.2 Training and Evaluation Setup

We use the HuggingFace-Transformers (Wolf et al., 2019) library to implement the RAG-end2end architecture. We initialize the DPR and BART models of using the open-source Hugging-Face checkpoints.[4] Prior to finetuning, we index and encode the external knowledge base using FAISS. We select HNSW FLAT as the indexing mechanism (with 128 bi-directional links). We use 100 words as the maximum passage length as suggested by the prior RAG work (Lewis et al., 2020a). During training, we use six Tesla V100 GPUs with 32 GBs of memory. Four of them are used for training, and two are used for re-encoding. We train each RAG model variant 4.2 for ten epochs and select the final checkpoint with the highest validation accuracy.

We use the Exact Match (EM), F1 score, and Top-k retrieval accuracy as evaluation metrics. The EM score computes the word-level exact match between the predicted answer and the real answer. The F1-score calculates the number of words in the predicted answer that are aligned with the real answer regardless of the order. The Top-k retrieval accuracy is calculated by matching the answer strings with the contents of the retrieved k passages.

We compare RAG and RAG-end2end in the following five scenarios.

1. **RAG-original**. This model is finetuned on the natural question dataset (Kwiatkowski et al., 2019) with the Wikipedia knowledge base and serves as the non-domain adapted baseline.[5] This model is not finetuned with domain-specific question-answer pairs, and we report the zero-shot performance.

2. **RAG-original-QA**. This is the original RAG model finetuned with only domain-specific question-answer pairs.

3. **RAG-end2end-QA**. This is the RAG model with our end2end retriever extensions and finetuned only with domain-specific question-answer pairs.

4. **RAG-original-QA** $+$ **R**. This is the RAG original model finetuned with both

---

[4]RAG-token-base checkpoint.
[5]Public RAG-token-nq checkpoint.

Model Name	EM	F1	Top-5	Top-20
COVID-19 Domain				
(1) RAG-original	0.0	4.73	10.56	15.69
(2) RAG-original-QA	2.95	12.01	12.29	18.43
(3) RAG-end2end-QA	8.08	18.38	19.85	26.91
(4) RAG-original-QA+R	3.66	12.20	12.79	18.45
(5) RAG-end2end-QA+R	8.32	19.57	23.05	31.23
News Domain				
(1) RAG-original	4.33	7.92	19.46	30.33
(2) RAG-original-QA	7.26	14.26	22.86	34.55
(3) RAG-end2end-QA	8.39	16.31	28.89	41.20
(4) RAG-original-QA+R	8.62	16.46	27.28	39.56
(5) RAG-end2end-QA+R	14.08	23.7	39.67	50.95
Conversation Domain				
(1) RAG-original	5.49	9.27	12.14	20.02
(2) RAG-original-QA	12.09	20.05	22.73	32.05
(3) RAG-end2end-QA	24.25	36.05	46.01	55.55
(4) RAG-original-QA+R	14.21	24.62	26.32	36.21
(5) RAG-end2end-QA+R	25.95	37.96	49.11	58.75

Table 1: Domain adaptation performance of different RAG models used in our experiments. We illustrate the results related to all three domains. Details about each model are described in Section 4.2.

domain-specific question-answer pairs and our reconstruction signal.

5. **RAG-end2end-QA + R**. This is the RAG model with our end2end retriever extensions and trained with both question-answer pairs and our reconstruction signal.

We present the results of each scenario in Table 1.

4.3 Effect of End-to-End Retriever Training on Domain Adaptation

We first test if finetuning both the passage encoder and question encoder of the RAG's retriever while updating the external knowledge base would improve domain adaptation. We compare the performance of RAG-original-QA and RAG-end2end-QA, isolating any performance improvement due to the reconstruction signal. The results in Table 1 illustrate that RAG-end2end-QA significantly outperforms RAG-original-QA on all metrics—EM, F1, Top-5, and Top-20—across all three domains. The improvements in the EM score varied from 1.13 points in the News domain to 12.16 points in the Conversation domain.

Evaluating the performance of passage retrieval using Top-5 and Top-20 scores, we see a

marked increase of around 25 points in the conversation domain, with the other domains showing improvements of between 4.7 to 6.6 points.

Above all, these results suggest that finetuning both the passage and question encoders of the RAG's retriever while updating the external knowledge base can improve domain adaptation.

4.4 Effect of Adding the Statement-Reconstruction Auxiliary Task

In this experiment, we test our next hypothesis: Adding the auxiliary training signal of statement reconstruction along with QA pairs improves domain adaptation. We compare the performance of RAG-end2end with and without the reconstruction signal by comparing the performance of RAG-end2end-QA + R and RAG-end2end-QA in Table 1. This shows that RAG-end2end-QA + R outperforms RAG-end2end-QA for all three domains. The range of increases in the EM scores varied from 1.7 points in the Conversation domain to an 8.39 points increase in the News domain. The top-20 retrieval accuracy also increased in a range between 3.2 to 8 points.

We further compare the effect of adding the reconstruction signal to RAG-original by comparing RAG-Original-QA with RAG-Original-QA + R. We find that even without the end2end extension, the reconstruction signal improves the performance moderately. This improvement in the EM score ranged from 0.84 points in the COVID-19 domain and 3.12 points in the Conversation domain.

Finally, we highlight the overall improvement of our contributions by comparing RAG-Original-QA with RAG-end2end-QA+ R. As the most significant improvement; we highlight the 13-point improvement of EM score for the Conversation domain. For retrieval performance, we highlight the 27-point improvement in the top 5 and 16 improvements in the top 20 for the Conversation domain.

To demonstrate the reconstruction statement generation, we provide an example of the generated reconstruction output of given a statement for each domain using the RAG-end2end-QAR + R model in Table 2. The second column contains the input statements with the special token $< p >$, the third column shows a snapshot of retrieved top-5 documents, and the final column shows

8

Domain	Input Statement	Most Similar Retrieved Document	Reconstructed Statement
COVID-19	\<p\>Cough (82.5%), fever (75%), and malaise (58.8%) were the most common symptoms, and crepitations (60%), and wheezes (40%) were the most common signs.	Most common signs and symptoms on admission included fever and cough. Of all children, 32% had complaint of difficulty in respiration. Other symptoms observed were myalgia, headache and vomiting. On examination, 66% cases had crepitations and 42% had wheezing. Hypoxemia was observed in 31% cases at admission	The most common signs and symptoms on admission were fever and cough, and 32% had complaint of difficulty breathing
News	\<p\>Capsule was carrying South Korea's first astronaut.	MOSCOW, Russia Russian space officials say the crew of the Soyuz space ship is resting after a rough ride back to Earth. A South Korean bioengineer was one of three people on board the Soyuz capsule. The craft carrying South Korea's first astronaut landed in northern Kazakhstan on Saturday, 260 miles off its mark, they said.	Soyuz capsule carrying South Korea's first astronaut lands 260 miles off its mark.
Conversation	\<p\>The Supreme Court will hear the case on the grounds of First Amendment protection of free speech.	(PETITIONER): Yes, Your Honor. Mr. Tory, who was appearing pro se in the trial court, from the very outset objected that he was being held liable for speech protected by the First Amendment. \<end\>$	Justice Souter and MR. CHEMERINSKY are arguing that the injunction in the case of Tory should not be applied

Table 2: Examples of reconstructed statements. Reconstructions generally capture the context of the retrieved documents and are similar to the input statement but are not always factually 100% correct (e.g., COVID-19 example). Input statement column shows the input to the model with the special \<p\> token. The Retrieved Documents shows a snap-shot of the top-retrieved document used to re-construct the statement.

the re-constructed statements. As the reconstruction statements demonstrate, we highlight that the model can generate statements close enough to the input.

4.5 Retriever's Domain Adaptation with RAG-end2end

An important part of our RAG-end2end extension is updating the entire DPR retriever during training. Previous work (Ma et al., 2020) has explored the importance of the domain adaptation of neural retrievers and highlighted the performance gains in domain-specific retrieval tasks. We argue, based on our above-mentioned RAG end2end's retriever performances and prior work, that when adapting RAG to various domains, having a domain-specific retriever plays a key role in achieving good performance. However, this end-to-end RAG finetuning can get computationally costly, especially with the number of passages in the external knowledge base where they should get re-encoded and re-indexed. Instead of fine-tuning DPR as a part of RAG-end2end, an alternative approach is to finetune DPR on domain-specific data separately on its vector similarity-based loss function (Karpukhin et al., 2020) and then initializing the RAG architecture prior to finetuning with the QA data. We explore if RAG-end2end can perform on par if we initialize a RAG model with an independent domain-adopted DPR model. This helps us further understand the ability of the RAG-end2end extension to finetune the retriever with domain-specific data.

Standalone DPR Finetuning with Domain-specific Data

The standalone DPR can be finetuned if we have access to gold-standard passages that contain the answers for given questions and hard negative passages that consist of similar details to the question but not the exact answers. DPR uses a dot-product-based similarity loss, capturing the similarity between the correct passage

9

for the question while comparing with some hard-negative examples (Karpukhin et al., 2020) (which are lexically similar but do not contain the answer) (Karpukhin et al., 2020). We use the deep-haystack framework[6] to finetune DPR for each domain using domain-specific data. We created finetuning datasets for all three domains. First, for the COVID-19 domain, we utilized the synthetic question-answer pairs and their relevant passages that consist of 100 words. The use of domain-specific synthetic QA pairs for DPR finetuning has already shown permanence improvements (Ma et al., 2020). For hard-negative examples, we used BM-25 lexical matching search as mentioned by the DPR authors, where we retrieved passages that do not contain the answer based on their lexical similarity with the question. Although for the News domain and the Conversation domain, we have a supervised dataset where we can map questions into the correct passage, we did not get better results after finetuning the original DPR using the supervised data. The main reason for the degradation of performance is the length of the correct passage related to the question. In both News and Conversation domains, most of the questions come from longer passages, whereas the pre-trained DPR only accepts 100-word passages. To mitigate this issue, we generated synthetic question-answer pairs with the external knowledge bases of news and Conversation domains similar to the COVID-19 domains by following the same procedure mentioned in Section 4.1. Then the hard-negative examples were also mined according to the above-mentioned BM-25 lexical matching method. After training, we evaluate the DPR retrieval accuracy using the test dataset and external knowledge base for each domain, similar to the RAG's retrieval evaluation we conducted in Section 4.2.

Table 3 compares **(1) DPR-original**, which is the publicly available checkpoint[7] trained on Wikipedia, with **(2) DPR-domain-adapted**, which is the finetuned model with DPR's original loss function. The **(3) DPR-RAG-end2end** is the retrieval part of RAG-end2end-QA + R from Table 1 for comparison. We include the DPR-RAG-end2end model to highlight the improvement of the DPR model as a result of

---

[6]deepset-ai.
[7]DPR-checkpoint.

Model Name	Top-5	Top-20
COVID-19 Domain		
(1) DPR-original	9.39	14.72
(2) DPR-domain-adapted	13.66	20.01
(3) DPR-RAG-end2end	20.64	28.21
News Domain		
(1) DPR-original	20.95	31.04
(2) DPR-domain-adapted	20.98	31.92
(3) DPR-RAG-end2end	39.67	50.95
Coversation Domain		
(1) DPR-original	15.15	23.95
(2) DPR-domain-adapted	23.15	34.53
(3) DPR-RAG-end2end	49.11	58.75

Table 3: Comparison of DPR models finetunned on domain specific data against publicly available DPR checkpoint, which is trained on Wikipedia domain for all three domains.

RAG-end2end training with both training signals. When comparing the DPR-RAG-end2end model with the other variants in Table 1, we observe that the RAG-end2end architecture significantly improves the DPR's domain adaptation for all three domains. Therefore, in future work, RAG-end2end could be used as a way to train a neural retriever, which could benefit even for retrieval-only applications.

As shown in Table 3, we observe that finetuning DPR models on the original DPR loss function using domain-specific data improves the overall retrieval performance for each domain. For the COVID-19 and Conversation domains, there's a clear improvement in the top-5 and top-20 retrieval accuracies. We observed almost the same results for the News domain compared to the original DPR. This could be due to similar kinds of data in Wikipedia, which were originally used to train the DPR and CNN/DM text.

Overall, as illustrated in Table 3, we highlight the fact that the RAG-end2end's loss function has the ability to adapt the DPR to specific domains better than finetuning the DPR with the passages and question pairs for each domain. The improvements for all three domains in top-5 and top-20 retrieval accuracies of DPR-RAG-end2end compared to DPR-original and DPR-domain-adapted is noticeable. These results further highlight the

Model Name	EM score	F1 Score	Top-5	Top-20
COVID-19 Domain				
(1) RAG-original-QA+R	3.66	12.12	12.79	18.45
(2) RAG-original-QA+R (DPR-adapted)	7.36	17.91	22.39	30.87
(3) RAG-end2end-QA+R	8.32	19.51	23.05	31.23
News Domain				
(1) RAG-original-QA+R	8.62	16.46	27.28	39.56
(2) RAG-original-QA+R (DPR-adapted)	10.92	19.44	30.72	41.9
(3) RAG-end2end-QA+R	14.08	23.7	39.67	50.95
Conversation Domain				
(1) RAG-original-QA+R	14.21	24.62	26.32	36.21
(2) RAG-original-QA+R (DPR-adapted)	15.78	25.47	29.01	40.03
(3) RAG-end2end-QA+R	25.95	37.96	49.11	58.75

Table 4: Comparing the effect of RAG-end2end extension, against initializing RAG-original models with domain adapted DPR models prior to the finetuning (please review Table 1). We use the independently domain-adapted DPR models illustrated in Table 3.

ability of RAG-end2end to finetune or improve its retriever.

Initializing RAG with Domain-adapted DPR Prior to Finetuning

Next, we investigate the performance of RAG models when initialized with a domain-adapted DPR. We initialize RAG's question encoder and the passage encoder with DPR-domain-adapted (from trained models illustrated in Table 3) and finetune RAG with the settings of RAG-original-QA+R. The objective is to compare how the RAG models initialized with domain adopted DPR models perform in comparison to using the RAG-end2end extension.

Table 4 demonstrates results from four models. (1) RAG-original-QA+R and (3) RAG-end2end-QA+R are taken from the main results (Table 1). The **(2) RAG-original-QA+R (DPR-adapted)** model was first initialized with a domain-adopted DPR model (from Table 3) before being finetuned with domain-specific QA pairs and re-construction signals with the RAG-original settings.

The results in Table 4 indicate that for all domains, finetuning the RAG-original with a domain-adapted DPR gives higher performance than finetuning the RAG-original with the usual DPR model checkpoint (compare (1) and (2) in Table 4). We highlight the performance improvements for both answer generation accuracy and retrieval recall scores, where the COVID-19 do-

main has the largest improvements. We also compare the finetuning *RAG-end2end* model with the RAG-original model, which was first initialized with the domain-adapted DPR models (compare (2) and (3) in Table 4). This comparison shows that *RAG-end2end* training mechanism can outperform the RAG-original mechanism that uses a domain-adapted DPR. The results further highlight the importance of the *RAG-end2end* mechanism in domain adaptation where we do not need to train the DPR model separately.

5 Discussion

5.1 Role of Retriever in Domain Adaptation

As the results suggest, the retriever plays an essential role in domain adaptation for open-domain QA. It is clear that *RAG-end2end* training improves the results since it can update the embeddings and the indexing of the knowledge base. Compared with the original RAG finetuning, *RAG-end2end* improves the performance in all datasets. The main reason for this could be that neural retrievers such as DPR, which are trained on public datasets, struggle to perform well on domain-specific datasets. Our results also highlight an important aspect related to the performance of the stand-alone DPR for document retrieval. It shows that *RAG-end2end* can improve the domain adaptation of DPR better that finetuning the DPR on its own mechanism.

5.2 Cost of end2end Retriever Adaptation

It is important to note that *RAG-end2end* fine tuning can be expensive if the number of passages in the external knowledge base is large. If there are millions of passages, it would be beneficial to have a dedicated number of GPUs that complete the re-encoding process. Re-indexing with the FAISS library also depends on the number of cores in the CPUs. When having access to strong enough computational power, it is better to use *RAG-end2end* since we can directly use passages in a knowledge base and question-answer pairs to train both the retriever and the reader. Then we also do not need to generate synthetic question-answer pairs related to passages that are required to train the DPR.

Although the RETRO (Borgeaud et al., 2021) authors claim that frozen BERT embedding is

11

Model Name	EM	F1	Top-5	Top-20
SQUAD Open-Domain				
(1) RAG-original	28.12	39.42	59.64	72.38
(2) RAG-end2end	40.02	52.63	75.79	85.57

Table 5: Open-Domain performance comparison between RAG-original and RAG-end2end. We used SQAD dataset in ODQA manner to conduct our experiments.

sufficient for retrieval augmented models, our results suggest that for domain-specific models to perform well, a domain-adapted retriever component is beneficial. In future work, it will be important to explore how the models like RETRO (Borgeaud et al., 2021) perform on domain-specific scenarios going beyond general-purpose datasets.

5.3 Comparing the RAG-original with RAG-end2end on an In-domain Dataset

Although our work is mainly focused on the domain adaptation of RAG for specific domains, we also explored whether the end2end training would improve the overall results of an in-domain dataset. Since the original RAG model uses a DPR model that is trained on a Wikipedia-based Natural Questions dataset, we consider this in-domain. Although the SQUAD (Rajpurkar et al., 2016) dataset is a machine comprehension dataset, we adapted the SQUAD dataset to perform ODQA. First, we extracted the contexts related to each question-answer pair and created an external knowledge base. Then we split the knowledge base into 100-word passages. Our final knowledge base consists of 30K passages. As illustrated in Table 5, we compared the performance of RAG-original and *RAG-end2end* on the tasks of answer generation and retrieving correct documents. As the results suggest, *RAG-end2end* performs better than RAG-original even in other Wikipedia-based datasets. This could be due to *RAG-end2end* updating the context encoder and embeddings during the training process.

6 Conclusion and Future Work

In this paper, we proposed a novel extension of RAG: *RAG-end2end*, which, unlike RAG,

does joint training of the retriever and generator for the end QA task and domain adaptation. We showed that *RAG-end2end* could improve DPR performance better than finetuning the DPR independently. This allows for the training of DPR models with QA pairs and eliminates the need for gold-standard passages related to questions. We also highlighted that the addition of a re-construction auxiliary signal further improves both the retriever and the final answer generation accuracies. We evaluate our approach with three datasets from different domains (COVID-19, News, and Conversations), showing that *RAG-end2end* achieves significant performance improvements in all three domains compared to the original RAG implementation. In addition, we conducted several other experiments to validate our approach comprehensively. Overall, our results show that our approach is stable and generalizable across different domains. Our experiments highlight the importance of the RAG's retriever component in domain-specific question answering.

Based on our findings, we suggest three directions for future research in domain adaptation of RAG models. First, we consider it worthwhile to explore *RAG-end2end* on other tasks like fact checking (Lewis et al., 2020b), summarization (Shuster et al., 2021), and conversational response generation (Xu et al., 2021) where the original RAG has shown interesting results. Second, it is important to explore generative capabilities with qualitative metrics (Figure 2 in the appendix illustrates retrieved text and answers generated by rag-end2end and rag-original). This could be aligned with research areas like measuring factual consistency (Kryściński et al., 2019; Cao et al., 2022) and hallucinations (Cao et al., 2022; Shuster et al., 2021; Nie et al., 2019) of generative language models. Future work could explore whether updating the retriever and document embeddings during the training phase could improve factual consistency and reduce hallucinations in final generations. Third, the improvement of RAG with our extension (*RAG-end2end*) highlights the importance of the retriever in the RAG architecture, which motivates us to improve the retriever part further in future work. Also, as the statement re-construction signal acts as a good auxiliary signal, we encourage exploring other auxiliary signals, which could improve the overall performance of RAG models.

A Appendix

Test Set Question	RAG-original retrieved passages	RAG-end2end retrieved passages	Label	RAG-original prediction	RAG-end2end prediction
Where does the Kiwi girl that Darren T Maloney spoke to on the phone commute from ?	PETER KENYON, BYLINE: MARY O'CONNOR, It will be very hard, apart, you know, to kind of ensure free travel. And we have got used to it, I mean, I remember the time when you traveled to Northern Ireland when there was two and three hours' wait, you know, over and back across the border, And I wouldn't like to see that happen, I don't think it would help trade. <end> PETER KENYON, BYLINE: But on this day, at least, the economic arguments weren't entirely persuasive, As Bernadette Wray finishes her lunch, she says she's got nothing against immigrants, being one PETER KENYON, BYLINE: MARY O'CONNOR, It will be very hard, apart, you know, to kind of ensure free travel. And we have got used to it, I mean, I remember the time when you traveled to Northern Ireland when there was two and three hours' wait, you know, over and back across the border, And I wouldn't like to see that happen, I don't think it would help trade. <end> PETER KENYON, BYLINE: But on this day, at least, the economic arguments weren't entirely persuasive, As Bernadette Wray finishes her lunch, she says she's got nothing against immigrants, being one MARTINA DELVIN: Pleasure. <end> STEVE INSKEEP, HOST: She is an author and columnist in Dublin, really appreciate it, <end> MARTINA DELVIN: Pleasure. <end> STEVE INSKEEP, HOST: So by emphasizing the divide in this historically divided island, it would threaten the reignition of violence on the island there, That's what you're saying, <end> MARTINA DELVIN: That's right, And there were 30 years of violence, and they impacted on all sorts of ways on daily life and on people's ability to earn a living - led to emigration - all sorts, I mean, I grew up in Northern Ireland during The Troubles, And the - you know, the island was divided in the most arbitrary way, So you'll get, for example, a petrol station with	Darren T Maloney Susan Did I mention Wilson Coming back to Houstonmeneor for any length given his love for NYC Yup 6 weeks Mixed be everytime I get dug in I get moved Actually I am establishing a London base now and am happy with that but my project is on hold and I am a bit anxious The Kiwi girl sits next to me and has to commute from what seems like N Africa to get to work so she is sick of the commute Well that's all for now I will see you Darren T Maloney Susan Did I mention Wilson Coming back to Houstonmeneor for any length given his love for NYC Yup 6 weeks Mixed be everytime I get dug in I get moved Actually I am establishing a London base now and am happy with that but my project is on hold and I am a bit anxious The Kiwi girl sits next to me and has to commute from what seems like N Africa to get to work so she is sick of the commute Well that's all for now I will see you PatriceLMms Hey what's Happening Just wanted to let you know that today is my Birthday We're going yo go to the KiCi and Jo Jo Concert tonight at the Arena Theater should be very good Also did you hear from Robin that Anna's daughter Dana is in the hospital I'm going to call up there this afternoon and see how she's doing I'm assuming her last name is Brown Also I hooked Cynthia Patterson up with one of my customers A divorced brother from S Carolina Girlm she called me this morning and she was so giddy ... children must move to cities	N Africa	New Zealand	what seems like N Africa to get to work
Why Brian Kerrigan apologize to Deutsche bank?	JUSTICE GINSBURG: And the remedies being injunctive and declaratory. <end> MR. LAMKEN (PETITIONER): A— a form of specific relief. Generally they have the authority to effectively go in and revise the decision below, but the remedies ordinarily do not include monetary or compensatory relief I should say, <end> JUSTICE GINSBURG: Are you saying that it's parallel to what APA review of an agency decision would be? <end> MR. LAMKEN (PETITIONER): It's very much like that, The remand rule that this Court normally requires in the APA context is not so strictly observed in the context of — of review JUSTICE GINSBURG: And the remedies being injunctive and declaratory. <end> MR. LAMKEN (PETITIONER): A— a form of specific relief. Generally they have the authority to effectively go in and revise the decision below, but the remedies ordinarily do not include monetary or compensatory relief I should say, <end> JUSTICE GINSBURG: Are you saying that it's parallel to what APA review of an agency decision would be? <end> MR. LAMKEN (PETITIONER): It's very much like that. The remand rule that this Court normally requires in the APA context is not so strictly observed in the context of — of review JUSTICE GINSBURG: That's why one has precision because you are giving respect, full faith and credit, to a decision elsewhere. That's what precision doctrine is all about, We respect the judgment of the court that rendered it, We, therefore, give it full faith and credit, That's what precision doctrine is about, is about respect and credit. Isn't that so? <end> MR. CASTANIAS (RESPONDENT): That's — that's — that is — that is generally right, Justice Ginsburg, but at the same time, there — we all agree — Exxon Mobil, SABIC, and the decisions of this Court — there here has JUSTICE GINSBURG: That's why one has precision because you are giving respect, full faith and credit, to a decision elsewhere. That's what precision doctrine is about, is about respect and credit. Isn't that so? <end> MR. CASTANIAS (RESPONDENT): That's — that's — that is — that is generally right, Justice Ginsburg, but at the same time, there — we all agree — Exxon Mobil, SABIC, and the decisions of this Court — there here has JUSTICE GINSBURG: Then he — then he can publish his dissent, just as a Tax Court judge can? <end> MR. HUNGAR (RESPONDENT): No, but he can preclude the Tax Court judge from doing what the Tax Court judge did in this case, which is simply adopting his report, If the — if the special trial judge refuses to change his report — <end> JUSTICE GINSBURG: But then we still won't know what his report is, Yes, he can say, I won't sign this, Tax Court says, fine, This rule says I can reject your findings	Brian Kerrigan PRIVILEGED AND CONFIDENTIAL ATTORNEY CLIENT COMMUNICATION I spoke with Deutsche bank immediately after my conversation with Dan Lyons I apologize as I must not have communicated clearly that I understood the importance of this issue from a legal point of view My intent on having language proposed was not to concede to any proposed language by Deustche but to determine if it was possible to be assured of bringing in a significant commitment to the syndication of the transaction as well as not create any potential legal issues As this was not the case I told Brian Kerrigan PRIVILEGED AND CONFIDENTIAL ATTORNEY CLIENT COMMUNICATION I spoke with Deutsche bank immediately after my conversation with Dan Lyons I apologize as I must not have communicated clearly that I understood the importance of this issue from a legal point of view My intent on having language proposed was not to concede to any proposed language by Deustche but to determine if it was possible to be assured of bringing in a significant commitment to the syndication of the transaction as well as not create any potential legal issues As this was not the case I told Specifically the current definition of Indemnifiable Tax covers both present and future tax as evidenced in the definition Tax and thus necessarily should protect the counterparty in the case of a Change in Tax Law Moreover Deutsche Bank's amendment to Indemnifiable Tax potentially subjects us to risk that I don't feel is in our best interest to assume because we don't have any control over the counterparty's activities Bottom line I would reject the amendment Please let me know if you would like me to speak to their tax counsel Rhett Jackson EB 4680 7138534718 been intercepted or amended please tell us as soon as possible and You're right ofthpicitis has got me again Sorry end Paul Davis now really but Freud would forgive the association of thought from Cynthia's posting can you empbease earnest smiles Paul and Charles McCathieNevile Naturally you are forgiven Actually I think there are some potentially interesting legal ramifications if passing the cost of accessibility to the customer is found to be legal then it suggests a particular way to treat what for appeared to be already legally necessary accessibility improvements at least in Australia and the	not have communicated clearly	Deutsche Bank apologize for putting unnecessary stress on the U.S, housing market	not have communicated clearly that he understood the importance of this issue from a legal point of view
Which person made his most famous speech on the steps of the Lincoln Memorial?	Oliver Cromwell, who beheaded him, on the other. I don't know if you have to endorse one or the other. <end> GEN. ABBOTT (RESPONDENT): Well, Justice Kennedy, I believe that there is a very meaningful difference between this Court's standards of an endorsement and what a State or the nation may do with regard to commemoration. As an example, on the National Mall, there is, of course, the Lincoln Memorial and in the Lincoln Memorial, there is text from the King James version of the Bible, The nation commemorates and acknowledges Lincoln and what he has said, But by individual will be engaged in a — in a situation where they will be endangered, And I think that was certainly true of William Lloyd in the Totten case, When he crossed southern lines, he was very much endangered, and that's something that wasn't lost on President Lincoln. In footnote 3 of our opening brief, we have a quotation from President Lincoln about the inherent dangers of spies crossing lines and the need for secrecy to protect that, So that's why I think that claim is properly understood as not being covered by the Tucker Act and not being required individual will be engaged in a — in a situation where they will be endangered. And I think that was certainly true of William Lloyd in the Totten case. When he crossed southern lines, he was very much endangered, and that's something that wasn't lost on President Lincoln. In footnote 3 of our opening brief, we have a quotation from President Lincoln about the inherent dangers of spies crossing lines and the need for secrecy to protect that. So that's why I think that claim is properly understood as not being covered by the Tucker Act and not being required not being required individual will be engaged in a — in a situation where they will be endangered, And I think that was certainly true of William Lloyd in the Totten case, When he crossed southern lines, he was very much endangered, and that's something that wasn't lost on President Lincoln. In footnote 3 of our opening brief, we have a quotation from President Lincoln about the inherent dangers of spies crossing lines and the need for secrecy to protect that, So that's why I think this claim is properly understood as not being covered by the Tucker Act and not being required with President Lincoln to engage in espionage activities in the south. And this Court held that when the estate of — of Mr. Lloyd came to seek compensation from a court, that there was no judicial remedy to enforce that alleged agreement, and the remedy, if any, lay with the President's contingent fund. <end> JUSTICE KENNEDY: I —I'd like your help on this, Your interpretation of Totten — does it say that there is just no actionable contract, or does it say there's no jurisdictions like political question? I mean, you win under any of those theories, if we	end NEAL CONAN HOST Many other's visited the new King memorial on the National Mall and millions around the country took part in a national day of service Every year on this program we return to the march on Washington and probably King's most famous speech Here's Martin Luther King Jr August 28 1963 on the steps of the Lincoln Memorial and NEAL CONAN HOST MARTIN LUTHER KING JR I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation And five score years end NEAL CONAN HOST Many other's visited the new King memorial on the National Mall and millions around the country took part in a national day of service Every year on this program we return to the march on Washington and probably King's most famous speech Here's Martin Luther King Jr August 28 1963 on the steps of the Lincoln Memorial and NEAL CONAN HOST MARTIN LUTHER KING JR I am happy to join with you today in what will go down in history as the greatest demonstration for freedom in the history of our nation Here's Martin Luther King Jr on the steps of the Lincoln Memorial August 28 1963 This is TALK OF THE NATION from NPR News I'm Neal Conan in Washington Thank God almighty we are free at last end NEAL CONAN HOST Martin Luther King Jr on the steps of the Lincoln Memorial August 28 1963 This is TALK OF THE NATION from NPR News I'm Neal Conan in Washington NEAL CONAN HOST He came and gave a public lecture That lecture was in Mary Dodd Brown Memorial Chapel And finally I had the great honor of being a the temporary guardian of his gift to Lincoln University He gave a collection of books manuscripts and other memorabilia which is housed in Langston Hughes Memorial Library today end NEAL CONAN HOST It must have been I've read that in fact he liked to go to the dorms and hang out with the students and tell stories It must have been extraordinary to have that opportunity end LANGSTON	Martin Luther King Jr.	Lincoln	Martin Luther King Jr
What library is recommended for sending mail?	NEAL CONAN, HOST: So if you wrote a check, for example, to the United Way, you have no idea where that money is necessarily going - for good causes, you suspect - but that good cause may be the heating bill at the headquarters. <end> LAURA VANDERKAM: Well, certainly that's true. I mean, all nonprofits do have some need for overhead to run their operations, But generally, you know, this is the way professional philanthropy has worked, is that you trust that the experts who are running these philanthropies know what the most urgent causes are, know what is the NEAL CONAN, HOST: So if you wrote a check, for example, to the United Way, you have no idea where that money is necessarily going - for good causes, you suspect - but that good cause may be the heating bill at the headquarters. <end> LAURA VANDERKAM: Well, certainly that's true. I mean, all nonprofits do have some need for overhead to run their operations, But generally, you know, this is the way professional philanthropy has worked, is that you trust that the experts who are running these philanthropies know what the most urgent causes are, know what is the NEAL CONAN, HOST: So if you wrote a check, for example, to the United Way, you have no idea where that money is necessarily going - for good causes, you suspect - but that good cause may be the heating bill at the headquarters. <end> LAURA VANDERKAM: Well, certainly that's true. I mean, all nonprofits do have some need for overhead to run their operations, But generally, you know, this is the way professional philanthropy has worked, is that you trust that the experts who are running these philanthropies know what the most urgent causes are, know what is the NEAL CONAN, HOST: So if you wrote a check to - the United Way, you suspect - but that good cause may be the heating bill at the headquarters. <end> LAURA VANDERKAM: Well, certainly that's true. I mean, all nonprofits do have some need for overhead to run their operations, But generally, you know, this is the way professional philanthropy has worked, is that you trust that the experts who are running these things recycled, get paper out of the landfill, and this is an enormous burden on cities and on taxpayers because much of this mail is never even looked at. <end> NEAL CONAN, HOST: Not even looked at, <end> ELISABETH ROSENTHAL: Yup, <end> NEAL CONAN, HOST: And as you look at what is - the trend is unmistakable. I mean, people do request that their - I guess their credit card bills, most of us, show up on paper, We tend to pay my usual beat. And I know that cities everywhere in the United States are struggling to get more things recycled, get paper out of the landfill, and this is an enormous burden on cities and on taxpayers because much of this mail is never even looked at. <end> NEAL CONAN, HOST: And as you look at what is - the trend is unmistakable. I mean, people do request that their - I guess their credit card bills, most of us, show up on paper. We tend to pay	Kaminski Vince Kaminski Celeste I am forwarding you a letter from Prof Duane Seppi from Carnegie Mellon University I have known Duane for many years and I know that he does not make his recommendations without very good reasons I would recommend looking at John Gordon as a very strong candidate I think he will make a terrific contribution to Enron Vince Kaminski Vince Kaminski Celeste I am forwarding you a letter from Prof Duane Seppi from Carnegie Mellon University I have known Duane for many years and I know that he does not make his recommendations without very good reasons I would recommend looking at John Gordon as a very strong candidate I think he will make a terrific contribution to Enron Vince Sheridan Tilman Dear Ben I enjoyed meeting with you I sent a letter recommending you to our Dean in charge of the MBA program Let me know if you hear anything from the people in the admissions office I hope things work out and I have the opportunity to have you participate in our energy finance program By the way I like the idea of bringing in various industry leaders I currently do this in my class but it might make sense to consider ways to get first year students involved in this activity regards Sheridan Sheridan Tilman Dear Ben I enjoyed meeting with you I sent a letter recommending you to our Dean in charge of the MBA program Let me know if you hear anything from the people in the admissions office I hope things work out and I have the opportunity to have you participate in our energy finance program By the way I like the idea of bringing in various industry leaders I currently do this in my class but it might make sense to consider ways to get first year students involved in this activity regards Sheridan Sheridan	smtplib	e-mail	Postal Service
How many payday loan stores are there?	usually work, Alex. They buy the right to collect outstanding debts at a discount and they get a share of the take. The difference here is that some of these hospitals have bailed to auction on a couple of online websites that have been developed just for this purpose and that turns the debt collection firms into bidders. The hospital may get more money but the flip side of that is that buying a debt can become more expensive for the collectors and this arrangement has the collectors either buying the debt outright or providing a guarantee usually work, Alex. They buy the right to collect outstanding debts at a discount and they get a share of the take. The difference here is that some of these hospitals have started to auction off their debt on a couple of online websites that have been developed just for this purpose and that turns the debt collection firms into bidders. The hospital may get more money but the flip side of that is that buying a debt can become more expensive for the collectors and this arrangement has the collectors either buying the debt outright or providing a guarantee also subscriber newsletters that we are giving to our paying customers that help them understand the back-story behind some of the more significant news pieces that we are reporting on. also subscriber newsletters that we are giving to our paying customers that help them understand the back-story behind some of the more significant news pieces that we are reporting on. from Florida. It's a huge number. It's obviously decreasing now, And so now drug addicts and abusers, when they can get their pills from doctors, or if they don't buy them on the street, they are going to have to turn to pharmacies, But with that said, at least one pharmacy chain is implementing some changes, CVS recently sent a letter to what it deemed high-dispensing doctors, telling them that it was no longer going to fill their prescriptions for pain killers and other medication	McDonald's and STACEY VANEK SMITH BYLINE Actually there are more payday loan stores than McDonald's or Starbucks There are nearly 18000 payday loan stores in this country right now end RONALD MANN So I think what you really have to see is to step back and say or ask why are there so many people in our economy that are struggling so hard end STACEY VANEK SMITH BYLINE People like Amy Marineau and AMY MARINEAU The turning point for me was having to at 43 live with my mother again and not being able to take care of our family RONALD MANN I mean these are products that are there's a fair chance people aren't going to be able to pay them back end STACEY VANEK SMITH BYLINE Ronald says that is exactly why about 20 states have either banned payday loans entirely or really restricted them end CARDIFF GARCIA BYLINE On the other hand more than 30 states don't really have restrictions at all on payday lending his gotten huge or you might say supersized end RONALD MANN The number of payday loan stores is about the same as the number of RONALD MANN I mean these are products that are there's a fair chance people aren't going to be able to pay them back end STACEY VANEK SMITH BYLINE Ronald says that is exactly why about 20 states have either banned payday loans entirely or really restricted them end CARDIFF GARCIA BYLINE On the other hand more than 30 states don't really have restrictions at all on payday lending And in those states payday lending has gotten huge or you might say supersized end RONALD MANN The number of NEAL CONAN HOST This is TALK OF THE NATION I'm Neal Conan in Washington Yesterday the United States Postal Service announced plans to close about 250 processing centers lay off about 28000 workers and accept slower first class service as a consequence As many as 2000 post offices could close and Saturday delivery could be on the block as well end NEAL CONAN HOST All those cuts could reduce losses maybe even put the USPS in the black but when your mailbox is stuffed with direct mail ads some raise what was once unthinkable Has the post office outlived its	18000	there are payday loan stores in every State in the Nation	nearly 18000
Which programming language implements provide transformers according to Kimery?	was not a single overlapping step between that new process and Morse's process. <end> JUSTICE BREYER: Yes. And we here apply the correlation to any homocysteine test, any one here, any one in the future, any one that would impend. What's the difference? <end> MR. HUNGAR (None): Well, the difference is between claiming a — claiming all methods of achieving a particular result and claiming one process for achieving that particular result and then as one claiming any means of doing one particular step of that process. <end> JUSTICE BREYER: I apply electricity to all methods of putting was not a single overlapping step between that new process and Morse's process. <end> JUSTICE BREYER: Yes. And we here apply the correlation to any homocysteine test, any one here, any one in the future, any one that would impend. What's the difference? <end> MR. HUNGAR (None): Well, the difference is between claiming a — claiming all methods of achieving a particular result and claiming one process for achieving that particular result and then as one claiming any means of doing one particular step of that process. <end> JUSTICE BREYER: I apply electricity to all methods of putting in an appropriate way, what was at issue there, There was a disagreement about the content of the allegations. <end> MS. ROBIN-VERGEER (RESPONDENT): I don't think it's important, for, maybe, purposes of this, to iron this out, but I — respectfully, I don't agree with that characterization, because, even in the resolution of the grievance internally, the — what they found in the grievance was that they took no adverse action against him because of what he said — <end> JUSTICE BREYER: That doesn't — <end> MS. ROBIN-VERGEER (RESPONDENT): — in connection with this case. <end> JUSTICE BREYER: That isn't in an appropriate way, what was at issue there, There was a disagreement about the content of the allegations. <end> MS. ROBIN-VERGEER (RESPONDENT): I don't think it's important, for, maybe, purposes of this, to iron this out, but I — respectfully, I don't agree with that characterization, because, even in the resolution of the grievance internally, the — what they found in the grievance was that they took no adverse action against him because of what he said — <end> JUSTICE BREYER: That doesn't — <end> MS. ROBIN-VERGEER (RESPONDENT): — in connection with this case. <end> JUSTICE BREYER: That isn't	Kimbery I looked at the source code for Racketâ€™s require and provide and I examined %require and %provide a little end Kimbery Hereâ€™s what I learned require transformers and provide pretransformers are implemented in Racket Provide transformers are implemented in C end Kimbery I think provide transformers are pretty restricted because they have access to the bindings exported by allfromout and things like that but in order to know that the module needs to know which bindings are shadowed by the module body So provide transformers are essentially the very last step of macro transformation in a moduleâ€™s expansion end Kimbery I looked at the source code for Racketâ€™s require and provide and I examined %require and %provide a little end Kimbery Hereâ€™s what I learned require transformers and provide pretransformers are implemented in C end Kimbery I think provide transformers are pretty restricted because they have access to the bindings exported by allfromout and things like that but in order to know that the module needs to know which bindings are shadowed by the module body So provide transformers are essentially the very last step of macro transformation in a moduleâ€™s expansion and Chante le right that makes sense end Kimbery yes require transformers are essentially just ordinary expanders that require looks up with syntaxlocalvalue and Kimbery require transformers arenâ€™t special in any way only provide transformers are end Chantelle right and localprovide is nonsensical so there's no worries there end Chantelle would it make sense to have some way of delaying provide expansion in the same way that %expression delays things so that provide transformers	c	Python	C
How much in debt does Gustavo Arellano claim Hillary Clinton is in and should leave?	I think she should go. We need to unite this party. And you know I've said it before, I think she's the better candidate. I think that Barack is a better movement, <end> MADELEINE BRAND, host: And Gustavo? <end> Mr. GUSTAVO ARELLANO (Columnist, AskaMexican.net): This campaign is in tatters. She has to leave, you know. She's what, 31 million dollars in debt, she's increasingly becoming more and more bickering, throwing jabs at anybody who doesn't want to believe in her gospel that she was meant to be president. And I'm tired of Hillary Clinton. And I know nowadays we expect our candidates I think she's the better candidate. I think that Barack is a better movement. <end> MADELEINE BRAND, host: And Gustavo? <end> Mr. GUSTAVO ARELLANO (Columnist, AskaMexican.net): This campaign is in tatters. She has to leave, you know. She's what, 31 million dollars in debt, she's increasingly becoming more and more bickering, throwing jabs at anybody who doesn't want to believe in her gospel that she was meant to be president. And I'm tired of Hillary Clinton. And I know nowadays we expect our candidates RACHEL MARTIN, HOST: Meanwhile, federal workers are expected to miss their second paycheck tomorrow, they're political pressure on both sides, The Senate's got these two bills that they're going to bring up. They're both expected to fail. So where's the opening to end this? <end> DOMENICO MONTANARO, BYLINE: Well, the president has said, so far, that he's not budging on a wall. Democrats say they have a reason for not caving either. Take a listen to what House Speaker Nancy Pelosi said about that yesterday. <end> NANCY PELOSI: There is RACHEL MARTIN, HOST: Meanwhile, federal workers are expected to miss their second paycheck tomorrow. I mean, people are really suffering in this moment. There is political pressure on both sides, The Senate's got these two bills that they're going to bring up, they're both expected to fail, So where's the opening to end this? <end> DOMENICO MONTANARO, BYLINE: Well, the president has said, so far, that he's not budging on a wall. Democrats say they have a reason for not caving either. Hey, Domenico. <end> DOMENICO MONTANARO, BYLINE: Good morning, Rachel. <end> RACHEL MARTIN, HOST: So President Trump, as we know, not someone who backs down easily, but I guess he didn't really have a choice - right? - if Nancy Pelosi said he is not invited to her House.	Dylan Windham JeffDasovich The entire phone bill is 300 pesos Yes please fax to me And please email me rest of expenses when you get a chance so I can write you a check for everything planephone golf etc etc You guys get back OK end UNKNOWN_SPEAKER you don't want to know and Dylan Windham JeffDasovich we get back ok I didn't know tulsa had such a neat airport end Dylan Windham JeffDasovich That's so weird Prentice left her car door ajar when she went to go to school yesterday morning it was dead and she Dylan Windham JeffDasovich The entire phone bill is 300 pesos Yes please fax to me And please email me rest of expenses when you get a chance so I can write you a check for everything planephone golf etc etc You guys get back OK end UNKNOWN_SPEAKER you don't want to know end Dylan Windham JeffDasovich we get back ok I didn't know tulsa had such a neat airport end Dylan Windham JeffDasovich That's so weird Prentice left her car door ajar when she went to go to school yesterday morning it was dead and she KEN RUDIN BYLINE But here's another thing about the scenario here If ObamaClinton ticket losses in 2012 why would Hillary Clinton be the frontrunner for 2016 Because you know you have the Andrew Cuomos and all the Democrats waiting in the wings And if they won do you think after eight years of keeping them on the ticket for all the excitement and Geraldine Ferraro that you knew Mondale still lost 49 out of 50 states So I don't know what changing the vice president does so much end KEN RUDIN BYLINE But here's another thing about the scenario here If ObamaClinton ticket losses in 2012 why would Hillary Clinton be the frontrunner for 2016 Because you know you have the Andrew Cuomos and all the Democrats waiting in the wings And if they won do you think after eight years of have any collateral The answer is no I'm dirt poor So they say well I could loan you this money but you and your family have to put up and be the collateral against that I will hold against this loan until you repay it end Dr KEVIN BALES Activist and Author Ending Slavery Now I appreciate for Americas that you actually have to kind of stretch your mind to get around that idea that you and your own work become collateral against a loan But because everything you do and all your work becomes	31 million dollars	31 million dollars	$20 billion

Figure 2: Predicted answers and retrieved passages for a set of questions from the conversational domain (Wu et al., 2021b).

13

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. *arXiv preprint arXiv:1906.05416.* https://doi.org/10.18653/v1/P19-1620

Petr Baudiš and Jan Šedivỳ. 2015. Modeling of the question answering task in the yodaQA system. In *International Conference of the Cross-language Evaluation Forum for European Languages*, pages 222–228. Springer. https://doi.org/10.1007/978-3-319-24027-5_20

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*, pages 1533–1544.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2021. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426.*

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

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but factual! Inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354. https://doi.org/10.18653/v1/2022.acl-long.236

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, 28:1693–1701.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734.*

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551.* https://doi.org/10.18653/v1/P17-1147

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906.* https://doi.org/10.18653/v1/2020.emnlp-main.550

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858.*

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue

14

generation. *arXiv preprint arXiv:2107.07566.* `https://doi.org/10.18653/v1/2022.acl-long.579`

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840.* `https://doi.org/10.18653/v1/2020.emnlp-main.750`

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466. `https://doi.org/10.1162/tacl_a_00276`

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300.*

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. *arXiv preprint arXiv:2006.15020.*

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461.* `https://doi.org/10.18653/v1/2020.acl-main.703`

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401.*

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020c. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637.* `https://doi.org/10.18653/v1/2021.eacl-main.86`

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *arXiv preprint arXiv:2102.07033.* `https://doi.org/10.1162/tacl_a_00415`

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. *arXiv preprint cs/0205028.* `https://doi.org/10.3115/1118108.1118117`

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2020. Zero-shot neural retrieval via domain-targeted synthetic query generation. *arXiv preprint arXiv:2004.14503.* `https://doi.org/10.18653/v1/2021.eacl-main.92`

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, Colin Raffel, Adam Roberts, Tom Kwiatkowski, Patrick Lewis, Yuxiang Wu, Heinrich Küttler, Linqing Liu, Pasquale Minervini, Pontus Stenetorp, Sebastian Riedel, Sohee Yang, Minjoon Seo, Gautier Izacard, Fabio Petroni, Lucas Hosseini, Nicola De Cao, Edouard Grave, Ikuya Yamada, Sonse Shimaoka, Masatoshi Suzuki, Shumpei Miyawaki, Shun Sato, Ryo Takahashi, Jun Suzuki, Martin Fajcik, Martin Docekal, Karel Ondrej, Pavel Smrz, Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Wen-tau Yih. 2021. NeurIPS 2020 EfficientQA competition: Systems, analyses and lessons learned. *arXiv preprint arXiv:2101.00133v1.*

Timo Moller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online. Association for Computational Linguistics.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi

15

Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. Efficient large-scale language model training on GPU clusters using Megatron-LM. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15. `https://doi.org/10.1145/3458817.3476209`

Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679. `https://doi.org/10.18653/v1/P19-1256`

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQUAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*. `https://doi.org/10.18653/v1/D16-1264`

Stephen Robertson and Hugo Zaragoza. 2009. *The Probabilistic Relevance Framework: BM25 and Beyond*. Now Publishers Inc. `https://doi.org/10.1561/1500000019`

Devendra Singh Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers for open-domain question answering. *arXiv preprint arXiv:2101.00408*.

Siamak Shakeri, Cicero Nogueira dos Santos, Henry Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. End-to-end synthetic data generation for domain adaptation of question answering systems. *arXiv preprint arXiv:2010.06028*. `https://doi.org/10.18653/v1/2020.emnlp-main.439`

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*. `https://doi.org/10.18653/v1/2021.findings-emnlp.320`

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.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. NewsQA: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*. `https://doi.org/10.18653/v1/W17-2623`

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. *arXiv preprint arXiv:2004.10706v2*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*. `https://doi.org/10.18653/v1/2020.emnlp-demos.6`

Chien-Sheng Wu, Linqing Liu, Wenhao Liu, Pontus Stenetorp, and Caiming Xiong. 2021a. Controllable abstractive dialogue summarization with sketch supervision. *arXiv preprint arXiv:2105.14064*.

Chien-Sheng Wu, Andrea Madotto, Wenhao Liu, Pascale Fung, and Caiming Xiong. 2021b.

16

QAconv: Question answering on informative conversations. *arXiv preprint arXiv:2105.06912*.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv: 2107.07567*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018. `https://doi.org/10.18653/v1/D15-1237`

17