StreamingQA: A Benchmark for Adaptation to New Knowledge over Time in Question Answering Models

Adam Liška [* 1]   **Tomáš Kočiský** [* ♠ 1]   **Elena Gribovskaya** [* ♠ 1]   **Tayfun Terzi** [* 1]
Eren Sezener [1]   **Devang Agrawal** [1]   **Cyprien de Masson d'Autume** [1]   **Tim Scholtes** [1]   **Manzil Zaheer** [1]
Susannah Young [1]   **Ellen Gilsenan-McMahon** [1]   **Sophia Austin** [1]   **Phil Blunsom** [1 2]   **Angeliki Lazaridou** [1]

Abstract

Knowledge and language understanding of models evaluated through question answering (QA) has been usually studied on static snapshots of knowledge, like Wikipedia. However, our world is dynamic, evolves over time, and our models' knowledge becomes outdated. To study how semi-parametric QA models and their underlying parametric language models (LMs) adapt to evolving knowledge, we construct a new large-scale dataset, StreamingQA, with human written and generated questions asked on a given date, to be answered from 14 years of time-stamped news articles. We evaluate our models quarterly as they read new articles not seen in pre-training. We show that parametric models can be updated without full retraining, while avoiding catastrophic forgetting. For semi-parametric models, adding new articles into the search space allows for rapid adaptation, however, models with an outdated underlying LM under-perform those with a retrained LM. For questions about higher-frequency named entities, parametric updates are particularly beneficial. In our dynamic world, the StreamingQA dataset enables a more realistic evaluation of QA models, and our experiments highlight several promising directions for future research.

1. Introduction

Question answering (QA) allows us to interrogate models for their language understanding, knowledge, and reasoning abilities, while also being useful in various knowledge-oriented applications such as personal assistants or web

Table 1. Example questions from StreamingQA (Eval-Written). In addition to a question date, each question has three reference answers, and an evidence document with its publication date.

Recent subset
Question Date: Monday, February 24, 2020
Question: How many countries have committed to the net zero target as of today's date?
Past subset
Question Date: Sunday, April 12, 2020
Question: In November 2016, which Netflix series set in the United Kingdom was said to be "the most expensive television series ever"?

search. The questions that people ask span all our knowledge and can be about any point in the history, although often they are about the most recent events that happened in the last few weeks or days. Consider examples in Table 1 that ask about events as distant as 4 years before or as recent as the day when the question was asked. As the world and knowledge evolve, we need our QA models to adapt to new information, to not forget the past, and to maintain an up-to-date world model to make our interaction with such systems more meaningful. To evaluate and improve models' ability to adapt, we need a dataset with temporal grounding of both questions and knowledge—dates when questions were asked and publication dates of documents. As the currently available QA datasets are not suitable for this, we propose a novel dataset and subsequently perform a systematic study of adaptation in the state-of-the-art QA models.

Previous research has often focused on answering questions about individual passages or books (Rajpurkar et al., 2016; Kočiský et al., 2018), answering about static structured (Jia et al., 2021), or unstructured knowledge corpora such as Wikipedia (Lee et al., 2019a). More recent work has considered answering questions about knowledge with temporal grounding of facts in a knowledge graph (Saxena et al., 2021; Jia et al., 2021) or news articles (Wang et al., 2021; Dhingra et al., 2021) but without grounding the questions with a question date; or have repurposed questions from other datasets and added question dates (Zhang & Choi,

---

[*]Equal contribution in random order ♠Project Leads [1]DeepMind, London, UK [2]Department of Computer Science, University of Oxford, Oxford, UK. Correspondence to: Tomáš Kočiský <tkocisky@deepmind.com>, Elena Gribovskaya <egribovskaya@deepmind.com>.

2021) but still answered over a static snapshot of Wikipedia.

We present a new dataset, StreamingQA[1], that provides temporal context of both, the questions and the knowledge required to answer them. The dataset contains questions written by annotators or generated with a large-scale LM. The questions are answerable from a *streaming* knowledge corpus of time-stamped English WMT news articles published between 2007 and 2020 (see Figure 1). Having temporal metadata for questions and articles enables us to ingest new knowledge periodically, in a streaming setup, and evaluate on questions asked during that period. We consider questions about recent and past knowledge separately to measure adaptation and forgetting. Moreover, question dates allow to ask questions with relative time specifications (e.g., "3 months ago"), which are under-represented in the existing QA datasets. Lastly, news domain, compared to often used Wikipedia, provides more realistic challenges for open-book retrieval with redundant, noisy, and sometimes conflicting information.

Previous work demonstrated that large LMs struggle with temporal generalization, a type of domain shift that occurs when a model at test time needs to understand new knowledge, named entities, and topics (Lazaridou et al., 2021; Röttger & Pierrehumbert, 2021). In this work, we leverage StreamingQA to quantify similar adaptation in existing parametric (closed-book) and semi-parametric (open-book) QA models that today are frequently based on such LMs.

Our findings suggest that parametric adaptation improves QA performance for open-book approaches like RAG (Lewis et al., 2020a) and FiD (Izacard & Grave, 2020) (Section 4.2). Moreover, a more granular, frequency-based analysis (Section 4.3) suggests that parametric and semi-parametric approaches to adaptation are complementary: parametric adaptation improves accuracy of questions about frequent knowledge/named entities, where semi-parametric, dense retrieval-based methods can under-perform (Liu et al., 2021). In contrast, semi-parametric adaptation helps with less frequent knowledge where retrieval is less confused by redundancy and ambiguity of information associated with more frequent names, where the parametric LMs struggle. In the closed-book setup, we find that incremental fine-tuning works reasonably well without causing catastrophic forgetting (Section 4.1), and it also results in substantially lower computational costs than full model retraining[2]. Lastly, we also establish benchmarks for less computationally intensive QA tasks (Section 4.5): one-step adaptation and the usual static open-book QA task.

---

[1] We aim to make this dataset publicly available.

[2] In addition to new knowledge being created incrementally over time, our interest in iterative fine-tuning is also due to the large computational cost savings. In our setup, fine-tuning on a single month requires 95% fewer steps than full retraining.
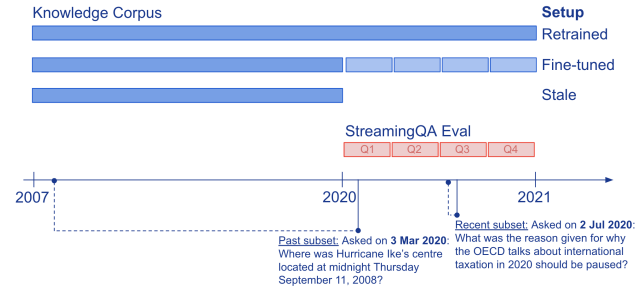

Figure 1. The StreamingQA task: we emulate a realistic scenario where a QA system needs to respond to user questions about a mix of recent and past events.

2. StreamingQA Dataset and Task

In this section, we introduce a new QA dataset and a task to evaluate models' adaptation and forgetting of knowledge. We require, in addition to questions, temporal metadata: a question date (when the question could have been asked) and a knowledge date (when an article that answers this question was published). Using this metadata enables us to evaluate how well the model understands new knowledge that becomes available incrementally at evaluation time (see Figure 1). We also use the timestamps to split the training and evaluation sets into non-overlapping historical periods. See Table 2 for examples of questions.

To construct the StreamingQA dataset, we consider 14 years (2007–2020) of English WMT news articles[3] (Akhbardeh et al., 2021), together with their publication dates, as our knowledge corpus (approx. 11M articles). Specifically, given an article, we first generate a question date, that is, the date when we want the question to be asked (Section 2.2), and subsequently, either (i) automatically generate questions (Section 2.3), or (ii) ask human annotators to write questions (Section 2.4). Lastly, to reduce noise, we apply automatic and human filtering to the questions, and collect additional reference answers (Section 2.5). We present additional statistics in Section 2.6.

We consider a streaming task (Section 2.1): we split questions into four quarterly sets over 2020 based on their question dates, where questions in each quarter are to be answered from articles published up to and including that quarter. For the adaptation and forgetting analysis, we further split the quarterly evaluation sets into the *recent subset* and the *past subset*. The dataset is constructed to have an approximately equal number of each. As recent questions cover 2020 and past questions cover all history uniformly, the overall distribution is biased towards the present, as we would expect it in an actual QA system. The recent subset contains questions about the most recent knowledge, i.e.,

---

[3] http://data.statmt.org/news-crawl/README

Table 2. Examples of written and generated questions from the recent and past subsets.

Question Date	Question	Answer	Passage Date	Subset	Added Time	Gen./Wri.
30.05.2020	What does Donald Trump, US president, call his 2020 plan to expedite the development of a COVID-19 vaccine?	Operation Warp Speed	19.05.2020	Recent	-	Written
18.04.2020	In the UK, as of Sunday, April 12, 2020, how much money in loans had been given to firms seeking cash to survive the coronavirus crisis?	£800 million	12.04.2020	Recent	-	Written
28.04.2020	Which soft drinks company was 2010 Australian favourite racing driver Mark Webber's team sponsored by?	Red Bull	29.03.2010	Past	-	Written
26.03.2020	Where did the research that claimed up to 50 per cent of the UK population may have already contracted the coronavirus, take place?	University of Oxford	25.03.2020	Recent	-	Generated
11.09.2020	Which player scored for St Mirren in November 2008?	Franco Miranda	16.11.2008	Past	Absolute	Generated
08.05.2020	Which hospital is Jamie Cooper receiving treatment at 13 years ago?	Selly Oak Hospital	01.11.2007	Past	Relative	Generated

articles published in the month before the question date; and the past subset contains questions asked during that quarter about articles published between 2007 and the question date. Furthermore, for training and validation, we use automatically generated questions asked during 2007–2019 with an analogous split between the recent and past subsets.

2.1. StreamingQA Task: QA with Temporal Metadata

Formally, our dataset consists of triples of question date, question, and answer, $\mathcal{Q} = \{(d_{q,i}, q_i, a_i)\}_{i=1}^N$, and a knowledge corpus of publication dates and documents, $\mathcal{C} = \{(d_{c,j}, c_j)\}_{j=1}^M$. For a given time period $t = [t_s, t_e]$ (e.g., January to March 2020), we consider questions asked during that period, $\mathcal{Q}_{=t} = \{(d_{q,i}, q_i, a_i) \in \mathcal{Q} : t_s \leq d_{q,i} \leq t_e\}$, about the corresponding subset of the knowledge corpus published until then, $\mathcal{C}_{\leq t} = \{(d_{c,j}, c_j) \in \mathcal{C} : d_{c,j} \leq t_e\}$. To answer a question in $\mathcal{Q}_{=t}$, we generate an answer using the corresponding knowledge corpus, $p(a_i | q_i, d_{q,i}, \mathcal{C}_{\leq t})$.
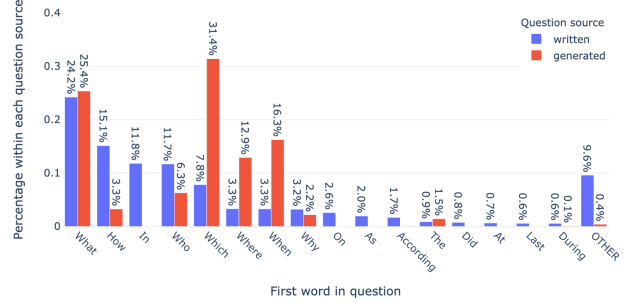
2.2. Questions Dates

We need to generate question dates that are plausible with respect to article dates and make sure that the dates and events are consistent. For evaluation sets, to create recent subset questions we sample a document with a publication date $d_c \sim \mathcal{U}[\text{Dec}2019, \text{Dec}2020]$ and a question date $d_q \sim \mathcal{U}[d_c, d_c + 30\text{days}]$. To create past subset questions, we sample a document with a publication date $d_c \sim \mathcal{U}[2007, \text{Dec}2020]$ and a question date $d_q \sim \mathcal{U}[\max(d_c, \text{Jan}2020), \max(d_c, \text{Jan}2020) + 365\text{days}]$. The articles are distributed uniformly within a month, or across all available history prior to the question date[4], for the two subsets respectively. For training and validation sets, we consider question dates in $[2007, \text{Dec}2019]$ and aim for an article distribution given a question date similar to above.

2.3. Automatically Generated Questions

We use automatic question generation as a scalable way to obtain questions grounded at different points in time. Questions are generated through few-shot prompting of a large LM (Rae et al., 2021), given an evidence document

[4]We filter out samples with question date not in 2020.

(a) Distribution of first word for written and generated questions.
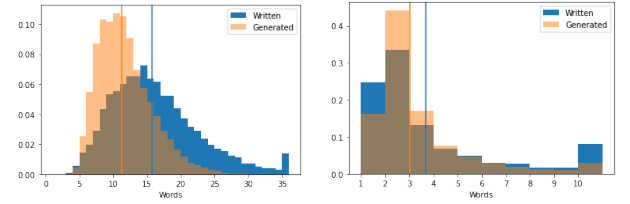
(b) Question length histograms and means. (c) Answer length histograms and means.

Figure 2. Details of the StreamingQA dataset.

and a target answer drawn from named entities, dates, and prepositional phrases contained as spans in the document. A challenge with creating questions for open-domain QA is that they need to be specific enough when considered in the context of *all* articles in the knowledge corpus. We first over-generate questions-answer pairs and then apply heuristic filters to eliminate trivial and/or low quality candidates (Appendix A.4). For questions included in the past subset, we append absolute or relative time specifications to question text, e.g., "3 months ago" or "in May 2017" (Appendix A.5), unless the text already contains such a specification or the answer is a date.

2.4. Human Written Questions

Human annotators were asked to write questions about a news article provided together with its publication date and desired question date. We chose annotators whose first language is English, who are in the US or the UK, and have a university education. Each annotator had to write up to five questions and answers about the details of the

Table 3. StreamingQA dataset statistics.

	Recent		Past		All	
	Questions	Articles	Questions	Articles	Questions	Articles
Train	49,765	49,653	50,235	50,065	100,000	99,466
Valid	5,000	4,999	5,000	4,998	10,000	9,993
Eval-Generated	15,618	15,452	12,122	11,899	27,740	26,969
Eval-Written	4,557	1,548	4,258	1,448	8,815	2,996

events described in the article, and framed these questions as if asking a friend. We explicitly asked the annotators to include enough context in the questions to make the questions as unambiguous as possible for the open-book setup.

2.5. Quality Filtering

We filtered both generated and human written questions for quality in two stages. First, we asked annotators to filter for good/bad question, similar to Kwiatkowski et al. (2019), filtering for factual, unambiguous, grammatical questions. To judge ambiguity, we have additionally provided the question date and asked not to assume a particular location (e.g., US). To include a question, we need 3 annotators to agree.

Secondly, we asked annotators to answer each question given the original passage, its publication date, and the question date. Annotators first selected parts of the passage that supported the answer[5], and then wrote a short answer in their own words. We did not require the answers to be sub-strings of the passage. We only kept questions where annotators could provide answers, and obtained additional references in the same way.[6]

2.6. Statistics

The StreamingQA dataset contains about 28k generated questions and about 8.8k human written questions for evaluation, and 100k and 10k questions for training and validation, respectively. See Table 3 for details. In Figure 2, we see that human written questions and answers tend to be somewhat longer. Based on the first question word distribution, we have a diverse set of both human written and generated questions, with the latter slightly biased to "Which"/"Where"/"When", likely due to prompting of the large LM. For written questions, many start with "In", which is often annotators providing temporal context so that questions stand without the article in the open-book setting. Examining the answer types on the written evaluation sets by automatic labeling, we have about 46.9%, 40.6%, and 12.4% of named entity, phrase, and date answers, respectively (see Appendix A.2). About 6% of evaluation reference answers

are seen among answers to questions in the training set, and 23% of training questions have an answer contained in evaluation reference answers.

3. Experimental Setup

To evaluate how well parametric and semi-parametric approaches to QA ingest and understand unseen information, we consider an auto-regressive, left-to-right Transformer-XL (TXL) (Dai et al., 2019) language model as our parametric, closed-book QA model (CB), and use it as the underlying LM for our RAG-style (Lewis et al., 2020b) semi-parametric, open-book QA model (OB). We also consider a more recent open-book model based on Fusion-in-Decoder (FID) that uses a T5 (Raffel et al., 2020) sequence-to-sequence model[7]. We use the standard metrics for QA: the F1 and the exact match (EM), after normalizing the answers, in the same way as Rajpurkar et al. (2016). These are suitable for the on average short answers in our dataset.

To include the temporal metadata, we prefix publication dates to articles ("Thursday, February 7, 2019. [article text]"), and for questions we add question dates ("Today is Wednesday, May 6, 2020. [question text]").

3.1. Language Model Pretraining and Finetuning

We consider three setups of TXL training: TXL_{STALE} model is trained on WMT articles until the end of December 2019 (approx. 10.1M articles), and is missing knowledge required for answering questions in the recent subset. $TXL_{RETR.}$ model is re-trained from scratch on all WMT articles until the end of December 2020, i.e., including the evaluation period (approx. 11.4M articles). Lastly, TXL_{FT} model is TXL_{STALE} that we additionally *iteratively fine-tune*[8] on the 2020 monthly article sets (each approx. 100k articles). TXL_{STALE} and $TXL_{RETR.}$ are trained for 200k steps on 32 TPUv3, whereas fine-tuning is performed for 10k steps.

Our TXL has 18 layers and 1,280 hidden units, resulting in 448M parameters, roughly 30% larger than GPT2-Medium and BERT-large. We set the TXL sequence length to 1,024, and the memory cache length to 384 during model pretraining, and use a SentencePiece vocabulary of 50,259 tokens (Kudo & Richardson, 2018).

Analogously, for T5 base, we fine-tune the pre-trained vanilla T5 model (Raffel et al., 2020) for 300k steps on WMT articles until the end of 2019, and subsequently iteratively fine-tune for 4k steps on the 2020 monthly splits. The

---

[5]We don't use this annotation in experiments.

[6]We have 3 reference answers for computing the metrics, first the answer used to generate the question or provided by the original question writer, and 2 additional provided by annotators.

[7]For T5, due to its existing pre-training, we control the knowledge of our model less precisely. However, both the T5 model and the BERT model (in DPR) were released in or before 2019, and so don't contain any knowledge from 2020, our evaluation period.

[8]The best checkpoint of month N is used as the starting point for fine-tuning on the articles of month $N+1$.
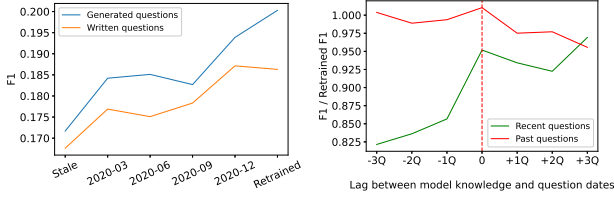
Figure 3. Left: F1 score on the whole evaluation dataset of CB$_{+\text{STALE}}$, CB$_{+\text{RETR.}}$, and CB$_{+\text{FT}}$ fine-tuned on articles published until the specified cut-off dates. Right: The effect of a temporal lag between the final training month of CB$_{+\text{FT}}$ and question dates for generated questions, relative to CB$_{+\text{RETR.}}$.

retrained version is fine-tuned for 300k steps on articles until the end of 2020 starting from the vanilla T5 checkpoint.

3.2. Closed-book QA

We use the closed-book QA task to examine a language model's knowledge. The task is to answer questions without any additional context provided, $p(a_i|q_i)$. We fine-tune each of the pre-trained TXL LMs (stale, fine-tuned, and fully retrained) for question answering on the StreamingQA training set, using 4 TPUv3, and select the best checkpoint, as measured by F1, on the validation set; both sets contain knowledge and questions asked in or before 2019. The QA models, CB$_{+\text{STALE}}$, CB$_{+\text{FT}}$, and CB$_{+\text{RETR.}}$, are then subsequently evaluated on the evaluation sets from 2020. Answers are sampled using greedy decoding.

3.3. Open-book QA

In this task, we answer questions given a knowledge corpus of articles, $p(a_i|d_{q,i}, q_i, \mathcal{C}_{\leq t})$. We use WMT news articles, sliced into 6-sentence chunks as our knowledge corpus, resulting in 30M (up to 2019) and 36M (up to 2020) passages.

The OB model is a variation of the Retrieval Augmented Generation model (RAG-sequence; Lewis et al. (2020b)) with a TXL-based generator, the same LMs as for our closed-book experiments. As a retriever we use Dense Passage Retrieval (DPR; Karpukhin et al. (2020)), trained on question/passage pairs from our training set (including the question and publication dates), with embedding size of 768. We retrieve 20 passages. We also consider the Fusion-in-Decoder model (FID; Izacard & Grave (2020)), which was shown to outperform RAG on a number of QA tasks, with the same pre-trained DPR retriever as OB but with 64 retrieved passages. In contrast to RAG, the FID's generator attends to all retrieved passages at the same time instead of individually. We train both models with a restricted search space that contains gold evidence of all training and validation questions; this helps to reduce computation time (we did not see a material performance degradation due to this).

4. Results and Analyses

In this section we analyse the performance of the closed-book and open-book models on StreamingQA. We have three closed-book models: CB$_{+\text{STALE}}$, the iteratively fine-tuned CB$_{+\text{FT}}$, and the retrained CB$_{+\text{RETR.}}$. In order to adapt open-book models to new information from 2020, we always include new articles in the search index and then either keep the stale generator (OB$_{+\text{IU}}$, FID$_{+\text{IU}}$), fine-tune the generator on new articles (OB$_{+\text{IU}+\text{FT}}$, FID$_{+\text{IU}+\text{FT}}$), or use a retrained generator (OB$_{+\text{IU}+\text{RETR.}}$, FID$_{+\text{IU}+\text{RETR.}}$).

4.1. Adaptation to New Knowledge and Forgetting in Closed-Book QA

Iterative LM fine-tuning improves performance on StreamingQA but lags retraining. We consider all questions in our evaluation sets (recent+past) and evaluate CB models, in Figure 3 (left). First, we observe that the CB$_{+\text{STALE}}$ is outperformed by each of the CB$_{+\text{FT}}$ models—the fine-tuning is able to incorporate new information for the half of the questions that are only answerable from 2020 documents. With each additional month of documents, the CB$_{+\text{FT}}$ models perform better for all answer types (named entities, phrases, dates; Appendix B.1). The improved performance is not simply due to more data, but is driven by better accuracy on the recent subset, while the performance on the past subset remains mostly unchanged (Appendix B.1). Secondly, we observe that CB$_{+\text{RETR.}}$ outperforms or is on par with all other models, and so vanilla adaptation that we consider for CB$_{+\text{FT}}$ should be improved to bridge the gap from fine-tuning to retraining.

Adaptation and forgetting We use question dates to split Eval-Generated and Eval-Written into quarterly sets. To understand how adaptation to new information is offset by forgetting of past information, we investigate the effect of a temporal lag between the question date and the end date of knowledge in the underlying LM. Note that the question date and the knowledge date is on average much closer for the recent subset (a few weeks) compared to the past subset (years), and so adaptation to new articles is more crucial for the recent subset performance.

When the lag is negative, the model knowledge is lagging behind a question date (QD)—the model is missing necessary information[9]. When the lag is positive, the questions are in the past with respect to the most recent information in the model, and in these settings, some previous information needed to answer these questions might have been overwritten—forgetting may occur.

---

[9]For example, an answer from a model trained until March 2020 for the question "What does Donald Trump, US president, call his 2020 plan to expedite the development of a COVID-19 vaccine?" asked on May 30, 2020 will be bucketed into -1Q.
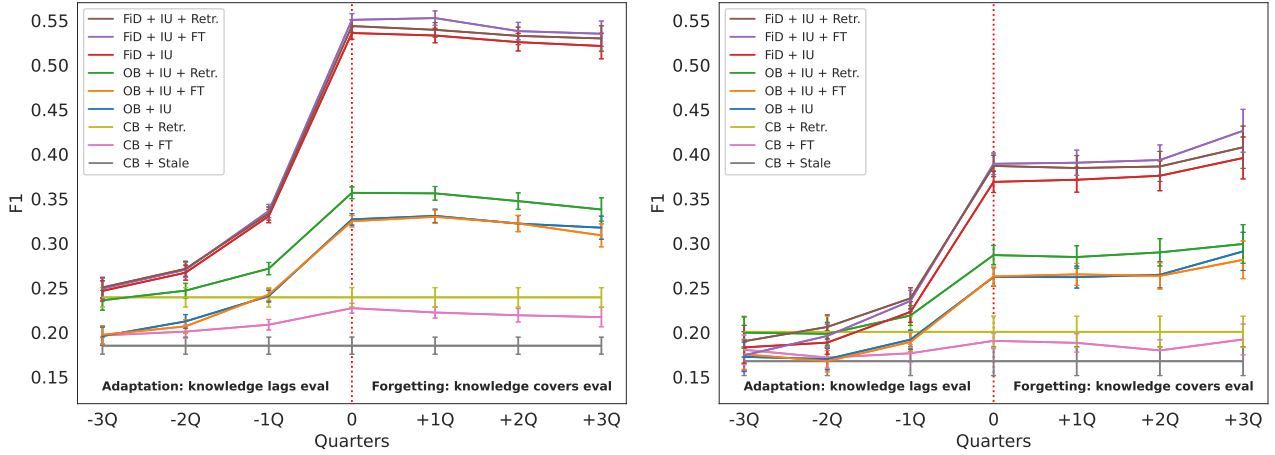
Figure 4. Adaptation and forgetting on recent subsets (generated, left; written, right). We observe that adapting the generator helps the FiD model, and helps the OB model when fully retrained, compared to index update only. Open-book models allow for much faster adaptation to recent knowledge than closed-book, with almost no forgetting. (IU = search index updated, FT = fine-tuned LM)
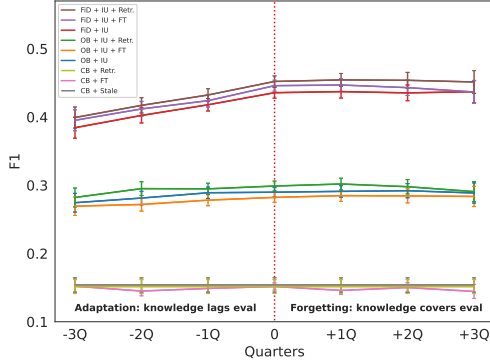

Figure 5. Adaptation and forgetting on past generated questions. We see only a slight improvement as the model acquires knowledge about 2020. We do not observe forgetting.

We aggregate the model answers for each lag, and plot the corresponding F1 relative to that of $CB_{+RETR.}$ in Figure 3 (right). As we fine-tune and the lag between the model knowledge and question month increases, the performance on the past subset slightly deteriorates until we under-perform $CB_{+RETR.}$ by about 5%. On the recent subset, the performance first improves significantly, and then as we pass the question quarter and continue fine-tuning on further data, we start seeing minor forgetting. Similar conclusions hold for the written questions (Appendix B.1).

4.2. Adaptation to New Knowledge and Forgetting in Open-Book QA

Adaptation and forgetting Similarly to the closed-book experiment, we examine adaptation and forgetting by aggregating model answers by a temporal lag between the evaluation set and the end of the model knowledge. We

consistently observe that on the recent subset of both generated and written questions (Figure 4), the open-book models (OB, FID) have a steep adaptation rate (from -1Q to 0Q) for all model variants, including just adding new articles into the search index without LM fine-tuning (OB_{+IU}, FID_{+IU}). For all of the models, we see almost no forgetting on the recent subsets. In Figure 5, for generated past questions, there is no forgetting (see Appendix B.2 for written questions). Note that a small fraction of questions from the past subset reference articles from 2020, so seeing 2020 knowledge slightly helps compared to a lagging model.

Is updating search space sufficient, or do we need to update underlying LMs? Recent open-book QA models consist of a search index, a retriever, and a generator. One strong argument in favour of the open-book models is that new information can be added directly into the search index, potentially without any additional training. Our results indicate that although the major performance gain comes from updating the search index, updating knowledge in the LM generator improves performance on recent subset questions too. We compare OB_{+IU}, FID_{+IU}, where the new documents are added only into the search index, with OB_{+IU+FT}, FID_{+IU+FT}, $OB_{+IU+RETR.}$, $FID_{+IU+RETR.}$, where both the index and the generator are updated. We observe this in Figure 4, for lags of 0Q, 1Q, 2Q, and 3Q, where the models have the required knowledge for answering. We see improvements for FID_{+IU+FT} (vs FID_{+IU}) on the recent subset of generated and written questions, and on the past subset (Figure 5) at 0Q followed by minor forgetting. Moreover, retraining the generator improves performance for all models and all subsets. For the FID/T5 models, we see FID_{+IU+FT} performing somewhat better than $FID_{+IU+RETR.}$ on the recent generated questions and performing worse on the past
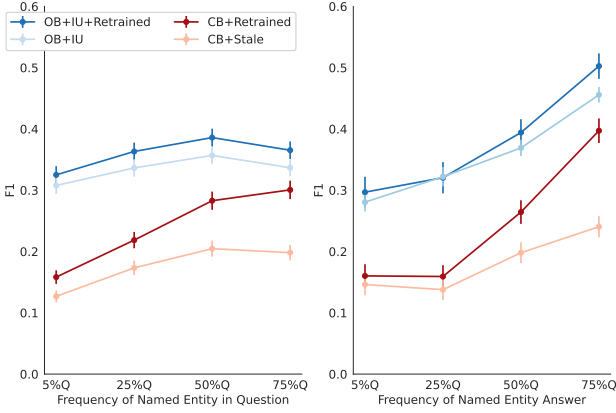
Figure 6. QA performance given question (left) or answer (right) named entity frequency quartiles.
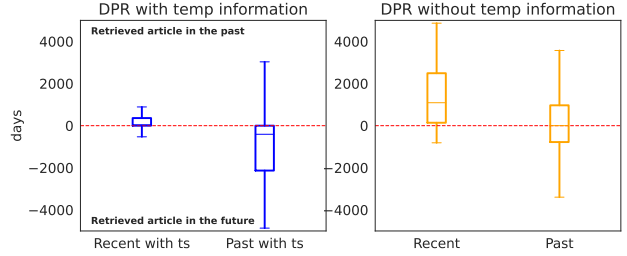

Figure 7. Temporal accuracy, measured in difference of days between gold passage timestamp (ts) minus retrieved passage ts, of the DPR trained on news for generated questions (recent vs past) with and without temporal information.

subset, suggesting fine-tuning on the recent data improved performance on the corresponding knowledge. Section 4.3 explores why fine-tuning the generator is helping.

4.3. Parametric vs Semi-parametric Adaptation

Seeing above that fine-tuning or retraining the generator helps, we want to understand for which questions the updated generator is particularly important compared to a stale generator. Lazaridou et al. (2021) previously demonstrated that one driving factor behind deteriorating temporal LM performance are changing frequencies of words, particularly named entities. We analyze QA performance by frequency of named entities appearing in (a) questions, and (b) answers, computed over the knowledge corpus up to 2019, and in 2020 only, respectively.

First, close-book performance is substantially better for questions that contain frequent named entities (see Appendix A.3 for examples): F1 is higher by absolute 10% (Figure 6, left), likely due to higher-frequency named entities in the knowledge corpus providing a stronger learning signal for the parametric model. Second, open-book performance does not show strong dependency on the frequency in the question, but this is due to two offsetting factors: DPR retrieval recall becomes worse with increasing frequency (recall@1 decreases by absolute 5%), while the generator performance improves. Therefore, at lower frequencies better performance is driven by non-parametric adaptation through the updated search space, and at higher frequencies by parametric adaptation.

Figure 6 (right) shows performance as a function of answer named entity frequencies in 2020: $CB_{+RETR.}$ outperforms CB_{+STALE} by 10-15% for more frequent answers, suggesting that one reason for better performance of updated generators is more accurate modeling of word frequencies in 2020.

4.4. Further Analyses

Temporal retrieval Using dates improves DPR performance for recent questions. In Figure 7, median temporal difference between retrieved and gold articles is 41 and 1101 days for DPR with and without dates, respectively. Improved temporal accuracy translates into better recall overall, recall@20 for the recent generated questions is 57% and 43% for DPR with and without dates, respectively. For past questions we do not see improvements, which suggests that the model may incorrectly interpret the two time specifications, i.e., the prepended question date and absolute or relative time specification in the question text. See Appendix B.3 for more.

Time specification in questions Generated questions in the past subset may contain an absolute or relative time specification in the question text, and we generally find that the open-book models perform best on questions without time specification[10], followed by absolute, and relative. For example, for FID_{+IU+FT}, F1 is 0.711, 0.469, 0.359, respectively, and 0.441 overall.

Static and updated questions For a preliminary analysis of "static" (knowledge that likely will not change) and "updated" (that might change) questions, we observed that the open-book models generally performed worse on "updated" questions (e.g., 0.435 vs 0.494 F1 for FID_{+IU+FT} on past, generated). The evaluation sets have 6.7%–11.2% of likely static questions based on majority agreement. The recall of retrieved documents is better for the static questions.

4.5. One-step Streaming and Static QA Benchmarks

StreamingQA dataset allows us to consider further two tasks, and we provide benchmarks to encourage research on these directions: one-step streaming setting[11] in Figure 8 (solid bars), and the usual static open-book QA setup (diagonal-

---

[10]These question ask about the most recent past, a few weeks before the question dates.

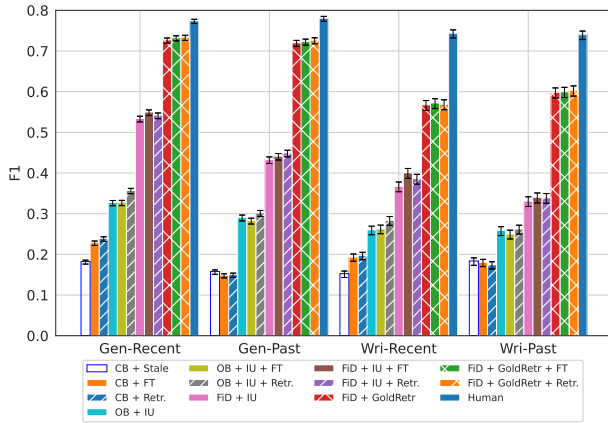[11]We report using the model fine-tuned iteratively on 12 months.

Figure 8. Static setup with all questions. Solid filled bars are models that had 2020 knowledge added incrementally.

line bars), evaluated on all 2020 questions. There is still a large gap to human performance; moreover the dataset creates challenges to retrieval and news articles reading comprehension, see models with gold evidence versus retrieved (cross-pattern bars). For the human benchmark we collected a fourth annotator answer. See Appendix B.4 for EM and a table with all metrics.

5. Related Work

QA datasets We summarize previous QA work on understanding of knowledge with temporal context in Section 1 and provide a dataset comparison table in Appendix A.1.

Question generation for QA Automatic question generation trained with supervision has been explored in QA for data augmentation (Alberti et al., 2019; Dong et al., 2019; Sultan et al., 2020), or as a way to enrich knowledge bases for QA-pair retriever models (Lewis et al., 2021). Here we instead leverage few-shot generation capabilities of large LMs (Rae et al., 2021; Brown et al., 2020) to generate questions and use them for both training and evaluation.

Open-domain QA Progress in neural information retrieval (Karpukhin et al., 2020; Lee et al., 2019b) enables open-domain QA models that are trained end-to-end as both the retriever and the reader are differentiable (Guu et al., 2020; Lewis et al., 2020b; Izacard & Grave, 2020; Sachan et al., 2021). Recent work in the domain has focused on improving performance by combining information from multiple documents efficiently (Sachan et al., 2021; Izacard & Grave, 2020) and on performance analysis of the dense retrievals, for instance, when dealing with named entities (Sciavolino et al., 2021; Liu et al., 2021).

Distribution shift in LM and QA There has been an increased interest in analyzing behaviour of Transformer-based language models under temporal shifts. Lazaridou et al. (2021) show that performance of Transformer-XL dete-

riorates when evaluated on data published after the training period, and use dynamic evaluation (Krause et al., 2017) to partially make up for this degradation. Röttger & Pierrehumbert (2021) study temporal adaptation of BERT models for downstream tasks, and Amba Hombaiah et al. (2021) propose new incremental methods for online BERT training. In the context of retrieval models, Khandelwal et al. (2020) and Lewis et al. (2020a) describe flexible model adaptation through knowledge source expansion or updating.

Continual learning Continual learning in language is a long-standing research topic (Carlson et al., 2010; Parisi et al., 2018) that has recently seen an increase in interest with researchers looking to create a Web-scale benchmark for studying continual learning (Hu et al., 2020); to propose new metrics for knowledge updates and to establish baselines (Jang et al., 2021).

6. Conclusion

In order to enable a more realistic evaluation of QA models, we introduced the first QA dataset and task for studying adaptation to new information over time in open and close-book settings with temporally non-overlapping training and evaluation sets.

As language models grow bigger, the cost of maintaining them up-to-date increases, and therefore adaptation ability of the models becomes more important. Our experimental results show that open-book QA models allow for fast and flexible adaptation through adding new articles into the search space, with fine-tuning or retraining generally further improving performance. The ability to inject new knowledge through the search space depends on retrieval accuracy and the more up-to-date parametric LMs are capable of compensating for retrieval errors. Additionally, our results show that iteratively fine-tuning the generator of the FiD QA model improves performance and costly retraining from scratch is not necessary. We leave for future work to better understand and close the performance gap between retrained and stale generators.

Future work StreamingQA highlights challenges of *temporal reasoning* and invites further research into this area: the past subset contains questions with the relative time specification, where retrieval struggles to extract relevant passages. For *fine-tuning*, we consider a vanilla setup without delving deeply into more sophisticated continual learning approaches. Future work should take in-depths look at how best to adapt QA models, and the problem of what to compress into weights or what to add to the search space. While we study adaptation to new knowledge, retrieving conflicting information due to *updated knowledge* (e.g., "How many seasons are in Game of Thrones?") is another important direction we did not tackle here.

References

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussà, M. R., España-Bonet, C., Fan, A., Federman, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C. M., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A., Turchi, M., Vydrin, V., and Zampieri, M. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, pp. 1–88, Online, 2021.

Alberti, C., Andor, D., Pitler, E., Devlin, J., and Collins, M. Synthetic qa corpora generation with roundtrip consistency, 2019.

Amba Hombaiah, S., Chen, T., Zhang, M., Bendersky, M., and Najork, M. Dynamic language models for continuously evolving content. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Aug 2021. doi: 10.1145/3447548.3467162. URL http://dx.doi.org/10.1145/3447548.3467162.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., and Mitchell, T. M. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

Chen, W., Wang, X., and Wang, W. Y. A dataset for answering time-sensitive questions, 2021.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2978–2988, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1285. URL https://aclanthology.org/P19-1285.

Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J., and Cohen, W. W. Time-aware language models as temporal knowledge bases, 2021.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H.-W. Unified language model pre-training for natural language understanding and generation, 2019.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pre-training, 2020.

Hu, H., Sener, O., Sha, F., and Koltun, V. Drinking from a firehose: Continual learning with web-scale natural language, 2020.

Izacard, G. and Grave, E. Leveraging passage retrieval with generative models for open domain question answering, 2020.

Jang, J., Ye, S., Yang, S., Shin, J., Han, J., Kim, G., Choi, S. J., and Seo, M. Towards continual knowledge learning of language models, 2021.

Jia, Z., Abujabal, A., Saha Roy, R., Strötgen, J., and Weikum, G. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pp. 1057–1062, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee. ISBN 9781450356404. doi: 10.1145/3184558.3191536. URL https://doi.org/10.1145/3184558.3191536.

Jia, Z., Pramanik, S., Saha Roy, R., and Weikum, G. Complex temporal question answering on knowledge graphs. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Oct 2021. doi: 10.1145/3459637.3482416. URL http://dx.doi.org/10.1145/3459637.3482416.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklBjCEKvH.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023.

Krause, B., Kahembwe, E., Murray, I., and Renals, S. Dynamic evaluation of neural sequence models, 2017.

Kudo, T. and Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liška, A., Terzi, T., Gimenez, M., de Masson d'Autume, C., Kočiský, T., Ruder, S., Yogatama, D., Cao, K., Young, S., and Blunsom, P. Mind the gap: Assessing temporal generalization in neural language models. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=73OmmrCfSyy.

Lee, K., Chang, M.-W., and Toutanova, K. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6086–6096, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1612. URL https://aclanthology.org/P19-1612.

Lee, K., Chang, M.-W., and Toutanova, K. Latent retrieval for weakly supervised open domain question answering, 2019b.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401, 2020a.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Lewis, P., Wu, Y., Liu, L., Minervini, P., Küttler, H., Piktus, A., Stenetorp, P., and Riedel, S. Paq: 65 million probably-asked questions and what you can do with them, 2021.

Liu, L., Lewis, P. S. H., Riedel, S., and Stenetorp, P. Challenges in generalization in open domain question answering. *CoRR*, abs/2109.01156, 2021. URL https://arxiv.org/abs/2109.01156.

Ning, Q., Wu, H., Han, R., Peng, N., Gardner, M., and Roth, D. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1158–1172, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.88. URL https://aclanthology.org/2020.emnlp-main.88.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual lifelong learning with neural networks: A review. 2018.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Röttger, P. and Pierrehumbert, J. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2400–2412, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.206. URL https://aclanthology.org/2021.findings-emnlp.206.

Sachan, D. S., Reddy, S., Hamilton, W., Dyer, C., and Yogatama, D. End-to-end training of multi-document reader and retriever for open-domain question answering, 2021.

Saxena, A., Chakrabarti, S., and Talukdar, P. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6663–6676, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.520. URL https://aclanthology.org/2021.acl-long.520.

Sciavolino, C., Zhong, Z., Lee, J., and Chen, D. Simple entity-centric questions challenge dense retrievers, 2021.

Sultan, M. A., Chandel, S., Fernandez Astudillo, R., and Castelli, V. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5651–5656, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.500. URL https://aclanthology.org/2020.acl-main.500.

Wang, J., Jatowt, A., and Yoshikawa, M. Archivalqa: A large-scale benchmark dataset for open domain question answering over archival news collections, 2021.

Zhang, M. and Choi, E. SituatedQA: Incorporating extralinguistic contexts into QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7371–7387, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.586.

Table 4. Related datasets overview. Abbreviations: Wri-other=filtered from other datasets; KC=knowledge corpus; QD=question date; APD=article publication date.

| Dataset		Knowledge Corpus (KC)	Closed/Open	KC Size	Train	Valid	Eval	QD	APD	Temporal Qs
StreamingQA (this work)	Wri+Gen	News (WMT07-20)	OB	11M / 36M	100k	10k	28k + 8.9k	Yes	Yes	Yes
TempLama (Dhingra et al., 2021)	Templ./Cloze	News (CustomNews)	CB	–	10k	5k	35k	X	Yes	Yes
ArchivalQA (Wang et al., 2021)	Gen	News (NYT86-07)	OB	1.8M	850k	100k	100k	X	Yes (transform relative time using APD)	Yes
SituatedQA-temporal (Zhang & Choi, 2021)	Wri-other	Wikipedia	OB	6M / 21M	6k	3.4k	2.8k	Yes	X	Yes
Time-Sensitive QA (Chen et al., 2021)	Templ./Wri	Wikipedia	OB	6M / 21M	29k	6.1k	6.1k	X	X	Yes
Nartural Questions (Kwiatkowski et al., 2019; Lee et al., 2019a)	Wri	Wikipedia	OB	6M / 21M	79.2k	8.8k	3.6k	X	X	X
PAQ (Lewis et al., 2021)	Gen	Wikipedia	OB	6M / 21M	65M	X	X	X	X	X
CronQuestions (Saxena et al., 2021)	Templ.	KG	KG		350k	30k	30k	X	KG with temporal information	Yes
TempQuestions (Jia et al., 2018)	Wri-other	KG	KG				1.2k			Yes
TimeQuestions (Jia et al., 2021)	Wri-other	KG	KG		9.7k	3.2k	3.2k	X	KG with temporal information	Yes
TORQUE (Ning et al., 2020)	Wri	3.2k short passages	Known-source		24.5k	1.5k	4.6k	X	X	Yes
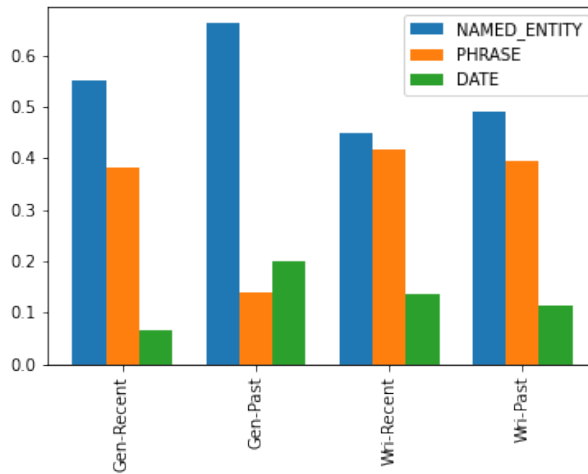
Figure 9. Answer type proportions in the StreamingQA evaluation sets.

A. StreamingQA Dataset

A.1. Comparison of StreamingQA with other related datasets

In Table 4, we present a comparison of StreamingQA with other related datasets.

A.2. Answer types in evaluation set

We provide further detail on the answer types in our evaluation set in Figure 9.

Table 5. Examples of questions with high and low frequency named entities.

| |
|---|
| *Questions with high frequency named entities* |
| When will Google have its annual I/O conference? |
| What is the name of the Greek Prime Minister's residence? |
| Which former New York City Mayor is developing mobile apps to help New York state trace coronavirus cases? |
| *Questions with low frequency named entities* |
| What is the name of the Managing Director and Chief Executive of Latitude Financial Services? |
| Which actress plays Vanessa Woodfield? |
| Why must John Momis step down? |
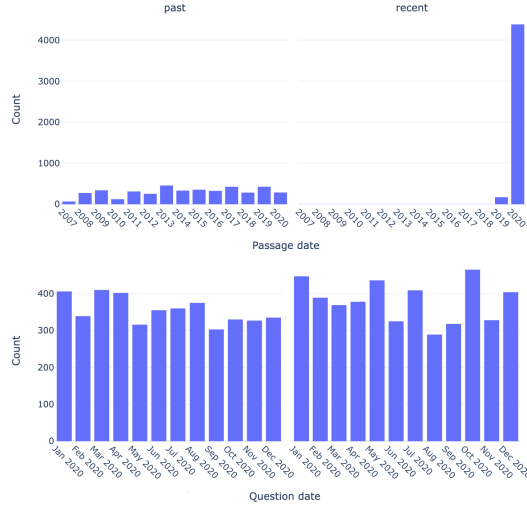
Figure 10. Question and article publication date distribution for Eval-Written. Eval-Generated is similar.

A.3. Examples of questions with high- and low-frequency named entities

See examples in Table 5 of question with high and low frequency named entities.

A.4. Automatic filtering of trivial and/or low-quality generated questions

In order to remove trivial and/or low-quality questions, we apply the following filters: (i) remove questions that contain their answer as a sub-span; (ii) few-shot prompt a large LM for QA and ensure it generated the original target answer given the evidence document and the generated question for named entity or date answers exactly, and for phrases with 40% words overlap; (iii) additionally we perform Google Search via the Google Search API[12] with the question text and evidence publication date as a query, and keep only questions for which the answer is present in the top 10 search results; and (iv) for phrase-answer questions, we only keep questions that contain a named entity in the question, hence eliminating questions that are too generic.

A.5. Adding time specification to past generated questions

For all questions we include an absolute or a relative time specification, choosing randomly between the two. For absolute time specification we include the month and the year of the article's publication date, for example, "in May 2017". For relative, we compute the difference between the question date and the article publication date, and include "" (< 7 days), "a week ago" (< 1 days), "N weeks ago" ($<= 8$ weeks), "N months ago" (< 2 years), or "N years ago".

B. Experiments and Results

B.1. Closed-book

We present the F1 scores broken down by answer type in Figure 11, and the effect of temporal lag between model knowledge and question dates for written questions in Figure 12.

Perplexity vs closed-book QA performance. An interesting point of comparison is between closed-book QA performance and the perplexity of the underlying LM on test documents. As TXL_{FT} is fine-tuned on more months, we expect its perplexity on evidence documents of the recent subset to reduce, while its perplexity on evidence documents of the past subset to either stay the same (in the optimal scenario) or increase, if the model forgets. Figure 13 shows these two effects.
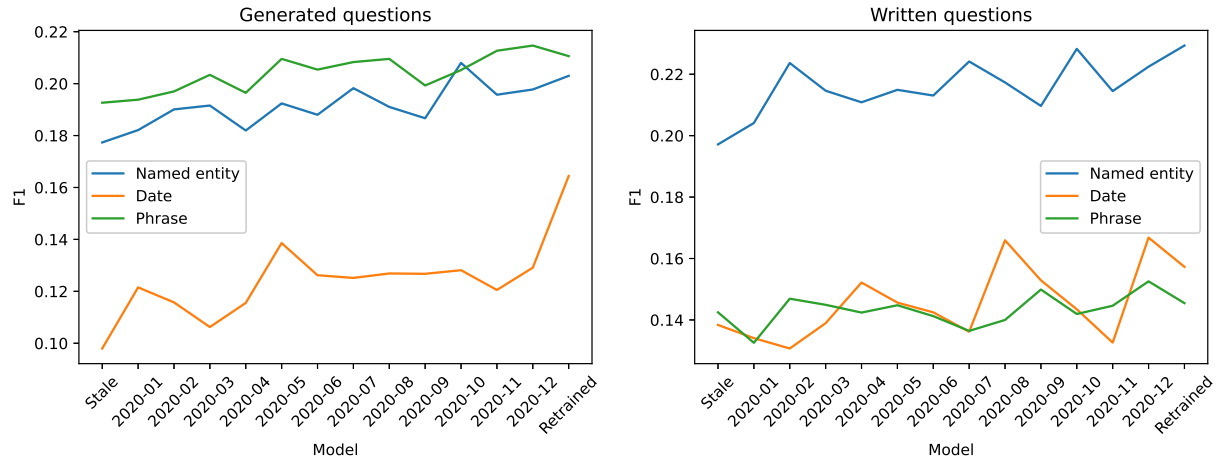
Figure 11. F1 score on the whole dataset of models fine-tuned on data until different cut-off dates, by answer type.
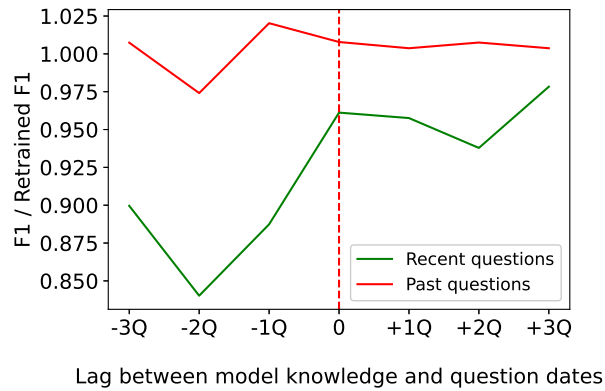

Figure 12. The effect of temporal lag between the final training month of CB_{+FT} and question dates for written questions, relative to $CB_{+RETR.}$.
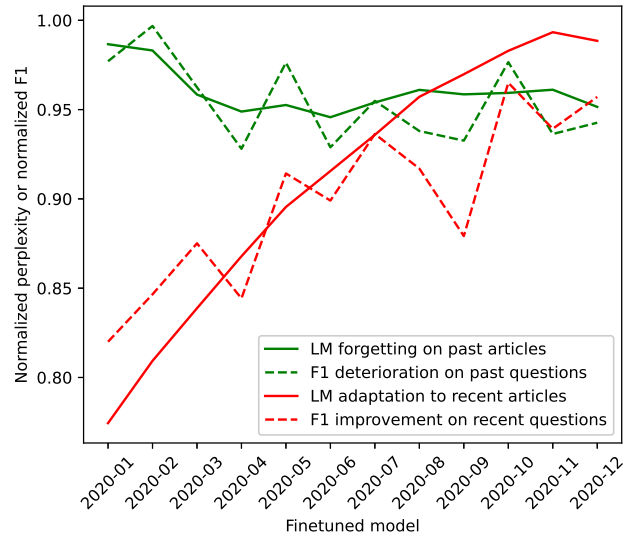
Figure 13. Relationship between adaptation and forgetting of LM (solid) and CB QA models (dashed). Red lines show fine-tuned vs stale adaptation/improvements to recent articles on the recent subset. Green lines show fine-tuned vs retrained forgetting of past articles on the past subset. LM forgetting is expressed as $\text{TXL}_{\text{STALE}}$ perplexity / TXL_{FT} perplexity and F1 deterioration as $\text{CB}_{\text{+FT}}$ F1 / $\text{CB}_{\text{+STALE}}$ F1, whereas LM adaptation is expressed as $\text{TXL}_{\text{RETR.}}$ perplexity / TXL_{FT} perplexity and F1 improvement as $\text{CB}_{\text{+FT}}$ F1 / $\text{CB}_{\text{+RETR.}}$ F1.
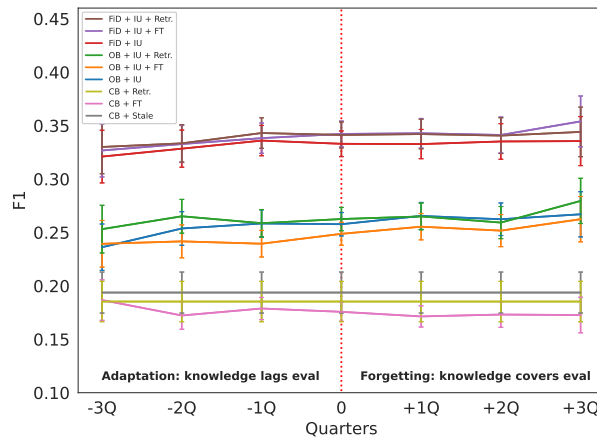

Figure 14. Adaptation and forgetting on past written questions: Open-book versus closed-book.
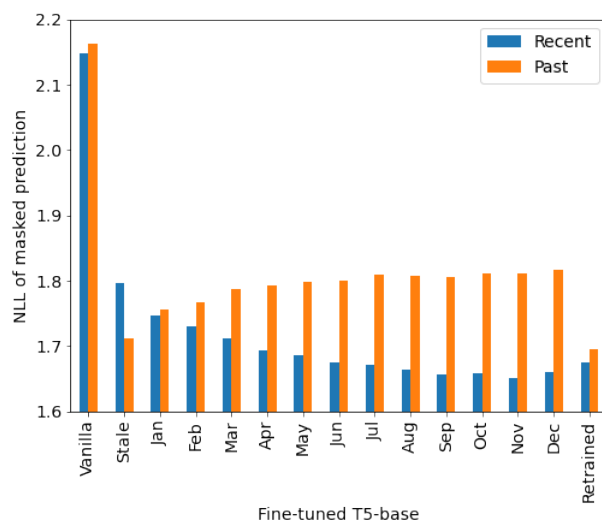
Figure 15. Negative log likelihood of masked span prediction for evaluation documents of the T5. We show the vanilla T5, T5 fine-tuned on WMT up to 2019 ("Stale"), monthly fine-tuned, and trained on all WMT up to 2020. We see forgetting on the past subset, and a slight recency bias on the recent subset, compared to retraining.

B.2. Open-book

We show adaptation and forgetting on past written questions in Figure 14, and the masked span prediction performance for evaluation documents of the T5 model in Figure 15.

B.3. Temporal accuracy of DPR trained on news

Figure 16 shows temporal distribution of gold and retrieved passages for recent questions for Q4'2020: DPR with timestamp matches the temporal distribution of the gold passages much closer. Interestingly, the model seems to get somewhat confused about the year—the second lower spike is in Q4'2019.

B.4. One-step Streaming and Static QA Benchmarks

We provide the EM in Figure 17 and all the metrics in Table 6.

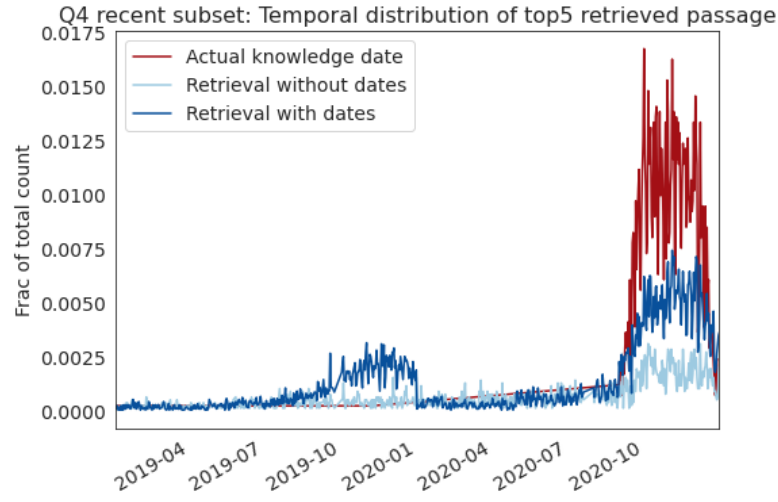[12]https://developers.google.com/custom-search

Figure 16. Temporal distribution of gold and retrieved passages of the recent questions for Q4'2020. DPR with timestamp matches the temporal distribution of the gold passages much closer. Interestingly, the model seems to get somewhat confused about the year - the second lower spike is in Q4'2019.
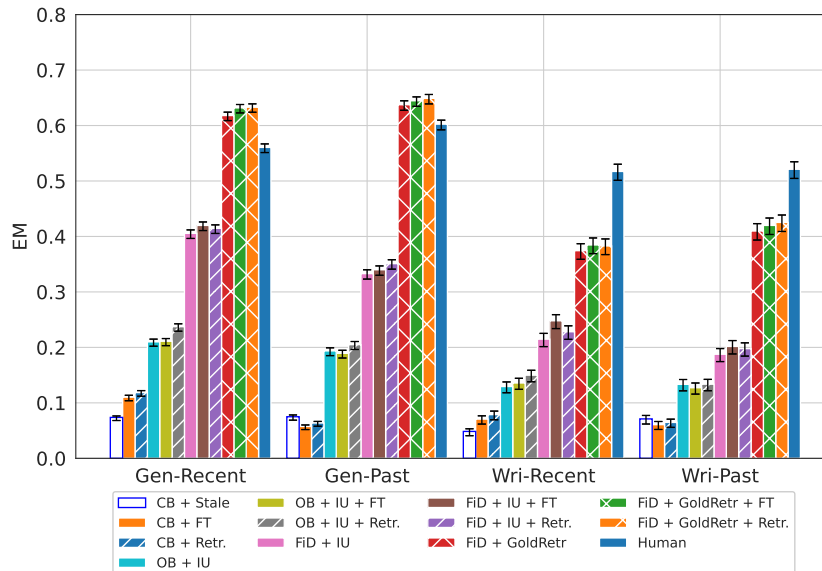

Figure 17. Static setup with all questions. Solid filled bars are models that had 2020 knowledge added incrementally.

Table 6. Static setup.

Model	Recent		Past	
	EM	F1	EM	F1
Generated				
CB + Stale	0.0738	0.1825	0.0750	0.1578
CB + FT	0.1103	0.2289	0.0575	0.1487
CB + Retr.	0.1185	0.2392	0.0637	0.1502
OB + IU	0.2100	0.3276	0.1937	0.2910
OB + IU + FT	0.2110	0.3278	0.1893	0.2832
OB + IU + Retr.	0.2374	0.3572	0.2049	0.3022
FiD + IU	0.4056	0.5341	0.3330	0.4330
FiD + IU + FT	0.4199	0.5498	0.3400	0.4414
FiD + IU + Retr.	0.4146	0.5421	0.3510	0.4494
FiD + GoldRetr	0.6180	0.7271	0.6375	0.7202
FiD + GoldRetr + FT	0.6318	0.7326	0.6446	0.7233
FiD + GoldRetr + Retr.	0.6330	0.7339	0.6490	0.7265
Human	0.5606	0.7742	0.6025	0.7806
Written				
CB + Stale	0.0485	0.1526	0.0711	0.1838
CB + FT	0.0707	0.1935	0.0607	0.1803
CB + Retr.	0.0788	0.1974	0.0647	0.1744
OB + IU	0.1295	0.2603	0.1334	0.2587
OB + IU + FT	0.1358	0.2629	0.1273	0.2501
OB + IU + Retr.	0.1499	0.2838	0.1336	0.2623
FiD + IU	0.2148	0.3675	0.1876	0.3312
FiD + IU + FT	0.2480	0.4005	0.2017	0.3402
FiD + IU + Retr.	0.2282	0.3862	0.1977	0.3390
FiD + GoldRetr	0.3744	0.5676	0.4098	0.5983
FiD + GoldRetr + FT	0.3847	0.5721	0.4199	0.5996
FiD + GoldRetr + Retr.	0.3829	0.5693	0.4253	0.6032
Human	0.5172	0.7435	0.5211	0.7401