How do QA models combine knowledge from LM and 100 passages?

Anonymous ACL submission

Abstract

Retrieval-based generation models achieve high accuracy in open retrieval question answering by assessing rich knowledge sources — multiple retrieved passages and parametric knowledge in the language model (LM). Yet, little is known about how they blend information stored in their LM parameters with that from retrieved evidence documents. We study this by simulating knowledge conflicts (i.e., where parametric knowledge suggests one answer and different passages suggest different answers). We find that retrieval performance largely decides which knowledge source models use, and a state-of-the-art model barely relies on parametric knowledge when given multiple passages. When presented with passages suggesting multiple answers, however, models use parametric knowledge to break the ties. We discover a troubling trend that contradictions in diverse knowledge sources affect model confidence only marginally. Together, our study helps interpreting answers from these models and suggests directions for future work.

1 Introduction

Traditionally, QA models have relied on retrieved documents to provide provenance for their answers (Chen et al., 2017). More recent studies (Petroni et al., 2019) have shown that large language models are able to retain vast amounts of factual knowledge seen during pretraining, and closed-book QA systems (Roberts et al., 2020) build upon this foundation by memorizing facts from QA finetuning. Retrieval-based generation approaches (Izacard and Grave, 2021; Lewis et al., 2020) emerge as the best of both worlds – generating free-form answers from the question paired with retrieved evidence documents. They further combine these parametric knowledge sources with a large number of retrieved evidence documents, achieving state-of-the-art performances on open retrieval QA datasets (Joshi et al., 2017; Kwiatkowski et al., 2019).

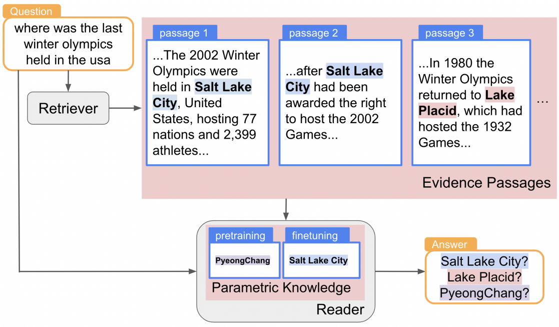
Question
where was the last winter olympics held in the usa
Retriever
passage 1
...The 2002 Winter Olympics were held in Salt Lake City, United States, hosting 77 nations and 2,399 athletes...
passage 2
...after Salt Lake City had been awarded the right to host the 2002 Games...
passage 3
...In 1980 the Winter Olympics returned to Lake Placid, which had hosted the 1932 Games...
...
Evidence Passages
pretraining
PyeongChang
finetuning
Salt Lake City
Parametric Knowledge
Reader
Answer
Salt Lake City?
Lake Placid?
PyeongChang?

Figure 1: Retrieval-based generation models use two main knowledge sources (indicated by the red background), the retrieved evidence passages and the parametric knowledge. In this example, the pretrained reader suggests *PyeongChang* as the answer, which frequently co-occur with Winter Olympics, while the evidence passages suggests two answers (*Salt Lake City/Lake Placid*). We investigate for a given question which knowledge source was the most influential to output an answer.

Understanding how retrieval-based generation models combine information from parametric and non-parametric knowledge sources is crucial for interpreting and debugging such complex systems, particularly in adversarial and complex real world scenarios where these sources may conflict with each other (see an example in Figure 1). This will be helpful for both developers to debug such models and for users to estimate how much they should trust an answer. Thus, we focus on the following core question: when provided with numerous evidence passages and a rich, pretrained and finetuned language model, which knowledge source do models ground their answers in?

A recent study (Longpre et al., 2021) investigated this in a limited single evidence document setting. We expand this study to consider a more realistic scenario, where models consider *multiple* evidence passages (e.g., 100 passages), and observe results diverging from their reported heavy reliance on parametric knowledge. We further simulate a setting where a *subset* of evidence passages are perturbed to suggest a different answer to reflect the realistic scenario where retrieval returns a mixed bag of information. Such scenarios are common in settings where some passages are updated with new information, while other passages remain outdated (Shah et al., 2020; Zhang and Choi, 2021). Such conflicts can also occur when passages are adversarially edited to contain false information, or when passages are authored by multiple people who have differing opinions about an answer (Chen et al., 2019). We find that retrieval-based generation models are primarily extractive and are heavily influenced by a few most relevant documents instead of aggregating information over a large set of documents.

Having identified that models mostly rely on evidence passages rather than parametric knowledge, we further evaluate how sensitive models are toward semantic perturbation to the evidence documents (e.g., adding negation). We simulate a scenario where a subset of passages suggest one answer, while the remaining passages reject the answer. We find that retrieval-based generation models behave similarly to extractive models, sharing their weakness of outputting answer candidates with high confidence, even after the context is modified to no longer support the answer (Ribeiro et al., 2020). Calibration analysis reveals that a model's confidence does not decrease despite contradictions between knowledge sources for a large subset of examples (30-40%).

To summarize, we empirically test how QA models use diverse knowledge sources in a multi-passage setting. Our findings are as follows: when provided with a high recall retriever, models rely almost exclusively on the evidence passages without hallucinating answers from parametric knowledge. When different passages suggest multiple conflicting answers, models prefer the answer that matches their parametric knowledge, and most of this parametric knowledge comes from finetuning rather than pretraining. Lastly, we identify various weaknesses of retrieval-based generation models, which tend to copy answers with high confidence even when there is insufficient support for them in the retrieved evidence. Furthermore, model confidence does not reflect the existence of conflicting answers between knowledge sources. We suggest that future modeling should focus on proper calibration of presenting a single answer in the presence of rich, potentially conflicting, knowledge sources.

Model	Generative	Retrieval-Based	Multi-Pass
DPR		✓	
REALM		✓	
T5	✓		
RAG	✓	✓	
FiD	✓	✓	✓

Table 1: Overview of recent open retrieval QA approaches. *Generative* indicates whether the model generates the answer and, therefore, can produce answers not found in the retrieved documents. *Retrieval-Based* indicates whether the model uses retrieval to find relevant passages to help produce an answer. *Multi-Passage* indicates whether the system is able to model interactions between separate evidence passages.

2 Background

We study open retrieval question answering, where the goal is to find an appropriate answer y^* for a given question q. Systems for open retrieval QA may also be provided with access to a knowledge corpus consisting of a large number of passages, p, which is used to help answer the question. For the remainder of this paper, we use the open retrieval split (Lee et al., 2019) of the NaturalQuestions dataset (NQ-Open) (Kwiatkowski et al., 2019) and use Wikipedia as our knowledge corpus.[1]

2.1 Model

We investigate two prominent retrieval-based generation QA models: Fusion-in-Decoder (Izacard and Grave, 2021) and Retrieval Augmented Generation model (Lewis et al., 2020). Both architectures have reader and retriever components. They both use the same dense phrase retriever (Karpukhin et al., 2020) which learns an embedding of question and passage, and retrieves a fixed number (N) of passages that are most similar to the query embedding. They mainly differ in their reader architecture and learning objective, which we describe below.

Fusion-in-Decoder (FiD) The reader model is based on pretrained language model (specifically, T5-large (Raffel et al., 2020)). Each retrieved passage, p_i $(i = [1, N])$, is concatenated with the question, q, before being encoded by T5 to generate representations, $[h_1^i, ..., h_m^i]$, where m is the length of the ith passage prepended with the question. All

[1] Following Lee et al. (2019), we use the English Wikipedia dump from Dec. 20, 2018.

N passages are then concatenated to form a single sequence, $[h_1^1, ..., h_m^1, ..., h_1^N, ..., h_m^N]$, which the decoder interacts with using cross-attention to generate the answer.[2]

We use trained FiD (large) checkpoint provided by the authors for most analysis.[3] When evaluating models with access to different number of passages, we re-train FiD model (pretrained weights loaded from T5-large) using 1, 5, 20 and 50 passages retrieved by DPR. Refer to Appendix A.2 for full model and training details.

Retrieval Augmented Generation (RAG) RAG conditions on each retrieved evidence document individually to produce an answer, marginalizing the probability of producing an answer over all retrieved evidence documents.[4] By applying this constraint, RAG is able to jointly train the reader and retriever during finetuning, at the cost of ignoring interactions between evidence documents. FiD, in contrast, is able to model such interactions during decoding while the reader and retriever is completely disjoint.

Recent followup work explored jointly training the reader and retriever in FiD (Izacard and Grave, 2020; Sachan et al., 2021; Yang and Seo, 2020), showing small gains. Table 1 summarizes different architectures, including two prominent open book approaches (Karpukhin et al., 2020; Guu et al., 2020), one closed book approach (Roberts et al., 2020) and two retrieval-based generation approaches. As FiD shows a strong performance and efficiently use multiple passages at decoding time, we focus most of our analysis (Section 4 and Section 5) on FiD model.

2.2 Model Confidence Study

We analyze the model confidence score, asking a more nuanced question: *is model's confidence on the gold answer decreased after we perturb knowledge sources?* We compare the model confidence on the same example before and after perturbation. We determine the confidence of the model using either (1) the generation probability of the answer or (2) the confidence score of separately trained answer calibrator, which provides a score indicating the probability of the model correctly predicting the answer for each example.

Model	retrieval success	data %	Extractive %	EM	Abstractive %	EM
FiD	Y	88.6	98.3	59.60	1.7	-
	N	11.4	82.9	-	17.1	21.3
	Total	100	96.6	53.9	3.4	12.4
RAG	Y	62.5	92.9	60.2	7.0	-
	N	37.5	57.9	-	42.1	11.2
	Total	100	79.8	43.9	20.2	9.6

Table 2: Proportion of model predictions on the NQ-Open development set that are extractive vs. abstractive along with their exact match accuracy (EM). Results are split based on whether the retrieval was successful (i.e., gold answer string is within the top K retrieved documents (Y), or if the answer string cannot be found in the top K passages (N)). For the FiD model, K = 100; for the RAG model, K = 5. Overall, we observe that retrieval based generation models are primarily extractive, predicting answers strings within the retrieved passages over 75% of the time. '-' means cells that have zero performance by definition.

We train a binary calibrator following prior work (Kamath et al., 2020; Zhang et al., 2021), using gradient boosting library XGBoost (Chen and Guestrin, 2016). The goal of the calibrator is to enable selective question answering – equipping models to decide when to abstain from answering. Given an input question q and learned model M_θ, the calibrator predicts whether the predicted answer $\hat{y} = M_\theta(q)$ will match the annotated answer $y*$. While model confidence is a good proxy, prior work showed that separately training a calibrator shows further gains.

The input to the calibrator is the concatenation of the generation probability and the encoder feature representation averaged across length, and the output is a score indicating the probability of the model correctly predicting the answer. We reserve 1K examples of NQ Open training set for validation, and trained our calibrator on the remaining data. Hyperparameters are selected based on AUROC on validation set and reported in Appendix A.1.

3 When retrieval-based generation models rely on parametric knowledge?

As an initial step investigating whether retrieval-based generation models ground their answers in the retrieval corpus or in the pretrained language model's parametric knowledge, we evaluate

[2] We use the version proposed in Izacard and Grave (2020), where knowledge distillation from reader to retriever is performed to enhance the retriever.

[3] https://github.com/facebookresearch/FiD

[4] RAG also presents a variant of a model that relies on each retrieved document to generate for each token, but shows worse performance. We use the version in https://huggingface.co/facebook/rag-sequence-nq

Model	# Pass. train / inf.	% ex.	Ans. R	Exact Match Orig.	 Sub.	M_R
FiD	1 / 1	-	-	17	47	27
FiD	1 / 1	17.8	48.5	10.1	61.1	14.1
RAG	5 / 1	16.3	62.5	10.3	65.9	13.5
RAG	5 / 5	20.4	62.5	11.6	63.7	15.3
FiD	5 / 1	19.4	72.9	2.8	74.9	3.6
FiD	5 / 5	24.5	72.9	2.5	68.8	3.5
FiD	20 / 1	18.4	83.1	1.1	73.0	1.4
FiD	20 / 20	25.1	83.1	1.4	67.0	2.1
FiD	50 / 1	17.8	86.8	0.3	82.0	0.4
FiD	50 / 50	27.7	86.8	1.0	71.5	1.4
FiD	100 / 1	18.1	88.7	0.6	81.5	0.8
FiD	100 / 100	29.5	88.7	2.5	65.3	3.8

Table 3: Exact Match / Memorization Ratio for FiD model with different amount of passages. The results in the first row are reported in Longpre et al. (2021), which uses MRQA version of NQ (Fisch et al., 2019) dataset. All other rows use NQ-Open split. We do not report results for RAG with 1 training and 1 inference passage, as in a single document setting it is equivalent to FiD with 1 / 1 passages with the exception of the choice in pretrained LM (T5 vs. BART). % ex. reports the number of examples in each set, after removing examples without an entity answer and where model made incorrect predictions.

Question: When was the last time the Bills won their division?

Type		Passage	Answer
None	Original Entity	. . . the **1995** Bills **won** the AFC East . . .	1995
Entity Sub.	Random (Same Type)	. . . the **1936** Bills won the AFC East . . .	1936
	Negation	. . . the 1995 Bills **did not win** the AFC East . . .	-
Semantic Pert.	Modality	. . . the 1995 Bills **might win** the AFC East . . .	-
	Future	. . . the 1995 Bills **will win** the AFC East . . .	-
	Text Infilling	. . . the 1995 Bills **lost** the AFC East . . .	-

Table 4: Example perturbations. Entity substitutions modify the passage by replacing the answer entity mention with another answer candidate of the same entity type. Given the modified passage, the new answer is the substitute entity. Semantic perturbation modifies the main verb of the answer sentence such that the answer to the question is no longer valid, without introducing a new answer candidate.

whether model generates novel answer this is not present in a set of evidence documents. Unlike extractive QA models (Seo et al., 2017), generation based approaches (Roberts et al., 2020; Izacard and Grave, 2021) do not require the evidence document to contain the gold answer span.

Table 2 reports how often models generate a span not found in the evidence passages, split by the retrieval performance on the NQ-Open development set (Kwiatkowski et al., 2019; Lee et al., 2019). We observe that models typically copy a span from the evidence passages, only generating novel spans for 3.4% of examples for FiD and 20.2% for RAG. Even for the small subset of examples where the retrieved documents do not contain answer string, FiD remains extractive for over 80% of such examples. In contrast, for RAG, where retrieved documents frequently miss the gold answer (37.5%), such copying behavior was less common, generating unseen text for 42.1% of examples. These results suggest that models rely heavily on the retrieved documents instead of parametric knowledge in LM only when the retriever performance is strong.

Revisiting knowledge conflict study in Longpre et al. (2021) This observation stands at odds with the study from Longpre et al. (2021), which showed that models frequently rely on parametric knowledge, generating answers not present in the evidence passage. This recent work took the first step into analyzing knowledge sources in retrieval-augmented generation models, investigating whether they rely on parametric knowledge from the language model or information from retrieved passages. They simulated knowledge conflicts by substituting the existing answer with a new answer candidate in the evidence passage (see Table 4 for an example). The original passage is minimally changed, yet now suggests an alternative, incorrect answer candidate that likely contradicts with knowledge from language model. Longpre et al. (2021) assumed the model has access to a single evidence document.

We identify that the main difference in their experimental setup is in using a **single** evidence passage rather than multiple evidence passages. We re-visit their study, as we find single document setting to be unrealistic. Most open-retrieval QA models (Lewis et al., 2020; Karpukhin et al., 2020; Izacard and Grave, 2021) are trained with multiple passage to make up for imperfect passage retrieval. When the model is provided with 100 passages, the correct span is available nearly 90% of the time (compared up to 50% when provided one passage), thus the model remains extractive.

Following their experimental setup, we only evaluate on examples that the model has correctly answered (as perturbing examples where models are already confused is unnecessary) and where the answer is an entity[5]. We then substitute every answer entity mention in *all evidence passages* with a random entity of same type.[6] All manipulation was done only at inference time.

We report the exact match score to the original annotated answer. Prior to perturbation, the exact match score against the original answer is 100%. We also report the exact match score to the substituted answer and memorization ratio (M_R):

$$M_R = \frac{p_o}{p_o + p_s}$$

where p_o is the fraction of examples where the model predicts the original answer, and p_s is the fraction of examples predicting the substitute answer.

Table 3 reports how models respond to entity-substituted contexts with a differing number of passages available at training and inference time. In congruence with our prior experiments, we observe higher reliance on parametric knowledge as answer recall in the retrieved evidence decreases. Departing from Longpre et al. (2021), we find that memorization in FiD is rare when reader is provided with multiple passages at training time, and FiD grounds its answers mostly in evidence passages instead of its parametric knowledge when answer recall is reliably high. Furthermore, when provided with multiple evidence passages with comparable answer recall, FiD exhibits far less memorization than RAG, suggesting that using a multi-passage reader that doesn't marginalize over passages also inhibits memorization.

4 Simulating Mixed Bag of Evidence Passages

Having identified that retrieval-based generation models rely heavily on evidence passages, especially when paired with a high-performance retriever, we study **how models make use of multiple evidence passages when different passages suggest different answers**. This happens frequently in real life, as questions can be ambiguous based on different, valid interpretations of question (Min et al., 2020) or different extra-linguistic contexts suggesting different answers (Zhang and Choi, 2021).

We introduce two perturbations – an entity substitution perturbation inspired by Longpre et al. (2021) and adversarial semantic perturbation (Jia and Liang, 2017) – both will dissuade model from returning the original answer in the evidence passage. Table 4 presents example perturbations. We will present entity substitution perturbation (Section 4.1), and then present results on semantic perturbations (Section 4.2). As in section 3, we only consider examples where the FiD model answers correctly with the original passages and analyze best FiD model trained with 100 passages.

4.1 Entity Substitution

Setting. To simulate a mixed bag of evidence passages, we perform partial entity substitution, changing answers to a subset of passages mentioning the answer entity. On average, the answer entity is mentioned in 16.7% of 100 retrieved evidence passages for NQ-open dataset. We substituted 25%, 50%, 75% and 100% of evidence passages that contains the original gold answer span with a new entity. We sample passages to substitute answer entity in three ways.

- random: randomly sample n% of passages
- top-retrieval: select top n% of passages with highest retrieval score
- top-attention: select top n% of passages where reader model pays the most attention to. The attention score for each passage is computed as the cross-attention score on the first decoded token averaged across layers, heads and the tokens in the passagage, as defined in Izacard and Grave (2020).

Results. Figure 2 reports the results with different amount of perturbation (i.e., how many evidence passages are perturbed) and different sampling of passages to substitute entity. The model refrain from predicting the original answer if none of the passage contains it. However, when we perturbed about half of randomly chosen passages, the model favors the original answer almost twice as frequently (52.5 vs. 25.1), indicating the model relies on parametric knowledge when there are more than one potential answer.

When we perturb the top scoring passages, either by retrieval score or attention score, the model changes its answer a lot more frequently. Here, even perturbing only 25% of passages changes the gold answer for about 30% of examples compared to 8% of examples for random sampling. This sug-

[5] Exact numbers of filtered examples are shown in Table 3.

[6] The entity type is coarsely defined as person (PER), date (DAT), numeric (NUM), organization (ORG), and location (LOC).

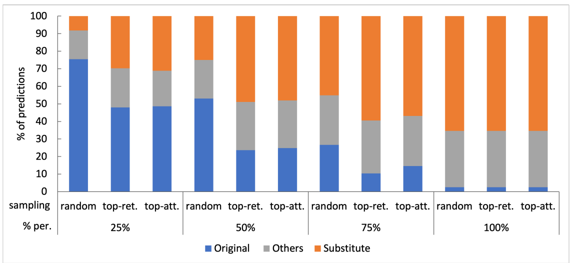

Figure 2: Substituting different proportion of retrieved passages containing gold answer spans on filtered NQ Open test set (N=2,581). We vary the sampling of retrieved passages to substitute answers.

% Perturbed	Gen. Prob.	Calibration
25	48.15%	57.07%
50	49.67%	56.30%
75	49.86%	56.84%
100	52.22%	56.95%

Table 5: The percentage of examples in which model confidence on the correct answer dropped after perturbation for partial substitution in NQ Open test set.

gests lowly ranked retrieved passages might contain the answer entity without being relevant to the query, which model successfully ignores.

Confidence Study. Table 5 reports the change in model confidence after the random entity substitution in the evidence passages. Repeating the results from Zhang et al. (2021), separately trained calibrator consistently outperforms the model's inherent confidence score. Surprisingly, there is no clear connection between the percentage of perturbed passages and model confidence. One possible explanation is that the model focus on a **single** passage which contains the most likely answer without aggregating information from multiple passages.

To further investigate this, we substitute answers in all passages except top K passages, ranked by the attention score from the reader. Table 6 presents the results. If you change the answer to all passages except for the top scoring article, the model outputs the substituted answer for about 30% of times, producing the original answer on about half of the articles (51.80%). As long as we preserve three top scoring articles, the model is less impacted by the changes in the rest of the retrieved passages, returning the original answer for almost 80% of examples. This suggests model might ignore many retrieved passages and focus on only a handful of most relevant passages.

k	Original	Substitute
1	51.80	28.94
3	79.66	8.80
5	87.29	4.03

Table 6: Substituting all the passages except top k passages (k=[1,3,5]), which are selected based on passage attention scores. On average, 16.7 passages out of 100 passages contained gold answer entity. Yet, with access of up to 3 passages containing the gold answer span, the FiD model can still generate the original answer 80% of the time.

4.2 Adversarial Semantic Perturbation

Semantic perturbation follows earlier work on counterfactual example generation with heuristics (Ribeiro et al., 2020) which perturbs the sentence containing the answer. We simulate four perturbations, and after each perturbation, the model should **refrain** from returning the original answer. We aim to test model's understanding of the passage with such perturbation.

Setting. We design the four perturbations applicable to question answering: negation, changing to future tense, adding modal verb and text infilling. We run a dependency parser on the sentence containing the gold answer span.[7] We filter examples where the root token of answer sentence is not a verb (about 40% of sentences, see Appendix A.3 for full statistics). Then, we apply simple rules (see Appendix A.4) to modify the verb. For text infilling, the only difference is that we convert the root token into “[blank]" and fill in the blank using language modeling (Donahue et al., 2020). When passages contain multiple gold answer spans, the changes are made to all of them as long as their root tokens are verbs.

Results. We report the exact match score to the original annotated answer. After the perturbation, the exact match score should drop significantly, as all edits invalidate the original answer. Table 7 showed the exact match score after semantic perturbations. The first two columns covers examples where we made at least one perturbation, but not all occurrence of answer string was perturbed (the coverage of perturbation is 67-86%). Thus, sometimes the gold answer span remains in the evidence passages, inducing models to return the original answer. To control for this, we further reports scores on examples where all evidence sentences

[7] We use StanfordNLP (Qi et al., 2018) toolkit.

	partial coverage		full coverage	
# passages	1	100	1	100
negation	82.49	86.80	74.71	71.26
modality	89.90	92.48	88.77	84.05
future	91.90	94.03	90.72	86.93
text-infilling	88.66	93.21	86.96	84.71

Table 7: Exact match score with the original answer after perturbation of each type: models largely disregard the perturbation and outputs the original answer.

containing the gold answer are perturbed (later two columns). We still observe models returning the original answer, similar to extractive models.[8]

Confidence Study. We repeat the calibration study with semantic perturbation. As separately trained calibrator was more robust than the model confidence score again, we report the ratio of calibration scores before and after the perturbation in Figure 3. The calibration score remain mostly steady after the perturbation (centered around 0). Slightly higher distribution lies in the negative x-axis, indicating the perturbations lowers the confidence for such examples (60-70% of examples). The exact numbers can be found in the Appendix A.8. Model was particularly less sensitive to temporal perturbation (future).

We observe that model behaves similarly to extractive model (Ribeiro et al., 2020), returning an entity answer matching the answer type with high confidence even when the passage no longer supports the answer.

5 Further Analysis

We further examine our results, focusing on the quality of substitute answer in entity substitution study and which parametric knowledge (pretraining vs. fine-tuning) was used.

Improving Substitute Entities Prior work (Longpre et al., 2021) substitutes answer entity with another entity with same coarse entity type. This makes substitute entities sometimes unreasonable, despite better than randomly sampling entities without type constraint. For example, “Heartbreak Hotel" was substituted as an answer to the following question “who did the lions play on thanksgiving last year”.

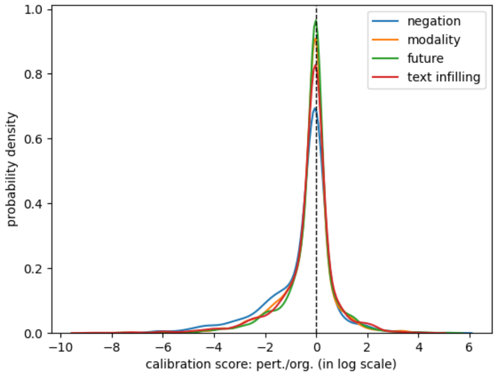

Figure 3: The ratio of calibration score after perturbation to that before perturbation, in log scale. The occurrences of examples of different ratio are plotted in terms of probability density (the area under curve is sum to 1). The distributions are bell-shaped, but shift slightly towards negative x-axis.

Entity source	AmbigQA (N=448)		SituatedQA (N=55)	
% per.	Ori.	Sub.	Ori.	Sub.
25	75.45	9.15	80.00	7.27
50	51.11	27.68	56.36	25.45
75	25.22	46.21	38.18	43.64
100	5.80	63.17	14.55	58.18

Table 8: Results of substituting different proportion of 100-retrieved passages on NQ Open where entities are derived from AmbigQA and SituatedQA dataset. The number next to the entity refers to the number of examples in this evaluation set after filtering.

We make perturbation more realistic by substituting with alternative answer from two datasets, AmbigQA (Min et al., 2020) and SituatedQA (Zhang and Choi, 2021), which augmented existing NQ open dataset. Both datasets annotated valid alternative answers for different interpretation of the same question (AmbigQA) and answers belonging to different temporal contexts (SituatedQA) for NQ Open dataset. We sample these additional answers as a new answer to inject (details in Appendix A.6).

Table 8 presents perturbation results with valid entities sourced from AmbigQA and SituatedQA. We identify a surprising trend – that model outputs original answers more frequently when substituted with better alternatives. This contradicts our intuition as model should be less hesitant to choose new substitute answer as they are also valid answer to the question, for different contexts. We further investigate this issue below.

Does parametric knowledge come from pretraining or fine-tuning? Some memorization (2–15%) remains even after all the evidence documents

[8] Further details about semantic perturbation (e.g., statistics of percentage of valid examples after each perturbation) can be found in the Appendix A.3.

% per.	Dataset	NAO	AO	AO%
50%	NQ (Random Entity)	62.32	68.54	85.93
50%	w/ AmbigQA Entity	54.16	67.61	78.35
50%	w/ SituatedQA Entity	50.00	69.77	94.55
100%	NQ (Random Entity)	0.00	4.43	85.93
100%	w/ AmbigQA Entity	1.59	10.16	78.35
100%	w/ SituatedQA Entity	0.00	21.05	94.55

Table 9: Memorization ratio (M_R of substituting different number of passages on NQ Open No Answer Overlap (NAO) / Answer overlap (AO) set. AO% signifies the percentage of examples that belong to AO set for each subset.

are perturbed, and model is biased toward the original answer under partial substitution. We aim to identify whether it comes from pretraining or fine-tuning of the reader model by using the evaluation data splits from prior work (Lewis et al., 2021): questions where answers were seen (Answer Overlap (AO)) and questions where answers were unseen (No Answer Overlap (NAO)). If memorization ratio is higher on AO set compared to NAO set, we can hypothesize that memorization mostly happens during fine-tuning compared to pre-training.[9]

Table 9 presents results for 50% and 100% substitution setting.[10] This study shed lights on mysterious trend: there were more examples with answer overlap in AmbigQA/SituatedQA subset. If we perturb all the evidence documents, the model exhibit little to no memorization on NAO portion. We can thus infer that memorization effect comes almost exclusively from fine-tuning. When accounting for different proportion of answer overlap examples in the subsets, memorization ratio is **lower** in AmbigQA/SituatedQA NAO set. This suggests that model uses parametric knowledge – which answer candidate is more reasonable – in a subtle way, even when behaving as a copying model.

6 Related Work

Recent analysis (Lewis et al., 2021; Krishna et al., 2021) pointed the overlap in training and evaluation dataset inflates question answering performances. Longpre et al. (2021) showed that the reader model tend to memorize entity answers despite the answer mentions are substituted by another entity. We showed that memorization do occur when the model can only have access to one passage, but can be reduced significantly if the model is trained with multiple passages. Concurrent work (Pan et al., 2021) investigates QA models' robustness to misinformation by providing contradicting contexts. Their finding echoes our finding that model is sensitive to frequency of answer span occurrences in the evidence passages. Our work further introduce alternative valid answers for the first time and careful sampling of evidence passage for perturbation.

Recent works evaluated robustness by minimally perturbing input examples (Kaushik et al., 2020; Gardner et al., 2020) to identify models that are invariant under distributional shift. Prior work explored automatically generating such perturbed input (counterfactual data) with heuristics (Ribeiro et al., 2020) or learned models (Wu et al., 2021; Bartolo et al., 2020; Paranjape et al., 2021). Our perturbation methods are rule-based similar to Ribeiro et al. (2020), but designed specifically for QA task.

7 Conclusion

We summarize our findings here:

1. Do models ground their answers from retrieved document or parametric knowledge? (Section 3) Current SoTA model ground its answer mostly from retrieved passages, **when** trained with a high recall retriever (Table 2, 3).
2. How does model use multiple passages when different passages suggest different answers? (Section 4.1) Model mostly use a few, most relevant passages (Table 6), and use parametric knowledge to break ties (Figure 2, Table 9).
3. How does model behave if some passages are perturbed **not** to support an answer? (Section 4.2) Model largely ignores semantic perturbations and outputs potential answer entity in the retrieved passages (Table 7).
4. How is the model's confidence score affected by knowledge conflicts? Confidence score is not very sensitive to knowledge conflicts (Table 5, Figure 3), and separately trained calibrator offers some improvements.
5. Does parametric knowledge come from fine-tuning or pre-training? Fine-tuning (Table 9).

We find retrieval-based generation models are primarily extractive even though they can generate novel answers. Being extractive also links to models' ignorance of semantic perturbations. Models show limited ability to aggregate conflicting information among its rich knowledge sources, encouraging future work in this domain.

[9] Earlier study (Longpre et al., 2021) in a single document setting also report memorization is more severe in AO set.

[10] See Appendix A.7 for 25% and 75% substitution setting.

References

Max Bartolo, A Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

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

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. *ArXiv*, abs/1906.03538.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. *arXiv preprint arXiv:2005.05339*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. *Proceedings of the International Conference on Machine Learning (ICML).*

Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, abs/1707.07328.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).*

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).*

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

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually augmented data. *International Conference on Learning Representations (ICLR).*

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. *ArXiv*, abs/2103.06332.

T. Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, C. Alberti, D. Epstein, Illia Polosukhin, J. Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics (TACL)*, 7:453–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, F. Petroni, V. Karpukhin, Naman Goyal, Heinrich Kuttler, M. Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS).*

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *EACL*.

Shayne Longpre, Kartik Kumar Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *ArXiv*, abs/2109.05052.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Liangming Pan, Wenhu Chen, Min-Yen Kan, and William Yang Wang. 2021. Contraqa: Question answering under contradicting contexts. *arXiv preprint arXiv:2110.07803*.

Bhargavi Paranjape, Matthew Lamm, and Ian Tenney. 2021. Retrieval-guided counterfactual generation for qa. *ArXiv*, abs/2110.07596.

Fabio Petroni, Tim Rocktäschel, A. H. Miller, P. Lewis, A. Bakhtin, Y. Wu, and S. Riedel. 2019. Language models as knowledge bases? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Marco Túlio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist.

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

Devendra Singh Sachan, Siva Reddy, William Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *NeurIPS*, abs/2106.05346.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. *International Conference on Learning Representations (ICLR)*, abs/1611.01603.

Darsh J. Shah, Tal Schuster, and Regina Barzilay. 2020. Automatic fact-guided sentence modification. In *AAAI*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. 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)*, pages 6707–6723, Online. Association for Computational Linguistics.

Sohee Yang and Minjoon Seo. 2020. Is retriever merely an approximator of reader? *ArXiv*, abs/2010.10999.

Michael J.Q. Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Knowing more about questions can help: Improving calibration in question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1958–1970, Online. Association for Computational Linguistics.

#Passages		neg. /modal. / fut.	negation-polyjuice	text-infilling
%Ex.	1	61.14%	57.38%	62.43%
	100	89.55%	88.18%	89.68%
%Cov.	1	85.77%	82.90%	86.07%
	100	66.93%	61.12%	68.25%
%Ex. (100% Cov.)	1	51.87%	46.39%	53.23%
	100.	15.13%	11.89%	16.06%

Table 10: Data statistics for different perturbations schemes. The first two rows are the numbers of examples, shown in percentage out of the examples that FiD can answer correctly. The third and fourth rows shows the percentage of gold answer span covered (valid for perturbation) in the chosen examples. The last two rows shows the percentage of valid examples we could get if all the gold answer spans are perturbed.

A Appendix

A.1 Calibrator Hyperparameter

We use 100 boosting rounds, subsample ratio of 0.5 and learning rate of 0.5. The same subsample ratio is applied for constructing each tree, for each level and for each split.

A.2 Model and Training Details

The Fusion-in-Decoder(FiD) model consist of a retriever and a reader module. The retriever (Karpukhin et al., 2020) is a BERT bi-encoder model, which calculate the similarity between the question q and each of the passages $\{p_i\}$ in the knowledge source and output the most similar ones. The similarity is computed as the dot product of the encoded vectors

$$E_Q(q)^T E_P(p_i)$$

where E_Q is the question encoder and E_P is the passage encoder.

The reader module is a pretrained T5-large (Raffel et al., 2020), an encoder-decoder model containing 770M parameters. Each passage is concatenated with the question and truncated to 250 word pieces. For our experiments finetuning FiD, we train the reader module with 1, 20, and 50 evidence passages. To train the reader, we use the AdamW optimizer (Loshchilov and Hutter, 2018) and a learning rate of $5 \cdot 10^{-5}$ with linear warmup of 8000 steps followed by linear decay to zero. The total training steps is 300k, and the final model checkpoint is selected based on exact match score on NQ Open development set. We only use batch size of 1 due to memory constraints. The models take roughly 7 GPU days to train on a Quadro RTX 8000 machine.

A.3 Perturbation Coverage

As mentioned in Section 4, if the root token of the answer sentence is not a verb, then we ignore that sentence, and thus some examples would be excluded. The first row shows the percentage of valid examples after applying the rules mentioned in Section 4. We consider it valid example if one of the gold answer span can be perturbed. The corresponding percentage of perturbed gold answer spans is shown in the third row. A small portion of gold answer spans remain unchanged after performing the perturbation. For the second and fourth row it shows the same except the model has access to 100 passages. The percentage of valid examples are much higher since we consider the example valid if one of the gold answer spans in any of the passages can be perturbed. The last two rows show the percentage of examples where all gold answer spans in all the retrieved passages can be perturbed.

A.4 Technical Details on Semantic Perturbations

For perturbation schemes except text infilling, we first identify the root token's part-of-speech tag. If it is in one of [VB, VBP, VBZ], then we treat it as the present tense, and modify the verb accordingly. (e.g. V → "does not V"/"do not V" for negation, V → "may V" for modality, V → "will V" for future tense) The lemmatized verb forms after "will" and "may" are obtained by the "WordNetLemmatizer" class in nltk[11]. We also identify ["is", "am", "are"] and modify the verbs into their corresponding forms. If the part-of-speech tag is VBD, then it is in past tense and the root token is modified similarly to present tense. Lastly, if the part-of-speech tag is VBN or VBG, then it is present/past participle or gerund. We then identify the be-verbs and/or ["had", "have", "has"], and perform modifications accordingly.

A.5 Model Tested on NQ Open Subset

Both AmbigQA and SituatedQA annotate subsets of NQ Open. To ensure identical data distribution and isolate the effect of different substitute answers, we report results of random entity substitution on AmbigQA set and SitutatedQA set respectively. We present the results in Table 11. For AmbigQA

[11] https://www.nltk.org/$_{m}odules/nltk/stem/wordnet.html$

% Perturbed	Random Entity (on AmbigQA set) Original	 Substitute	AmbigQA Entity Original	 Substitute	Random Entity (on SituatedQA set) Original	 Substitute	SituatedQA Entity Original	 Substitute
25	74.90	6.37	75.45	9.15	76.74	4.65	80.00	7.27
50	51.79	24.70	51.11	27.68	55.81	16.28	56.36	25.45
75	27.88	43.03	25.22	46.21	46.51	13.95	38.18	43.64
100	2.39	65.34	5.80	63.17	4.65	39.53	14.55	58.18

Table 11: Entity substitution results on subsets of NQ-Open. We perform random entity substitution on the AmbigQA and SituatedQA sets for fair comparisons between different sources of substitute answers.

% perturbed	NQ Open Original	 Substitute	AmbigQA Original	 Substitute	SituatedQA Original	 Substitute
25	67.35	9.51	72.16	7.21	66.67	0.00
50	45.50	27.51	40.20	34.02	33.33	33.33
75	21.85	48.84	22.68	41.23	0.00	66.67
100	0.00	68.12	1.03	63.92	0.00	66.67

Table 12: Exact match score of substituting different number of passages on NAO sets.

% perturbed	Exact Match Original	 Substitute
25	80.00	7.27
50	60.00	25.45
75	41.82	43.64
100	18.18	60.00

Table 13: Results of substituting different number of passages on SituatedQA. The substitute answer is randomly selected from the SituateQA answer set and is not in the original ansewr set.

Change Type	Gen. Prob.	Calibration
negation	65.94%	70.28%
modality	62.75%	66.34%
future	58.87%	62.92%
text-infilling	60.56%	64.36%

Table 14: The percentage of examples in which model confidence dropped after perturbation; i.e., the model confidence when predicting the original example is higher than the perturbed example. Model confidence is measured with generation probability/calibration.

subset, different substitute entity types (random or alternative valid entity) do not seem to affect the results too much. However, the model seems to bias toward the substitute answer more with valid alternative entity substitutions on SituatedQA subset, indicating the parametric knowledge of model do know which answers are more likely to be correct. One possible explanation is that AmbigQA answers do not always take the same form as the original ones (e.g. *76th season* and *1995* in Table 4).

A.6 Answer Entity Sampling Details

When substituting with AmbigQA answers, we consider only the examples with multiple valid answers. For each example, we randomly sample one answer not in the original answer set of NQ as the substitute answer. For substitution with SituatedQA answers, we select the most recent answer as substitute answer. We also include the result of randomly sample an answer from SituatedQA answer set in Table 13.

A.7 Full Results on No Answer Overlap Set

Table 12 contain the full results on NAO set for NQ Open, AmbigQA, and SituateQA.

A.8 Confidence Study Full Results

Table 14 contains the full results for confidence study on adversarial semantic perturbation.