Mitigating False-Negative Contexts in Multi-Document Question Answering with Retrieval Marginalization

Ansong Ni[♢*] **Matt Gardner**[♣] **Pradeep Dasigi**[♣]
[♢]Department of Computer Science, Yale University
[♣]Allen Institute for AI
ansong.ni@yale.edu
{mattg,pradeepd}@allenai.org

Abstract

Question Answering (QA) tasks requiring information from multiple documents often rely on a retrieval model to identify relevant information for reasoning. The retrieval model is typically trained to maximize the likelihood of the labeled supporting evidence. However, when retrieving from large text corpora such as Wikipedia, the correct answer can often be obtained from multiple evidence candidates. Moreover, not all such candidates are labeled as positive during annotation, rendering the training signal weak and noisy. This problem is exacerbated when the questions are unanswerable or when the answers are Boolean, since the model cannot rely on lexical overlap to make a connection between the answer and supporting evidence. We develop a new parameterization of set-valued retrieval that handles unanswerable queries, and we show that marginalizing over this set during training allows a model to mitigate false negatives in supporting evidence annotations. We test our method on two multi-document QA datasets, IIRC and HotpotQA. On IIRC, we show that joint modeling with marginalization improves model performance by 5.5 F1 points and achieves a new state-of-the-art performance of 50.5 F1. We also show that retrieval marginalization results in 4.1 QA F1 improvement over a non-marginalized baseline on HotpotQA in the fullwiki setting.[1]

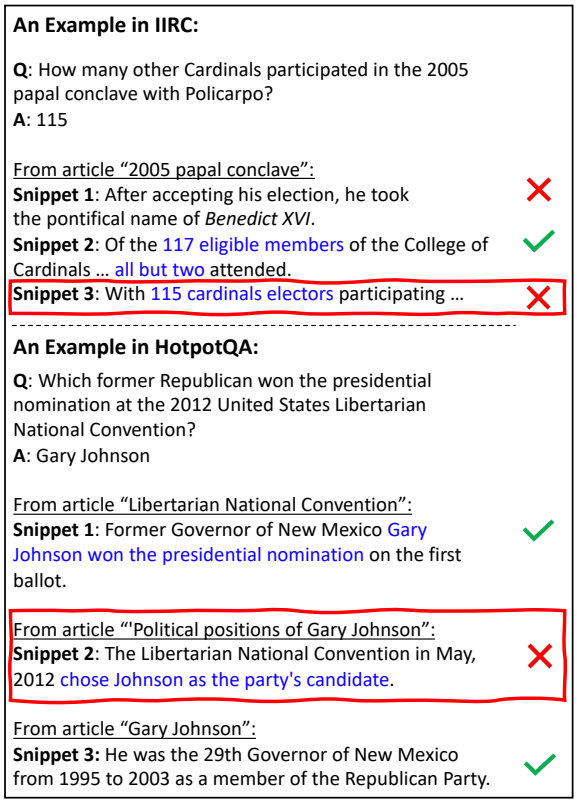

Figure 1: Examples of false-negative contexts in multi-document QA. Equivalent information is marked in blue, and only the snippets with "✔" are annotated as gold evidence. False negatives are outlined in red and both are retrieved by our proposed framework.

1 Introduction

Multi-document question answering refers to the task of answering questions that require reading multiple documents, extracting relevant facts, and reasoning over them. Systems built for this task typically involve retrieval and reasoning components that work in tandem. The retrieval component needs to extract information from the documents that is suitable for the reasoning model to perform the end-task effectively. Recent advances in reading comprehension have resulted in models that have been shown to answer questions requiring complex reasoning types such as bridging, comparison (Asai et al., 2020; Fang et al., 2020) or even arithmetic (Ran et al., 2019; Gupta et al., 2020), *given adequate context*. However, when the context needs to be retrieved from a large text corpus (*e.g.,* Wikipedia), the performance of such reading comprehension models is greatly affected by the quality of the retrieval model. Given supervision

[*]Majority of the work done as an intern at AI2.
[1]Code available at https://github.com/niansong1996/retrieval_marginalization.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6149–6161
November 7–11, 2021.

at all stages (*i.e.,* document, supporting evidence and answer), it is common to build retrieval and reasoning models independently and connect them as a pipeline at test time. In this case, the retrieval and reasoning models are usually trained to maximize the likelihood of labeled supporting evidence snippets and the answer given the gold context respectively.

However, in a multi-document QA setting, it is common to have some relevant snippets not marked as gold. Two such examples are shown in Figure 1. In the first example, only snippet 2 is marked as gold evidence, and consequently snippets 1 and 3 are treated as negative examples during retrieval. This is problematic because unlike snippet 1 which is actually irrelevant, snippet 3 is not only useful, but provides an even more direct way to derive the correct answer since it does not require a subtraction. Similarly, in the second example two evidence snippets from different documents contain the same information, thus at least two contexts can be used to answer this question, yet only one of them is labeled as being a positive example for the training objective. We define these contexts that contain non-gold snippets and can still be used to answer the questions as *alternative contexts*. Alternative contexts are inevitable when datasets are created from large corpora, because it is prohibitively expensive to exhaustively annotate all possible contexts. These alternative contexts are false negatives during training and lead to a noisy and weak learning signal, even with this "fully-supervised" setup.

We design a training procedure for handling these false negatives, as well as cases where retrieval should fail (*i.e.,* when the question is unanswerable). Specifically, we assign probabilities to documents, evidence candidates, and potential answers with parameterized models, and marginalize over a set of potential contexts by combining top retrieved evidence from each document, allowing the model to score false negatives highly. To make the marginalization feasible, we decompose the retrieval problem into document selection and evidence retrieval and show how we can still model contexts as sets. We evaluate our model on two multi-document QA datasets: IIRC (Ferguson et al., 2020) and HotpotQA (Yang et al., 2018). We see 2.8 and 4.8 F1 point improvement on IIRC and HotpotQA respectively by jointly modeling our proposed set-valued retrieval and the reasoning steps, and a further 2.7 and 4.1 F1 point improvement respectively by using retrieval marginalization. Our final result of 50.5 F1 on the test set of IIRC represents a new state-of-the-art.

2 Multi-Document QA

Here we formally describe the multi-document QA setting and highlight the two main challenges in this setting that our work attempts to address.

Problem Definition Multi-document question answering measures both the retrieval and reasoning abilities of a QA system. Given a question q and a set of documents $\mathcal{D} = \{d^1, d^2, ..., d^n\}$, each document containing a set of evidence snippets $d^i = \{s^i_1, s^i_2, ..., s^i_{n_i}\}$, the goal of the model is to output the correct answer a. This task is typically modeled with a retrieval step, which locates a set of evidence $C = \{s^{i_1}_{j_1}, s^{i_2}_{j_2}, ..., s^{i_k}_{j_k}\}$ to formulate a context, and a reasoning step to derive the answer from such context C. Though such models can be learned with or without annotations on supporting evidence, we focus on the fully-supervised setting and assume supervision for all stages. It is also common for such documents to have some internal structure (*e.g.,* hyperlinks in Wikipedia, citations for academic papers), which can be used to constrain the space of retrieval.

Inevitable False-negatives in Context Retrieval Annotations Even when supporting evidence is annotated, we claim that the learning signal provided by those labels may be weak and noisy when retrieving from a large corpus such as Wikipedia. This is due to the redundancy of information in such large corpora: it is common to have multiple sets of evidence snippets that can answer the same question, as in Figure 1. To quantify how often alternative contexts exist for the multi-document QA problem, we analyzed IIRC (Ferguson et al., 2020), an information seeking multi-document QA dataset. We sampled 50 answerable questions with their annotated gold context and manually checked if equivalent information can be found in sentences not labeled as supporting evidence, in the same document. We found that more than half of the questions have at least one alternative evidence, and on average [2] there is more than one sentence we can find in the same document that contains the

[2]We count up to 5 alternative evidence snippets per document when measuring the average. Some documents contain many more alternative evidence snippets. An example would be when a question seeks information on the "winner of World War I" in relevant Wikipedia pages.

same information as the gold evidence. Note that it is also possible to have alternative evidence in a different document, which would further increase the frequency of questions with false-negative contexts.

Due to the prevalence of such false-negative contexts, simply training the retrieval model to maximize the likelihood of the labeled supporting evidence would result in the models ignoring or even being confused by other unlabeled relevant information that could benefit the reasoning process. The problem is more severe when considering questions with Boolean answers or those that are unanswerable since the answers have no lexical overlap with corresponding evidence, making it harder to identify unlabeled yet relevant evidence snippets. Such false-negative context annotations are also inevitable in the data creation process, since the annotators will have to exhaustively search for evidence snippets from all relevant documents, which is rather impractical. As solving this problem is not typically feasible during data collection, we instead deal with it during learning.

Learning to Reason with Noisy Context Given retrieved supporting evidence as context, the second step of the problem is reading comprehension. Recently proposed models have shown promise in answering questions that require multi-hop (Asai et al., 2020; Fang et al., 2020) or numerical (Ran et al., 2019; Gupta et al., 2020) reasoning given small and sufficient snippets as contexts. However, the performance of such models degenerates rapidly when they are evaluated in a pipeline setup and attempt to reason with retrieved contexts that are potentially noisy and incomplete. For instance, Ferguson et al. (2020) found that the performance of reasoning models dropped 39.2 absolute F1 when trained on gold contexts and evaluated on retrieved contexts. This is mainly because the model is exposed to much less noise at training time than at test time.

Handling Retrieval for Unanswerable Questions It is especially challenging when we consider unanswerable questions since it is possible to have seemingly relevant documents that are missing key information.[3] This is common for an information-seeking setting such as IIRC, where the question annotators are only given an initial paragraph to generate questions, with the actual content of linked documents being unseen. Thus it raises the question of how to make use of such learning signal and correctly model the retrieval step for unanswerable questions.

[3] An example would be looking for the birth year of some person when such information is not presented even in the Wikipedia article titled with their name.

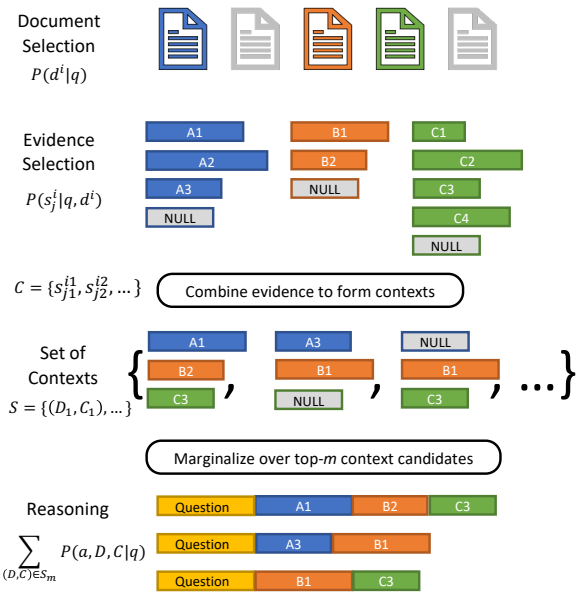

Figure 2: Our proposed framework of set-valued retrieval for handling context of unknown size.

3 Learning with Marginalization over Retrieval

To address these challenges, we decompose the retrieval problem into document selection and evidence retrieval, leaving the handling of unanswerable questions entirely to the evidence retrieval component. These two components together produce a probability distribution over sets of retrieved contexts, which we marginalize over during training to account for false negatives. Our general framework is illustrated in Figure 2.

3.1 Modeling Multi-Document Retrieval

As described in Section 2, the end product of retrieval for multi-document QA should be a set of evidence snippets from different documents $\{s_{j1}^{i_1}, s_{j2}^{i_2}, ..., s_{jk}^{i_k}\}$ which are combined to be the reasoning context C. When a question is not answerable, C is empty, and supervision of the model is not entirely straightforward. The decision of where to look for evidence is separate from whether the necessary information is present, and a naive supervision that nothing should be retrieved risks erroneously telling the model that the place it chose

to look was incorrect, possibly leading to spurious correlations between question artifacts and answerability judgments. For this reason, we separate document selection from evidence retrieval, and we leave answerability determinations entirely to the evidence retrieval step.

Document selection Given the question q and a set of documents $\mathcal{D} = \{d^1, d^2, ..., d^n\}$, to evaluate the relevance of each document d^i and the question q, we first jointly encode them with a transformer-based model. To model the selection of documents as a set variable, a Sigmoid function (σ) is used to compute the document probability:

$$x_{d^i} = \text{Encode}(d^i, q)$$
$$P(d^i|q) = \sigma(w_d^\top \cdot x_{d^i} + b_d)$$

For simplification, we assume the selection of each document is independent, thus the joint probability of selecting a set of documents $D = \{d^{i_1}, d^{i_2}, ..., d^{i_k}\}, D \subseteq \mathcal{D}$ can be computed with:

$$P(D|q) = \prod_{d^i \in D} P(d^i|q) \cdot \prod_{d^{i'} \in \mathcal{D}-D} (1 - P(d^{i'}|q)) \tag{1}$$

Evidence retrieval Given the set of selected documents D, the goal for the evidence retrieval model is to select the evidence snippets $s_j^i \in d^i$ that are relevant to question q for each document $d^i \in D$. To model the relevance between an evidence snippet and the question, we first use pretrained language models to obtain a joint embedding of the concatenated question-evidence input. To simplify the problem while approximating the set of evidence snippets as context, we only take one evidence snippet from each document. In addition, we allow the evidence retrieval model to retrieve nothing from a document by predicting `NULL`, a special token which is artificially added to the end of every document. This is essential for our modeling since it places the responsibility of determining sentence-level relevance solely on the evidence retrieval step and allows it to reject the proposal from the document model by selecting the `NULL` option, which is useful especially for unanswerable questions. Finally, we model the probability of an evidence snippet being retrieved given its document as:

$$x_{s_j^i} = \text{Encode}([s_j||q])$$
$$P(s_j^i|q, d^i) = \underset{s_j^i \in d^i}{\text{Softmax}}(w_s^\top \cdot x_{s_j^i} + b_s)$$

Here we can derive the joint probability of a set of evidence snippets $C = \{s_{j_1}^{i_1}, s_{j_2}^{i_2}, ..., s_{j_k}^{i_k}\}$ being retrieved as context:

$$\begin{aligned} P(D, C|q) &= P(D|q) \cdot P(C|D, q) \\ &= P(D|q) \prod_{s_j^i \in C, s_j^i \in d^i, d^i \in D} P(s_j^i|d^i, q) \end{aligned} \tag{2}$$

3.2 Joint Modeling with Marginalization

With the retrieved context C, the final step is to predict the answer. For this part, we use existing reading comprehension models that take a question and relatively small context and output a probability distribution of its answer predictions. The retrieved sentences in the context are simply concatenated and treated as context for the reading comprehension model (RC). Given the context C and question q, the probability of the answer is defined as:

$$P(a|q, D, C) = \text{RC}(q, [s_{j_1}^{i_1}||s_{j_2}^{i_2}||...s_{j_k}^{i_k}]) \tag{3}$$

Now we can derive the joint probability of the answer and the retrieved context:

$$\begin{aligned} &P(a, D, C|q) \\ =&P(a|D, C, q) \cdot P(D, C|q) \\ =&P(D|q) \prod_{s_j^i \in C, s_j^i \in d^i, d^i \in D} P(a|D, C) \cdot P(s_j^i|d^i, q) \end{aligned}$$

With the objective to maximize the likelihood of the training set with supervision on gold $\bar{C}$, $\bar{D}$ and $\bar{a}$, the loss function is as in Equation 4:

$$\min_\theta \mathcal{G}_D + \mathcal{G}_C + \mathcal{G}_a \tag{4}$$

where

$$\begin{aligned} \mathcal{G}_D =& -\sum_{d^i \in \bar{D}} \log P(d^i|q) \\ & -\sum_{d^{i'} \in \mathcal{D}-\bar{D}} \log(1 - P(d^{i'}|q)) \\ \mathcal{G}_C =& -\sum_{d^i \in \bar{D}} \sum_{s_j^i \in d^i, s_j^i \in \bar{C}} \log P(s_j^i|d^i, q) \\ \mathcal{G}_a =& -\log P(\bar{a}|\bar{D}, \bar{C}, q) \end{aligned}$$

Marginalization over Retrieved Evidence As mentioned in Section 2, the learning signals for $\{\mathcal{G}_D, \mathcal{G}_C, \mathcal{G}_a\}$ may be noisy and weak because the objectives in Equation 4 assume that given a question-answer pair (q, a) there is only one set

of gold context $\bar{C}$ that can derive the correct answer. To augment the learning signal, we propose to add the weakly-supervised objective with marginalization over a set of alternative context $S = \{(D'_1, C'_1), (D'_2, C'_2), ..., (D'_m, C'_m)\}$ given the selected documents D:

$$\begin{aligned} \mathcal{G}_M &= -\log P(\bar{a}|q) \\ &= -\log \sum_{(D'_i, C'_i) \in S} P(\bar{a}, D'_i, C'_i|q) \end{aligned} \tag{5}$$

Ideally, we want the marginalization set S to be all possible combinations of sentences in different documents, but this is infeasible for large text corpora. So here, we approximate the marginalization set by: 1) using only the top-ranked document set D, and 2) selecting only the top-m contexts from each d^i in D.

However, not all of the contexts in the top-m set S are good alternative contexts, especially when the retrieval model is under-trained and performs poorly. We use a set of answer-type-dependent heuristics to determine whether a context C is *valid*: (1) *Span*: when the context has at least one span that matches the gold answer string; (2) *Number*: when the answer can be derived from the numbers in the context with the arithmetic operation supported by our RC model, or it is a span in the context; (3) *Yes/No/Unanswerable*: all contexts are considered valid. Using these heuristics, we can divide the top m retrieved context S into two subsets $S_1, S_2 \subseteq S$, where S_1 contains all *valid* contexts while the contexts in S_2 are invalid.

Auxiliary Loss for Invalid Context Because contexts in S_2 are not valid alternative contexts for obtaining the correct answer, we do not marginalize over contexts in this set. We can still use them during training, however, by formulating an auxiliary loss that encourages the RC model to predict the question as unanswerable (*i.e.,* $a = a_N$) given these invalid contexts:

$$\mathcal{G}_{a_N} = -\sum_{(D^*, C^*) \in S_2} \log P(a_N|D^*, C^*, q) \tag{6}$$

Note that here we do not use joint probability $P(a_N, D^*, C^*|q)$ since doing so would also encourage the retrieval models to retrieve irrelevant context for answerable questions. In this way, this auxiliary loss can also be viewed as augmenting the dataset with extra unanswerable question-context pairs for the RC model.

Our final learning objective of the joint modeling is the result of combining the objectives in Equation 4 , Equation 5 and Equation 6:

$$\min_{\theta} \mathcal{G}_D + \mathcal{G}_C + \mathcal{G}_a + \mathcal{G}_M + \alpha \cdot \mathcal{G}_{a_N} \tag{7}$$

The only weight we tune in the objective is α to regulate the contribution of the loss from the invalid contexts the RC model encounters at training time.

4 Experiments

4.1 Datasets and Settings

We test our method on two multi-document question answering datasets: IIRC and HotpotQA.

IIRC (Ferguson et al., 2020) is a dataset consisting of 13K information-seeking questions generated by crowdworkers who had access only to single Wikipedia paragraphs and the list of hyperlinks to other Wikipedia articles, but not the articles themselves. Given an initial paragraph, a model needs to retrieve missing information from multiple linked documents to answer the question. Since the question annotators can only see partial context, the questions and contexts containing the answers have less lexical overlap. The questions in IIRC may have one of four types of answers: 1) *span*; 2) *number* (resulting from discrete operations); 3) *yes/no*; 4) *none* (when the questions are unanswerable).

HotpotQA fullwiki (Yang et al., 2018) consists of 113K questions and the contexts for answering those questions are a pair of Wikipedia paragraphs. In the *fullwiki* setting, the model has access to 5.2M paragraph candidates and needs to retrieve relevant information from this corpus. We believe this open domain QA setting would provide a different perspective for studying false-negative contexts, especially across different documents.

4.2 Model Details

Transformer-based pretrained language models are used to encode questions and contexts for retrieval and reasoning in our experiments. To be as comparable to previous methods as possible, we use `RoBERTa-base` for IIRC and `BERT-large-wwm` for HotpotQA fullwiki. For document selection in IIRC, we only rely on the initial paragraph with hyperlinks and not the content in linked documents themselves. On HotpotQA fullwiki, we follow Asai et al. (2020) and bootstrap our document selection model from their recurrent graph retriever. For the reasoning part,

Model	Dev				Test			
	Retrieval		Reasoning		Retrieval		Reasoning	
	Doc-F1	Rt-Recall	QA EM	QA F1	Doc-F1	Rt-Recall	QA EM	QA F1
Baseline (NumNet+)	-	-	29.6	33.0	-	-	27.7	31.1
Pipeline (NumNet+)			41.7	45.8			41.3	44.3
†Pipeline (T5-Large)	**88.7**	**62.8**	44.2	47.4	90.8	**67.9**	37.8	41.0
†Pipeline (PReasM-Large)			46.5	50.0			42.0	45.1
Ours (NumNet+)	87.3	62.0	**46.9**	**50.6**	**91.0**	67.5	**47.4**	**50.5**

Table 1: **Main results on IIRC.** "Baseline" refers to the performance reported in Ferguson et al. (2020) and "-" denotes that no results are available. Work marked with † is by Yoran et al. (2021), which appeared after our initial submission. All pipeline models shares the same retrieval model and its output thus the same retrieval performance.

we use NumNet+ (Ran et al., 2019) for IIRC and `BERT-wwm` (Devlin et al., 2019) for HotpotQA. Note that our general modeling framework is agnostic to the choices of specific models for retrieval and reasoning. We choose these models to experiment with since they are easy to use and present strong results as shown in previous work (Groeneveld et al., 2020). More implementation details can be found in Appendix A.

4.3 Evaluation Metrics

For HotpotQA, we follow previous work (Yang et al., 2018; Asai et al., 2020; Xiong et al., 2020) and use F1 score and Exact Match (EM) for the answer (QA) and supporting facts (SP) prediction. Similarly, we report QA F1 and EM for IIRC as in Ferguson et al. (2020). In addition, we define the following metrics for understanding the retrieval performance: *(1) Document selection F1 (Doc-F1)* measures the performance of the document retrieval model given the documents marked as gold; *(2) Overall retrieval recall(Rt-Recall)* measures the retrieval ability of the overall retrieval system given the annotated set of evidence snippets. Over the metrics above, our main goal is to improve question answering performance, which is best measured by QA F1.

4.4 Training Settings

Since documents can contain up to hundreds of sentences, for efficient training of our evidence retrieval model, we downsample the negative examples to 7 for IIRC and 3 for HotpotQA. But no downsampling is done during inference. For IIRC, we take $m = 4$ for the top-m context marginalization and take $m = 5$ for HotpotQA. For the weight for invalid context loss, we use 0.5 for IIRC[4] and 0 for HotpotQA since it does not have unanswerable questions. For memory and storage efficiency, we tie the pretrained language model weights among all the components in our joint model. The models are trained for 30 epochs for IIRC and 5 epochs for HotpotQA fullwiki. Our most expensive experiment takes about 1.5 days to run on two RTX 8000 (48GB) GPUs or one A100 (40GB) GPU, while a typical experiment takes about half of that computing power.[5]

5 Main Results on IIRC

Table 1 shows our main results on IIRC. We can see that our proposed joint model with marginalization outperforms the pipeline model by 5.2 and 4.8 points for QA exact match and F1 score, respectively. While the 17.6 point improvement over the baseline system seems large, the correct point of comparison for our contribution in this work is our pipeline system, which is simply an improved version of the pipeline used by the baseline system.[6] Another comparison worth noticing is that despite the large improvement on the QA side, the retrieval performance is slightly lower than its pipeline counterpart. Our hypothesis is that our trained joint model with marginalization can better utilize the alternative contexts that are not marked as gold and derive the correct answers based on them. Since the retrieval performance is compared with only the annotated evidence, the gain from alternative contexts cannot be reflected in these numbers. We explore this hypothesis in Section 5.1 and some examples are shown in Section 7 and Appendix B.

To further understand the effectiveness of joint modeling with marginalization comparing to the

[4]We performed a simple binary search and found 0.5 to work better than 0 or 1.

[5]Note that our models are jointly trained, thus this is the total training cost for both retrieval and reasoning.

[6]The main differences include using RoBERTa instead of BERT during retrieval, adding a `NULL` option for evidence retrieval and having a binary classifier for binary questions.

Answer Types (%)	Pipeline	Joint Model
None (26.7%)	49.6	62.2 (+12.6)
Yes/No (9.8%)	52.3	62.5 (+10.2)
Span (45.6%)	48.4	47.2 (-1.2)
Number (17.9%)	30.0	35.6 (+5.6)
All (100%)	45.8	50.6 (+4.8)

Table 2: **IIRC answer F1 breakdown by answer types.** Number in the brackets refer to the absolute improvement from the pipeline model.

Model	Doc-F1	Rt-Recall	QA F1
full model	87.3	62.0	**50.6**
- invalid context loss	87.5	**62.5**	49.2
- marginalization	**87.8**	62.2	47.9
- joint modeling	**87.8**	62.4	45.1

Table 3: **Marginalization and other ablations on IIRC.** Note that the removal of parts (noted by "-") from the full model is *accumulative*[7]. The last setting is equivalent to a pipeline model with shared RoBERTa weights.

pipeline model, we breakdown the QA performance by different answer types in Table 2. Our proposed method yields large performance gains on unanswerable questions, and those with binary and numerical answers. As we discussed in Section 2, since the retrieval model cannot rely on lexical overlap between contexts and correct answers for these types of questions, it is harder for it to learn from the false negatives, and the reasoning model trained in a pipeline is more susceptible to noise in the retrieved context. We also notice that the QA F1 on span-type questions drops 1.2 points; we think this is because the auxiliary loss we have on invalid contexts slightly altered the distribution to favor the unanswerable questions. To confirm this, we removed the auxiliary loss and its QA F1 on span questions went back up to 48.7 points.

5.1 Analysis

Effectiveness of retrieval marginalization Table 3 shows that training with marginalization improves the final QA F1 performance by 2.7 points, while doing slightly worse in terms of retrieving annotated context. To go beyond pure numbers and explain why modeling with retrieval marginalization results in better final QA performance, we analyzed 50 questions where the model with marginalization correctly answered a question that the model without marginalization missed, and 50 questions where the opposite was true. We found that in 24% of the cases where the marginalization model was correct, it relied on non-gold evidence to make its prediction, while this was only true 4% of the time for the model without marginalization. This suggests that marginalization over retrieval improves the QA performance by retrieving alternative contexts that can help reasoning. We show some specific examples of this in Section 7 and Appendix B.

Effectiveness of invalid context loss Without the auxiliary loss on the subset of invalid contexts during marginalization, we observe a 1.4 points decrease in the QA performance. On further inspection, we found that the main reason was the performance on the unanswerable questions, which decreases 8.9 points in F1 (not shown in the table).

Effectiveness of joint modeling We also explore the setting where both marginalization and joint modeling are taken away from our model. This is similar to the pipeline setting but we minimize the sum of fully supervised losses from all three models and the pretrained language model weights are shared. The difference between rows 3 and 4 in Table 3 illustrates the performance improvements from joint modeling alone, which is 2.8 QA F1. We believe this is largely due to the fact that when the model is jointly trained, the reasoning model is dynamically adapting to the noisy retrieval results, which makes it more resilient to noise at test time.

6 Results on HotpotQA

The general effectiveness of retrieval marginalization and joint modeling on HotpotQA fullwiki is displayed in Table 4. We observe a 4.8 QA F1 improvement with joint modeling and a further 4.1 points improvement with retrieval marginalization. Similar to IIRC, joint modeling and retrieval marginalization improve final QA performance despite inferior retrieval scores evaluated against annotated supporting evidence. To better understand how our proposed methods alter the retrieval and reasoning steps, we compare the output

[7]It is also a must because marginalization depends on joint modeling and the auxiliary loss depends on having a marginalization set.

[8]Here we compare with Multi-hop RAG, an adaptation of RAG (Lewis et al., 2020) to HotpotQA fullwiki by Xiong et al. (2020).

Model	SP EM	SP F1	QA EM	QA F1
full model	14.3	60.7	**58.6**	**71.2**
- marginalization	18.2	63.1	54.4	67.1
- joint modeling	**28.0**	**66.0**	50.2	62.3
Previous work				
SR-MRS (Nie et al., 2019)	39.9	71.5	46.5	58.8
T-XH (Zhao et al., 2019)	42.2	71.6	50.2	62.4
HGN (Fang et al., 2020)	50.0	76.4	56.7	69.2
Asai et al. (2020)	49.2	76.1	60.5	73.3
MDR (Xiong et al., 2020)	**57.5**	**80.9**	**62.3**	**75.3**
RAG[8](Lewis et al., 2020)	-	-	51.2	63.9

Table 4: **HotpotQA answer F1 on the development set.** Best performance in our ablation and from previous work are in **bold**. "-" denotes that no results are available.

from our full model as well as the version without marginalization and joint modeling. Interestingly, while we found most alternative sources of evidence are from the same document in IIRC, the alternative evidence our full model locates for HotpotQA fullwiki is mostly from different documents. We believe this is because HotpotQA fullwiki considers far more documents (*i.e.,* 5.2M in the whole Wikipedia) than IIRC (*i.e.,* less than 20 linked documents) for each question, but HotpotQA typically use the introductory paragraph instead of the whole document, giving less chance for inner-document alternatives. Concrete examples are shown in Section 7 and Appendix B

While the purpose of these experiments is to show the effectiveness of our proposed methods for mitigating the false negatives in HotpotQA rather than competing with state-of-the-art models, we do list some results from previous work in Table 4 to help put ours in context. With joint modeling and retrieval, our full model achieves a QA F1 of 71.2 and EM of 58.6, surpassing several strong baselines presented by previous work, though there is still a non-trivial gap with state-of-the-art models. Potential improvements that are orthogonal to our contributions include modeling document selection as a sequence and condition selection based on previous selected documents as done by Asai et al. (2020) and Xiong et al. (2020), and using stronger language models (*e.g.,* ELECTRA (Clark et al., 2019)). Note that all three modules of our framework are simple classification models on top of BERT, and that the formulation of our proposed set-valued retrieval and marginalization are model-agnostic.

7 Qualitative Analysis

As mentioned in Section 5.1 and Section 6, the reason for our model to achieve higher QA scores though having lower scores in matching annotated evidence is that our model locates false-negative evidence that can be used as alternatives for reasoning. In Table 5, we show two of such alternatives our model found in the development set for IIRC and HotpotQA fullwiki. From the IIRC example, we can see that our model is able to find alternative evidence in different sentences of the same gold document (*i.e., "Alexander Hamilton"*) or even in a non-gold document (*i.e., "Aaron Burr"*), mitigating false-negatives annotations in sentence or document retrieval. The HotpotQA example shows that in multi-hop reasoning, key evidence is not exclusively in documents titled with bridging entity (*i.e., "E Street band"*), but also sometimes included in a document for related third entity (*i.e., "David Sancious"*) as well. This indicates that such false-negative contexts can be more prevalent in multi-hop QA settings.

8 Related Work

False-negative contexts in QA In terms of dealing with false negatives in retrieved texts for question answering, the most similar prior work to ours is by Clark and Gardner (2017). However, they focused only on span-type answers while we apply similar methods to more complex reasoning types. False negative contexts are not exclusive to IIRC or HotpotQA, but are a rather common issue when scaling up QA to document or multi-document level. In prior analysis of TyDiQA (Clark et al., 2020) and Natural Questions (Kwiatkowski et al., 2019), it is suggested that humans typically have low recall on finding supporting evidence, though they focused on dealing with such issues in evaluation while we focus on altering the training process. Such issue was also found by previous work on HotpotQA, as Xiong et al. (2020) noted that many of the "errors" in their document retrieval model are actually valid alternative contexts. The false-negative contexts can also lead to false-negative annotations of answer spans. In some previous work (Chen et al., 2017; Hu et al., 2019; Asai et al., 2020), it was shown effective to manually add distantly-supervised examples in the reasoning model's training, while we try to solve this problem from the retrieval part.

IIRC EXAMPLE
Question: Who won the famous duel that took place in Weehawken? **Answer**: Aaron Burr **Evidence in initial paragraph**: Hudson County, New Jersey: Weehawken became notorious for duels, including the nation's most famous between Alexander Hamilton and Aaron Burr in 1804. **Gold evidence**: Alexander Hamilton: Burr took careful aim and shot first, and Hamilton fired while falling, after being struck by Burr's bullet. **Predicted alternative evidence**: 1. Alexander Hamilton: Taking offense, Burr challenged him to a duel on July 11, 1804, in which Burr shot and mortally wounded Hamilton, who died the following day. 2. Aaron Burr: Both men fired, and Hamilton was mortally wounded by a shot just above the hip.
HOTPOTQA FULLWIKI EXAMPLE
Question: What band did both Delores Holmes and an inductee to the Rock and Roll Hall of Fame sing in? **Answer**: E Street Band **Gold evidence**: Garry Tallent : Tallent was inducted as a member of the E Street Band into the Rock and Roll Hall of Fame. **Predicted alternative evidence**: 1. David Sancious : In 2014, Sancious was inducted into the Rock and Roll Hall of Fame as a member of the E Street Band. 2. E Street Band : The band was inducted into the Rock and Roll Hall of Fame in 2014. **Other evidence in common**: Delores Holmes : She was best known for her years as backup singer for the Bruce Springsteen Band during 1969 to 1972, the last grouping before the E Street Band.

Table 5: Examples of alternative evidence retrieved for IIRC (up) and HotpotQA fullwiki (down) by our model. Document titles are underlined and key information is marked in blue.

Joint models for retrieval and reasoning More recent work such as ORQA (Lee et al., 2019) and Dense Passage Retrieval (Karpukhin et al., 2020) focus on modeling retrieval for question answering. But their main focus is to develop efficient neural retrieval systems that help to scale up QA to a large corpus, which is orthogonal to our contribution in mitigating false-negative contexts. Some recent works also investigate the possibility of modeling retrieval as a latent task and enabling end-to-end training. REALM (Guu et al., 2020) used a neural retriever to augment language models with external knowledge for question answering. While REALM also uses a maximum marginal likelihood objective, it only marginalizes over retrieved documents. However, our model marginalizes over a set of contexts consisting of evidence snippets from different documents to account for set-valued retrieval of variable sizes, which is crucial for multi-document QA. RAG (Lewis et al., 2020) adopted a similar method as REALM, but for sequence generation tasks. RAG can be adapted to perform multi-hop question answering (Xiong et al., 2020), and we directly compared our results with multi-hop RAG on HotpotQA fullwiki in the experiments. Finally, our work leverages marginalization over latent variables to deal with weak and noisy supervision signals, which is reminiscent of using maximum marginal likelihood for training weakly supervised semantic parsers (Berant et al., 2013; Krishnamurthy et al., 2017, among others).

9 Conclusion

We proposed a new probabilistic model for retrieving set-valued contexts for multi-document QA and show that training the QA model with marginalization over this set can help mitigate the false negatives in evidence annotations. Experiments on IIRC and HotpotQA fullwiki show that our proposed framework can learn to retrieve unlabeled alternative contexts and improves QA F1 by 5.5 on IIRC and 8.9 on HotpotQA.

Acknowledgements

The authors would like to thank Tushar Khot, Hannaneh Hajishirzi, James Ferguson, Dragomir Radev, and the anonymous reviewers for the useful discussion and comments.

Ethical Considerations

We address the ethical considerations of this work from the following perspectives:

Misuse potential Our models are built specifically for answering factoid questions on Wikipedia. We think it is unlikely for our method to be misused for other domains.

Failed modes In our case, failing to answer a user-issued question may result in incorrect or misleading information. Thus we should be careful when put our systems into practical use.

Computation power Our most expensive experiment takes about 1.5 days to run on two RTX 8000 (48GB) GPUs or one A100 (40GB) GPU, while a typical experiment takes about half of that computing power. While conducing experiments, we made effort to take advantage of technologies such as mixed-precision training to shorten our training time under the same experiment setting and save power consumption.

References

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

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

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.

Christopher Clark and Matt Gardner. 2017. Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online. Association for Computational Linguistics.

James Ferguson, Matt Gardner, Hannaneh Hajishirzi, Tushar Khot, and Pradeep Dasigi. 2020. Iirc: A dataset of incomplete information reading comprehension questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1137–1147.

Dirk Groeneveld, Tushar Khot, Mausam, and Ashish Sabharwal. 2020. A simple yet strong pipeline for HotpotQA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8839–8845, Online. Association for Computational Linguistics.

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2020. Neural module networks for reasoning over text. In *International Conference on Learning Representations (ICLR)*.

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

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. Retrieve, read, rerank: Towards end-to-end multi-document reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2285–2295, Florence, Italy. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Jayant Krishnamurthy, Pradeep Dasigi, and Matt Gardner. 2017. Neural semantic parsing with type constraints for semi-structured tables. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1526.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob 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*, 7:452–466.

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

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2553–2566.

Qiu Ran, Yankai Lin, Peng Li, Jie Zhou, and Zhiyuan Liu. 2019. NumNet: Machine reading comprehension with numerical reasoning. In *Proceedings of EMNLP*.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. 2020. Answering complex open-domain questions with multi-hop dense retrieval.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Ori Yoran, Alon Talmor, and Jonathan Berant. 2021. Turning tables: Generating examples from semi-structured tables for endowing language models with reasoning skills.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2019. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *International Conference on Learning Representations*.

A Model Implementation Details

Since the two datasets where we conduct our experiments are different in terms of document length and structure, reasoning types from questions and possible answer types, here we describe the dataset-specific implementation details for IIRC and HotpotQA.

A.1 Document Selection

Though IIRC and both settings from HotpotQA can be seen as multi-document QA problems, the documents are structured differently. Accordingly we handle the document selection part differently for the two datasets, namely using different Encode$(\cdot)$ functions mentioned in section Section 3.1. But note that the outputs are all a set of documents (or paragraphs for HotpotQA) with their probabilities, so other parts of the framework remain the same.

Link prediction for IIRC. For IIRC, an initial paragraph p is given and we need to follow certain links in it, so the document selection problem can be translated into a link prediction problem given p and q. So for IIRC, we define Encode$(\cdot)$ as the BERT embedding of the concatenated question-paragraph sequence at the position of the link l_i to the document d_i:

$$\text{Encode}_{\text{IIRC}}(d^i, q) = \text{BERT}_{[l_i]}([q||p])$$

For IIRC, since we do not know how many links needs to be followed to answer the questions, we use $P(d^i|q, p) = 0.5$ as a threshold for document selection.

HotpotQA fullwiki. To select paragraphs for HotpotQA, we simply concatenate the question and the first 64 tokens of the candidate paragraph d^i, run it through BERT and take the embedding of the separation token:

$$\text{Encode}_{\text{Hotpot}}(d^i, q) = \text{BERT}_{[\text{SEP}]}([q||d^i])$$

We further follow Asai et al. (2020) and apply their trained recurrent retriever to retrieve a small subset of relevant paragraphs $\mathcal{D}'$ with the highest score. We choose to use this model because they are the best performing model on HotpotQA fullwiki setting with public code. For a better learning signal, we manually add the paragraphs marked as gold if they are not already included in $\mathcal{D}'$ but we only use $\mathcal{D}'$ at test time for a fair comparison under the fullwiki setting.

A.2 Reading Comprehension Models

We use existing reading comprehension models for both IIRC and HotpotQA. NumNet+ (Ran et al., 2019) is used for IIRC since it can handle numerical reasoning. We augment the model by adding binary and unanswerable as two additional question types to its question type classification model and further introduce a binary classification model for outputting "Yes" and "No" when a question is classified as binary type. For HotpotQA, to handle questions with binary answers, we append "yes or no" to the start of the retrieved context to transform its reasoning part to a pure span-prediction problem. Then we follow Devlin et al. (2019) and append two linear layers to the contextualized embeddings from transformer-based language models, and they are used to separately model the starting and ending position of the span.

B Retrieved Alternative Contexts

Here we show more examples of alternative contexts retrieved by our proposed methods, for the development set for IIRC in Table 6 and for HotpotQA fullwiki in Table 7.

Question: In what year was the football club joined by Haycock for one season in 1910 first founded?
Answer: 1884
Evidence in initial paragraph:
Fred Haycock: After a couple of years he ... and then Southern League clubs Luton Town and Portsmouth, before returning to the Football League with Lincoln City in 1910.
Gold evidence:
Lincoln City F.C.: After the disbanding of Lincoln Rovers (formerly Lincoln Recreation) in 1884, Lincoln City FC was formed.
Predicted alternative evidence:
Lincoln City F.C.: Previously, Lincoln City had played at the nearby John O'Gaunts ground since the club's 1884 inception.

Question: Where was the representative who gave approval to read excerpts from McCaughey's op-ed born?
Answer: Waterloo, Iowa
Evidence in initial paragraph:
Death panel: On July 27, excerpts from the McCaughey's op-ed were read, with approval, by Representative (Rep.) Michele Bachmann (R-MN) on the floor of the U.S. House of Representatives.
Gold evidence:
Michele Bachmann: Bachmann was born Michele Marie Amble in Waterloo, Iowa.
Predicted alternative evidence:
Michele Bachmann: Michele Marie Amble was born in Waterloo, Iowa on April 6, 1956, to Norwegian-American parents David John Amble (1929–2003) and "Arlene" Jean Amble (née Johnson) (born c. 1932).

Question: Did the singer Nicholls opened for at the De La Warr Pavilion release any studio albums?
Answer: Yes
Evidence in initial paragraph:
Danni Nicholls: Nicholls also opened as support for Lucinda Williams at the De La Warr Pavilion Bexhill-on-Sea for one of only two shows in the UK and completed ...
Gold evidence:
Lucinda Williams: In 1988, she released her self-titled album, Lucinda Williams.
Predicted alternative evidence:
Lucinda Williams: The American folk/rock band Augustana references the musician in the song "Meet You There," on their studio album Can't Love, Can't Hurt.

Table 6: Examples of alternative evidence retrieved for IIRC by our model. Document titles are underlined and key information is marked in blue.

Question: What race track in the midwest hosts a 500 mile race every May?
Answer: Indianapolis Motor Speedway
Gold evidence:
1957 Indianapolis 500 : The 41st International 500-Mile Sweepstakes was held at the Indianapolis Motor Speedway on Thursday, May 30, 1957.
Predicted alternative evidence:
1. 1974 Indianapolis 500 : The 58th 500 Mile International Sweepstakes was held at the Indianapolis Motor Speedway in Speedway, Indiana on Sunday, May 26, 1974.
2. 1977 Indianapolis 500 : The 61st International 500 Mile Sweepstakes was held at the Indianapolis Motor Speedway in Speedway, Indiana on Sunday, May 29, 1977.
Other evidence in common:
Indianapolis Motor Speedway : The Indianapolis Motor Speedway is an automobile racing circuit located in Speedway, Indiana, (an enclave suburb of Indianapolis) in the United States.

Question: What is the middle name of the actress who plays Bobbi Bacha in Suburban Madness?
Answer: Ann
Gold evidence:
Bobbi Bacha : Bobbi Bacha is a Texas Private Investigator portrayed in 2004 TV Sony Pictures Movie "Suburban Madness" played by actress Sela Ward.
Predicted alternative evidence:
Suburban Madness : Suburban Madness is an American crime drama television film, based on a true story, starring Sela Ward as PI Bobbi Bacha of Blue Moon Investigations.
Other evidence in common:
Sela Ward : Sela Ann Ward (born July 11, 1956) is an American actress, author and producer, best known for her roles on television beginning in the early '80s.

Table 7: Examples of alternative evidence retrieved for HotpotQA fullwiki by our model. Document titles are underlined and key information is marked in blue.