Simple and Efficient ways to Improve REALM

Vidhisha Balachandran[1] **Ashish Vaswani**[2] **Yulia Tsvetkov**[3] **Niki Parmar**[2]
[1] Language Technologies Institute, Carnegie Mellon University
[2] Google Research
[3] Paul G. Allen School of Computer Science & Engineering, University of Washington
vbalacha@cs.cmu.edu
yuliats@cs.washington.edu
{avaswani, nikip}@google.com

Abstract

Dense retrieval has been shown to be effective for Open Domain Question Answering, surpassing sparse retrieval methods like BM25. One such model, REALM, (Guu et al., 2020) is an end-to-end dense retrieval system that uses MLM based pretraining for improved downstream QA performance. However, the current REALM setup uses limited resources and is not comparable in scale to more recent systems, contributing to its lower performance. Additionally, it relies on noisy supervision for retrieval during fine-tuning. We propose REALM++, where we improve upon the training and inference setups and introduce better supervision signal for improving performance, without any architectural changes. REALM++ achieves ∼5.5% absolute accuracy gains over the baseline while being faster to train. It also matches the performance of large models which have 3x more parameters demonstrating the efficiency of our setup.

1 Introduction

Open-domain question answering (ODQA) (Voorhees et al., 1999) is a task that aims to answer questions directly using a large set of documents without being given a specific document. These systems generally employ a "retriever-reader" based approach where a *document retriever* first retrieves a subset of evidence documents and a *document reader* processes the documents to identify the correct answer (Chen et al., 2017). Recently, dense retrieval methods (Seo et al., 2018, 2019; Das et al., 2019; Karpukhin et al., 2020) have improved over sparse retrievers like BM25 (Robertson and Zaragoza, 2009) and made training these systems end-to-end by leveraging approximate MIPS search (Shrivastava and Li, 2014). REALM is an end-to-end model, pre-trained on masked language modeling, that can be finetuned for QA tasks without relying on external sources like BM25 for supervision like DPR (Karpukhin et al., 2020). Hence, it is simple and easier to train but is not competitive to pipeline alternatives like DPR. When finetuning, it uses a single GPU making it not directly comparable in scale to DPR which uses more resources for better optimization. Due to limited resources it is inefficient, taking more than a day to train. Additionally, it uses distant supervision for the retriever in the form of passages containing the target answer leading to ambiguous supervision for training.

In this paper, we present a study of REALM aimed at understanding and improving its limitations. We find that REALM is significantly underoptimized and improve the training by scaling the system through (i) using exact MIPS search, (ii) introducing larger batch training, and (iii) scaling the reader to process more documents. We further address the noisy distant retrieval supervision by augmenting the training sets with human-annotated evidence passages. Since such human annotations are not available for every dataset and is expensive to obtain, we show that models trained with strong supervision transfer well to other datasets where such annotations are not available, indicating the benefits beyond a single annotated dataset.

Incorporating our best findings, we show that an improved version of REALM, which we call REALM++ achieves ∼5.5% absolute accuracy improvements over the baseline on multiple ODQA benchmarks while processing 4x more examples/sec and outperforms all prior methods of similar parameter regime. Further, it shows comparable performance to models with 3x more parameters. Our results demonstrate that scale and supervision play an important role in ODQA systems highlighting the need for careful comparisons across systems in ODQA and for taking scale and efficiency into account in addition to performance when reporting results.

Proceedings of the 3rd Workshop on Machine Reading for Question Answering , pages 158–164
Nov 10th, 2021

Experiments	Test EM	Dev EM	Dev R@10
REALM (Guu et al., 2020)	40.4	38.2	-
REALM (Ours)	39.4	35.6	68.8
+Scale	42.8	37.9	69.5
+Scale+PS (10 docs inf)	43.2	38.6	69.9
+Scale+PS (100 docs inf)	44.8	38.6	69.9
+Scale+Rerank	42.3	37.4	67.5

Table 1: Answer Span EM Accuracy and Answer Recall@10. **Improving training setup improves Test EM Acc.** Test = test set, Dev = development set

Model	R@10	DevEM
REALM	68.8	35.6
ScaledR (FixedRet)	59.6	33.1
ScaledR+Rerank (FixedRet)	67.9	35.8
ScaledR+Rerank+PS (FixedRet)	67.5	37.1
ScaledR (TrainedRet)	**69.5**	**37.9**
ScaledR+Rerank (TrainedRet)	67.5	37.4

Table 2: Answer Recall and Span EM for fixed v/s finetuned retriever. ScaledR = Scaled REALM, PS = Passage Supervision. **Reranking is useful when retriever is fixed** but is not effective when the retriever is trained.

2 Exploring Limits of REALM

Open Domain QA is typically modeled as a Machine Reading Comprehension (MRC) model which answers a question using a large corpus of text documents/passages by employing a "retriever-reader" approach. REALM specifically uses dense retrieval to identify c ($c = 5000$) relevant passages and a BERT based reader to process a smaller set of top-k ($k = 5$) passages and find answer spans. When finetuning[1], the retriever is trained using distant supervision with passages containing the target answer as positive and the reader is supervised using human annotated short answer spans (Lee et al., 2019). We follow the same design and optimization setup of finetuning REALM on QA.

We explore the limits of various experiment choices by introducing simple changes to the training and inference setup. Table 1 compares results from our replicated experiments of REALM to prior published results and shows that our experiments produce similar results on the Natural Questions (NQ) (Kwiatkowski et al., 2019) dataset. Detailed analysis across other metrics is in in A.2.

2.1 Scaling the Training Setup

REALM performs an approximate MIPS for retrieving the top c relevant documents based on a retrieval score, $S_{retr}(p_i, q) = h_q^\top h_{p_i}$ where h_q and h_{p_i} are question and passage representations respectively. The system is finetuned in practice on a single machine with a 12GB GPU with batch size 1. While this is modest use of resources, we show that this results in suboptimal training. We begin by scaling the REALM system during training. We perform *exact* MIPS search by leveraging the efficiency of large matrix multiplications of TPUs (Wang et al., 2019) to compute the retrieval score, S_{retr}, for ∼13M passages of corpus and extract c passages having the highest scores. We further increase the training batch size to 16 by leveraging 8 TPUv3 cores on Google Cloud for distributed training. Finally, we increase the number of documents passed to the reader to $k = 10$ during training.

Scaling training setup improves QA results: From Table 1 we observe that simple experiment choices like larger batch training and exact MIPS search significantly improve the Exact-Match Accuracy by 3.4% without introducing any model design changes. This shows that the original REALM setup was under-optimized and has much better performance than previously reported.

2.2 Introducing Strong Passage Supervision

To finetune the retriever, REALM relies on distant supervision in the form of passages containing the target answer. However, such a signal can lead to noisy and unrelated documents to be given a positive signal (Lin et al., 2018) as examples in Table 5 of Appendix A show. We address this by introducing supervision from human annotations similar to Yang et al. (2015); Nguyen et al. (2016), to train the retriever by updating the retrieval scores by optimizing their marginal log-likelihood.

$$P(p_i|Q) = \frac{\exp(S_{retr}(p_i, Q))}{\sum_{p_j \in \{p_i\}_{1:c}} \exp(S_{retr}(p_j, Q))}$$

$$L(Q, LA) = -\log \sum_{p_j \in \{p_i\}_{1:c}, p_i \in LA} P(p_i|Q)$$

where LA is a list of human annotated evidence passages (e.g. Long Answers in Natural Questions), $L(Q, LA)$ denotes the passage supervision loss that is augmented to the existing retriever distant supervision and span prediction loss, and $p_i \in LA$ indicates whether the passage was in the annotated passages. Here, we assume that the passages in corpus and the annotated evidence passages in the dataset are from the same source (e.g.

[1]REALM is pretrained using MLM on CC-News corpus

Wikipedia). Since corpus passages and annotated passages in the dataset can differ (e.g. due to different Wikipedia versions), we consider any passage in the retrieved set that has 50%[2] word overlap[3] with the target passages as a positive match.

Supervision through evidence passage annotations improves performance: From Table 1 we see an improvement of 0.5% over the scaled REALM model leading to 3.8% improvement over the prior baseline REALM model, showing the benefit of providing better supervision. In §A.3 we present qualitative examples where the improved passage supervision leads to answer spans being extracted from the most relevant document to the question. While noisy distant supervision has been shown as effective for dense retrieval, our work experimentally shows that it can be limiting and simply introducing better supervision through gold evidence passages is beneficial.

2.3 Reranking

Table 4 in A.2 shows that though the retrieved 5000 documents has high answer recall ($\sim 95\%$), the recall significantly drops ($\sim 77\%$) in the top 10 documents processed by the reader. Readers are computationally intensive and memory limits makes scaling them to process more documents difficult. We explore an approach to rerank the retrieved documents to improve recall@10 and end accuracy.

Our Document Reranker has L layers of cross-document and query-document interactions to learn rich document representations. For each layer, the output passage representations from the previous layer, $\{u^{l-1}\}_{1:c}$ are first passed through a Transformer block (T) with multi-headed self-attention (Vaswani et al., 2017) which allows for interaction between the passage representations and produces cross-document aware representations u_i^l.

$$u_i^l = \text{T}(Q{=}u_i^{l-1}, K{=}\{u^{l-1}\}_{1:c}, V{=}\{u^{l-1}\}_{1:c})$$

where Q, K, V represent the query, key and value respectively in the transformer attention module. To model interaction between passages and query, the attended passage representation u^l and query representations from the previous layer v^{l-1} are passed through a multi-head cross-attention Transformer to produce query aware representations v_i^l.

$$v_q^l{=}\text{T}(Q{=}v_q^{l-1}, K{=}\{u^l\}_{1:c}, V{=}\{u^l\}_{1:c})$$

For the first layer we consider the dense retriever's query and document representations as the input u^0 and v^0 to the reranker. The rich document and query representations from the final layer ($\{u^L\}_{1:c}$,v_q^L) are used to compute the retriever score, $S_{retr}(p_i, q)$ to find the top-k documents for the reader.

Document Reranking does not significant gains when retriever is jointly trained but is highly effective when retriever is fixed: In Table 1 We observe that the accuracy of the system drops by 0.5% and the recall@10 drops by 0.2% when augmenting the reranker. We further study the role of the reranker in a fixed retriever setting where the top 5000 documents are retrieved once and kept constant during training. While, such a setting is a more efficient since documents are not retrieved at every training step, the retriever's zero-shot performance can be quite low, potentially hurting end accuracy. From Table 2, we see that the scaled REALM model with a fixed-retriever has very low recall@Top-10 and EM. Here, augmenting the model with the Document Reranker significantly improves recall and EM performance, where recall@Top-10 improves by 8.3% and EM by 2.7%. Further introducing passage supervision during training improves performance by increasing the end accuracy by $\sim 1.3\%$ making the fixed retriever setting very competitive in performance to a jointly trained retriever-reader setting.

2.4 Scaling the Reader at Inference

Due to memory constraints the reader cannot be scaled to process more documents during training without architectural changes. Such constraints do not apply during inference, since optimization based weights and parameters are not saved, the memory usage of the model reduces, allowing for the reader to process more documents. We experiment with scaling the reader to process more documents only during inference.

Scaling the reader during inference significantly boosts performance: In Table 1 we see that the reader processing $k = 100$ documents significantly improves accuracy, achieving 44.8% on NQ which surpasses the baseline REALM by 4.4%. This shows that such systems can leverage a small number of documents ($k = 10$) for faster training and gain the benefits of scaling the reader ($k = 100$) at inference. Further from Figure 1 we

[2]We experimented with thresholds=(0.3, 0.5, 0.75) and used threshold with best performance based on validation set

[3]We also experimented with ngram overlap which was similar in performance but computationally expensive.

Model	Size	NQ (79k/4k)	WQ (3k/2k)	CT (1k/1k)
ORQA (Lee et al., 2019)	Base	33.3	36.4	30.1
PathRetriever (Asai et al., 2020)	Base	32.6	-	-
$\text{REALM}_{\text{News}}$ (Guu et al., 2020)	Base	40.4	40.7	42.9
DPR (Karpukhin et al., 2020)	Base	41.5	42.4	49.4
$\text{ReConsider}_{\text{Base}}$ (Iyer et al., 2020)	Base	43.1	44.4	49.3
REALM++ (10 docs)	Base	43.2	44.5*	47.2*
REALM++ (100 docs)	Base	**44.8**	**45.6***	**49.7***
$\text{DPR-BERT}_{\text{Large}}$ (Iyer et al., 2020)	Large	44.6	44.8	53.5
$\text{RAG}_{\text{Large}}$ (Lewis et al., 2020b)	Large	44.5	45.5	52.2
$\text{ReConsider}^{\dagger}_{\text{Large}}$ (Iyer et al., 2020)	Large	**45.5**	**45.9**	**55.3**

Table 3: Test QA (Exact Match) Accuracy on Open-QA benchmarks showing **REALM++ improving over prior methods of similar size.** The number of train/test examples are shown in parentheses next to each benchmark. *indicates models finetuned on trained NQ model, as proposed in (Karpukhin et al., 2020). †Though $\text{ReConsider}_{large}$ has higher accuracy, their approach of using answer span focused reranking model is orthogonal can be directly applied to our output.

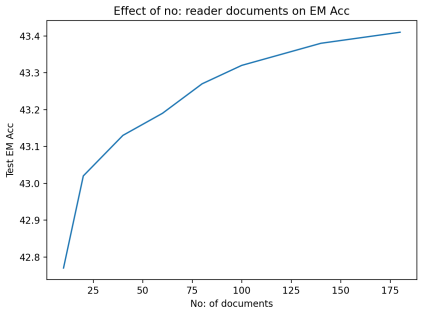

Figure 1: Test QA EM Acc v/s No: Reader Documents on NQ. **QA Span EM increases when more documents are processed by the reader during inference.**

see that the gains increase with increasing number documents with a slight saturation beyond 120 documents. This is potentially due to increased answer recall in documents[4].

3 REALM ++

Based on the findings in §2, we incorporate the best working components: (i) scaling at training §2.1 (ii) better passage supervision §2.2 and (iii) scaling reader during inference §2.4 to establish an improved REALM model, which we call REALM++. We study the effect of REALM++ on three datasets NQ, Web Questions (WQ) (Berant et al., 2013), and Curated Trec (CT) (Baudiš and Šedivỳ, 2015). As WQ and CT do not have evidence passage annotations, we use them to study the transfer capabilities of the passage supervised NQ model.

REALM++ outperforms models of similar size and is comparable to large models: Table 3 shows that REALM++ outperforms prior methods of similar size (models based on BERT_{base}) with no modifications to the model design. When transferred to WQ and CT which do not have human annotation for evidence passages, REALM++ shows an improvement of 3.8% on WQ and 4.3% on CT over base REALM showing benefit beyond a single dataset. REALM++ produces state-of-art results on extractive ODQA among models of similar size in all three datasets using a single end-to-end model. Additionally REALM++, which uses BERT_{base} ($\sim$ 110M params), performs comparable to large models based on BERT_{large} and BART_{large} (Lewis et al., 2020a) ($\sim$ 340M params) with 3x lesser params.

Discussion of speed and memory usage: By using 8 TPUv3 cores and increased batch size for training our REALM++ model, we can process 4x more examples/sec as compared to REALM and reduce training time from 2 days to 12hrs. REALM++ maintains the same number of parameters as the base REALM model and the entire model fits within 12GB memory which is the equivalent of an Nvidia Titan X. This demonstrates that our REALM++ model is efficient and can improve training time by leveraging distributed training.

4 Conclusion

In this work, we present a study of a dense-retrieval QA system, REALM, and identify key limitations in its experimental setup. We find that REALM is significantly undertrained and we improve REALM by introducing simple changes to its training, supervision, and inference setup. We

[4]For the rest of the experiments in Table 3, we use 100 documents for fair comparison to other methods

propose REALM++ which incorporates our best findings and show that it can achieve significant improvement over prior methods and perform comparably with models with 3x more parameters.

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 *ICLR*.

Petr Baudiš and Jan Šedivỳ. 2015. Modeling of the question answering task in the yodaqa system. In *International Conference of the Cross-Language Evaluation Forum for European Languages*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *EMNLP*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *ACL*.

Rajarshi Das, S. Dhuliawala, M. Zaheer, and A. McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. *ICLR*.

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

Srinivasan Iyer, Sewon Min, Yashar Mehdad, and Wentau Yih. 2020. Reconsider: Re-ranking using span-focused cross-attention for open domain question answering. *arXiv preprint arXiv:2010.10757*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *TACL*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *ACL*.

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

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *neurips*.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset.

S. Robertson and H. Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Minjoon Seo, T. Kwiatkowski, Ankur P. Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2018. Phrase-indexed question answering: A new challenge for scalable document comprehension. In *EMNLP*.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. In *ACL*.

Anshumali Shrivastava and P. Li. 2014. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). In *NIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Ellen M Voorhees et al. 1999. The TREC-8 question answering track report. In *TREC*.

Yu Emma Wang, Gu-Yeon Wei, and David Brooks. 2019. Benchmarking tpu, gpu, and cpu platforms for deep learning. *arXiv preprint arXiv:1907.10701*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

A Example Appendix

A.1 Datasets

For our study, we use three open-domain QA datasets following Guu et al. (2020):

Natural Questions (NQ) (Kwiatkowski et al., 2019) contains real user queries from Google Search. We consider questions with short answers (<=5 tokens) and the long answers for passage supervision.

WebQuestions (WQ) (Berant et al., 2013) is a collection of questions collected from Google Suggest API, with Freebase entity answers whose string forms are the target short answers.

CuratedTREC (CT) (Baudiš and Šedivỳ, 2015) contains curated questions from TREC QA track with real user questions and answer as regular expression matching all acceptable answers.

A.2 REALM baseline analysis

Experiments fairly reproduce results: Table 4 reports the results from our experiments and compares them to published results from REALM (Guu et al., 2020). We find that our experiments produce similar results on NQ and WQ with slightly lower results on CT on the test set. We believe that this could be due to varying checkpoints due to early stopping. For fair evaluation, we use results from our experiments as a comparison for the remainder of the study.

Answer recall drops significantly with reducing documents: We additionally present a breakdown of REALM's retriever and reader performances on the development set across the three datasets. While REALM retrieves $c = 5000$ documents for distantly supervising the retriever, only the top $k = 5$ documents are processed by the reader for finding the right answer. Comparing the recall of answers in the retrieved documents at different subsets of documents we observe very high ($> 90\%$) recall@5000 for all three datasets but the recall@5 effectively drops to $\sim$70%, showing that the document that contains the answer is not necessarily present in the top-5 highlighting limitations in the retriever.

Wide margins exist for improving reader performance: Comparing the Exact Match accuracy of the system with the upper bound (the system is right if the passage contains the answer) shows wide gaps in the performance of the reader. While $\sim$63% of the questions from NQ have the answer in the top retrieved documents, REALM is only able to get the exact span of the answer for $\sim$36% of them showing the limitations of the reader in identifying the exact answer span in the document.

Metric	NQ	WQ	CT
Test EM (Guu et al., 2020)	40.4	40.7	42.9
Test EM	39.4	40.8	39.3
Dev EM	35.6	45.4	42.9
Dev EM Upper Bound	63.3	75.9	70.8
R@5	63.9	78.5	70.0
R@10	69.4	85.7	76.4
R@100	80.7	94.0	85.6
R@1000	86.8	97.7	91.7
R@5000	91.1	99.2	93.3

Table 4: **Experiments reproduce results of REALM on NQ dataset**. First section compares Test EM from our experiments with previous published results from (Guu et al., 2020). The bottom section compares Dev EM with Upper Bound performance of the Reader.

A.3 Qualitative Analysis

In §2.2, we introduced strong passage supervision from annotated evidence passages to enable the model to distinguish misleading passages that might contain the target answer. Table 6 shows examples of questions where using passage supervision helps retrieve correct passages for the QA task. For Questions 1 and 3, the baseline model incorrectly retrieves a wrong passage of a similar genre or topic as the question, while for Question 2 the baseline model retrieves a completely incorrect, irrelevant passage. The model trained with passage supervision identifies the right context for answering the question, which aligns with the human annotation for each question.

Question	Incorrect Passage	Correct Passage from Human Annotations
Where did the idea of a unicorn come from?	Unicorn is a privately held startup company whose name was coined in 2013 by venture capitalist Aileen Lee.	Unicorns are not found in Greek mythology, but rather in the accounts of natural history, for Greek writers of natural history were convinced of the reality of unicorns.
What type of reproduction do whiptail lizards use?	MLB Whiparound is an American baseball television show on Fox Sports 1 hosted by Chris Myers and Kevin Burkhardt.	The New Mexico whiptail lizard is a crossbreed of a western whiptail and the little striped whiptail. The lizard is a female-only species that reproduces asexually by producing an egg through parthenogenesis.
Which president supported the creation of the Environmental Protection Agency(EPA)?	Some historians say that President Richard Nixon's southern strategy turned the southern United States into a republican stronghold, while others deem economic factors more important in the change.	The Environmental Protection Agency (EPA) is an agency of the federal government of the United States created for the purpose of protecting human health and the environment. President Richard Nixon proposed the establishment of EPA and it began operation on December 2, 1970, after Nixon signed an executive order.

Table 5: Examples of Questions from Natural Questions with incorrect retrieved passages from Guu et al. (2020) with the correct human annotated relevant passages showing the necessity for human annotation based supervision.

Question	Incorrect Ret Passage	Correct REALM++ Ret Passage
Where did the Battle of Issus take place?	The Battle of Alexander at Issus is a 1529 oil painting by the German artist Albrecht Altdorfer, a pioneer of landscape art and a founding member of the Danube School .	The Battle of Issus occurred in southern Anatolia, on November 5, 333 BC between the Hellenic League led by Alexander the Great and the Achaemenid Empire, led by Darius III.
Who played bubba in the Heat of the Night?	A late Stevan Ridley touchdown run set up by a 23 - yard Deangelo Peterson run on a fourth - down play gave LSU the upset victory and effectively ended the opportunity for an Alabama repeat of the national championship.	Carlos Alan Autry Jr. is an American actor, politician, and former National Football League player. He played the role of Captain Bubba Skinner on the NBC television series, "In the Heat of the Night", starring Carroll O'Connor.
Actress who plays penelope garcia on criminal minds?	How to get away with Murder is an American television series created by Peter Nowalk and produced by Shonda Rhimes and ABC Studios.	Kirsten Simone Vangsness is an American actress, currently staring as FBI Analyst Penelope Garcia on the CBS series "Criminal Minds".

Table 6: Qualitative Analysis of questions from NQ showing questions where baseline REALM retrieved incorrect passages and **training with passage supervision helped retrieve the right passage**.