Limitations of Open-Domain Question Answering Benchmarks for Document-level Reasoning

Ehsan Kamalloo*
Univesity of Waterloo
Waterloo, Ontario, Canada
ekamalloo@uwaterloo.ca

Charles L. A. Clarke
Univesity of Waterloo
Waterloo, Ontario, Canada
claclark@gmail.com

Davood Rafiei
University of Alberta
Edmonton, Alberta, Canada
drafiei@ualberta.ca

ABSTRACT

Many recent QA models retrieve answers from passages, rather than whole documents, due to the limitations of deep learning models with limited context size. However, this approach ignores important document-level cues that can be crucial in answering questions. This paper reviews three open-domain QA benchmarks from a document-level perspective and finds that they are biased towards passage-level information. Out of 17,000 assessed questions, 82 were identified as requiring document-level reasoning and could not be answered by passage-based models. Document-level retrieval (BM25) outperformed both dense and sparse passage-level retrieval on these questions, highlighting the need for more evaluation of models' ability to understand documents, an often-overlooked challenge in open-domain QA.

CCS CONCEPTS

• **Information systems** → **Test collections**; **Retrieval effectiveness**; **Question answering**; *Document representation.*

KEYWORDS

open-domain question answering, document-level reasoning, long context question answering

ACM Reference Format:
Ehsan Kamalloo, Charles L. A. Clarke, and Davood Rafiei. 2023. Limitations of Open-Domain Question Answering Benchmarks for Document-level Reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3539618.3592011

1 INTRODUCTION

Retrieving candidate documents that contain potential answer(s) is at the core of *Open-domain Question Answering (QA)* whose main goal is to find answers to information-seeking questions over a massive collection of long documents. A broad range of retrieval models have been adopted for this purpose, from sparse retrievers such as BM25 [45, 53] to dense retrievers [27], retrieval-augmented models [6, 34], and more recently, large language models [7, 12, 36]. Shared among all these models is the granularity of retrieval. In particular, the standard practice is to split long documents — e.g., Wikipedia articles — into fixed-length passages [51].

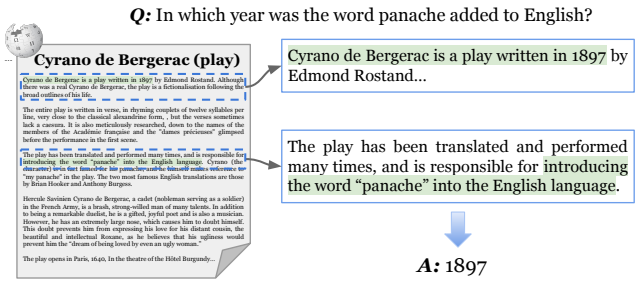


Figure 1: An example question that highlights the importance of document-level reasoning in retrieving passages.

The ubiquity of passage retrieval is rooted in the struggles of QA models, even large language models [7], in capturing long-range dependencies within documents [28]. In the deep learning era, the limited context size of neural models [2, 19] is another contributing factor to the popularity of passage retrieval. Nonetheless, passage retrieval has repeatedly proven to be effective on a range of open-domain QA benchmarks [14, 27, 51].

The key assumption in passage retrieval is that the knowledge source is substantial or has enough redundancy such that the answer can be found somewhere in a localized context [14]. Documents are written in a logically-structured manner and follow a cohesive narrative [25]. By carving their discourse into passages, the relationship among different parts of documents (e.g., coreferences) is not maintained. These issues introduce additional challenges that can impede models from answering some questions. For example, in Figure 1, the key information to answer the question is dispersed in two paragraphs that are distant from each another.

In this paper, our goal is to examine (1) how much the current open-domain QA benchmarks test for document-level reasoning, and (2) whether passage-based models are able to identify answers in the absence of document-level evidence. To this end, we conduct a pilot study over a small set of manually-crafted questions — e.g., the question in Figure 1 — to verify if passage-based pipelines fail when document-level evidence is required. Motivated by this observation, we explore the prevalence of such questions in existing open-domain QA datasets by carrying out document-level retrieval on three widely adopted open-domain QA datasets — i.e., Natural Questions-OPEN [33], TriviaQA [26], and WebQuestions [5]. We find 325 questions for which document-level retrieval outperforms

*Work done while at University of Alberta.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3592011

passage-level retrieval. We manually audit these questions and find only 82 questions that require document-level reasoning to answer. These questions are factoid questions with no composite structure, unlike multi-hop questions that may also require reasoning over long contexts. Our analysis reveals that these benchmarks are heavily skewed toward questions where passage-level information is sufficient, thus lacking long-context reasoning. Our data is released at https://github.com/ehsk/DocReasoning-OpenQA.

Our contributions include: (1) providing an in-depth analysis of three widely adopted open-domain QA benchmarks to identify questions for which document-level evidence is critical, and (2) curating a subset of 82 questions from existing benchmarks to underscore the limitations in evaluation of document-level reasoning, where models often fail. Our goal is to encourage further research and benchmarks that require more than passage-level reasoning.

2 RELATED WORK

Retrieval in open-domain QA. In the deep learning era, open-domain QA models are streamlined to retriever-reader pipelines. DrQA [10] pioneered two-stage pipelines where retrieval is carried out using TF-IDF at the document level. Subsequent work [13, 32, 44, 50] added a re-ranking step that recalibrates retrieval scores for paragraphs or sentences, derived from retrieved documents. Yang et al. empirically showed that retrieval at paragraph-level yields better results [54]. Wang et al. found that fixed-length overlapping passages, in contrast to paragraphs as actual units of discourse, work best for retrieval [51]. With the rise of dense retrievers [9] and retrieval-augmented models [34], recent models [22, 27, 29, 33, 52] have switched to passage retrieval outright.

Document modelling in QA. Document retrieval is generally approximated via evidence from passages [8], also known as *passage-based document retrieval* [30], because the ability to identify localized contexts that can be captured via passage retrieval is instrumental in QA. In this avenue, many studies [4, 18, 40] aggregated passage scores to rank documents, with recent models leveraging neural models for aggregation [1, 21, 35, 46]. A simple aggregation method is to use the score of the first passage, or the best passage as a proxy for document scores [15]. In closed-domain QA, numerous efforts [11, 39, 49, 56] leveraged document structure to answer questions. In line with this work, we suggest that a combination of two granularity levels — i.e., document and passage — should be considered for open-domain QA.

Reasoning in QA. Datasets with different types of reasoning [16, 20, 38, 43, 55] are prolific in QA. Among the various types of complex questions, multi-hop questions may require document-level reasoning if the multiple pieces of evidence to answer a question are taken from the same document. Multi-hop questions are generally written with a composite structure to connect multiple statements together. In contrast, our questions in this paper are simple factoid questions with no composite structure. The need for document-level reasoning does not hinge on how questions are written and in fact, is a byproduct of how documents are written. Another related line of work is long context QA in closed-domain settings [17] where questions are asked given a long context.

3 METHOD

3.1 Materials

Datasets. We use the following three widely-used information-seeking QA datasets:

(1) **Natural Questions-OPEN (NQ-OPEN) [33]:** Originally derived from Natural Questions (NQ) [31], this dataset serves as an established benchmark for factual question answering. NQ is curated from Google search queries for which answers can be found in Wikipedia. Since the original test set in NQ is hidden, the original dev set that contains 3,610 questions is used as an unseen test set [27, 33].
(2) **TriviaQA (TQA) [26]:** TQA is comprised of trivia questions mined from a variety of quiz-league websites. Similar to NQ-OPEN, the dev set of TQA with 11,313 questions is used as the test data since the original test set is hidden [27].
(3) **WebQuestions (WQ) [5]:** Consisting of 2,032 questions, this dataset was collected for QA over knowledge bases. In WQ, questions are obtained from the Google Suggest API and the answers are entities whose corresponding Freebase IDs were annotated.

Retrieval Models. BM25 is a widely employed sparse retriever for open-domain QA, which treats text as a bag of words. We employ BM25 for both passage retrieval and document retrieval. For dense retrieval, we adopt ANCE [52], a prominent dense retriever for open-domain QA, whose model checkpoints are publicly available. In summary, we use three retrievers throughout the paper: ANCE and BM25 for passage retrieval, and BM25 for document retrieval.

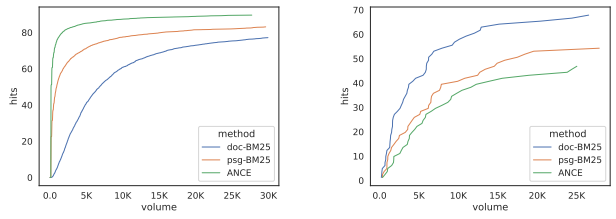
3.2 Document retrieval vs. Passage retrieval

To understand when document-level reasoning is appropriate, we first need to compare the output of document retrieval against passage retrieval results. However, the ranking of candidate documents is not directly comparable to the ranking of candidate passages because of the disparity in their granularity levels, passage vs. document. To overcome this problem, we compute the *text volume*, which we define as the minimum number of tokens that must be read to find an answer in the retrieved results, equalizing the two granularity levels. More specifically, given that each document/passage is a sequence of tokens, we accumulate the number of tokens from the top of the retrieved list until an answer is found. We compute the *hit ratio* (*hits@vol*), i.e., the percentage of questions for which an answer document/passage is found, with respect to text volume.

The left plot in Figure 2 illustrates hit ratios for the three retrievers vs. text volume. Both passage retrievers outperform document retrieval by a high margin because (1) identifying localized information is crucial in answering information-seeking questions [14], and (2) documents are long and are likely to contain extraneous information that introduce noise into the ranking [35]. The rare questions for which document retrieval may be required are lost among the questions for which passage retrieval is sufficient.

3.3 Data Collection

We aim to identify questions for which the evidence is spread between different parts of a document. Table 1 summarizes the



(a) NQ-OPEN (Full) (b) NQ-OPEN (Doc wins)

Figure 2: Hits ratio vs. text volume for the full NQ-OPEN collection (left), and for the subset on which document retrieval outperforms passage retrieval (right). Although both passage retrievers outperform document retrieval by a high margin on the full dataset, there exists questions for which document retrieval outperforms both passage retrievers on the selected questions. We observed similar outcomes on TQA and WQ.

Table 1: Number of questions for which document retrieval surpasses passage retrieval

Dataset	#Questions that Doc wins vs.		
	Psg-BM25	**ANCE**	**Psg-Oracle**
NQ-OPEN	462 (12.8%)	177 (4.9%)	81 (2.2%)
TQA	595 (5.3%)	551 (4.9%)	192 (1.7%)
WQ	245 (12.1%)	113 (5.6%)	52 (2.6%)
Total	1,302	841	325

statistics for such questions on the three datasets. *Psg-Oracle* indicates the best passage retriever out of the two, BM25 as a sparse retriever or ANCE as a dense retriever. This oracle, which picks the best retriever in prior to retrieval provides an estimate of the upper bound for passage retrieval, making the comparison with document retrieval more robust. In total, for 325 questions (4.5%), passage retrieval fails after retrieving the same volume of text at which document retrieval succeeds.

We plot *hits@vol* varying text volume only on the selected questions, depicted in the right plot of Figure 2 for NQ-OPEN. We observed similar outcomes on TQA and WQ. Interestingly, ANCE struggles the most on these questions. It is on par with passage-level BM25 on TQA and WQ, while falling behind on NQ-OPEN.

4 ASSESSMENT OF FAILURE MODES

When document retrieval outperforms passage retrieval for a question, it does not necessarily mean that the question requires document-level reasoning to be answered. To this end, we conducted a human assessment, completed by one of the authors, to identify the failure modes of passage retrieval in our 325 questions.

The goal of our human annotation is twofold: (1) determining if the retrieved documents legitimately answer the question, and (2) determining whether passages are sufficient for answering the question. Thus, the annotation procedure was conducted in two steps. First, for each question, the annotator checked if the top-2 documents returned by the document retriever contain an official answer. Then, for each passage retrieval model, the annotator inspected top-5 passages. When the question was annotated as unanswerable, the top-2 passages containing an official answer were also examined, if they were not already among top-5 passages. To select a question as a candidate , the annotator scanned the documents to ensure document-level information is required to answer the question. This procedure took approximately 5 minutes per question on average, revealing three broad types of failure modes:

(A) Question-related problems (32.3%): Questions that are impossible to answer given the knowledge source [3] or are so ambiguous that they cannot be answered without clarification [42]. For example, the question "*The lyric 'Always sunny in a rich man's world', is from which song?*" cannot be answered based on textual content extracted from Wikipedia. The question "*where will the first round of march madness be played?*" misses the competition year and if it is asking about men's or women's basketball.

(B) Answer-related problems (42.5%): Questions for which annotated answers are incorrect. For the question "*how many times has psg won champions league?*" the official answer is 46, but the actual answer is 0. Other questions miss variations of answers that are acceptable. For example, for the question"*In the mid 1990s what major fossil discovery was made in Liaoning, China?*" the official answer is "*Well-preserved fossils of feathered dinosaurs*," but the phrase "*feathered dinosaur fossils*" should also be acceptable.

(C) Lack of document-level understanding (25.2%): Questions that require document-level reasoning in order to determine correct answers from the knowledge source. A detailed example of such questions is illustrated in Figure 1.

Overall, we find that for nearly 25% of the questions — 82 questions in total — document-level cues are critical. These clues include an understanding of the core topic of documents or of the document structure. In the remainder of the paper, we use these questions as a benchmark to explore possible solutions to the problem posed by the requirement for document-level retrieval.

5 EXPERIMENTS

In this section, our goal is to evaluate simple strategies that are known to work for document-level modelling on our benchmark to find whether document-level understanding presents a challenge in open-domain QA that is not captured by passage-level models.

5.1 Predicting Retrieval Granularity Level

We first investigate if a requirement for document-level retrieval is a characteristic of questions alone. We build a classifier to predict the granularity level of retrieval, which takes a question as input and predicts if retrieval should be conducted at document-level. Such a classifier is reminiscent of *a priori* answerability prediction via the question alone, which can achieve an accuracy of 73% [3].

Our knowledge source is Wikipedia articles from the snapshot of 20-Dec-2018, following [27, 33]. We used Wikipedia passages, provided by DPR [27]. Specifically, Wikipedia articles were split into non-overlapping passages of 100 words [51] along with the article title that is concatenated to the start of each passage. We use Pyserini [37] for constructing inverted indices and for obtaining pre-encoded index files for dense retrieval.

Table 2: Exact-match accuracy of mixed granularity vs. when only document retrieval and only passage retrieval is used.

Retriever	Full NQ-OPEN	Benchmark
mixed granularity	39.7	8.8
only passage granularity	**41.4**	8.6
only document granularity	33.5	**12.1**

For passage retrieval, we use the best reported BM25 parameters from DPR, $k_1 = 0.9$ and $b = 0.4$. For document retrieval, k_1 and b were tuned on the dev set of each dataset separately. We bootstrap k_1 and b by repeatedly resampling from $[0, 3]$ and $[0, 1]$ (ranges are taken from [48]), 100 times with replacement. We use $k_1 = 2.5$ and $b = 0.3$ on NQ-OPEN, $k_1 = 1.5$ and $b = 0.2$ on TQA, and $k_1 = 2.9$ and $b = 0.3$ on WQ.

Training data is constructed by computing text volume of BM25, as explained in Section 3.2, for both passage retrieval vol_{psg} and document retrieval vol_{doc} on the training set of NQ-OPEN. The label of each question is determined by $argmin(vol_{\text{psg}}, vol_{\text{doc}})$. The dataset consists of 69,896 questions in the training set with 9,848 (14.1%) labelled as document, and 3,610 questions in the test set with 858 (5.6%) for the document-level granularity. We fine-tuned $\text{RoBERTa}_{\text{base}}$ [41] on this dataset for 5 epochs , and with a weighted cross entropy loss to account for data imbalance. Our classifier achieves an accuracy of 65.7% (AUC=0.665, and recall=58.7%) on the test set.

We plug in our classifier into an open-domain QA pipeline with BM25 as the retriever and Fusion-in-Decoder (FiD) [24] as the reader, providing a *mixed granularity* strategy. As shown in Table 2, this baseline achieves an exact match (EM) accuracy of 39.7% and 8.8% on the full NQ-OPEN, and our benchmark, respectively. While the mixed granularity pipeline outperforms document granularity on the full dataset and outperforms passage granularity on our benchmark subset of 82 questions, our classifier is insufficient to solve the problem posed by our benchmark by predicting the appropriate retriever on an *a priori* basis.

5.2 End-to-End Results

In this section, we measure the end-to-end performance of various models on our benchmark of 82 questions. We pair our retrievers with FiD [24] as the reader, similar to Section 5.1. In addition, we test two state-of-the-art passage-level baselines, FiD-KD [23] and EMDR^2 [47], on our benchmark. For document retrieval, retrieved documents are split into passages as FiD accepts only passages. We restrict the number of passages that are fed to the reader to 100, similar to previous work [24, 27]. This restriction puts document retrieval at disadvantage since some parts of documents may be cut off. We also consider a simple dense document retrieval, known as `MaxP` [15], that estimates document relevance from a ranked list of passages by taking the score of the top-ranking passage in the document.

As shown in Table 3, document retrieval with our naive approach substantially underperforms on the full dataset, whereas it beats passage-level BM25 and ANCE on our benchmark. FiD-KD and EMDR^2 work best on our benchmark even though they operate at the passage level. Yet, the performance of both models significantly drops, compared to the full dataset, thus underlining the lack of document-level reasoning capabilities in these models. Moreover, the poor performance of `MaxP` demonstrates that document-level reasoning cannot be approximated using passage-level heuristics. Overall, these results highlight that document-level information is central to answering the questions in our benchmark.

Table 3: EM accuracy of various open-domain QA models on NQ-OPEN and on our identified subset.

Pipeline	Granularity	Full	Our Subset
Doc-BM25 + FiD	document	33.5	14.9
Psg-BM25 + FiD	passage	41.4	11.9
ANCE `MaxP` + FiD	document	38.4	7.5
ANCE + FiD	passage	46.6	9.0
FiD-KD	passage	49.6	**17.9**
EMDR^2	passage	**52.5**	**17.9**

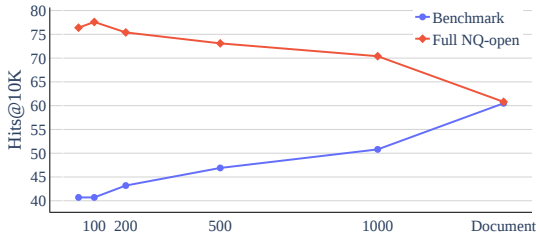


Figure 3: Hit ratio at volume 10K for various passage lengths on NQ-OPEN and our benchmark.

5.3 Varying Passage Length

One hypothesis is that increasing the passage length can be helpful when passages are not long enough to reflect the document discourse. To investigate this, we vary passage length within {50, 100, 200, 500, 1000} and perform BM25 retrieval for each passage length. To this end, we construct a separate index for each passage length and tune BM25 parameters as explained in Section 5.1. Then, we retrieve passages using BM25 over each index and measure hits ratio at volume 10K. The results are visualized in Fig. 3 for the full NQ-OPEN as well as our evaluation benchmark. Even though the performance declines overall with longer passage lengths, the hits ratio actually increases on our document-level benchmark. These results indicate that more context is indeed required to locate plausible candidates on our benchmark.

6 CONCLUSION

Passage retrieval is not sufficient for open-domain QA models especially when answering questions requires document-level reasoning. We show that this phenomenon is largely overlooked in existing benchmarks. To this end, we introduce a novel small benchmark, carefully curated from three well-known open-domain QA datasets. Our evaluation of the state-of-the-art models on this benchmark confirms our hypothesis that these models are not fit for document-level reasoning questions.

REFERENCES

[1] Qingyao Ai, Brendan O'Connor, and W Bruce Croft. 2018. A neural passage model for ad-hoc document retrieval. In *Advances in Information Retrieval*. Springer International Publishing, 537–543.

[2] Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: Encoding Long and Structured Inputs in Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 268–284. https://aclanthology.org/2020.emnlp-main.19

[3] Akari Asai and Eunsol Choi. 2021. Challenges in Information-Seeking QA: Unanswerable Questions and Paragraph Retrieval. 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)*. Association for Computational Linguistics, Online, 1492–1504. https://aclanthology.org/2021.acl-long.118

[4] Michael Bendersky and Oren Kurland. 2010. Utilizing passage-based language models for ad hoc document retrieval. *Information Retrieval* 13, 2 (2010), 157–187. https://doi.org/10.1007/s10791-009-9118-8

[5] 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*. Association for Computational Linguistics, Seattle, Washington, USA, 1533–1544.

[6] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*. PMLR, 2206–2240.

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

[8] James P. Callan. 1994. Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Springer-Verlag, 302–310.

[9] Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training Tasks for Embedding-based Large-scale Retrieval. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rkg-mA4FDr

[10] 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)*. Association for Computational Linguistics, Vancouver, Canada, 1870–1879. https://aclanthology.org/P17-1171

[11] Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. 2017. Coarse-to-Fine Question Answering for Long Documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 209–220. https://aclanthology.org/P17-1020

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint* (2022). arXiv:2204.02311

[13] Christopher Clark and Matt Gardner. 2018. Simple and Effective Multi-Paragraph Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 845–855. https://aclanthology.org/P18-1078

[14] Charles LA Clarke, Gordon V Cormack, and Thomas R Lynam. 2001. Exploiting redundancy in question answering. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 358–365.

[15] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 985–988. https://doi.org/10.1145/3331184.3331303

[16] Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning. 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)*. Association for Computational Linguistics, Hong Kong, China, 5925–5932. https://aclanthology.org/D19-1606

[17] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4599–4610. https://doi.org/10.18653/v1/2021.naacl-main.365

[18] Ludovic Denoyer, Hugo Zaragoza, and Patrick Gallinari. 2001. HMM-based passage models for document classification and ranking. In *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research*. 126–135.

[19] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://aclanthology.org/N19-1423

[20] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2368–2378. https://aclanthology.org/N19-1246

[21] Yixing Fan, Jiafeng Guo, Yanyan Lan, Jun Xu, Chengxiang Zhai, and Xueqi Cheng. 2018. Modeling diverse relevance patterns in ad-hoc retrieval. In *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval* (Ann Arbor, MI, USA). Association for Computing Machinery, 375–384. https://doi.org/10.1145/3209978.3209980

[22] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*. PMLR, 3929–3938.

[23] Gautier Izacard and Edouard Grave. 2021. Distilling Knowledge from Reader to Retriever for Question Answering. In *International Conference on Learning Representations*. https://openreview.net/forum?id=NTEz-6wysdb

[24] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 874–880. https://aclanthology.org/2021.eacl-main.74

[25] Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic Text Matching for Long-Form Documents. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, 795–806. https://doi.org/10.1145/3308558.3313707

[26] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1601–1611. https://aclanthology.org/P17-1147

[27] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://aclanthology.org/2020.emnlp-main.550

[28] Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 284–294. https://aclanthology.org/P18-1027

[29] Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided Supervision for OpenQA with ColBERT. *Transactions of the Association for Computational Linguistics* 9 (2021), 929–944. https://aclanthology.org/2021.tacl-1.55

[30] Koichi Kise, Markus Junker, Andreas Dengel, and Keinosuke Matsumoto. 2001. Passage-based document retrieval as a tool for text mining with user's information needs. In *Proceedings of the 4th International Conference on Discovery Science*. Springer-Verlag, 155–169.

[31] 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 (2019), 452–466. https://aclanthology.org/Q19-1026

[32] Jinhyuk Lee, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang. 2018. Ranking Paragraphs for Improving Answer Recall in Open-Domain Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 565–569. https://aclanthology.org/D18-1053

[33] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6086–6096. https://aclanthology.org/P19-1612

[34] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. 9459–9474.

[35] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PARADE: Passage Representation Aggregation for Document Reranking. *arXiv* abs/2008.09093 (2020). https://arxiv.org/abs/2008.09093

[36] Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. 2021. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs* (2021).

[37] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2356–2362. https://doi.org/10.1145/3404835.3463238

[38] Kevin Lin, Oyvind Tafjord, Peter Clark, and Matt Gardner. 2019. Reasoning Over Paragraph Effects in Situations. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Hong Kong, China, 58–62. https://aclanthology.org/D19-5808

[39] Dayiheng Liu, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, and Nan Duan. 2020. RikiNet: Reading Wikipedia Pages for Natural Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6762–6771. https://aclanthology.org/2020.acl-main.604

[40] Xiaoyong Liu and W Bruce Croft. 2002. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*. Association for Computing Machinery, 375–382. https://doi.org/10.1145/584792.584854

[41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692* (2019). https://doi.org/10.48550/arXiv.1907.11692

[42] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 5783–5797. https://aclanthology.org/2020.emnlp-main.466

[43] Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. 2021. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4582–4598. https://aclanthology.org/2021.naacl-main.364

[44] 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)*. Association for Computational Linguistics, Hong Kong, China, 2553–2566. https://aclanthology.org/D19-1258

[45] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995).

[46] Eilon Sheetrit, Anna Shtok, and Oren Kurland. 2020. A passage-based approach to learning to rank documents. *Information Retrieval Journal* 23, 2 (2020), 159–186. https://doi.org/10.1007/s10791-020-09369-x

[47] Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer, and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 25968–25981.

[48] Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*. 58–65.

[49] Hui Wan, Song Feng, Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis Lastras. 2021. Does Structure Matter? Encoding Documents for Machine Reading Comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 4626–4634. https://aclanthology.org/2021.naacl-main.367

[50] Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R^3: Reinforced ranker-reader for open-domain question answering. In *AAAI*.

[51] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering. 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)*. Association for Computational Linguistics, Hong Kong, China, 5878–5882. https://aclanthology.org/D19-1599

[52] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *International Conference on Learning Representations*. https://openreview.net/forum?id=zeFrfgyZln

[53] Mostafa Yadegari, Ehsan Kamalloo, and Davood Rafiei. 2022. Detecting Frozen Phrases in Open-Domain Question Answering. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, 1990–1996. https://doi.org/10.1145/3477495.3531793

[54] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-End Open-Domain Question Answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Association for Computational Linguistics, Minneapolis, Minnesota, 72–77. https://aclanthology.org/N19-4013

[55] 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*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380. https://aclanthology.org/D18-1259

[56] Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu. 2020. Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6708–6718. https://aclanthology.org/2020.acl-main.599